



Approaching the human in the loop – legal perspectives on hybrid human/algorithmic decision-making in three contexts

Therese Enarsson, Lena Enqvist & Markus Naarttijärvi

To cite this article: Therese Enarsson, Lena Enqvist & Markus Naarttijärvi (2022) Approaching the human in the loop – legal perspectives on hybrid human/algorithmic decision-making in three contexts, Information & Communications Technology Law, 31:1, 123-153, DOI: [10.1080/13600834.2021.1958860](https://doi.org/10.1080/13600834.2021.1958860)

To link to this article: <https://doi.org/10.1080/13600834.2021.1958860>



© 2021 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



Published online: 27 Jul 2021.



Submit your article to this journal [↗](#)



Article views: 1067



View related articles [↗](#)



View Crossmark data [↗](#)

Approaching the human in the loop – legal perspectives on hybrid human/algorithmic decision-making in three contexts

Therese Enarsson, Lena Enqvist and Markus Naarttijärvi

Department of Law, Umeå University, Umeå, Sweden

ABSTRACT

Public and private organizations are increasingly implementing various algorithmic decision-making systems. Through legal and practical incentives, humans will often need to be kept in the loop of such decision-making to maintain human agency and accountability, provide legal safeguards, or perform quality control. Introducing such human oversight results in various forms of semi-automated, or *hybrid* decision-making – where algorithmic and human agents interact. Building on previous research we illustrate the legal dependencies forming an impetus for hybrid decision-making in the policing, social welfare, and online moderation contexts. We highlight the further need to situate hybrid decision-making in a wider legal environment of data protection, constitutional and administrative legal principles, as well as the need for contextual analysis of such principles. Finally, we outline a research agenda to capture contextual legal dependencies of hybrid decision-making, pointing to the need to go beyond legal doctrinal studies by adopting socio-technical perspectives and empirical studies.

KEYWORDS

Hybrid decision-making; artificial intelligence; human in the loop; policing; social welfare; online moderation

1. Introduction

1.1. Background

The ambitions of integrating artificial intelligence (AI) in diverse public and private sectors are becoming increasingly apparent, with the European Commission spearheading a commitment to furthering the use of AI in public and private sectors.¹ The Commission has, however, also highlighted that these ambitions may be hampered by legal uncertainties which limit the willingness of private enterprises to invest in AI ventures, while undermining trust in AI as individuals fear an adverse impact on their rights.² In response, the Commission, as well as the EU High-level expert group on AI both stress that the allocation of functions across humans and

CONTACT Markus Naarttijärvi  markus.naarttijarvi@umu.se

¹European Commission, 'White Paper on Artificial Intelligence: A European Approach to Excellence and Trust' <https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en>, this ambition is reflected in national ambitions as well, such as the Swedish government strategy of 2020, see Swedish Government, En väl fungerande ordning för val och beslutsfattande i kommuner och regioner [SOU 2021:16].

²European Commission (n 1) 9.

AI systems should follow *human-centric* design principles and leave meaningful opportunity for human choice.³ This has further been highlighted in the proposed EU regulation of AI, through a risk-based approach requiring increased levels of human oversight in higher risk systems.⁴

In a decision-making environment, introducing human oversight of AI-based or algorithmic work processes results in various forms of semi-automated, or *hybrid* decision-making – where algorithmic and human agents interact.⁵ While the risks and promises of AI have brought increased attention to this interaction, decision-making supported by ICT in the specific context of the public sector has, well in advance of the current discourses on AI, been described as a move from *street-level bureaucracies* to *system-level bureaucracies* resulting in a reduction of human discretion. The impetus to implement hybrid decision-making may vary. In some cases, it may be driven by ambitions of increased efficiency where reducing human discretion is a specific goal which cannot fully be realized due to technical or legal constraints.⁶ In other areas, such as online moderation, the need for human contextual analysis is well known, but the sheer scope of the task facing moderators and external pressures calls for further automation.⁷ However, in many cases, keeping a *human in the loop* is a deliberate attempt to maintain human agency and accountability, and to provide legal safeguards and quality control. Hybrid decision-making can thus be said to operate in-between somewhat counterbalancing ambitions, where the wish for effectivization and automation may require a reduction of human discretion at the same time as legal requirements of maintaining human oversight and agency may necessitate such discretion.

Consequently, hybrid decision-making environments raise issues beyond the traditional understanding of pure automation. This implies that issues such as data-protection, accountability, and transparency cannot form the end-point of a discussion on hybrid decision-making. In this article, we approach the ambitions of implementing algorithmic decision making in three legal contexts; policing, social welfare systems, and online moderation.

These environments are chosen as they are currently subject to intense efforts of automation due to both external pressure and internal ambitions and necessities. While disparate regarding many of the legal rules affecting them as well as the practicalities facing decision-makers within them, they also have commonalities beyond automation ambitions. There is a core of these environments which implicates individual rights of those subject to hybrid decisions. Many of the legal principles operate across these environments, and the scope of decision-making can potentially impact a great number of individuals. We will approach each environment from a mainly European

³EU High Level Expert Group on AI (AI HLEG), 'Policy and Investment Recommendations for Trustworthy Artificial Intelligence' <<https://digital-strategy.ec.europa.eu/en/library/policy-and-investment-recommendations-trustworthy-artificial-intelligence>>; European Commission, 'White Paper on Artificial Intelligence: A European Approach to Excellence and Trust' <https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en>.

⁴European Commission, 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. 2021 [COM(2021) 206 final], article 14.

⁵On this terminology, see further below in Section 1.2.

⁶See Section 2.1.

⁷See Section 4.4.

legal standpoint, using the Swedish legal system as a sample when necessary to put legal requirements in a jurisdictional context.

Through these examples, we highlight how legal dependencies are likely to make each context dependent on different forms of hybrid decision-making systems. Drawing on previous literature and research, this article makes three primary contributions: First, it highlights the need to situate hybrid decision-making in a wider legal environment including not only data protection rules relating to hybrid decision-making, but also constitutional and administrative legal principles. Second, we illustrate how general principles and rules operating on higher normative levels – such as constitutional principles and human rights – will need to be contextually situated and interpreted taking into account specific circumstances and implications of hybrid decisions in each context. Third, it outlines a research agenda to capture contextual legal dependencies of hybrid decision-making through a focus on the *human in the loop*. Our analysis highlights how a wider set of legal principles permeates this hybrid decision-making, influencing the degree to which they can be implemented in line with legal requirements. It also points to the need for research into hybrid decision-making environments to go beyond legal doctrinal studies, by the implementation of a socio-technical perspective and the use of empirical studies.

1.2. A brief note on terminology

Previous research has discussed hybrid (human/algorithmic) decision-making through a diverse terminology, which is also indicative of the variety of research fields having approached the issue. In this article, ‘hybrid decision-making’ is seen as a form of *semi-automated decision making*. It includes what Smit and Zoet describe as decisions executed by decision-making processes featuring both human and machine actors,⁸ and observations like those made by Morison and Harkens on humans not being replaced in these processes but taking on ‘a different role, as the overseer and correcting mechanism for the algorithmic predictions’.⁹ ‘Hybrid decision-making’, as we use it here, is therefore agnostic to the degree of automation of a process – and is thus inclusive towards a range of different types of technologically mediated systems of support, risk evaluation and investigation. We will, however, situate the term in a wider set of overlapping and competing concepts.

Hybrid decision-making can involve the use of *algorithmic support* for human decision-makers, with algorithmically generated knowledge systems used both to execute or inform decisions.¹⁰ Importantly, hybrid decisions signify a degree of interaction between algorithmic and human agents. Consequently, hybrid systems are not fully automated, *human-out-of-the-loop*, systems. Hybrid decision-making instead comprises a range of systems. It includes those systems where human agents retain full decision-making autonomy but rely on algorithmic or automated aspects of information gathering,

⁸Koen Smit and Martinj Zoet, ‘A Governance Framework for (Semi) Automated Decision-Making’, *Proceedings of the Tenth International Conference on Information, Process, and Knowledge Management (eKNOW)* (2018).

⁹John Morison and Adam Harkens, ‘Re-Engineering Justice? Robot Judges, Computerised Courts and (Semi) Automated Legal Decision-Making’ (2019) 39 *Legal Studies* 618, 626.

¹⁰Michael Veale, Max Van Kleek and Reuben Binns, ‘Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making’, *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (ACM, 2018) <<https://dl.acm.org/doi/10.1145/3173574.3174014>> accessed 7 June 2021.

as well as the range of *recommendation*¹¹ or *recommender*¹² systems. It also includes those systems where humans are included as a primarily rubber-stamping mechanism, with only nominal control or responsibility for decisions (termed ‘quasi-automation’ by Wagner).¹³

Also worthwhile is relating hybrid decision-making to the degree to which the decision-making is overseen through *humans-in-the-loop* (HITL), *humans-on-the-loop* (HOTL), or *humans-in-command* (HIC), a terminology both commonly used in research,¹⁴ and included in the EU ethics guidelines for trustworthy AI.¹⁵ According to the EU high-level expert group definitions, HITL requires ‘capability for human intervention in every decision cycle of the system’, while HOTL instead takes aim at human intervention through design and monitoring of the system as such.¹⁶ In this article ‘hybrid decision-making’ includes both HITL and HOTL systems, although excludes those systems where the human control and oversight is limited to deciding when to implement an otherwise fully automated system (and to evaluate its output on a more systematic level). The latter type of systems may be referred to as HIC systems,¹⁷ or *meta-autonomy*.¹⁸

While some of the above-mentioned terms and concepts have made imprints in legal discussions, it is important to note that they, as of yet, have no fixed legal meaning or effect. And that the terminology we have covered, with a few exceptions, has not primarily been established or used in research of legal nature. Our reference to ‘hybrid decision-making’ is therefore not an attempt to suggest any legal definition or fixed legal implications of such systems (which we argue still cannot be done). Our aspiration is rather to delineate a problem area that will help us focus our analysis on the interaction between human and algorithmic agents in hybrid systems. By doing so, our aim is to further highlight that work processes and practical interactions between human and algorithmic agents must be considered, similar to what has been described by Endert and others as *human-is-the-loop* analytics.¹⁹ Otherwise, tensions between human values and statistically focused algorithms may be lost.²⁰ Also of interest is the impact of technological mediation and the sedimentation of choices made in the decision-making system design processes for human decision-making agents using those systems.²¹

¹¹Christian Djeflal, ‘AI, Democracy and the Law’ in Andreas Sudmann (ed), *The Democratization of Artificial Intelligence* (Transcript Verlag, 2019), 265.

¹²Karen Yeung, ‘Algorithmic Regulation: A Critical Interrogation: Algorithmic Regulation’ (2018) 12 *Regulation & Governance* 505, 507.

¹³Ben Wagner, ‘Liable, but Not in Control? Ensuring Meaningful Human Agency in Automated Decision-Making Systems: Human Agency in Decision-Making Systems’ (2019) 11 *Policy & Internet* 104.

¹⁴See for instance; Morison and Harkens (n 9) 625–26; Claudio Coletta and Rob Kitchen, ‘Algorithmic Governance: Regulating the “Heartbeat” of a City Using the Internet of Things’ (2017) 4 *Big Data & Society* 1–16.

¹⁵EU High Level Expert Group on AI (AI HLEG), ‘Ethics Guidelines for Trustworthy AI’ <<https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>>, 16.

¹⁶*ibid.*

¹⁷*ibid.*

¹⁸Luciano Floridi and others, ‘AI4People – an Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations’ (2018) 28 *Minds and Machines* 689, 698.

¹⁹Alex Endert and others, ‘The Human Is the Loop: New Directions for Visual Analytics’ (2014) 43 *Journal of Intelligent Information Systems* 411, 413.

²⁰Morison and Harkens (n 9) 626.

²¹Cf. Vlad Niculescu-Dincă, ‘Towards a Sedimentology of Information Infrastructures: A Geological Approach for Understanding the City’ (2018) 31 *Philosophy & Technology* 455.

