Defining the scope of AI ADM system risk assessment

Lee, M.S.A.; Cobbe, J.; Janssen, H.; Singh, J.

DOI
10.4337/9781800371682.00025
Publication date
2022
Document Version
Final published version
Published in
Research handbook on EU data protection law
License
Article 25fa Dutch Copyright Act (https://www.openaccess.nl/en/policies/open-access-in-dutch-copyright-law-taverne-amendment)
Link to publication

Citation for published version (APA):
Lee, M. S. A., Cobbe, J., Janssen, H., & Singh, J. (2022). Defining the scope of AI ADM system risk assessment. In E. Kosta, R. Leenes, & I. Kamara (Eds.), *Research handbook on EU data protection law* (pp. 405-434). (Research Handbooks in European Law). Edward Elgar Publishing. https://doi.org/10.4337/9781800371682.00025

16. Defining the scope of AI ADM system risk assessment

Michelle Seng Ah Lee, Jennifer Cobbe, Heleen Janssen and Jatinder Singh

1. INTRODUCTION

There is a growing range of guidance documents that relate to governance of technical systems, in areas including privacy and data protection, fundamental rights, and ethical considerations. These materials, issued by regulators,[1] governments,[2] legislative bodies,[3] and international organisations,[45] are often framed as specifically targeted at a specific technology. In particular, artificial intelligence (AI) and automated decision making (ADM) have been a key area of focus for guidance. Given that both AI technologies and ADM processes entail the processing of data—in many contexts, personal data—many guidance documents elaborate on the relevance of such technologies in the context of the General Data Protection Regulation (GDPR).[6] And as AI and ADM continue to grow in prominence, we see a growing body of publications

[1] Information Commissioner's Office, 'Big data, artificial intelligence, machine learning and data protection' [2017] Data Protection Act and General Data Protection Regulation https://ico.org.uk/media/fororganisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf accessed 15 February 2021.

[2] Government of the Netherlands, 'Strategisch Actieplan voor Artificiele Intelligentie (Strategic action Plan¨ on AI, Policy Brief of 18 October 2019, Government of The Netherlands)' (18 October 2019) https: //www.rijksoverheid.nl/documenten/beleidsnotas/2019/10/08/strategisch- actieplan- voor-artificieleintelligentie accessed 18 November 2020; OECDAI Policy Observatory, 'National strategies, agendas and plans' (2020) https://oecd.ai/dashboards/policy-instruments/National strategies agendas and plans accessed 28 November 2020.

[3] European Parliament, 'European Parliament resolution on automated decision-making processes: ensuring consumer protection and free movement of goods and services (2019/2915(RSP))' (6 February 2020) https: //www.europarl.europa.eu/doceo/document/B-9-2020-0094 EN.html accessed 18 November 2020.

[4] European Commission Independent High Level Expert Group on Artificial Intelligence, 'A Definition of Artificial Intelligence: main capabilities and scientific disciplines. Report — study of 8 April 2019' (18 April).

[5] https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai accessed 18 November 2020; European Commission, 'Communication on Artificial Intelligence. Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions on Artificial Intelligence for Europe (COM/2018/237 final)' (25 April 2018) https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM %5C%3A2018%5C% 3A237%5C%3AFIN accessed 24 November 2020.

Council of Europe Commissioner for Human Rights, 'Unboxing Artificial Intelligence: 10 steps to protect Human Rights' (1 May 2019) https://rm.coe.int/ unboxing-artificial-intelligence-10-steps-to-protect-human-rights-reco/1680946e64 accessed 25 November 2020.

[6] EU General Data Protection Regulation (GDPR): Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the

from various organisations on associated topics, such as transparency and ethics of AI and ADM systems, highlighting the concern around these systems' impact on fundamental human rights and societal inequalities.[7]

However, terms such as AI and ADM are broad, and what is meant by such terms can be unclear. That is, there can be different interpretations of what 'AI' and 'ADM' constitute, and the sometimes-blurred boundaries between AI and ADM means it is often unclear to what extent guidance described as being for AI is also relevant for non-AI systems (and vice versa).

Moreover, guidance and recommendations that target typical concerns regarding technology specifics, e.g. the algorithm's type (AI) or some degree of automation of an organisation's processes (ADM), will often only partially capture a system's risk profile.[8] In other words, it will generally be unsuitable for organisations to solely rely on definitions of AI to interpret the recommendations in guidance documents without accounting for other relevant factors. Therefore, the proliferation of guidance specific to AI may, for instance, give rise to the mistaken assumption that all AI systems are higher risk than non-AI and require exceptional and separate risk management. In this context, risk management better entails a more holistic assessment of system risk, rather than based on some 'top-down' categorisation of the technologies employed.

Organisations have the responsibility to ensure governance processes are fit for purpose for each algorithmic system. In a data-protection context, the GDPR makes this clear.[9] In order to apply the recommendations of guidance documents, and indeed, better account for system risk, organisations should take care in appropriately interpreting the scope of the guidance given the inherent ambiguity in overloaded terms such as AI. In this chapter, we argue that further work is needed to close the disconnect between the terminology-based guidance and the risk-based governance. Whether an organisation classifies a system as AI or non-AI, the specific risk factors mentioned in the guidance should be considered in the application of its suggestions. For example, an AI-specific guidance on the usage of non-traditional, alternative data sets should apply to a non-AI algorithm if it also uses such data.

A system (defined in section 2.3) should be subjected to governance and control processes appropriate for its risk, yet guidance has disproportionately been specific to AI. This may give a misleading impression that *all* AI systems are higher risk than non-AI systems. In practice, however, a non-AI system may well entail an algorithmic process of greater risk.[10] Guidance documents that focus on AI[11] may therefore encourage practitioners to create an organisational

processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ 2016 L 119/1 2016.

[7] See, e.g.: Jessica Morley and others, 'From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices' (2020) 26(4) *Science and Engineering Ethics* 2141.

[8] Jennifer Cobbe, Michelle Seng Ah Lee, and Jatinder Singh, 'Reviewable Automated Decision-Making: A Framework for Accountable Algorithmic Systems' [2021] arXiv preprint arXiv: 2102.04201.

[9] Arts 24, 25, 32, 35 GDPR.

[10] Michelle Seng Ah Lee, 'Context-conscious fairness in using machine learning to make decisions' (2019) 5(2) AI Matters 23.

[11] Information Commissioner's Office, 'Big data, artificial intelligence, machine learning and data protection' [2017] Data Protection Act and General Data Protection Regulation https://ico.org.uk/media/fororganisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf accessed 15 February 2021; Government of the Netherlands, 'Strategisch Actieplan voor Artificiele Intelligentie

inventory of AI systems and subject them to exceptional or more stringent risk management processes, potentially overlooking non-AI systems that may share similar risks. For example, a complex rules-based system with thousands of variables may face similar challenges to an AI system in its opacity and explainability.