2. Hybrid decisions – the view beyond the law

2.1. Bureaucrats everywhere!

The issue of human discretion in decision-making where ICT and automation are deployed has been given considerable attention within public administration research (broadly defined). Starting off by drawing on this research, we will highlight those features that may also benefit legal research by aiding a more theoretical understanding of hybrid decision-making. As will be seen, the perspectives and insights from this public administration research may also inspire and parallel a broader socio-technical discussion on decision-making in private organizations and the foundations of algorithmic decision-making through big data.

In public administration research, Lipsky's 'street-level bureaucrats' is often used as a stepping off point, which describes public servants' capacity of exercising discretion in the decision-making process – essentially making policy through their interactions with citizens.²² Lipsky's theory led to developments in the field of public administration, and later the rise of public management as a research field focusing on managerial decisions.²³ The next theoretical step towards a focus on automation came from Bovens and Zouridis who in the digital context observed a shift from street-level, through 'screen-level', towards 'system-level' bureaucracies, where judgment is increasingly delegated to computer systems with automated decision-making as a result.²⁴

ICT has come to play a decisive role in the organizations' operations. It is not only used to register and store data, as in the early days of automation, but also to execute and control the whole production process. Routine cases are handled without human interference. Expert systems have replaced professional workers. Apart from the occasional public information officer and the help desk staff, there are no other street-level bureaucrats as Lipsky defines them.²⁵

The result of this shift towards automation has 'led to reconsiderations of bureaucrats' role and of digital discretion'.²⁶ Connected to this is a shift in management from NPM to 'digital era governance',²⁷ and a decreasing face-to-face human interaction of administrative decision-makers.²⁸ To what extent discretion is reduced in system level bureaucracies is debated. Some suggest that the increased use of system-level bureaucracy could mean the end of decision-making discretion.²⁹ Our point of departure aligns with Buffat,

²²Michael Lipsky, *Street-Level Bureaucracy: Dilemmas of the Individual in Public Services* (30th anniversary expanded ed, Russell Sage Foundation, 2010).

²³Justin B Bullock, 'Artificial Intelligence, Discretion, and Bureaucracy' (2019) 49 *The American Review of Public Administration* 751, 751.

²⁴Mark Bovens and Stavros Zouridis, 'From Street-Level to System-Level Bureaucracies: How Information and Communication Technology Is Transforming Administrative Discretion and Constitutional Control' (2002) 62 *Public Administration Review* 174. See also Bullock (n 23) 751.

²⁵Bovens and Zouridis (n 24) 180.

²⁶Agneta Ranerup and Helle Zinner Henriksen, 'Digital Discretion: Unpacking Human and Technological Agency in Automated Decision Making in Sweden's Social Services' [2020] *Social Science Computer Review* 1, p. 2. See also Bullock (n 23) 751; Peter André Busch and Helle Zinner Henriksen, 'Digital Discretion: A Systematic Literature Review of ICT and Street-Level Discretion' (2018) 23 *Information Polity* 3.

²⁷Patrick Dunleavy and others, 'New Public Management Is Dead--Long Live Digital-Era Governance' (2005) 16 *Journal of Public Administration Research and Theory* 467.

²⁸Ida Lindgren and others, 'Close Encounters of the Digital Kind: A Research Agenda for the Digitalization of Public Services' (2019) 36 *Government Information Quarterly* 427.

²⁹Stavros Zouridis, Marlies van Eck and Mark Bovens, 'Automated Discretion' in Tony Evans and Peter Hupe (eds), *Discretion and the Quest for Controlled Freedom* (Springer International Publishing, 2020) 326–27.

who suggests that there is no unilateral effect of technology, instead pointing to 'the inability of ICT tools to capture the whole picture of frontline work and choices, limited resources for managers to control time and attention, and work organization or the skills possessed by street-level agents' which leads to discretion continuing to exist.³⁰ As we will highlight however, this analysis also needs to take into account legal aspects influencing decision-makers' discretion.

Despite its many important contributions to understanding discretion, it is worth noting that public administration research rarely grapples in-depth with the legal issues surrounding decision-making. The need to interpret law within the context of concrete cases is one of the reasons for the existence of street-level bureaucrats' discretion in Lipsky's theory. This is as the bureaucrats – including judges, lawyers and police – operate with a need for 'sensitive observation and judgement, which are not reduceable to programmed formats'.³¹ Building on this, Lindgren and others state that '[i]t remains unclear if this type of public service can and should be digitized', while still pointing to how legal frameworks in Scandinavia and other places hindering the implementation of automation are being rewritten to enable it.³² The actual legal rules and principles that may determine the *scope* of discretion or *establish obstacles* for limiting discretion through automation are, however, rarely explored, though the reasons for this may vary.

In some case-studies, furthering the adherence of rules through minimizing differing interpretations or personal factors are viewed as part of the primary objectives of IT-systems, where human discretion is considered costly, inefficient and prone to error.³³ We argue that this approach entangles the analysis of the law with ICT issues, and therefore runs the risk of obscuring complex legal issues relating to discretionary determinations and interpretations of law, rather than acknowledging the need for sensitive observation and judgement.

Overall, the goal of achieving effective administration seems to be a primary focus of the field of public administration and public management when discussing digitalization. This approach also largely assumes the perspective of the administrative agency. Bullock for example states that '[q]uality of administration can be characterized, in part, by how effectively public administrators use their discretion to achieve policy goals'.³⁴ As a consequence, quality is not equated with legal compliance, but rather policy throughput, and discretion thus seems to be defined as the sphere where the administrator is legally unbound and can – ideally – pursue a more effective implementation of policy goals.

From a legal perspective however, the question of discretion is tied to the function of public administrators in a wider rule of law construct. From this perspective, discretion may exist where there is no clear legal rule deciding the outcome of a decision. This discretion does, however, not function as a *carte blanche* as it is tempered by legal principles that condition the outcome. The utilization of discretion, from a legal perspective, thus implies that account should be taken to values inherent in the legal system as such, irrespective of whether they challenge policy goals and the effectiveness of administration.³⁵

³⁰Aurélien Buffat, 'Street-Level Bureaucracy and E-Government' (2015) 17 *Public Management Review* 149.

³¹Lipsky (n 22) 14–15.

³²Lindgren and others (n 28) 431.

³³Ranerup and Henriksen (n 26) 4.

³⁴Bullock (n 23) 753.

³⁵Ronald Dworkin, *Taking Rights Seriously* (Harvard Univ Press, 1978).

Legal principles, such as the principle of proportionality, may also require that legal rules are disappplied or set aside despite being clear and precise. And, notably, this circumstance challenges the possibility to automate even contexts where rules are detailed enough to minimize legal discretion.³⁶ Although not generally of any primary concern, these potential effects of automation are not lost on public administration research. It has been recognized that AI could effectively carry out values and policies of political actors, but also eliminate ‘the chance for the personal ethics or professional competence of human bureaucrats to curb the excesses of these same politicians’.³⁷ The law may perhaps be implicitly included in ethical, democratic, professional or relational values discussed.³⁸ However, the lack of references to the legal norms that may form a stronger normative basis for curbing policy aims than personal ethics, is still noteworthy. One could argue that the professional and relational values of a public servant are exactly those that can serve to strengthen legal values, which in turn are founded in the formalization or codification of democratic and ethical values.³⁹ On this note, it should be pointed out that automation efforts are not unchallenged in the administrative context. Arguing from a legal perspective, Calo and Keats Citron claim that by engaging in automation, administrative agencies ‘undermine the premise of the administrative state’. They reason that agencies deserve their possessed powers based on their expertise, flexibility, and nimbleness – and that this is true both at a pragmatic level and on the level of first order (legal) principles. Their claim is therefore that ‘[a]gencies that automate throw away expertise and discretion with both hands’.⁴⁰

The broader issue here seems to be the potential conflict between legal rules and legal principles, where the latter implies a greater need for human judgment to make individual assessments irrespective of whether they are ineffective and costly, as this serves the preservation of those higher order legal requirements ultimately based on long term democratic foundations. In their seminal work identifying the phenomenon of system-level bureaucracies, Bovens and Zouridis highlighted how new discretionary powers granted to systems designers, and the digital rigidity to individual circumstances that such systems establish, both challenges the constitutional state.⁴¹

2.2. From public to private and beyond

So far, we have touched upon academic discussions relating to automation and the use of ICT in *public* decision-making. Similar issues may, however, also arise in relation to the use of automation and hybrid decision-making in *private organizations*. There, the use of automation may often serve similar aims of effective management and speedy decision-

³⁶The case of *Riggs v Palmer*, New York Court of Appeals 115 NY 506 (1889) is a common example, whereby clear and simple heritance rules were set aside to prevent the inheritance of a Francis B. Palmer to pass to his grandson, who murdered him to receive his inheritance. The ruling is used by legal philosopher Ronald Dworkin to prove that principles, not only rules, must be taken into account when analyzing the question of ‘what is law’, see Dworkin (n 35).

³⁷Bullock (n 23) 758.

³⁸Cf. Busch and Henriksen (n 26) 18.

³⁹Markus Naarttijärvi, ‘Legality and Democratic Deliberation in Black Box Policing’ (2019) 1 Technology and Regulation 35.

⁴⁰Ryan Calo and Keats Citron Danielle, ‘The Automated Administrative State: A Crisis of Legitimacy’ (2021) 40 Emory Law Journal 797, 835.

⁴¹Bovens and Zouridis (n 24) 174. Bullock on the other hand argues that AI systems may improve the second concern by allowing more personalized judgements than previous ICT systems, while minimizing human biases, see Bullock (n 23) 757 Of course, AI systems can also exasperate issues with supervision and public accessibility through the ‘black box’ problem, which has led to the focus on transparency of AI decision-making. see also Calo and Keats Citron (n 40) 835.

making, while in many situations not facing the same obstacles in terms of constitutional and administrative principles or demands of democratic accountability. On the other hand, the use of automation in private organizations may exhibit similar dynamics to public decision-making in areas relating to fundamental rights, or to *rule of law* concerns such as non-discrimination. The demands flowing from familiar public law ideals such as transparency, proportionality and accountability may in fact seep into the private law sphere through a multitude of channels – such as the application of non-discrimination law and principles, data protection, or the delegation of responsibilities or functions from public bodies.⁴²

Such dynamics can be seen in the area of social media moderation, where difficulties regarding the compliance with, for instance, national laws or human rights have been noted by NGOs, resulting in overly extensive moderation. This is since social media platforms lack the tools and knowledge to perform an advanced legal balancing of interests.⁴³ The use of AI in moderation has been seen as both necessary, due to the challenge with seemingly endless masses of content to moderate, and problematic due to difficulties in making contextual decisions without a human in the loop.⁴⁴ At the same time, social media research has clearly shown that, in our modern society, participation on social media or partaking in discussions on news platforms is a way of exercising freedom of speech. Striking a balance between removing hateful and abusive speech that silence people (and thus diminishing their possibilities to exercise free speech) while not infringing on others' rights to express themselves, can also be an important part of maintaining a free democratic society.⁴⁵ The responsibility of striking that balance has increasingly been placed on social media and news platforms, demanding them to make balanced decisions regarding the deletion of content or blocking of users. To ensure compliance and enable accountability in relation to both individual users and national as well as international laws, they are also required to provide transparency in their decision-making.⁴⁶

Connecting both private and public hybrid decision-making are also the wider structures that afford automation efforts, namely the accumulation and processing of large quantities of data, i.e. *big data*.⁴⁷ While not all hybrid decision-making employ big data, the current ambitions of automation in both private and public decision-making are still largely driven by the presumptions and affordances of big data and ICT, namely

⁴²See for instance Nicholas P Suzor and others, 'What Do We Mean When We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation' (2019) 13 International Journal of Communication 1526; Alexandre de Streel and others, 'Online Platforms' Moderation of Illegal Content Online Law, Practices and Options for Reform' (European Parliament 2020) Study for the committee on Internal Market and Consumer Protection, Policy Department for Economic, Scientific and Quality of Life Policies.

⁴³de Streel and others (n 42) 43.

⁴⁴Thomas Davidson and others, 'Automated Hate Speech Detection and the Problem of Offensive Language', *Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017)*; Sean MacAvaney and others, 'Hate Speech Detection: Challenges and Solutions' (2019) 14 PLOS ONE e0221152.

⁴⁵See for instance Anita Bernstein, 'Abuse and Harassment Diminish Free Speech' (2014) 35 Pace Law Review 1; Danielle Keats Citron, 'Civil Rights in Our Information Age' in Saul Levmore and Martha Craven Nussbaum (eds), *The Offensive Internet: Speech, Privacy, and Reputation* (Harvard Univ Press, 2010); Michael Salter and Chris Bryden 'I Can See You: Harassment and Stalking on the Internet' (2009) 18 Information & Communications Technology Law 99.

⁴⁶Suzor and others (n 42).

⁴⁷The difficulty of defining big data as a technology has led Zuboff to adopt a more social definition of big data which can be useful here, where 'big data' is a part of a new logic of accumulation aimed at predicting and modifying human behavior, see Shoshana Zuboff, 'Big Other: Surveillance Capitalism and the Prospects of an Information Civilization' (2015) 30 Journal of Information Technology 75, 76.

the capacity to *informate* and *automate*.⁴⁸ The developments in big data have led Zuboff to develop the term *surveillance capitalism* to describe the resulting extraction, commodification and control of individuals, establishing ‘a new form of power in which contract and the rule of law are supplanted by the rewards and punishments of a new kind of invisible hand’.⁴⁹ Working in surveillance studies, Lyon has further connected this surveillance capitalism back to established theories of *surveillance as social sorting*, where big data may risk reinforcing existing inequalities and marginalization.⁵⁰ Here, Zuboff’s research into Silicon Valley giants capitalistic power ties in with studies in political science such as Eubanks’ study of data-based discrimination,⁵¹ and in sociology through for example Brayne’s study of algorithms and prediction in policing.⁵² This illustrates the common impacts and logics underpinning the use of big data in a variety of public and private settings, to accumulate data and predict, sort, or modify human behavior.

The connection between big data and automation of decision-making is likely why significant efforts within legal science have been directed towards analyzing the associated data protection issues of automation, especially since the entry into force of the General Data Protection Regulation (GDPR)⁵³ in Europe. This strain of research seems to indicate that the requirements of transparency and the right to explanations of automated decisions under the GDPR are unlikely to offer a complete remedy to algorithmic harms given the limited and unclear scope of such rights.⁵⁴ On the other hand, other data protection principles may offer a better way forward, such as the right to be forgotten and privacy by design.⁵⁵

The shortcomings of the GDPR as a remedy for all the potential ailments of automation have led legal researchers to approach issues relating to the automation of decision-making through either wider or more narrow approaches. The wider approach tends to use the *rule of law* as a collection of fundamental legal principles against which automation and the use of algorithmic governing can be analyzed.⁵⁶ The narrower, more contextually sensitive approach instead situates decision-making in a specific legal setting using, for example, relevant administrative law rules, procedures and principles or specific rights of individuals subject to decisions which are highlighted in particular decision-making circumstances.⁵⁷

⁴⁸*ibid.*

⁴⁹*ibid.* 82.

⁵⁰Didier Bigo, ‘Surveillance Capitalism, Surveillance Culture and Data Politics’ in Didier Bigo, Engin F. Isin and Evelyn Sharon Ruppert (eds), *Data Politics: Worlds, Subjects, Rights* (Routledge, Taylor & Francis Group, 2019).

⁵¹Virginia Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor* (1st edn St Martin’s Press, 2017).

⁵²Sarah Brayne, *Predict and Surveil: Data, Discretion, and the Future of Policing* (Oxford University Press, 2021).

⁵³Regulation (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) [2016] OJ L119/1.

⁵⁴See for an overview, Lilian Edwards and Michael Veale, ‘Slave to the Algorithm? Why a “Right to an Explanation” Is Probably Not the Remedy You Are Looking For’ (2017) 16 *Duke Law & Technology Review* 18, with references.

⁵⁵Edwards and Veale (n 54).

⁵⁶See for examples Monika Zalnieriute, Lyria Bennett Moses and George Williams, ‘The Rule of Law and Automation of Government Decision-Making’ (2019) 82 *The Modern Law Review* 425, who have focused on rule of law implications on automated government decision-making, arguing that converting rule of law values into design specifications that can be understood by system designers, and enforced through regulation, professional standards, contracts, courts, or other mechanisms, represents a formidable technical and legal challenge; and Naarttijärvi (n 39) who uses qualitative legality as a rule of law ideal to analyse implications of algorithms in policing.

⁵⁷See for example Marion Oswald, ‘Algorithm-Assisted Decision-Making in the Public Sector: Framing the Issues Using Administrative Law Rules Governing Discretionary Power’ (2018) 376 *Philosophical Transactions of the Royal Society*

We argue that the need to keep humans in the loop is likely to challenge ambitions of automation in many legal contexts, in particular where the fundamental rights of those subject to decisions may be affected. We also argue that the particular challenges will vary between specific legal settings, and that hybrid decision-making therefore must take different forms, and consequently actualize different types of legal considerations. In the following, we will illustrate this through an analysis of both general rules affecting the automation of decision-making as such, and the specific dynamics existing within our three decision-making contexts. Our analysis will indicate that understanding the resulting hybrid decision-making from a legal point of view will require (to borrow a concept from data science) a *full stack analysis*, combining micro-level analysis of the legal environment surrounding each decision-making context, with wider macro-level considerations of fundamental rights, constitutional principles and rule of law concerns. It will also show that such an analysis will benefit from a socio-technological understanding of the implications algorithmic technologies carry for human agents acting within a hybrid system, as it helps to capture the important interactions between law, technology, decision-makers and the subject of decisions.

3. Facing the legal framework

3.1. The General Data Protection Regulation – a look at the trees

As indicated above, existing legal research in the context of automation of decision-making in Europe has to date tended to place significant focus on the GDPR. This is hardly surprising, as this regulation, like the earlier Data Protection Directive (DPD), contains specific rules relating to automated decisions. This provision in Article 22 GDPR relates, however, only to decision-making processes based *solely* on automated processing (including profiling), which has been given a rather narrow interpretation.⁵⁸ These circumstances do, however, not mark that the GDPR only recognizes the risks involved with fully automated decision-making processes. As we will see, the rationale for Article 22 was based also on the risks involved in hybrid decision-making. Importantly, the GDPR also provides other important and generally applicable safeguards that may impact the implementation of, and underlying data-processing necessary for, hybrid decision-making to various extents. So, while our point is that the GDPR should not be the end-point of any legal analysis of either fully or hybrid decision-making, it is therefore still a useful point of departure regarding automation in the European legal context.

Tracing the rationale of Article 22 GDPR requires a closer look at the background to the DPD, which carried a similar though not identical wording in its Article 15.⁵⁹ In this context, the Commission pointed to an intention to protect ‘the interest of the data subject in participating in the making of decisions which are of importance to him’ and to avoid the ‘data-shadow’ of the individual becoming the sole basis for decisions as

A: Mathematical, Physical and Engineering Sciences 1; Markku Suksi, ‘Administrative Due Process When Using Automated Decision-Making in Public Administration: Some Notes from a Finnish Perspective’ (2021) 29 Artificial Intelligence and Law 87.

⁵⁸Article 29 Data Protection Working Party, ‘Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679’, 8; Lee A Bygrave, ‘Article 22 Automated Individual Decision-Making, Including Profiling’ in Lee A Bygrave, *The EU General Data Protection Regulation (GDPR)* (Oxford University Press, 2020) 530.

⁵⁹Bygrave (n 58) 526.

that would ‘deprive the individual of the capacity to influence decision-making processes’.⁶⁰ In its amended proposal, the Commission seemed to raise issues relating to hybrid decisions in its explanatory memorandum:

The danger of the misuse of data processing in decision-making may become a major problem in future [*sic*]: the result produced by the machine, using more and more sophisticated software, and even expert systems, has an apparently objective and incontrovertible character to which a human decision-maker may attach too much weight, thus abdicating his own responsibilities.⁶¹

The *travaux* thus points to an awareness of the risks of hybrid systems narrowing the autonomy of human decision-makers in a line of reasoning that is echoed in much of the current debate on algorithmic decision-making. As introduced, the proposal and the data protection framework would, however, come to directly address only a smaller subset of this issue by taking aim at decisions taken *solely* by automatic processing strictly applied by the user of a system. As the commission stated later in the explanatory memorandum, ‘[d]ata processing may provide an aid to decision-making, but it cannot be the end of the matter; human judgment must have its place’.⁶²

Automated processes thus fall outside of the field of application of Article 22 when they remain decisional support tools, ‘provided the human decision-maker considers the merits of the result rather than being blindly or automatically steered by the process’.⁶³ The degree of human involvement in the decision to avoid triggering Article 22 has been described by the European Data Protection Board (EDPB) as:

The controller cannot avoid the Article 22 provisions by fabricating human involvement. For example, if someone routinely applies automatically generated profiles to individuals without any actual influence on the result, this would still be a decision based solely on automated processing.

To qualify as human involvement, the controller must ensure that any oversight of the decision is meaningful, rather than just a token gesture. It should be carried out by someone who has the authority and competence to change the decision. As part of the analysis, they should consider all the relevant data.⁶⁴

The EDPB guidelines indicate the need for a contextual analysis, considering assessments such as the avoidance of ‘routine application’ of profiles, the ‘authority’ and ‘competence’ of human decision-makers, as well at what stage this human involvement takes place.⁶⁵ However, as noted by Mendoza and Bygrave, as long as the final decision is subject to such human control of the merits of a decision, ‘the fact that a large or even

⁶⁰European Commission, Proposal for a Council Directive concerning the protection of individuals in relation to the processing of personal data 1990 [COM(90) 314 final – SYN 287 90/C 277/03], Explanatory memorandum, 29, see also Bygrave (n 58) 526.

⁶¹European Commission, Commission of the European Communities amended proposal for a Council Directive on the protection of individuals with regard to the processing of personal data and on the free movement of such data 1992 [COM(92) 422 final – SYN 287], Explanatory memorandum, 26.

⁶²*Ibid.*

⁶³See Bygrave (n 58) 533, basing his analysis on the European Commission’s Amended Proposal of 1992, 26; see also Article 29 Data Protection Working Party (n 58) 20–21; Maja Brkan, ‘Do Algorithms Rule the World? Algorithmic Decision-Making and Data Protection in the Framework of the GDPR and Beyond’ (2019) 27 International Journal of Law and Information Technology 91, 101–02; Lee A Bygrave, ‘Minding the Machine v2.0: The EU General Data Protection Regulation and Automated Decision-Making’ in Karen Yeung & Martin Lodge (ed), *Algorithmic Regulation* (Oxford University Press, 2019) 253.

⁶⁴Article 29 Data Protection Working Party (n 58) 21.

predominant part of the decisional process is automated will not attract the application of Article 22'.⁶⁶

Arguably, the interpretation of Article 22 highlighted so far accounts for some, but not all the concerns of the European Commission in shaping Article 15 of the DPD back in 1992. It is likely that the difficulty of defining a broader area of semi-automated decisions, and the vast area of a potential application of doing so, influenced the relatively limited approach in the EU data protection framework. There is a considerable space between blind and automatic application and 'attaching too much weight' to expert systems as the commission put it in the *travaux*.⁶⁷ The contextual factors mentioned by the EDBP still suggest the need for an analysis of the institutional, practical, and legal environment the human decision-maker operates in. But the potential issue of human decision-makers abdicating their decision-making responsibilities may remain even if they have formal authorization to review and reexamine the results of automated processes. This may, in particular be the case if decision-making processes are based on presumptions of the accuracy of automated processes, or where factors of time, resources or effort favor the acceptance of automated recommendations.

The EDPB, in adopting the guidelines of the Article 29 Data Protection Working Party, has seen fit to distinguish between the rules of the GDPR that applies to 'solely automated individual decision-making, including profiling' on the one hand, and profiling and automated decisions that are 'not solely automated' on the other.⁶⁸ The latter category includes *inter alia* the requirements of purpose limitation, data protection by design and by default, data minimization, proportionality, accuracy, and transparency.⁶⁹ These principles are such that require a contextual analysis as they are dependent on the purposes and procedures of each data processing and the organizational, legal and operational context in which they are carried out. Hybrid decision-making is likely to make up the majority of deployed automation efforts, given the narrow interpretation of fully automated decision-making. The need for contextual analysis of these general principles is likely to carry a significant impact on the extent and shape of hybrid decision-making systems in practice. As these principles are common between private and public decision-making and carry over to the law-enforcement context through their implementation in the Data Protection Law Enforcement Directive,⁷⁰ they are also relevant across a wide field

⁶⁵*ibid.*, 'the controller should identify and record the degree of any human involvement in the decision-making process and at what stage this takes place'.