In other words, guidance materials typically take a ‘top-down’ approach, focusing on AI or ADM as technical categories under which some commonly associated risks are elaborated. We argue that organisations should not base their risk management based solely on terminology-driven technology classification, but rather, take an approach that *specifically aims at the risks and issues of any algorithmic system*, whether AI or non-AI, remaining sensitive to the nuances and context of its operation. Many risks attributed to AI are not specific to or exceptional to AI, so it is important that the recommendations are technology and technique-agnostic, focusing on the specific risk factors that must be mitigated. Organisations should not only look to guidance that is AI specific, but rather, consider any recommendations as part of a broader and more holistic risk-based governance process.

This chapter explores (i) the inherent ambiguity in definitions in AI and its relation to ADM, and the potential implications for an organisational understanding of a system’s risk, (ii) the limitations of the terminology-driven, top-down categorisation in framing systems on the understanding of system risk; and (iii) how more holistic, bottom-up risk management processes offer potential beyond those driven by the technical mechanisms a system might employ. Given the fast-moving developments in AI, we shift the discussion away from the terminology (the top-down) towards a more practically workable approach (the bottom-up approach).

Specifically, we first disentangle the intersections of AI-related terminology, using a case study to demonstrate the limitations of these classifications. We then show that whether a system uses AI or is used for ADM is only partially relevant in assessing its risk. In this context, the risks include both risks at the organisational level (e.g., strategic, operational, cyber/security, regulatory, legal risks) and the risks at a societal level (e.g., impact on fundamental human rights, societal inequalities, sustainability, market efficiency, etc.). As the societal impact carries with it potential legal, regulatory, and reputational risks for the organisation, these are closely intertwined. For example, AI risks are typically classified into traditional enterprise risk categories: model, technology, regulatory/compliance, conduct, people,

(Strategic action Plan¨ on AI, Policy Brief of 18 October 2019, Government of The Netherlands)’ (18 October 2019) https: //www.rijksoverheid.nl/documenten/beleidsnotas/2019/10/08/strategisch - actieplan- voor- artificieleintelligentie accessed 18 November 2020; OECD Legal Instruments, ‘OECD/LEGAL/0449, OECD Recommendation of the Council on Artificial Intelligence’ (22 May 2019), https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449%5C%20 accessed 25 November 2020; European Commission, ‘Communication on Artificial Intelligence. Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions on Artificial Intelligence for Europe (COM/2018/237 final)’ (25 April 2018) https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM %5C%3A2018%5C% 3A237%5C%3AFIN accessed 24 November 2020; European Commission, ‘White Paper on Artificial Intelligence: a European approach to excellence and trust’ (19 February 2020) https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-europeanapproach-excellence-and-trust en accessed 25 November 2020; European Commission Independent High Level Expert Group on Artificial Intelligence, ‘A Definition of Artificial Intelligence: main capabilities and scientific disciplines. Report — study of 8 April 2019’ (18 April) https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai accessed 18 November 2020.

market and supplier, tying in the ethical risks in their discussion of market, compliance, and conduct risks.[12]

In arguing for a more detailed and holistic view of the potential impacts of algorithmic systems, we then propose an illustrative set of dimensions that organisations might encompass to help firms understand and assess their risks: (1) the context (domain and potential impact), (2) process (technical, business), and (3) the technology (technique, data). We present these as an indicative starting point, towards enabling more comprehensive and precise application of the AI-driven guidance documents that accounts for the specific nature of a system and its risks.

This chapter is composed of three sections. Section 1 explores the definitions of key terms common in algorithmic systems, disentangling their relationships. Section 2 highlights the limitations of framing a system's risk specifically around terminologies such as AI. Section 3 presents a range of a system's risk factors across several dimensions. We elaborate these through a use case, showing how a more holistic view of a system's technology, processes, and context better assists risk identification and mitigation. Overall, while guidance documents may justifiably focus on the new trends and technologies and their governance, the organisational interpretation of their scope should not rely on the categorisation of systems as AI or non-AI to determine their governance process.

Our contribution is three-fold: (1) demonstrating the ambiguity in definitions of AI, especially in its overlap with ADM process, (2) using four illustrative examples to show the limitations of terminology-driven framing in reflecting a system's risk, and (3) arguing, through an exemplar risk framework, the need for a risk-based approach on algorithmic system governance that is separate from its categorisation as AI or non-AI. In all, we make the case for organisations to take a holistic, risk-based approach, encompassing technology-agnostic considerations, in order to better align with the multi-faceted nature of the risks algorithmic systems.

2. INHERENT AMBIGUITIES IN AI/ADM DEFINITIONS

Organisations refer to guidance documents for interpretation of regulations such as GDPR, so how they define terms is relevant in understanding the practical implications. However, guidance documents often use terms such as AI and ADM in in different ways (sections 2.1, 2.2). AI is associated with the technique used in the algorithmic system because it refers to the technology. ADM is associated with to what extent the algorithm influences a decision process. For instance, some guidance documents on data protection and governance specifically focus on AI but with references to ADM,[13] while some others target ADM but with recommenda-

[12] Tom Bigham and others, 'AI and risk management' [2018] Deloitte Centre for Regulatory Strategy EMEA 1 https://www2.deloitte.com/content/dam/Deloitte/uk/Documents/financial-services/deloitte-uk-ai-and-risk-management.pdf accessed 15 February 2021.

[13] Information Commissioner's Office, 'Big data, artificial intelligence, machine learning and data protection' [2017] Data Protection Act and General Data Protection Regulation https://ico.org.uk/media/fororganisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf accessed 15 February 2021; Government of the Netherlands, 'Strategisch Actieplan voor Artificiele Intelligentie (Strategic action Plan¨ on AI, Policy Brief of 18 October 2019, Government of The Netherlands)' (18 October 2019) https: //www.rijksoverheid.nl/documenten/beleidsnotas/2019/10/08/strategisch- actieplan- voor-

tions for AI.[14] The varied and subjective nature of these terms can result in confusion as to the scope, relevance and applicability of the guidance materials, cited in sections 2.1 and 2.2, resulting potential non-compliance for organisations. This section will explore the nuances between some key, commonly used terms, with the aim of showing the limitations of framing guidance materials and recommendations on the concepts of AI and ADM. A system may be AI, ADM, both, or neither, and if a compliance requirement is indeed predicated on this classification, it is important to understand the potential overlap of these two terminologies.