⁶⁶Isak Mendoza and Lee A Bygrave, 'The Right Not to Be Subject to Automated Decisions Based on Profiling' in Tatiana-Eleni Synodinou and others (eds), *EU Internet Law: Regulation and Enforcement* (Springer International Publishing, 2017) 88.

⁶⁷European Commission (n 61) 26.

⁶⁸Article 29 Data Protection Working Party (n 58) 9.

⁶⁹Article 29 Data Protection Working Party (n 58) 9; See also Lee A Bygrave, 'Minding the Machine v2.0: The EU General Data Protection Regulation and Automated Decision-Making' in Karen Yeung and Martin Lodge (ed), *Algorithmic Regulation* (Oxford University Press, 2019) 260.

⁷⁰Directive (EU) 2016/680 on the protection of natural persons regarding processing of personal data connected with criminal offences or the execution of criminal penalties, and on the free movement of such data. See in particular Article 11 which holds that 'Member States shall provide for a decision based solely on automated processing, including profiling, which produces an adverse legal effect concerning the data subject or significantly affects him or her, to be prohibited unless authorised by Union or Member State law to which the controller is subject and which provides appropriate safeguards for the rights and freedoms of the data subject, at least the right to obtain human intervention on the part of the controller'. Such automated decisions may not, under the second point of the article be based on sensitive categories of data 'unless suitable measures to safeguard the data subject's rights and freedoms and legitimate interests are in place', and may not under the third point of the article result in discrimination based on any of the sensitive categories of data.

of decision-making. Before turning to this more contextual analysis, we will however touch upon a few more general notes on decision-making beyond the field of data protection.

3.2. Beyond data protection – a brief look at the forest

Requirements such as legality, transparency and proportionality exist far beyond data protection, and significantly predate it. They stem from wider ideals of law which are multi-leveled and multi-faceted. On a normative level, it is well established that rule of law principles, jurisprudential ideals, and qualitative legality as elaborated in constitutional and human rights case law establish standards to uphold in terms of clarity and foreseeability of the legal rules as such.⁷¹ These normative requirements have the potential to impact technologically mediated rules clouded in vague or technology-neutral language.⁷² Transparency-related legal requirements also impact the delegation of decision-making authority, for example through requirements to provide clear and defined limits to decision-making authority and establish effective legal safeguards for individuals.⁷³ At the decision-making level, rule of law requirements requiring a material and procedural basis in law for decisions rather than predictions may impact the possibilities of using machine learning algorithms trained on historical data.⁷⁴ The need to provide intelligible and individualized reasons for such decisions may also prevent applying certain machine learning algorithms, barring significant advances in explainable AI.⁷⁵

While there is an interconnection between the GDPR and general principles of law operating outside of the data protection area, it is also worth noting what the GDPR does not capture. The GDPR, by its nature and logic, takes aim at the processing of personal data. It captures a very specific aspect of the relationship between data subjects and data processors and is an attempt to ensure that data subjects retain a degree of autonomy in relation to their data.⁷⁶ The new AI regulation proposed by the European Commission aims to fill in some of the gaps left by the GDPR by taking into account the specific systemic risks and issues that AI brings.⁷⁷ It places a greater focus on the actors developing or implementing AI systems ('providers'), but appears to function primarily preventatively, by focusing on the specific risk of an AI system and the need for specific safeguards to counter those risks, including the need for humans in the loop.⁷⁸

There is of course a lot yet to be done in analyzing the intersection between the GDPR and the future European AI regulation. However, at this early stage, both the GDPR and

⁷¹See Lon L Fuller, *The Morality of Law* (Yale University Press, 1969) 63–64; Joseph Raz, *The Authority of Law: Essays on Law and Morality* (2nd ed Oxford University Press, 2009) 214–19; Venice Commission, 'Report on the Rule of Law' (Venice Commission 2011) 003rev-e.

⁷²See Naarttijärvi (n 39).

⁷³See Geranne Lautenbach, *The Concept of the Rule of Law and the European Court of Human Rights* (Oxford University Press, 2013) 97–101.

⁷⁴See Suksi (n 57) 104, concluding 'If machine-learning [automated decision-making] were used, it is likely that the connection to the law and to the principle of legality would be broken, whereupon the rule of law would turn into the rule of algorithm'.

⁷⁵Cf. Jenna Burrell, 'How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms' (2016) 3 *Big Data & Society* 1, 10; Emre Bayamlioğlu and Ronald Leenes, 'The "Rule of Law" Implications of Data-Driven Decision-Making: A Techno-Regulatory Perspective' (2018) 10 *Law, Innovation and Technology* 295, 306–11.

⁷⁶Cf Orla Lynskey, 'Deconstructing Data Protection: The "Added-Value" of a Right to Data Protection in the EU Legal Order' (2014) 63 *International and Comparative Law Quarterly* 569.

⁷⁷See European Commission (n 4) Title III, Chapter 1.

⁷⁸See European Commission (n 4) Article 14.

the proposed new AI regulation seem to leave gaps in terms of two significant dynamics, the power dynamics between the decision-making authority and individuals' subject to decisions, as well as the legal role and institutional context facing the human in the loop. Both the GDPR and the proposed rules on AI largely fail to capture those aspects of decision-making which relates to power and the exercise of power (whether government power or economic and social power).⁷⁹ Subjects to decisions may have rights, and decision-makers may have duties, that flow from a much wider set of legal frameworks relating to the exercise of power – in particular when individual rights are involved. These frameworks include the EU Charter of Fundamental Rights ('the Charter'), the European Convention of Human Rights ('the ECHR'), and national constitutional and administrative rules and principles. Also, if keeping humans in the loop is so important for the legal rules targeting automation efforts, then understanding the legal role and decision-making context facing this *human-in-the-loop* is key.

4. Context matters – diving into three legal environments

4.1. Introduction

As we have shown above, the legal implications of, and preconditions for, hybrid decision-making are largely contextual. We have also argued that analyzing hybrid decision-making requires an analysis which is mindful not only of the wider legal environment it takes place in, but also the role and responsibilities of human decision-makers. In this section, we will therefore turn to three sample environments, policing, social welfare benefits, and social media moderation, in order to illustrate how the need for a human-in-the-loop may surface in hybrid decision-making contexts. As mentioned at the outset of this article, these environments are currently subject to intense efforts of automation. Illustrating how these efforts have led to different types of hybrid decision-making due to a combination of legal, organizational and socio-technical factors, we will outline some specific concerns that may face the respective human decision-making agents when navigating their interactions with algorithmic agents.

4.2. The officer in the loop – algorithmic prediction and risk assessment in policing

The automation and datafication of law enforcement operations seems to be a continuous process which in some ways have defined the last decades of police innovation. As in many other fields, the advances in machine learning and algorithmic prediction have been driving forces in the application of various types of profiling and processing tools based on big data and statistical analysis. As pointed out by Brayne, these developments – while perhaps accelerated in recent years – are extensions of a development towards more data-driven policing taking place since at least the 1960s in the US.⁸⁰ In Europe, and more specifically in Scandinavia, these developments begun in the late 1960s and early 1970s through the increased influence of

⁷⁹Arguably, data protection can however address some *informational* power asymmetries, see Lynskey (n 76).

⁸⁰Brayne (n 52) 20–21.

criminology research on policing,⁸¹ but arguably accelerating through concepts such as problem-oriented policing, community policing and evidence-based policing from the 1990s and onwards.⁸² So while the ‘newness’ of data-driven policing should not be exaggerated, the potential impact of such policing should not be underestimated either as the scale and technological underpinnings of data-driven policing have developed as well,⁸³ not the least through the influence of private actors.⁸⁴

Internationally, there are abundant examples of algorithmic automation taking on an expanded role in policing. In recent years, significant attention has been focused on automated facial recognition (AFR), which has been attributed to the availability of image databases and technological advances in relevant algorithms and validation techniques.⁸⁵ As a technology, AFR relies on machine-learning algorithms and biometric analysis to match subjects in still or video footage with subjects in other footage, such as in police photo databases. This can be used retroactively, to attempt to compare faces from surveillance tapes with photos of known subjects in police databases for instance – thereby automating a time-consuming human comparison, which in certain contexts, such as child abuse investigations, can also be traumatizing for human investigators.⁸⁶ More proactive uses exist as well, with ‘live facial recognition’ (LFR) having the potential to identify persons of interest in live surveillance footage of streets,⁸⁷ or to assist in biometrically identifying individuals at border checkpoints,⁸⁸ thereby sorting them for potential further scrutiny or police intervention. Here, the automation provides a trigger for attention or intervention that will feed into the decision-making discretion of human officers.⁸⁹ The ‘human in the loop’ aspect of such systems has been seen as an important legal safeguard when justifying interventions in this context.⁹⁰ Meanwhile, researchers have criticized the lack of awareness of the influence that algorithmic systems may have on

⁸¹ Most clearly evident in the creation of different Crime Prevention Councils in Sweden and Denmark in the early 1970s, which later influenced similar constructions in Norway and Finland. The Swedish Crime Prevention Council began a research committee under the Ministry of Justice in Sweden and later as an independent authority tasked with following crime statistics, make prognoses and inform decision making, see *Brottsförebyggande kunskapsutveckling 2004* [SOU 2004:18], 29–34.

⁸² Swedish Crime Prevention Council, ‘Hur – Var – Närpolis – En Granskning Av Närpolisreformen’ (2001) 2001:5, 23. Swedish Crime Prevention Council, ‘Svensk Polis I Förändring – En Granskning Av Närpolisreformen’ (1999), 9.

⁸³ Elizabeth E Joh, ‘Policing the Smart City’ (2019) 15 *International Journal of Law in Context* 177, 178.

⁸⁴ See Brayne (n 52) 19.

⁸⁵ See Paramjit Kaur and others, ‘Facial-Recognition Algorithms: A Literature Review’ (2020) 60 *Medicine, Science and the Law* 131.

⁸⁶ Commercially available law enforcement software offers this type of facial recognition solutions, see for example Cellebrite pathfinder, see Ariel Watson, ‘Reduce Trauma for Child Abuse Investigators Using Digital Intelligence’ (*Cellebrite*, 24 July 2018) <<https://www.cellebrite.com/en/reduce-trauma-for-child-abuse-investigators-using-digital-intelligence/>>. The Swedish police authority, having received approval from the Swedish data protection agency, have begun using automated facial recognition to cross-reference suspects in visual evidence with existing police databases, see Swedish Police Authority, ‘Ansiktsgenkänning Får Användas För Att Utreda Brott’ (*polisen.se*, 24 October 2019) <<https://polisen.se/aktuellt/pressmeddelanden/2019/oktober/ansiktsgenkanning-far-anvandas-for-att-utreda-brott/>>, and Integritetsskyddsmyndigheten, ‘Förhandssamråd om Polismyndighetens planerade användning av programvara för ansiktsgenkänning mot signalementsregistret’ (2019) Dnr DI-2019-10508.

⁸⁷ Such ‘live’ automated facial recognition has been tested by the Metropolitan Police Service (MPS) in the UK, Pete Fussey and Murray Daragh, ‘Independent Report on the London Metropolitan Police Service’s Trial of Live Facial Recognition Technology’ (University of Essex 2019). The independent report of this test highlighted (p. 5) that ‘it is highly possible that the LFR trial process adopted by the MPS would be held unlawful if challenged before the courts’.

⁸⁸ Testing of such systems are under way in Sweden, see Swedish Government, *Behandling av känsliga personuppgifter i testverksamhet enligt utlänningsdatalagen 2020* [Prop. 2020/21:5].

⁸⁹ See Kyriakos N Kotsoglou and Marion Oswald, ‘The Long Arm of the Algorithm? Automated Facial Recognition as Evidence and Trigger for Police Intervention’ (2020) 2 *Forensic Science International: Synergy* 86.

⁹⁰ *R (Bridges) v Chief Constable of the South Wales Police* [2019] High Court of Justice, Queen’s bench division EWHC 2341 (Admin).

human decision-makers in this context.⁹¹ Given the ‘live’ aspect of LFR, human decisions in that context are likely to be more time-pressed than with retroactive AFR, as possible interventions will need to take place when subjects are passing by specific locations.

Automated analysis of large data sets also forms the basis for profiling,⁹² which can be applied for a variety of purposes. Automated statistical analysis of communications data can provide the police with sociograms and potentially illustrate the chain-of-command within a network of individuals in intelligence or crime investigation efforts.⁹³ This information is likely to mainly inform human analysis and decisions, but carries the implicit potential to act as a trigger for future interventions. Similarly, analysis of passenger name registration (PNR) data from airlines forms the basis for risk analysis selecting passengers for further border checks or custom controls.⁹⁴ Unlike LFR systems identifying individuals already of interest to police authorities, PNR records are used to identify unknown individuals who, based on their travel records, could warrant further scrutiny.⁹⁵ Current EU-legislation requires human review of automated decisions,⁹⁶ ‘to ensure that no decisions having an adverse effect on an individual (such as being subject to further checks on arrival or departure) are taken without human intervention’.⁹⁷ Of course, there are also the ambitions and applications of different types of predictive policing systems, attempting to indicate what locations, or increasingly, which persons, could be involved in future crime.⁹⁸

While the use-cases outlined above are among the commonly discussed uses of algorithms in the policing context, automation efforts can take more subtle forms as well. As software systems connect previously discrete databases or integrate private sources of data into police frontends, they implicitly automate what was previously seen as discrete investigatory processes and measures.⁹⁹ In doing so, such integrations present

⁹¹See Kotsoglou and Oswald (n 89) 88.

⁹²Profiling is defined in article 3 (4) of Directive (EU) 2016/680 as

any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person’s performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements.

This mirrors the definition in the GDPR article 4 (4).

⁹³See Cellebrite, ‘Cellebrite Pathfinder’ <<https://www.cellebrite.com/en/pathfinder/>>, a software allowing automated analysis assistance for both communication patterns and location data.

⁹⁴See for example the measures implemented through the PNR-directive, Directive (EU) 2016/681 of the European Parliament and of the Council of 27 April 2016 on the use of passenger name record (PNR) data for the prevention, detection, investigation and prosecution of terrorist offences and serious crime, and the system which was the purpose behind the EU-Canada PNR exchange programme scrutinized by the CJEU in *Opinion 1/15 of the Court* [2017] Court of Justice of the European Union (Grand Chamber) ECLI:EU:C:2017:592.

⁹⁵The EU Commission highlights how ‘The analysis of PNR data can provide the authorities with important elements from a criminal intelligence point of view, allowing them to detect suspicious travel patterns and identify associates of criminals and terrorists, in particular those previously unknown to law enforcement’. See European Commission – Migration and Home Affairs, ‘Passenger Name Record’ (What we do) <https://ec.europa.eu/home-affairs/what-we-do/policies/law-enforcement-cooperation/information-exchange/pnr_en> accessed 10 June 2021.

⁹⁶See Article 6.5 of the PNR-directive.

⁹⁷Commission Staff Working Document accompanying the ‘Report from the Commission to the European Parliament and the Council on the Review of Directive 2016/681 on the Use of Passenger Name Record (PNR) Data for the Prevention, Detection, Investigation and Prosecution of Terrorist Offences and Serious Crime’ COM(2020) 305 final, p. 20.

⁹⁸See for example Vicki Sentas and Camilla Pandolfini, ‘Policing Young People in NSW: A Study of the Suspect Targeting Management Plan.’ (Youth Justice Coalition NSW 2017); See also Dylan J Fitzpatrick, Wilpen L Gorr and Daniel B Neill, ‘Keeping Score: Predictive Analytics in Policing’ (2019) 2 Annual Review of Criminology 473, 482, outlining the Chicago Police Department’s ‘strategic subject list’.

⁹⁹Cf. Brayne (n 52).

a more holistic (and privacy invasive) picture and analysis of individuals subject to searches.¹⁰⁰ All the while, decisions made in the design and implementation of the system will mediate the human perception of the data, sedimenting earlier decisions in the organization.¹⁰¹

Understanding the implications for these hybrid systems for human decision makers requires awareness of the particulars of policing as a field of public power. It is a field which is heterogenous in the sense that it covers, *inter alia*, a multitude of loosely regulated service functions; wide and often discretionary powers of maintaining public order; as well as the more strictly regulated powers and procedures relating to criminal law in formal investigations.¹⁰² On top of those functions, there is the increasing role played by police intelligence gathering and crime prevention efforts which furthers a focus on *risk* and considerations of future dangers.¹⁰³ Throughout these functions exist a considerable degree of discretion, which in some ways have come to define academic discussions on policing.¹⁰⁴ Even in areas of strictly regulated powers, such as in the criminal procedure, discretion can come into play at early stages, for example through what Joh has labeled 'surveillance discretion', i.e. the decision about *who* to focus police attention on.¹⁰⁵ Discretion can also be expressed through decisions of *non-intervention*. Such decisions to not apply the full extent of police powers are low-visibility decisions,¹⁰⁶ but may – as early and influential research has pointed out – be expressions of humanity, common-sense, or the exercise of the spirit, rather than the letter of the law.¹⁰⁷ On the other hand, within this discretionary area exist risks for less desirable phenomena, such as corruption, discrimination and resistance to management and leadership.¹⁰⁸ Consequently, the implementation of certain technological innovations in policing has also been aimed at limiting the discretion of individual officers and enabling closer management.¹⁰⁹

Compounding the legal complexity of the field, issues and data can move rather fluidly between policing contexts subject to differing levels of regulation. An interaction with the public may begin as a public order measure or traffic measure, transition into formal investigatory measures in the criminal law context, while simultaneously being fed into

¹⁰⁰In this sense, the integration of discrete data sources about the individual plays into what Ericsson and Haggerty have labeled the 'surveillance assemblage', see Richard Victor Ericson and Kevin D Haggerty, 'The Surveillant Assemblage' (2000) 51 *The British Journal of Sociology* 605.

¹⁰¹Niculescu-Dincă (n 21) 468.

¹⁰²Lena Landström and Markus Naarttijärvi, 'Gränser För Polisiär Innovation – Rättssäkerhet, Enhetlighet Och Demokratisk Legitimitet' (2020) 107 *Nordisk Tidsskrift for Kriminalvidenskab* 268, 270.

¹⁰³Cf. Richard Victor Ericson and Kevin D Haggerty, *Policing the Risk Society* (University of Toronto Press, 1997); Markus Naarttijärvi, *För din och andras säkerhet Konstitutionella proportionalitetskrav och Säkerhetspolisens preventiva tvångsmedel*. (Iustus förlag 2013) 494–504; David L Carter and Jeremy G Carter, 'Intelligence-Led Policing: Conceptual and Functional Considerations for Public Policy' (2009) 20 *Criminal Justice Policy Review* 310.

¹⁰⁴Michael Rowe, 'Rendering Visible the Invisible: Police Discretion, Professionalism and Decision-Making' (2007) 17 *Policing and Society* 279; Mireille Hildebrandt, 'Proactive Forensic Profiling: Proactive Criminalization?' in R Anthony Duff and others (eds), *The Boundaries of the Criminal Law* (Oxford University Press, 2010).

¹⁰⁵Elizabeth E Joh, 'The New Surveillance Discretion: Automated Suspicion, Big Data, and Policing' (2016) 10 *Harvard Law and Policy Review* 15, 16.

¹⁰⁶Joseph Goldstein, 'Police Discretion Not to Invoke the Criminal Process; Low-Visibility Decisions in the Administration of Justice' (1959) 69 *Yale Law Journal* 543.

¹⁰⁷Herman Goldstein, 'Police Discretion: The Ideal versus the Real' (1963) 23 *Public Administration Review* 140, 143.

¹⁰⁸*Ibid.* 144–45.

¹⁰⁹See Brayne (n 52); Sarah Brayne and Angèle Christin, 'Technologies of Crime Prediction: The Reception of Algorithms in Policing and Criminal Courts' [2020] *Social Problems* preprint; Rowe M, 'Rendering Visible the Invisible: Police Discretion, Professionalism and Decision-Making' (2007) 17 *Policing and Society* 279.

an intelligence pipeline.¹¹⁰ Furthermore, specific events or individuals can, for instance, simultaneously be subject to both intelligence activities and formal investigations with information moving between activities, databases, teams – challenging the separation between the two law-enforcement contexts. To be sure, in certain areas and contexts policing can reflect administration of cases within more common public administrative bodies, with reports coming in and being investigated, ending with a formal decision. In comparison however, policing is often significantly more open-ended and officers tasked not only with responding to incoming requests or reports, but to proactively discover and prevent crime, maintaining public order in the face of ever-changing events and circumstances, and also responding to various and frequent political prioritizations and public concerns of the day.¹¹¹

Previous research in the US and UK context has indicated that current legal safeguards, primarily taking aim at the criminal process, may be ill equipped to address the particular issues of discretion, prediction, and biases that automation may imply.¹¹² As implementation of these systems tend to influence primarily pre-investigatory stages of policing, such as in intelligence and preventive functions, they are rarely assessed by courts.¹¹³ This disconnect between traditional safeguards of criminal procedures and systems of algorithmic automation has led to the suggestion of relying more on administrative law to address these issues.¹¹⁴ Meanwhile, legal scholars working in the European context have highlighted the potential of data protection and non-discrimination frameworks,¹¹⁵ both of which have strong foundations in the EU-law context. In the recent proposal for AI-regulation in the EU the risks involved in algorithmic policing are broadly acknowledged, as well as the need to keep humans in the loop.¹¹⁶ The Data Protection Law Enforcement Directive also emphasizes the need for human intervention and oversight of automated decisions, in particular where profiling or the underlying data use involves sensitive categories of data.¹¹⁷ Less attention has however been given to the actual decision-making situation facing that human in the loop in a policing context, and the complex hybrid environment in which they are to act.

In the end, capturing the complexity of the interaction between the law and hybrid decision-making in policing is likely to require a broader approach which can capture both the underlying data use, the legal rules relevant for human (and algorithmic) decision-making, and the implicit effects on both the balance of power within the state and the fundamental rights of individuals subject to police decisions. This highlights the need for a combination of constitutional and human rights law, administrative law

¹¹⁰This can be highlighted by how the influential concept of Intelligence-led policing is 'envisioned as a proactive practice driven by information sharing and analysis integrated across organizational functions', see Jeremy G Carter, 'Institutional Pressures and Isomorphism: The Impact on Intelligence-Led Policing Adoption' (2016) 19 *Police Quarterly* 435. See also, regarding the fluidity of intelligence operations, Carter and Carter (n 103).

¹¹¹See Lena Landström, Niklas Eklund and Markus Naarttijärvi, 'Legal Limits to Prioritisation in Policing – Challenging the Impact of Centralisation' (2020) 30 *Policing and Society* 1061.

¹¹²See Brayne (n 52) 118–35; Andrew D Selbst, 'Disparate Impact in Big Data Policing' (2018) 52 *Georgia Law Review* 109, 144–54; Danielle Keats Citron, 'Technological Due Process' (2008) 85 *Washington University Law Review* 1249.

¹¹³See Brayne (n 52) 118–25; Naarttijärvi (n 39).

¹¹⁴Oswald (n 57).

¹¹⁵Frederik J Zuiderveen Borgesius, 'Strengthening Legal Protection against Discrimination by Algorithms and Artificial Intelligence' (2020) 24 *The International Journal of Human Rights* 1572.

¹¹⁶European Commission, (n 4) recitals 38, 48; article 14.

¹¹⁷See Directive (EU) 2016/680 on the protection of natural persons regarding the processing of personal data connected with criminal offences or the execution of criminal penalties, and on the free movement of such data, Article 11.

and data protection. However, as noted by Brayne, capturing the actual impact of algorithmic systems will require moving from conceptual studies on the level of legal principles and future concerns, to case-studies and empirical work which captures the day-to-day interactions between human and algorithmic decision-making agents.¹¹⁸ In other words, it requires legal scholars to engage with the humans inside the loop.