2.1 Many Definitions of AI

AI is a loaded term that can be defined in different ways depending on the context. Applied AI in industry often refers to (1) *narrow AI*, the use of software to study or accomplish specific problem solving or reasoning tasks, in contrast to (2) *general AI*, which is able to reason, plan, and solve problems autonomously for cross-domain tasks beyond its original design, or to (3) *superintelligence*, a system that can outperform humans in every field.[15]

Some definitions focus on the perceptions of intelligence, while others highlight the interaction with the environment. In 1955, the Dartmouth Research Project defined AI as the problem of 'making a machine behave in ways that would be called intelligent if a human were so behaving'.[16] AI then became embedded into the infrastructure across industries in the 2000s with greater availability of large data sets.[17] Stuart Russell and Peter Norvig, in their highly-prominent textbook *Artificial Intelligence: A Modern Approach*, define AI as 'the designing and building of intelligent agents that receive percepts from the environment and take actions that affect that environment'.[18] Percepts and actions are ultimately data flows, and the algorithm learns patterns from data, in contrast to systems in which a human designer explicitly hard-codes the 'rules'. Machine learning (ML) refers to a set of techniques used in AI to detect and extrapolate patterns from data.[19] Because these algorithms often process data for the purpose of deriving insights for decision-making, AI is linked to ADM as one of the techniques that may be used.

artificieleintelligentie accessed 18 November 2020; OECDAI Policy Observatory, 'National strategies, agendas and plans' (2020) https://oecd.ai/dashboards/policy-instruments/National strategies agendas and plans accessed 28 November 2020.

[14] European Parliament, 'European Parliament resolution on automated decision-making processes: ensuring consumer protection and free movement of goods and services (2019/2915(RSP))' (6 February 2020) https: //www.europarl.europa.eu/doceo/document/B-9-2020-0094 EN.html accessed 18 November 2020; Article 29 Data Protection Working Party, 'Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679' (WP 251rev.01 of 6 February 2018) https://ec.europa.eu/ newsroom/article29/item-detail.cfm?item id=612053 accessed 18 November 2020.

[15] Andreas Kaplan and Michael Haenlein, 'Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence' (2019) 62(1) *Business Horizons* 15.

[16] John McCarthy and others, 'A proposal for the Dartmouth summer research project on artificial intelligence, august 31, 1955' (2006) 27(4) *AI magazine* 12.

[17] Stuart Russell and Peter Norvig, *Artificial intelligence: a modern approach* (2002).

[18] Stuart Russell and Peter Norvig, *Artificial intelligence: a modern approach* (2002), 5.

[19] Stuart Russell and Peter Norvig, *Artificial intelligence: a modern approach* (2002), 5.

The European Commission's *Communication on Artificial Intelligence* (2018) gives a similarly broad definition:

> Artificial Intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal.[20]

This definition was adopted by the European Union's High Level Expert Group on AI (HLEGAI).[21] National governments in member states have adopted this broad definition, such as the Netherlands in their policy brief on a national strategy on AI.[22] The Council of Europe's Steering Committee on Media and Information Society (CDMSI) tasked an expert committee with *A study of the implications of advanced digital technologies (including AI systems) for the concept of responsibility within a human rights framework*.[23] In their work, the experts borrowed their definition of AI from the European Commission's Communication of 2018.[24]

However, the scope of what constitutes AI is unclear in the above definitions, particularly when such definitions are being considered in the context of organisational risk management. The definitions of some technical terms are generally agreed; for example, there is a clear definition of what is a regression model. However, there is disagreement on whether AI encompasses relatively simple algorithms, such as logistic regression, but others limit the definition of AI to more complex models, such as deep neural networks.[25] This debate can be summarised in a frequently misquoted 'AI is whatever has not been done yet,' which the original author

[20] European Commission, 'Communication on Artificial Intelligence. Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions on Artificial Intelligence for Europe (COM/2018/237 final)' (25 April 2018) https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%5C%3A2018%5C% 3A237%5C%3AFIN accessed 24 November 2020.

[21] European Commission Independent High Level Expert Group on Artificial Intelligence, 'A Definition of Artificial Intelligence: main capabilities and scientific disciplines. Report — study of 8 April 2019' (18 April) https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai accessed 18 November 2020.

[22] Government of the Netherlands, 'Strategisch Actieplan voor Artificiele Intelligentie (Strategic action Plan¨ on AI, Policy Brief of 18 October 2019, Government of The Netherlands)' (18 October 2019) https: //www.rijksoverheid.nl/documenten/beleidsnotas/2019/10/08/strategisch- actieplan- voor-artificieleintelligentie accessed 18 November 2020).

[23] Karen Yeung, 'A study of the implications of advanced digital technologies (including AI systems) for the concept of responsibility within a human rights framework, Council of Europe study DGI(2019)05, (MSI-AUT)' (2018) https://rm.coe.int/responsability-and-ai-en/168097d9c5 accessed 26 November 2020.

[24] European Commission, 'Communication on Artificial Intelligence. Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions on Artificial Intelligence for Europe (COM/2018/237 final)' (25 April 2018) https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%5C%3A2018%5C% 3A237%5C%3AFIN accessed 24 November 2020.

[25] Andreas Kaplan and Michael Haenlein, 'Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence' (2019) 62(1) *Business Horizons* 15.

corrected to be 'intelligence is whatever machines haven't done yet'.[26] As such, the definition of AI varies without a clear consensus.

The recently proposed draft AI Regulation[27] defines AI as: 'software that is developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with'. Annex I lists:

(a) ML approaches: including supervised, unsupervised and reinforcement learning, using a wide variety of methods including deep learning;
(b) logic- and knowledge-based approaches, including knowledge representation, inductive (logic) programming knowledge bases, inference/deductive engines, (symbolic) reasoning and expert systems;
(c) statistical approaches, Bayesian estimation, search and optimization methods.[28]

While the exact definition is still subject to change given the regulation is still a draft, this demonstrates a tension between how AI is often defined in academic literature and in regulations. Statistical approaches and logic-based approaches that do not use machine learning would not have fallen under AI in a traditional academic definition of AI.[29]

2.2 Nuances of Defining ADM

ADM involves automating business processes, in contrast to AI, which describes the technique used in the system. ADM, such as credit scoring systems, often has requirements to involve a human reviewer, a common safeguard found in various regulatory instruments and provisions.[30] ADM is directly referenced in data protection guidance documents, not only in the context of GDPR, but in other countries as well, e.g., Canada[31] and New Zealand.[32] Given

[26] Karen Zita Haigh, 'AI technologies for tactical edge networks' [2011], 2.

[27] European Commission, 'Proposal for a Regulation of the European Parliament and of the Council laying down harmonized rules on artificial intelligence (Artificial intelligence Act) and amending certain union legislative acts (COM/2021/206 final)' (21 April 2021) https://eur-lex.europa.eu/legal-scontent/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206 accessed 26 June 2021; European Commission, 'Communication on Artificial Intelligence. Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions on Artificial Intelligence for Europe (COM/2018/237 final)' (25 April 2018) https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%5C%3A2018%5C% 3A237%5C%3AFIN accessed 24 November 2020.

[28] European Commission, 'Proposal for a Regulation of the European Parliament and of the Council laying down harmonized rules on artificial intelligence (Artificial intelligence Act) and amending certain union legislative acts (COM/2021/206 final)' (21 April 2021) https://eur-lex.europa.eu/legal-scontent/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206 accessed 26 June 2021, Annex I.

[29] Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach* (2002), 5.

[30] Reuben Binns, 'Human Judgement in Algorithmic Loops; Individual Justice and Automated Decision-Making' [2019] *Individual Justice and Automated Decision-Making* (September 11, 2019).

[31] Government of Canada, 'Directive on Automated Decision-Making' (5 February 2019) https://www.tbssct.gc.ca/pol/doc-eng.aspx?id=32592 accessed 25 November 2020.

[32] Digital Council for Aotearoa New Zealand, 'Trust and Automated Decision-Making: an interim report on the Digital Council's 2020 research project' (2020) https://www.digital.govt.nz/digital -government/leadership/ digital-council-for-aotearoa-new-zealand/digital-council-reports/trust-an d-automated-decision-makinginterim-report/ accessed 26 November 2020.

GDPR's prominence in discussions around ADM, we will focus on and reference it in our examples.

Although GDPR does not explicitly define ADM, it implies that there are at least three categories (relevant in the context of GDPR) of ADM processing.[33] The first of these is decision-making which *is not* solely automated; i.e., decision-making where an automated system produces information on which a final determination is subsequently made by a human, or where the system's decisions are subject to meaningful review by a human (i.e., *partially* automated, where there is a 'human-in-the-loop').[34] The second is decision-making which *is* solely automated and which *does not* produce legal or similarly significant effects for the data subject; i.e., where a system directly produces a decision, or where any human review is cursory or superficial rather than meaningful, but where the effects of the decision are relatively inconsequential.[35] The third category is decision-making which *is* solely automated and which *does* produce legal or similarly significant effects for the data subject; i.e., decision-making that is solely automated and which affects some kind of legal right or entitlement or otherwise produces some non-legal effect which is as potentially significant for the data subject. Any of these three categories of ADM may or may not involve profiling (and profiling may or may not be part of ADM[36]).

GDPR prohibits the third category of ADM—that which is solely automated and which produces legal or similarly significant effects—except on the basis of a strictly limited number of exceptions (which vary depending on whether special category data is involved or not).[37] These are where (a) where the decision is authorised by law or is necessary for a contract (for 'ordinary' personal data); (b) where it is necessary for reasons of substantial public interest (for special category data); or (c) the data subject has given their explicit consent. Regardless of the kind of personal data or the legal basis, suitable safeguards must be in place to protect data subjects. For either of the first two categories of ADM, the more general legal bases for processing apply.[38]

The question of when a decision is *solely* automated is not necessarily straightforward. The former Article 29 Data Protection Working Party (WP29) suggests that a decision can be solely automated even where a human is involved (for instance, where a human routinely applies the outputs of a system without first reviewing them).[39] The WP29 suggested that, to qualify as a non-solely automated decision, a human intervener's oversight of the decision must be meaningful rather than just a 'token gesture.' They should have the authority and com-

[33] Art 22(1), Recital 71, GDPR.

[34] Article 29 Data Protection Working Party, 'Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679' (WP 251rev.01 of 6 February 2018) https://ec.europa.eu/ newsroom/article29/item-detail.cfm?item id=612053 accessed 18 November 2020.

[35] Article 29 Data Protection Working Party, 'Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679' (WP 251rev.01 of 6 February 2018) https://ec.europa.eu/ newsroom/article29/item-detail.cfm?item id=612053 accessed 18 November 2020.

[36] Article 29 Data Protection Working Party, 'Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679' (WP 251rev.01 of 6 February 2018) https://ec.europa.eu/ newsroom/article29/item-detail.cfm?item id=612053 accessed 18 November 2020.

[37] Art 22 (2) GDPR; Article 29 Data Protection Working Party, 'Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679' (2018) WP251rev.01 https://ec.europa.eu/newsroom/article29/items/612053, 19.

[38] Arts 6 and 9, GDPR.

[39] Art 29, GDPR.

petence to change the decision, and in reviewing the system's decision they should consider all the relevant data.[40] Where this is not the case, a decision may be solely automated even where a human is in-the-loop.[41]

Given the classification of ADM varies, guidance documents on data protection focusing on ADM may leave the applicability and scope of their recommendations open to interpretation. If guidance refers to 'ADM', it may not be clear whether it applies *only* to solely ADM with legal or similarly significant effects (the third category, prohibited by GDPR in all but a small number of categories) or also to solely ADM without such effects (the second category), or, indeed, also to ADM with a substantial automated component but which is not solely automated (the first category).

Formulating an ADM risk management strategy entails not only a consideration of these sub-categories, which is important in understanding the legal and regulatory obligations, but also the nuances of how other facets of the system may influence its overall risk. While GDPR is technology-agnostic, ADM risk management should be sensitive to other factors beyond the definition of ADM associated with human-in-the-loop and legally significant effects. For example, the risks of solely ADM may be exacerbated by the introduction of AI, more so than in non-solely ADM; the opacity and complexity of data processing, coupled with a lack of human oversight, hinders error detection. As stated in AI definitions discussion (§2.1), a lot of AI models are complex and highly non-linear, making it difficult to produce a human-readable explanation. Conversely, having a human-in-the-loop is not necessarily a sufficient safeguard; a partially ADM system could pose more of a risk than a solely ADM system, depending on the context and purpose of its use. Where any guidance document refers to both AI and ADM, there is yet further ambiguity in the scope for organisational interpretation. A system may have both AI and ADM elements (particularly where profiling is involved in ADM), or AI but not ADM, or ADM but not AI, or neither AI nor ADM. If a legal or regulatory compliance requirement, such as to GDPR, is indeed predicated on the class of a system as AI and/or ADM, it is important to understand the potential overlap of these two terminologies.

2.3 Our Definitions: AI and ADM

As the previous discussion indicates, it is difficult to achieve consensus on the meaning of the terms AI and ADM—there are different opinions on each term's scope and delineation. Our aim here in disentangling these definitions is to open the conversation, highlighting the limitations of using specific terminology where considerations are broadly relevant. That is, we seek to show the grey areas between the categories, though we certainly do not purport to prescribe our definitions as the only definition, nor the most 'correct'. Instead we illustrate the nuances through examples, and use these definitions to demonstrate the limitations of framing guidance in these terms in the case study.

[40] Art 29, GDPR.

[41] 'It's Reducing a Human Being to a Percentage'; Perceptions of Justice in Algorithmic Decisions' by Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, Nigel Shadbolt, In Proceedings of ACM Conference on Human Factors in Computing Systems, April 21–26, Montreal, Canada (2018).

For the purpose of this chapter, we use the following definitions to aid our illustration of the limitations of relying on specific terminology, given that the risks may not be associated with the system categorisation:

- Model: a formal, usually quantitative, representation of a real-life phenomenon by which a prediction, decision, or recommended action is derived, given known factors and assumptions. A model can be based on data and/or expert knowledge, by humans and/or by automated tools like machine learning algorithms;
- Algorithm: computational method, formula, or procedure;
- Technique: method used in the technical design, build, and testing of the algorithm;
- Process: a series of logical / ordered operations involved in decision-making, encompassing both the technical and business actions taken;
- Data processing: any operation or set of operations which is performed on personal data;
- System: a set of interacting data, algorithm(s), and/or model(s) to form a technical workflow or product, e.g., a facial recognition algorithm that triggers an identity verification model;
- Machine learning (ML): statistical techniques used for prediction and classification;
- Artificial intelligence (AI): models to mimic intelligent human behaviour, using ML techniques;
- Automated decision-making (ADM): decision-making process that involves a substantial automated component by technological means. Solely ADM has no human involvement,[42] to be distinguished from non-solely (or partial) ADM; and
- Profiling: automated personal data processing with the objective of evaluating personal aspects about a natural person, including to analyse or predict aspects of their performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements.[43] A system may perform profiling of people as a part of ADM and /or using AI, which will be further discussed in the next section.