4.3. The administrator in the loop – the implementation of automated decision support systems in Swedish public sickness benefits administration

The reallocation of public resources through the issuance of various social benefits is an important but sizeable and costly public task in most countries.¹¹⁹ Efforts of effectivization through automation have therefore been initiated or implemented in many different jurisdictions. This is also true in Sweden, which though a combination of its sizeable and comparatively generous social welfare system and pioneering ambitions of automation have come to serve as a frequent case-study in this administrative context.¹²⁰ An interesting, and less explored example, can be found in The Swedish Social Insurance Agency's (SSA) efforts to deploy automated decision support systems in its sickness benefits administration. This example will here be used to illustrate how the substantive arrangement of specific legal provisions may affect the prospect to fully automate decision making processes.

The SSA handles many types of benefits, and has increasingly been deploying different types of fully or semi-automated decision-making systems to help case management since the 1970s.¹²¹ But even if sickness benefits are among the costliest within the administration, their complexities have made them less suitable as a front-runner for the SSA's automation efforts.¹²² This standpoint has, however, begun to change in the last few years, and today the SSA even considers sickness benefits to be of highest priority for future automation efforts within the agency.¹²³

As already indicated, the legal conditions for eligibility to sickness benefits are not solely based on objectively verifiable facts, and therefore include complex legal assessments (including evaluation of evidence). Already following from the basic conditions – the applicant (covered by Swedish social security) must have an illness or injury that reduces his or her ability to work by at least 25 percent. This means that an assessment must be made on whether there is illness or injury and whether there is reduced ability to work, as well as if there is a causal link between these two conditions. Also, these basic conditions are relational in the sense that their contents

¹¹⁸Brayne (n 52) 119.

¹¹⁹There are many different types of benefits, which are funded and administered in various ways. Regularly responsible for this type of administration are, however, public authorities.

¹²⁰E.g. Ranerup and Henriksen (n 26). See also the government assignment to a number of Swedish government agencies to increase the capabilities to use AI, Swedish Government, Uppdrag att främja offentlig förvaltnings förmåga att använda artificiell intelligens 2021 [I2021/01825].

¹²¹The Swedish Social Insurance Inspectorate, 'Individuell Eller Standardiserad Socialförsäkring – En Diskussion För Mer Rättssäker Handläggning' (2015) Arbetsrapport 2015:3 <<https://isf.se/download/18.6ce5045216a58f96d2f56007/1565330432377/Individuell%20eller%20standardiserad%20socialförsäkring%202015-3ar.pdf>>, p. 16.

¹²²Försäkringskassan, 'Socialförsäkringen i Siffror 2020' (2020) <<https://forsakringskassan.se/wps/wcm/connect/dae19b87-ace6-4cda-a577-05af925b0317/socialforsakringen-i-siffror-2020.pdf?MOD=AJPERES&CVID=>>>, 15.

¹²³Försäkringskassan, 'Slutrapport. Förbättrat Beslutsstöd – En Del Av Försäkringskassans Digitala Agenda' (2019).

are ultimately linked to the conditions of the labor market and medical science developments.¹²⁴ Assessing eligibility to sickness benefits, especially in more complicated or long-term cases, may thus involve examinations of comprehensive investigation data (which in addition to the claimant's application and submitted medical certificates may include other information, such as other in-depth medical documentation and information from the employer *etcetera*). And above all, as illness or injury may affect different claimants in different ways (including their work ability) the assessment must be made on an *individual basis*.

The SSA has for a long time deployed various types of (analogue) decision support manuals for commonly recurring diagnoses known to cause absence from work due to sickness. The aim has been to facilitate faster and more equal case management. In addition to internal guidelines on case management, especially the guidelines issued by the Swedish National Board of Health and Welfare¹²⁵ which contain recommendations on sick leave periods for certain diagnoses, have impacted the SSA's practises. As these manuals have become increasingly detailed, they have been debated from a legal perspective, as the added detail might mean that substantial insets are made on how the law is being applied. 'Explication' therefore raises concerns about democratic accountability and how compliance will be safeguarded.¹²⁶ Notably, these concerns could be extended to automation of sickness benefits, as detail is an enabler of automation. It is also evident that these detailed manuals have been used by the SSA to facilitate the development and deployment of automated decision support systems, and is part of the agency's digitalization strategy.¹²⁷

The SSA's efforts to automate Swedish sickness benefit administration has so far been largely influenced by considerations on whether/when or not a human, or rather an administrator, need to be 'in the loop' to guarantee that individual assessments are being made when this is required. In 2016, for example, the SSA stopped its plans to *fully* automate issuances of sickness benefits in cases of simpler nature. The decision was made after recommendations from the agency's own legal unit after an internal audit, and the main reason referenced the legal requirement to make individual assessments (as it was considered to limit the possible extent of lawful automation on sickness benefits).¹²⁸ The legal unit's reasoning was also influenced by the Swedish E-delegation's report from 2014 'Automated decisions – fewer rules provide clearer regulation', which had argued that automation can *only* be relevant for decisions based on such 'hard' criteria that can unequivocally be translated to program code.¹²⁹ Rather than fully terminated, the project was, however, restructured and re-aimed at developing a *partially* automated decision-making support system instead. Thus, in September 2017, a system designed for handling select types of sickness benefit cases, generally considered of simpler nature from a case management perspective, was introduced. Initially all

¹²⁴Ch 27 § 25, 45–49 The Swedish Social Insurance Code [*Socialförsäkringsbalken 2010:110*].

¹²⁵[*Socialstyrelsen*].

¹²⁶Lotta Vahlne Westerhäll, Stefan Thorpenberg and Magnus Jonasson, *Läkarintyget i sjukförsäkringsprocessen: styrning, legitimitet och bevisning* (Santérus Förlag, 2009); Ruth Mannelqvist and Lena Enqvist, 'Myndighetsnormering Eller När Rätt Blir Orätt' (2013) 2013/14 Juridisk tidskrift 324.

¹²⁷Försäkringskassan, 'Försäkringskassans Digitala Agenda' (2017) dnr 052606-2017; Försäkringskassan (n 123).

¹²⁸Försäkringskassan, 'Sjukpenningärenden Med Förenklade Läkarintyg' (2016) Rättslig kvalitetsuppföljning 2016:3, 5.

¹²⁹*ibid* 17, citing Swedish Government Official Inquiry, *Automatiserade beslut – färre regler ger tydligare reglering* 2014 [SOU 2014:75], 16 and 40 f.

applications regarding diagnoses statistically likely to render case closure within 60 days in at least 70 percent of cases were included. Certain diagnostic groups of more ambiguous nature, such as mental diagnoses, or cases where the cause of sick leave was related to manifestations of symptoms or factors of importance to the state of health and contact with healthcare, were left out. Yet, already in October 2018 the number of included diagnoses was expanded. Since then, diagnoses that are statistically likely to render case closure closed within 80 days in at least 50 percent of cases, as well as mental diagnoses, are included.¹³⁰

The automation routine is applied to all new (select) cases where the application has been made via the SSA's web application form. It is created around so-called rule-based algorithms, and carried out through checks (of the application) against a number of rules which have been translated and pre-programmed into code. If any of the checked conditions are not met, the automatic processing is interrupted and the case is transferred to manual handling. An administrator then receives an automatically generated documentation on why it was interrupted, intended to inform the continued manual handling. What information this documentation contains depends on how many conditions the system checked until that point. Importantly, and distinguishing for a decision support system, *not all* steps of the process are automated. Manual assessments are always made by an administrator on whether the claimant's working ability is reduced as well as on the final decision on eligibility. In this respect, the case is handled in the same way as those subject to fully manual handling. Responsible are, however, only specially appointed administrators.¹³¹ In an internal audit, the SSA did not direct overall criticism to the automated decision support process as a whole, but found that it was coupled with various problems. Importantly, insufficient documentation about what information the system had collected and on how it had decided whether legal conditions had been met, was revealed. The report also expressed that there was 'potential for improvement' on how the individual work ability was investigated and assessed. The case investigations were found to be insufficient in as much as 67 percent of the cases. Out of these, the internal auditors deemed that a correct decision still had been made in 76 percent despite these shortcomings, but also that only about one-fourth of the cases were both sufficiently investigated and correctly assessed.¹³² The investigative insufficiencies included, among other things, whether the individual was covered by the insurance at all, when the period of illness started and whether there were previous sickness-periods of significance, as well as whether the employer had opportunities to offer other temporary work which the claimant was able to perform.¹³³ These findings led to several internal proposals for change, which have not yet been evaluated.

As seen, an 'administrator in the loop' has, clearly, been considered important to guarantee that an individual assessment of the illness or injury's impact on work ability is made. And this has affected what aspects of the administrative procedure that so far have been subject to automation. Interestingly, the SSA is also currently working to incorporate more advanced technology through an AI-based decision

¹³⁰Försäkringskassan, 'Sjukpenningärenden Som Handläggs Delvis Automatiserat' (2019) Rättslig kvalitetsuppföljning 2019:3, 12.

¹³¹Försäkringskassan (n 130) 7, 11.

¹³²Ibid 26 f.

¹³³Ibid 32 ff.

support system, intended to supplement the already deployed rule-based system described above. This new AI support system is developed and tested through the SSA's ongoing pilot project 'Skosa'. There is not yet much information available, but the aim is that AI will assist administrators by supporting their assessment on individual work ability – by letting it analyze medical certificates in relation to guidelines from the Swedish National Board of Health and Welfare¹³⁴ and the World Health Organizations' various classifications for health conditions.¹³⁵ In other words, this new decision support system is aimed at supporting (but not executing) precisely those aspects of the legal assessment that must be individually made (and therefore considered particularly difficult to automate).

The development above shows that there is a fairly strong driving force within the SSA to increase the extent to which the administration of sickness benefits should be automated. Its latest pilot programme 'Skosa' demonstrates that the ambition goes as far as to combine rule-based decision support systems with supplementing machine-learning support systems. This is noteworthy, especially considering that the latter type of system is not limited to pre-programmed code. To what extent it will be generally deployed in sick leave cases probably hinges on a successful turnout of the pilot. Yet, there clearly is a wish to utilize AI and machine learning within this type of administration. In sum, all the described developments serve to demonstrate that the question of whether or when to keep an 'administrator in the loop' in (semi-)automated decision-making processes is just the first step towards ensuring that individual assessments can be made when required by law. Even if administrators are allowed to depart from recommendations made by a decision support system, they will probably to a greater extent meet the specific cases as a 'set table', where large parts of the investigation may have already been executed – and collected data maybe even evaluated to some extent. As the administrator (in the Swedish sickness benefit context) must consider that illness and injury affect claimants and their work ability in different ways, a balanced examination of all relevant information available is imperative to make the assessment individual. And although a specific example, it spotlights that legal requirements of 'contextualisation' may not only stem from higher order principles of constitutional, administrative or human rights character – but may also result from specific lower-order provisions. In such cases, ensuring that automated decision support systems do not supplant individual assessments, may require administrators to play a particularly active role in assessing the outputs and recommendations of such systems. It is therefore not only of interest what data the various automated processes can examine and collect in the case. Of interest is also what type of information is being omitted along the way, before the case reaches the administrator. There is else a risk that recommendations made by the system fetter the administrator in a way that ultimately risks distancing case management from the application of the law. As an example, a lack of documentation about what information the system has collected and how it has decided whether the conditions were met can result in an information deficit for the administrator when taking over a case at a certain stage. That these circumstances can call the case administrator's role as guarantor

¹³⁴[Socialstyrelsen].

¹³⁵Delegationen för korrekta utbetalningar från välfärdssystemen, 'Digitalisering Och AI För Korrekta Utbetalningar Från Välfärdssystemen' (2019) Rapport 5, 49; Karin Lindström, 'Försäkringskassan Laddar För AI-Stöd till Handläggarna' [2019] *Computer Sweden* <<https://computersweden.idg.se/2.2683/1.720144/forsakringskassan-ai-stod>>.

of individual legal assessments into question is therefore an important aspect to consider in the development and deployment of such technology in case administration.¹³⁶

4.4. The moderator in the loop – staying nuanced in a tsunami of content

Another context where an increased use of AI has been proven important is in the detection and moderation of illegal and offensive material on social media platforms and news comment sections. Moderation can be done in various ways, and could – roughly – be divided into three different categories; *human to human* (a person reporting material that is later reviewed by a moderator), *fully automated* (material being detected and instantly removed by an AI)¹³⁷ and *semi-automated*. The latter is of course the focus of our study. Semi-automated moderation combines content flagged by an AI, such as hate speech, terrorism propaganda, disinformation or nudity, with a human assessment of the context surrounding the flagged material, and decides if the content should be removed or not.