With these definitions, we will now discuss the unclear intersections between AI, ADM, and profiling.

2.4 Overlap between Terminologies: AI, ADM, and Profiling

If we consider Figure 16.1, AI is associated with *what a system uses*, profiling is associated with *what a system does*, and ADM is associated with *what a system is used to do*. The overlaps among these three terminologies require a closer examination of the nuances of each system. Whether AI is being used for ADM, i.e., a system using AI to make automated decisions, depends on 1) its purpose and 2) the extent to which human intervention is applied. A retail chatbot, e.g., may interact with a customer without human input but provide only

[42] Article 29 Data Protection Working Party, ‘Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679’ (WP 251rev.01 of 6 February 2018) https://ec.europa.eu/ newsroom/article29/item-detail.cfm?item id=612053 accessed 18 November 2020.

[43] Art 4(4) and Recital 71, GDPR; Article 29 Data Protection working Party, ‘Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679’ (WP 251rev.01 of 6 February 2018) https://ec.europa.eu/ newsroom/article29/item-detail.cfm?item id=612053 accessed 18 November 2020, 6.

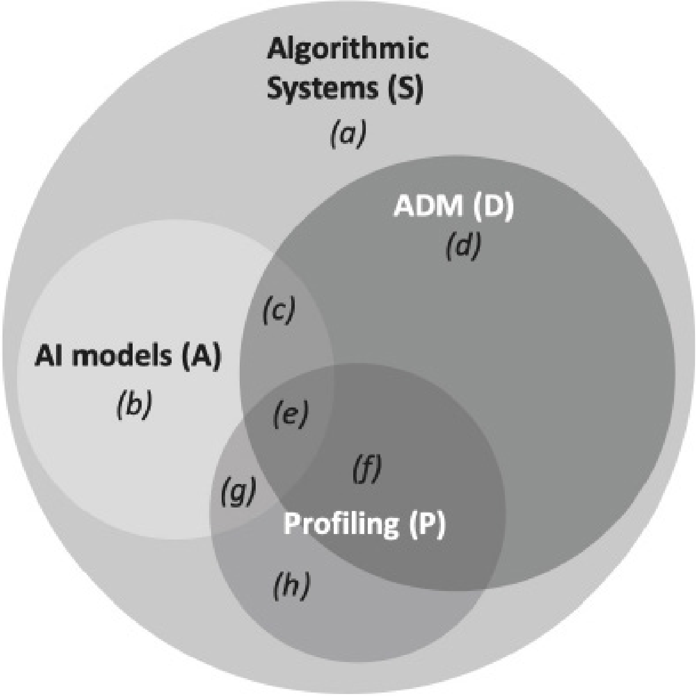
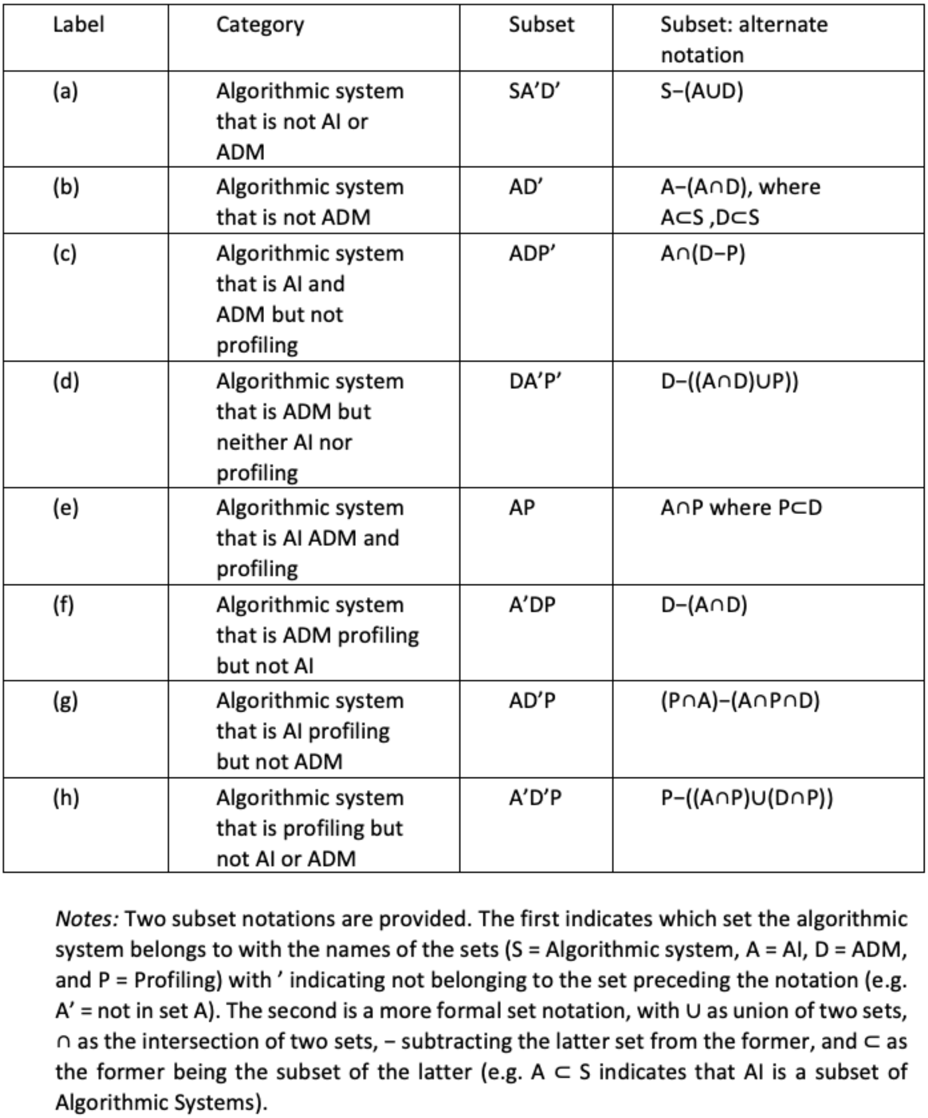


Figure 16.1 *Legal definition of AI, ADM, and profiling*

informational support, e.g., whether a particular product is in stock (label *(b)*). The chatbot's intent prediction algorithm arguably would not be considered ADM.

As Table 16.1 below outlines, a system may involve any combination of AI *techniques*, ADM *processes*, and profiling *activities*. An ML algorithm that analyses the text in job applications to automatically reject underqualified candidates would use AI in an ADM process for the purpose of profiling people (label *(e)*) due to its use of machine learning (ML) techniques and its processing of personal data to make automated decisions with respect to individuals. However, an ML algorithm trained on the same data to highlight parts of the CV amenable to available roles could be considered as using AI for profiling but not necessarily used for ADM (label *(g)*). This is because of the role that the algorithm is playing in the decision-making process; in the former example, the algorithm is the decisive agent, while in the latter, the algorithm merely facilitates and accelerates the human decision-making on who is hired.

Conversely, not all ADM would involve AI. To illustrate this, we consider the example of a scorecard to apply pre-defined criteria to screen out unqualified candidates; e.g., having obtained a postgraduate degree adds five points to the total score, and each year of job experience adds two points to the score. This is an ADM system that involves profiling but is rules-based in its nature and therefore does not use AI (label *(f)*). As a real-life example, the



Label	Category	Subset	Subset: alternate notation
(a)	Algorithmic system that is not AI or ADM	SA’D’	S−(A∪D)
(b)	Algorithmic system that is not ADM	AD’	A−(A∩D), where A⊂S ,D⊂S
(c)	Algorithmic system that is AI and ADM but not profiling	ADP’	A∩(D−P)
(d)	Algorithmic system that is ADM but neither AI nor profiling	DA’P’	D−((A∩D)∪P))
(e)	Algorithmic system that is AI ADM and profiling	AP	A∩P where P⊂D
(f)	Algorithmic system that is ADM profiling but not AI	A’DP	D−(A∩D)
(g)	Algorithmic system that is AI profiling but not ADM	AD’P	(P∩A)−(A∩P∩D)
(h)	Algorithmic system that is profiling but not AI or ADM	A’D’P	P−((A∩P)∪(D∩P))

Notes: Two subset notations are provided. The first indicates which set the algorithmic system belongs to with the names of the sets (S = Algorithmic system, A = AI, D = ADM, and P = Profiling) with ’ indicating not belonging to the set preceding the notation (e.g. A’ = not in set A). The second is a more formal set notation, with ∪ as union of two sets, ∩ as the intersection of two sets, − subtracting the latter set from the former, and ⊂ as the former being the subset of the latter (e.g. A ⊂ S indicates that AI is a subset of Algorithmic Systems).

Table 16.1 Venn diagram label definitions

UK’s EU citizens settlement scheme[44] involved solely ADM with legally significant effects that did not use either profiling or AI (label *(d)*): as it was an algorithm to check the information on people’s applications matched records held by other organisations and that it met

[44] Joe Tomlinson, ‘Quick and uneasy justice: An administrative justice analysis of the EU Settlement Scheme’ (2019) https://publiclawproject.org.uk/wp-content/uploads/2019/07/Joe-Tomlinson-Quick-and-UneasyJustice-Full-Report-2019.pdf accessed 25 November 2020.

a five-year residency criterion, automatically accepting if these two criteria were met. If an application only passed one, then a human reviewer took over and asked the applicant for more information. Because it is not evaluating personal characteristics but rather, classifying people by known information, it is not profiling.[45]

An algorithmic system may neither use AI nor be used for ADM, such as a rules-based program that highlights keywords in an application prior to human review (label *(a)*). Examples of systems that use AI for ADM but not profiling (label *(c)*) include those in algorithmic trading that do not involve personal information.

Profiling—automated personal data processing to evaluate personal aspects—may also be a part of ADM process or involve AI as a technique. GDPR characterises profiling as sometimes being part of solely automated decision-making (see, e.g., Art 22(1), Recital 71) on which other decisions may in turn be based (e.g., Art 35(3)(a), Recital 73), but profiling is itself not necessarily automated decision-making (e.g., Recital 24 or 70).[46] Even when it *is* a part of solely automated decision-making, it is not necessarily solely automated decision-making *with legal or similarly significant effects* (Art 22).[47] For example, general profiling may be performed without links to any decisions, e.g., high-level insights into the customer base distribution (label *(h)*). This could include AI techniques, e.g., clustering customers using unsupervised machine learning techniques into categories (label *(g)*), and if decisions are made, e.g., individualised pricing, it would then be a part of ADM (label *(e)*).

The above examples indicate that the differences in terminological categories *(a-h)* represented in Table 16.1 are complicated and open to interpretation. Due to the inherent ambiguity and nuance around these terminologies, the *framing of guidance around these terms may inadvertently mislead or confuse as to the document's intended or applicable scope*. While the guidance documents target their recommendations around 'AI' and/or 'ADM', the organisations must ensure they appropriately interpret to what extent the guidance is applicable to non-AI and/or non-ADM systems with similar risk profiles. Organisational governance should be framed around specific risks associated with a system, rather than depending solely on a system's classification. This would facilitate a more targeted documentation and logging for greater auditability, testing, and reviewability.

3. LIMITATIONS OF SYSTEM CATEGORISATION

Even if one were to agree on terminologies, and thus the scope of the notions of AI, ADM, and profiling, whether a system uses AI is only partially relevant in assessing its risk. Therefore, applying any of the guidance documents mentioned in sections 2.1 and 2.2 solely to what an organisation considers AI would be misaligned to the true risk profile of each system. For example, regulators responsible for GDPR enforcement have released guidance framed on

[45] See: Article 29 Data Protection Working Party, 'Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is 'likely to result in a high risk' for the purposes of Regulation 2016/679' (4 April 2017) http://ec.europa.eu/newsroom/article29/item-detail.cfm?item id=611236 accessed 26 November 2020.

[46] Article 29 Data Protection Working Party, 'Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679' (WP 251rev.01 of 6 February 2018) https://ec.europa.eu/ newsroom/article29/item-detail.cfm?item id=612053 accessed 18 November 2020.

[47] Art 22 GDPR.

AI. In the UK, the ICO released a report (ICO Report)[48] specifically on the implications of AI, big data, and machine learning on the enforcement of GDPR. While the guidance describes itself as generally applicable, its framing around AI, ML, and big data makes it ambiguous to what extent the recommendations apply to systems that do not employ AI techniques. The guidance justifies its focus on AI by claiming that it presents distinct challenges: the use of ML algorithms, the opacity of the processing, the tendency to collect 'all the data,' the re-purposing of the data, and the use of new types of data. However, these considerations are not unique to AI; similar risk factors that may be present in non-AI algorithms. A hiring scorecard may have hundreds of criteria with a complex logic flow, using third-party data sets and applicants' social media profiles.[49] The fact that it is rules-based and not using machine learning techniques does not detract from the potential regulatory risks, including GDPR and EU non-discrimination laws.