The fact that online platforms now use AI to detect hateful or abusive information on their platforms is mirrored by the increase in regulation regarding content moderation over the last few years, in attempts to counter online abuse such as hate speech or terrorism propaganda. One such attempt is the ‘EU Code of Conduct’, that is directed at the largest social media platforms, and that was signed in 2016 by Facebook, Microsoft, Google (incl. Youtube) and Twitter, later to be joined by other major platforms such as Instagram and Snapchat. By signing the Code of Conduct the companies have agreed to implement rules and standards on their platforms prohibiting the occurrence of racist and xenophobic hate speech, and to establish both teams and systems for reviewing content reported as violations of the set rules and standards.¹³⁸ The companies have agreed to review (most) of the reported hate speech of that nature within 24 hours, and to remove or prevent access to said material if necessary.¹³⁹ This process must also take into account the protection of free speech.¹⁴⁰ Compliance with the Code of Conduct is regularly checked by organizations located in different EU-countries, where the organizations send requests for removal of material and track the response time and whether the content is removed or not.¹⁴¹ The companies meet the criteria for speed as well as level of removal. The evaluation of the Code of Conduct suggests that 70–80 percent of reported material have been removed, which is deemed reasonable since not all reported material is illegal. Direct incitements of violence towards certain individuals or

¹³⁶In the Swedish constitutional and administrative context, administrators are personally responsible for their decision-making in individual cases. This specific national setting does not, however, obscure the merits of our observations regarding the role that administrators may play in ensuring that decisions are based on contextual assessments.

¹³⁷See Greyson K Young, ‘How Much Is Too Much: The Difficulties of Social Media Content Moderation’ [2021] Information & Communications Technology Law 1, 4, regarding material flagged by users and then removed by moderators, or such potentially problematic material flagged by an algorithm and sometimes immediately removed. The use of algorithmic detection is said to be used more frequently by larger and more resourceful companies, see Adriana Stephan, ‘Comparing Platform Hate Speech Policies: Reddit’s Inevitable Evolution’ (Freedman Spogli Institute for International Studies, 8 July 2020) <<https://fsi.stanford.edu/news/reddit-hate-speech>>.

¹³⁸EU Code of Conduct on Countering Illegal Hate Speech Online’ <https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en>.

¹³⁹ibid 2.

¹⁴⁰ibid 1.

¹⁴¹ibid 3.

groups have been removed to a larger degree than defamatory statements, pictures and discussions about such individuals or groups.¹⁴² Such distinctions could surely be part of a balancing between the need for removal of hateful and violent messages, versus not removing material that could be protected under international instruments and national laws on free speech.¹⁴³

On a national level, returning to the Swedish example, regulatory measures have been taken to counter illegal content in comment sections, targeting not only people who post illegal content, making them subject to legal action, but also those individuals responsible for providing the comment section, if they fail to monitor and delete unlawful content when necessary.¹⁴⁴ Under Swedish law, legal responsibility for what is posted in comment sections can thus be imposed on the person posting (through criminal provisions against, for instance, hate speech), and a person providing a forum on – for instance – Facebook. While the law assigning responsibility on forum providers has rarely been used, it was recently applied in a notable judgement from a Swedish court of appeals in late 2020. The case showed that administrators of Facebook groups can be held responsible under Swedish law for not removing material written by others that constitutes obvious hate speech.¹⁴⁵ This responsibility is limited to certain cases of *obviously* illegal material (such as child pornography, threats or hate speech) that even non-legal professionals should be able to detect. Worth noting is also that the responsibility to monitor must only be reasonable in relation to the size and type of forum. In large groups a notice and take down-system can be enough to fulfill the responsibility, if the administrator acts on others' notifications and also monitors to some extent. This is a way to protect individuals and societal values in a balanced way in relation to freedom of expression, since there is a risk that fewer people dare administer and facilitate discussion forums if the legal responsibilities are disproportionate.¹⁴⁶

It is interesting to note that, at least within this Swedish example, the responsibilities to monitor, detect and delete hateful or abusive material decrease when smaller actors, like individuals, are responsible for it. Reasonably, they do not have the same possibilities to monitor, or use automated tools for monitoring, as larger actors. And, as mentioned, more attention has been paid to the large social media platforms, and their responsibilities to moderate content, or at least remove certain content when necessary. Beside the EU Code of Conduct, the reasonably newly adopted EU regulation on addressing the dissemination of terrorist content online is also worth mentioning.¹⁴⁷ The regulation imposes a responsibility for (social) media providers to remove or disable material flagged as terrorism propaganda by a competent authority, within one hour. If the provider has knowledge that their platform is exposed to terrorism propaganda, they must also take actions to prevent

¹⁴²See European Commission, Directorate General for Justice and Consumers, '5th Evaluation of the Code of Conduct' (European Commission 2020) <https://ec.europa.eu/info/sites/default/files/codeofconduct_2020_factsheet_12.pdf>

¹⁴³This is also noted by the European Commission in their comments on the evaluation of the Code of Conduct, see European Commission, 'Questions and Answers on The Code of Conduct on Countering Illegal Hate Speech Online' (22 June 2020) <https://ec.europa.eu/commission/presscorner/detail/en/qanda_20_1135>.

¹⁴⁴This is regulated in Lag (1998:112) om ansvar för elektroniska anslagstavlor (also known as 'the BBS-act' in Sweden, where BBS stands for *bulletin board system*).

¹⁴⁵[2020] Svea Hovrätt B 8432-19.

¹⁴⁶Swedish Government, Ansvar för elektroniska anslagstavlor 1998 [Prop. 1997/98:15], 17; Swedish Government, Ett starkt straffrättsligt skydd för den personliga integriteten 2017 [Prop. 2016/17:222], 77–79.

¹⁴⁷Regulation (EU) 2021/784 of the European Parliament and of the Council, of 29 April 2021 on addressing the dissemination of terrorist content online.

this. Some form of moderation is therefore likely necessary, but there is no stated obligation to use moderation, or use automated tools for this. The social media provider can choose what measures to take, as long as the requirements of the regulation are met. Beyond this limited type of content, the general legal responsibilities of social media platform providers, like Facebook, remain quite uncertain however. One notable example of national legislation specifically targeting such providers can be found in the German *Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken* (*Netzwerkdurchsetzungsgesetz*, NetzDG, or eng. The Network Enforcement Act). This law is a clear step towards holding social media platforms accountable when not removing clearly illegal content within 24 hours (or in complex cases, a week). The enactment of The Network Enforcement Act has not been without controversy, mainly concerning issues as those mentioned above – if demands for quick and effective removal will risk leading to censorship and the infringement of free speech.¹⁴⁸

As mentioned, the protection of free speech is a vital aspect when moderating comment sections of news platforms, as well as the balancing between free speech (and the protection of an open public debate) and other interests, such as the protection of privacy of others. The responsibility of news platforms regarding third-party comments (made by individuals on their sites) has been discussed by the European Court of Human rights. Both in the case of *Delfi AS v Estonia* and *Magyar Tartalomszolgáltatók Egyesülete and Index.hu Zrt v Hungary* – concerning news sites that did not moderate content in advance but reviewed material reported by others – the court emphasized the need for exercising control of published comments.¹⁴⁹ This implies both that states need to establish rules protecting individuals through moderation requirements for platforms,¹⁵⁰ and that news platforms online consequently need to exercise control over published content, making swift decisions on whether or not the content should stay up. As a result, some news portals have simply shut down their comments sections because of the difficulties of controlling abusive speech, and others have adopted stricter guidelines for comments and more extensive moderation.¹⁵¹ One example of such stricter moderation is semi-automated moderation with *humans in the loop*, which can be seen as an attempt to achieve high levels of oversight over a large amount of published material while providing balanced decisions on what to allow or delete. In such moderation, AI agents can detect and flag potentially unwanted content, while human moderators can add context (with due regard to the surrounding discussion, the user itself, cultural aspects, etc.) for a more nuanced view of the situation. All this while – ideally – not excessively infringing on free speech or conveying a sense of censorship, thus avoiding legal

¹⁴⁸See Imara McMillan, 'Enforcement Through the Network: The Network Enforcement Act and Article 10 of the European Convention on Human Rights' (2019) 20 *Chicago Journal of International Law* 252; BBC News, 'Germany Starts Enforcing Hate Speech Law' (BBC, 1 January 2018) <https://www.bbc.com/news/technology-42510868>; Toor Amar, 'Germany Passes Controversial Law to Fine Facebook over Hate Speech' (The Verge, 30 June 2017) <<https://www.theverge.com/2017/6/30/15898386/germany-facebook-hate-speech-law-passed>>].

¹⁴⁹*Delfi AS v. Estonia* [GC], (App no. 64569/09), ECHR 2015 § 159; *Magyar Tartalomszolgáltatók Egyesülete and Index.hu Zrt v. Hungary* (App no. 22947/13), ECHR 2 February 2016 § 91.

¹⁵⁰Cf. *Magyar Tartalomszolgáltatók Egyesülete and Index.hu Zrt v. Hungary* (App no. 22947/13), ECHR 2 February 2016 § 55–57.

¹⁵¹This was also a factor discussed in *Delfi*, regarding the demands on news platforms to exercise control over speech, in relation to their opportunities to do so, leading to a potential loss of free speech if the comments sections are closed instead. See also Clothilde Goujard, 'Why News Websites Are Closing Their Comments Sections' (Medium, 8 September 2016) <<https://medium.com/global-editors-network/why-news-websites-are-closing-their-comments-sections-ea31139c469d>>.

liability for the news site for failures to remove hate speech and other legally problematic content.¹⁵²

It is also clear that social media platforms have recently seen an increased need for the use of AI, due to both political and social factors, such as widespread disinformation regarding the COVID-19 virus – and practical factors as an effect of the pandemic itself – leading to massive amounts of content being posted on the platforms with fewer moderators psychically in place to moderate.¹⁵³ The use of full automation has also often been deemed unsuitable given the importance of highly contextual and culturally dependent judgments and the risk for excessive limitations on freedom of expression.¹⁵⁴ This was acknowledged by Twitter when, in the beginning of the COVID-19 pandemic, they were forced to implement more automatic moderation while asking for understanding if the AI lacked the context human moderation can bring:

[We will increase] our use of machine learning and automation to take a wide range of actions on potentially abusive and manipulative content. We want to be clear: while we work to ensure our systems are consistent, they can sometimes lack the context that our teams bring, and this may result in us making mistakes. As a result, we will not permanently suspend any accounts based solely on our automated enforcement systems. Instead, we will continue to look for opportunities to build in human review checks where they will be most impactful.¹⁵⁵

This statement highlights the fact that these private actors – social media platforms and also news platforms – must tread lightly when exercising the responsibilities that, in a sense, have been delegated to them, regarding the delicate assignment of protecting democratic values and individual rights through moderation.

The use of AI and automation will in many cases be necessary to handle the massive amount of online content, simply for the fact that regardless of the number of moderators, they can never oversee the same amount of content and users as an AI, nor find patterns or detect potential abuse or unlawful material like a trained algorithm. However, as mentioned, humans are crucial for adding context, and also for providing *transparency* and *providing of reasons* regarding how and why a decision was made. Users need insight into why the content was removed, or why their account has been blocked to understand how and why this relates to community guidelines and – ideally – legal requirements. However, transparency into content moderation is also important in order to hold the social media or news platform accountable for their role in upholding legal standards and commitments.¹⁵⁶

¹⁵² Julian Risch and Ralf Krestel, 'Delete or Not Delete? Semi-Automatic Comment Moderation for the Newsroom', *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying* (2018) 166–76.

¹⁵³ Facebook described in 2020 that many moderators, who could only do their job on-site, would be sent home during the pandemic, making Facebook rely more on automated screening for high-severity content. Kelly Earley, 'Facebook Plans to Increase Automated Content Moderation' (Silicon Republic, 12 May 2020) <<https://www.siliconrepublic.com/companies/facebook-content-moderation-automated/>>; See also Facebook AI, 'Using AI to Detect COVID-19 Misinformation and Exploitative Content' (Facebook, 12 May 2020) <<https://ai.facebook.com/blog/using-ai-to-detect-covid-19-misinformation-and-exploitative-content/>>.

¹⁵⁴ Thomas Davidson and others, 'Automated Hate Speech Detection and the Problem of Offensive Language', *Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017)*; MacAvaney (n 44).

¹⁵⁵ Twitter, 'Twitter Company Update' (Twitter blog, 16 March 2020) <https://blog.twitter.com/en_us/topics/company/2020/An-update-on-our-continuity-strategy-during-COVID-19.html>. See also Hutchinson A, 'Twitter Will Increase Its Use of Automation Tools as It Looks to Ensure Accuracy in COVID-19 Discussion' (Social Media Today, 17 March 2020) <<https://www.socialmediatoday.com/news/twitter-will-increase-its-use-of-automation-tools-as-it-looks-to-ensure-acc/574263/>>.

¹⁵⁶ Suzor and others (n 42).

Nonetheless, to what extent transparency is provided in different social media and platforms varies, and the nature of it as well. In contrast to our previous examples the need for one specifically appointed administrator to be responsible (as is the case for decisions concerning sickness benefits), is not as relevant when discussing private actors. The individual decision maker acts under the company flag. It has also been noted that the identity of these moderators, how they work and what guidelines they follow and why, are often intentionally kept under wraps:

Of course, commercial content moderators are not literally invisible; indeed, if anyone should seek them out, they will be there [...]. But the work they do, the conditions under which they do it, and for whose benefit are all largely imperceptible to the users of the platforms that pay for and rely upon this labor. In fact, this invisibility is by design.¹⁵⁷

Another challenge, for human moderators, is the mental challenge of moderating abusive material¹⁵⁸ – not unlike child abuse investigations being traumatizing for human investigators within law enforcement. The human moderator must be kept in the loop in order to make contextual decisions, but the use of AI may help ease the burden of moderating some aspects of hateful, violent or disturbing material. To fully understand how moderation balances these issues, further research about human moderators, combining legal challenges and demands with empirical work, is important.

5. Evaluating the loop

5.1. Looking back

Our contribution so far has highlighted how hybrid decision-making have been discussed and researched from a wide range of disciplines and perspectives. This also reflects the complexity of hybrid decisions, as they are an amalgamation of legal, social, technical and organizational issues. As we have shown, the ambitions of automation are accelerating and the scope of hybrid decision-making systems are increasing, not the least through recent technological developments of AI and machine-learning, but also through the influx of new legal rules which stress keeping a human in the loop. This human in the loop has increasingly become a standard solution for solving the issues of transparency, bias, legal security and systemic risks relating to automation. This is not without its issues. As put by Yeung, ‘although human agency might on occasion act to overcome or mitigate the procedural and substantive concerns associated with the use of algorithmic decision-making systems, it cannot be systematically relied upon to do so’.¹⁵⁹ The role facing these human decision-makers can also be conflicting. On the one hand, they are often expected to (or legally required to) exercise *actual autonomy*, maintaining control of decisions and

¹⁵⁷Sarah T Roberts, *Behind the Screen: Content Moderation in the Shadows of Social Media* (Yale University Press, 2019) 2–3.

¹⁵⁸Elizabeth Dwoskin, ‘A Content Moderator Says She Got PTSD While Reviewing Images Posted on Facebook’ (*Washington Post*, 2018) <<https://www.washingtonpost.com/technology/2018/09/24/content-moderator-says-she-got-ptsd-while-reviewing-images-posted-facebook/>> ; Business telegraph, ‘Facebook, YouTube Content Moderators Asked to Sign PTSD Forms – Gadgets Now’ *BusinessTelegraph* (25 January 2020) <<https://www.gadgetsnow.com/tech-news/facebook-youtube-content-moderators-asked-to-sign-ptsd-forms/articleshow/73611173.cms>>; Anita Singh, ‘Facebook Moderators “Develop PTSD Because They Are Exposed to the Worst Content on the Internet”’ *The Telegraph* (31 May 2017) <<https://www.telegraph.co.uk/news/2017/05/31/facebookmoderators-develop-ptsd-exposed-worst-content-internet/>> accessed 4 March 2020.

¹⁵⁹Karen Yeung, ‘Algorithmic Regulation: A Critical Interrogation: Algorithmic Regulation’ (2018) 12 Regulation & Governance 505, 507 and 516.

keeping an eye on lawfulness, proportionality, accuracy and quality of decisions or recommendations generated by the machine (and potentially the underlying data). On the other hand, they are expected to do this in relation to decision-making systems designed for making decisions at scale, faced with restrictions concerning both time and resources. They also, as indicated, face a complex legal environment with a need to interpret the law and (to again borrow a term from Lipsky) exercise sensitive observation and judgment. Importantly, this legal environment requires both a wide and holistic approach and careful contextual analysis at the same time.

We have indicated the need to consider rules and principles flowing from both general principles of law (and the rule of law), human rights norms, constitutional norms, as well as lower-order provisions. Often forgotten in discussions on hybrid decision-making, is that legal principles and rights operate across and throughout legal systems, and therefore also across decision-making environments. These principles and rights not only modify the implications of specific rules, but may also require that clear rules (otherwise well adapted for automated decisions) should be set aside due to individual circumstances. While our intention has not been to discuss in detail here the specific issues raised by such legal requirements, the importance of taking this normative environment into account in future studies of hybrid decisions must be stressed.

While many of these rules and principles operate on levels which are common to a diverse set of hybrid decision-making environments, the actual implications and limits they place on such decisions will remain nebulous and abstract if not applied and interpreted within a specific context, mindful of the particulars influencing decision-making within it.

As illustrated above, the increased influence of algorithms in policing is a clear tendency and they are implemented into an enterprise which is resource-constrained and subject to strong external pressures for results. They are also implemented into a legal context where the exercise of discretion is often described as a defining characteristic. Policing thus contain discretionary, opaque and vaguely regulated areas – such as the pre-investigatory phase – where there is a risk of hybrid decision-making to be implemented with only limited legal safeguards beyond the human operating the system and checking the output. As the potential implications for the rights of individuals are extensive, the need for effective human control and oversight is urgent. However, given the role of discretion in traditional policing, analyzing the actual implications of hybrid systems on the discretion of the human (officer) in the loop should be a priority. The increased emphasis on preventive dimensions of policing, and the associated reduced role of traditional legal safeguards such as courts and other external controls tied to the formal investigatory stages of crime investigations, also highlights the importance of ensuring that automation efforts are subject to effective alternative safeguards.

And, as shown through our example of decisional support systems deployed within the Swedish public sickness benefits system, requirements of contextualized assessments may as well stem from specific lower-order provisions. Notably, such requirements do not hinder either full or partial automation as such, as they may be changed and remodelled to enable automation. However, as seen in our example, they may prompt the introduction of hybrid decision-making processes to enable automated assistance without necessitating substantial changes to the conditions under which benefits are granted. From this perspective, ‘keeping an administrator in the loop’ is key to ensuring

that nuanced and (legally) knowledgeable assessments can be made. Introducing more advanced and granular support systems, such as in our case the suggested AI-based system that will make more specific recommendations based on cross examinations of general recommendations on sick leave with individual medical certificates, could offer administrators more individualized decision support. Highlighted by the possible promises of present and future technical developments, they do, however, also raise the importance of administrators making substantive assessments rather than being overly reliant on system ‘recommendations’. Otherwise, their involvement in decision-making processes may (as expressed by the Commission) be characterized as ‘token gestures’ – and thus not provide any effective means to counterbalance the negative effects that automation may have on the possibilities to ensure that context matters.¹⁶⁰

As in our previous examples, the contextual nature of online moderation is also highly dependent on keeping humans in the loop, thereby providing platforms with tools for avoiding actual or perceived censorship or undue infringements of free speech. The massive amount of content uploaded on social media platforms and in comment sections, together with developing stricter legal responsibilities to quickly detect and delete problematic content, will make semi-automated solutions a necessity. Such solutions could also, potentially, assist moderators in prioritizing the order in which to moderate content, while the moderator can retain the contextual decisions. Unlike both law enforcement and the sickness benefit system, the actual individual decision maker, or moderator, does not have an equally prominent position as a formal decision maker in relation to the individual targeted by moderation decisions. The need for transparent, nuanced decisions is nonetheless important to maintain an open democratic space online, where both individual rights and societal values are protected. To fully understand and be able to legally assess this moderation – what platform providers and moderators face when being delegated the assignment of protecting democratic values and individual rights through moderation – a legal analysis needs to be combined with empirical studies.

5.2. Looking forward – the human in the machine

The ideal role of the humans in the loop in hybrid systems mirror many ideals traditionally placed on administrators and officers in the exercise of government power. Relying on expertise, exercising judgement, reasonableness, ethics and human values when translating general norms into individual action. However, within the loops of a hybrid system, human decision-makers are faced with not only navigating these legal norms, but also the norms and values embedded in the algorithmic agents. While the expectations placed on human agents in these systems are increasingly emphasized through legal developments relating to AI and automation, the question if they are actually authorized, equipped, and given the opportunity to fulfill this role, must be a key question for research into hybrid systems to investigate going forward. These human and social aspects of hybrid decisions have been highlighted by a Council of Europe Expert Committee, stressing

¹⁶⁰See Section 3.1.

that ‘algorithms and data processing techniques are produced by human beings and operated by human beings. Their implications can therefore not be understood without acknowledgement of the social constructs that exist around them’.¹⁶¹

Consequently, we have repeatedly stressed that socio technical perspectives are beneficial to inform legal analyses in hybrid decision-making environments. Our argument holds that the conditions under which ‘humans in the loop’ operate are not only important to ensure a human-centric approach to automation, but also to safeguard that it is ‘the law’ that is being applied in hybrid decision-making. ‘The law’ gives directions but does not in itself answer whether these directions are actionable at scale in hybrid decision-making environments. Hybrid automation does compound the normative operation that is ‘applying the law’ with its technical aids. This explains why much administration research tend to approach hybrid decision-making as diverse although composite practises, often focusing on whether automation allows for an appropriate ‘discretionary space’ and whether that space is utilized by ‘humans in the loop’ or not. Our argument is that the merits and risks of hybrid decision-making environments in safeguarding the Rule of Law more clearly reliefs when legal and socio-technical aspects of such procedures are separated analytically. This enables identification of what risks are present, when they may arise and, importantly, their causes. A traditional legal analysis will help establishing what ‘the law is’. It may also, as previously discussed, show that perceived ‘discretionary spaces’ may in fact be narrowed when an account is taken to all hierarchal (national, European and international) levels of the legal system in relation to the specific context of the case at hand. Even in more loosely regulated hybrid-environments, our example being online moderation, human moderators must balance values which are inherently tied to broader legal rights and principles. Here, the importance of the institutional and technical surroundings of decision-making are likely to further determine the scope and discretion of human decision-making agents.

In any case, a traditional legal analysis might, arguably, fall short of identifying how specific tasks and assessments, and authorizations are distributed between humans and machines, or what measures have been taken to ensure that system recommendations align with the law. It may therefore also fall short of identifying those legal concerns that arise due to the intersection between humans and machines in hybrid environments, making it evident that legal and socio technical perspectives often need to be combined. Such analyses are better equipped to identify discrepancies between the detailed and abstract across legal hierarchies, and therefore may also better serve both legal science, legislators as well as human agents tasked with staying ‘in the loop’.

Acknowledgements

All authors contributed to the design and implementation of the research, to the analysis of the results and to the writing of the manuscript.

¹⁶¹Council of Europe, ‘Algorithms and Human Rights – Study on the Human Rights Dimensions of Automated Data Processing Techniques and Possible Regulatory Implications’ (Council of Europe 2018) DGI(2017)12.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the Swedish Research Council under Grant number 2020-02278.