There are challenges in any algorithmic system, but not many of them can be attributed to the technique (AI vs. non-AI) selected. Accordingly, the ICO rightly emphasises in the ICO report that it is not relevant how an organisation defines AI: 'If you are processing this data in the context of statistical models and using those models to make predictions about people, this guidance will be relevant to you regardless of whether you classify those activities as ML (or AI).'[50]

All algorithmic systems, AI or non-AI, may be under scrutiny for any applicable legal and regulatory violations, including GDPR, and organisations are expected to ensure that the governance processes are fit for purpose for each algorithm, in accordance with GDPR[51] (and, indeed, other applicable law). However, the term 'AI' is overloaded, and the challenge lies in interpreting to what extent the guidance is applicable to non-AI algorithms, especially in cases where it has some of the characteristics (e.g., use of alternative, non-traditional data sets) that are typical of an AI algorithm. A more holistic risk assessment is needed where organisations take into consideration a broader set of risk factors than the selected technology.[52]

3.1 Illustrative Examples: Nuances of Risk Factors in Hiring Systems

This section uses illustrative examples of hiring systems to demonstrate that whether a system uses AI, and/or is used for ADM or profiling, does not give full information on the system risk. Given a terminology-driven approach to risk is limited, it points to the need for organisations

[48] Information Commissioner's Office, 'Big data, artificial intelligence, machine learning and data protection' [2017] Data Protection Act and General Data Protection Regulation https://ico.org.uk/media/fororganisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf accessed 15 February 2021.

[49] Anja Bechmann and Geoffrey C. Bowker, 'Unsupervised by any other name: Hidden layers of knowledge production in artificial intelligence on social media.' *Big Data & Society* 6.1 (2019): 2053951718819569.

[50] Information Commissioner's Office, 'Big data, artificial intelligence, machine learning and data protection' [2017] Data Protection Act and General Data Protection Regulation https://ico.org.uk/media/fororganisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf accessed 15 February 2021.

[51] Arts 24, 25, 32, 35 GDPR.

[52] Jennifer Cobbe, Michelle Seng Ah Lee, and Jatinder Singh, 'Reviewable Automated Decision-Making: A Framework for Accountable Algorithmic Systems.' In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (2021] 598–609. 2021.

to more directly address the risk factors of all models and systems, regardless of their technology. The top-down categorisation of systems as AI and non-AI is not helpful in understanding how it potentially applies to non-AI systems, and organisations taking a bottom-up approach focusing on risks found in any system would better inform the governance process.

There is a rise in algorithmic hiring systems. Third-party AI companies have started specialising in hiring, and non-traditional data sources and ML models are being used for employment across the process of sourcing, screening, interviewing, and selection/rejection in hiring stages.[53] In line with this, consider four decision-making systems for hiring new employees: Systems A, B, C, and D. These represent different systems in the same domain area (hiring), and, though hypothetical, are indicative of the approaches used in practice. We use these systems to demonstrate the nuances in what drives their overall risk profile and leverage the definitions of section 2.3 to show the limitations of the system classification into these terminologies. Again, we do not intend to prescribe our own definitions, which would only add to the confusion around the ambiguities in their interpretation.

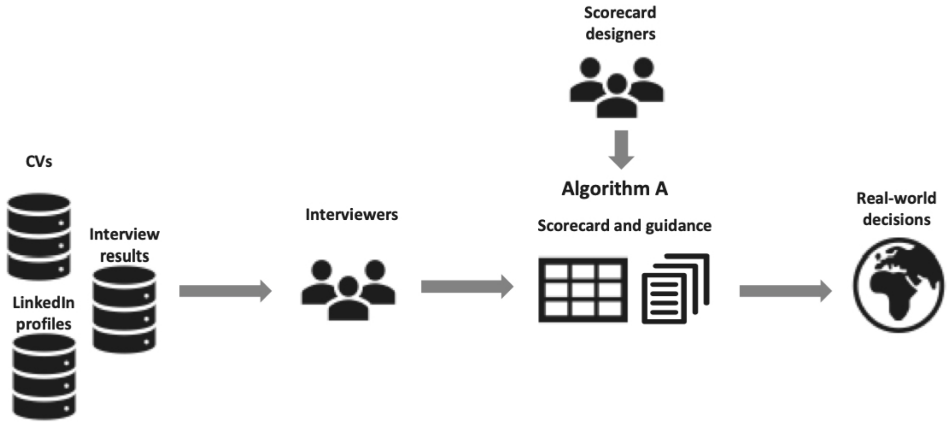


Figure 16.2 System A workflow

3.1.1 Hiring System A

Human interviewers score all candidates using a provided scorecard and guidance to measure their aptitude and fit across 20 categories (Figure 16.2). They also review the CVs and the candidates' LinkedIn profiles. For example, they rate the candidate's leadership potential from a scale of 1 to 10 based on their observation and the candidate's answers to questions. The 20 categories of metrics and their scoring mechanisms are previously disclosed to the candidates preparing for the interviews. This model is used as a one-off process for hiring senior leaders in a UK-based financial services company. Top 10 candidates are accepted.

[53] Javier Sanchez-Monedero, Lina Dencik, and Lilian Edwards, 'What does it mean to 'solve' the problem of discrimination in hiring? social, technical and legal perspectives from the UK on automated hiring systems' (2020), 458; Manish Raghavan and others, 'Mitigating bias in algorithmic hiring: Evaluating claims and practices' (2020), ACM Conference on Fairness, Accountability, and Transparency 2020, 469.

System A involves the processing of personal data, but it does not use AI nor is it ADM given it is conducted entirely by human decision-makers. It will be used to compare the risks across the different dimensions to the other three systems.

3.1.2 Hiring System B

Candidates' data sets are processed through a rules-based model (Algorithm B) to calculate a score, e.g., five points added for having a postgraduate degree and additional two points added if it is in a list of pre-defined subjects (Figure 16.3). This differs from System A in that the scores are calculated systematically based on known variables rather than subjective judgement of the interviewer. The candidates are then ranked by their total score, and the top 10 per cent of candidates are shortlisted for an interview, with any ties in the score broken by the interviewers.

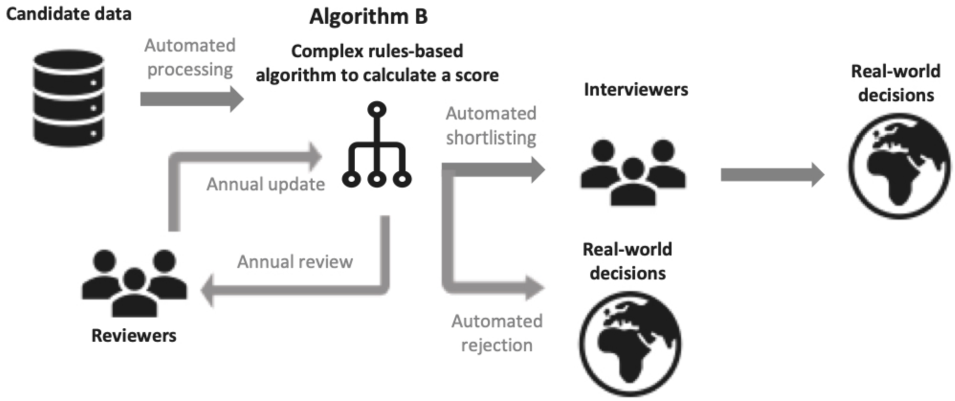


Figure 16.3 System B workflow

The scorecard considers over 1,000 variables, reviewed and updated annually. Algorithm B is in live environment and used on a continuous basis to make decisions on rolling applications for a global manufacturing company to hire 10,000 low-income, low-skilled workers per year. System B can be classified as *solely ADM involving profiling with a significant effect* because the algorithm short-lists the top candidates and automatically rejects those below the required threshold but does not use AI.

3.1.3 Hiring System C

System C (Figure 16.4) is the same as System B in the domain and use case. Algorithm C.1 and C.2 operate automatically on a continuous basis to make decisions on rolling applications for a global manufacturing company to hire 10,000 low-income, low-skilled workers per year.

However, unlike System B, data on existing employees and their original application are used to predict the new candidates' job performance. These predicted performance scores, along with automatically generated model explanations, are shared with the interviewers to assist their decision. The ML algorithm (C.1) considers only 20 features from the candidates' CV and application. A rules-based algorithm (C.2) flags any data that are outside the expected

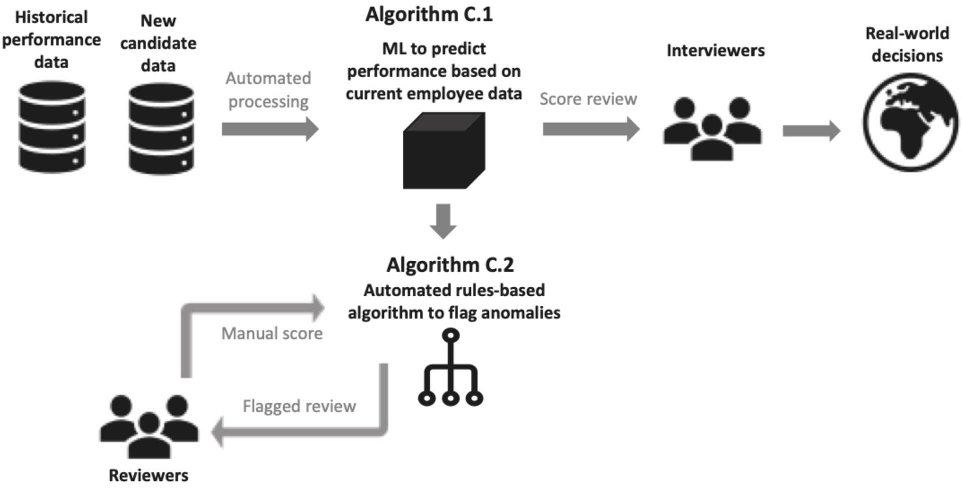


Figure 16.4 System C workflow

values for manual review, e.g., any candidates with salary expectations below or above the range or any candidates with non-traditional experiences. These flagged points are manually reviewed and scored with explanations provided by the reviewers. System C could be considered a *non-solely (partially) ADM* version of Hiring System B with AI, less data, and manual review of anomalies.

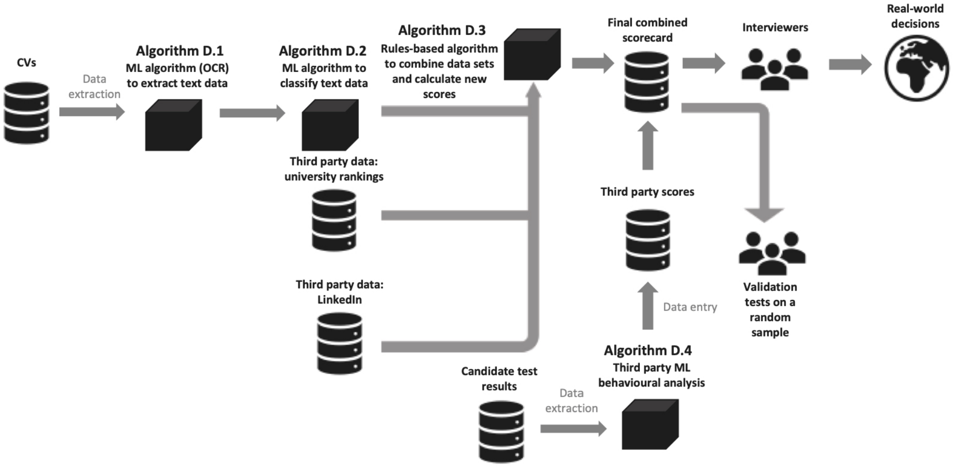

CVs
Data extraction
Algorithm D.1
ML algorithm (OCR) to extract text data
Algorithm D.2
ML algorithm to classify text data
Algorithm D.3
Rules-based algorithm to combine data sets and calculate new scores
Final combined scorecard
Interviewers
Real-world decisions
Third party data: university rankings
Third party data: LinkedIn
Third party scores
Validation tests on a random sample
Data entry
Algorithm D.4
Third party ML behavioural analysis
Candidate test results
Data extraction

Figure 16.5 System D workflow

3.1.4 Hiring System D

System D (Figure 16.5) involves three models, two of which are using machine learning (ML), which recognises patterns from data without explicit hard-coding by a programmer (vs.

a rules-based system), including in fields such as computer vision and language. Algorithm D.1 uses optical character recognition (ML) to extract information from the candidate's CV. Algorithm D.2 uses natural language processing (ML) to classify the information into relevant categories, e.g., detecting 'MSc' in the text is recorded as 'yes' in having a postgraduate degree. Algorithm D.3 combines the data with third-party data sets on university rankings and social media (LinkedIn) in a pre-defined rules-based algorithm to rate the importance of past experience. A third-party special firm conducts proprietary ML-driven behavioural tests (Algorithm D.4) designed based on business psychology research to score the candidates' aptitude and sends the scores and explanations to the human interviewers. A random sample of the combined data set is reviewed manually by a validation team to catch and correct any errors. The interviewers use the outputs of these algorithms as a part of the assessment process.

The algorithm is used once a year to hire recent university graduates for a multi-national European professional services company with 100 intake of new graduate hires per year. The contract with the third-party ML company has undergone the required approval process for new vendor onboarding, including a review of conflicts of interest and approvals from senior stakeholders.

System D uses AI but is not *solely* ADM given the meaningful human involvement in the decision-making process. D.1, D.2, and D.4 are AI algorithms but aim to facilitate the human assessment process through automation rather than being a decision-making engine on their own.

3.2 Risk Factors of Systems

The above shows that such terminological distinctions do not themselves give sufficient information for classifying the risk factors of each of the systems. Rather than focusing on the categorisation of systems as AI and/or ADM, organisations should address the risk factors holistically, including those of third-party data sets, privacy risks, security risks, which helps account for any relevant legal and regulatory requirements and associated guidance and identify the appropriate mitigation strategies for the risks.

GDPR requires that data controllers assess the risks posed by their processing to the rights and freedoms of individuals,[54] including risks of discrimination and other forms of economic and social disadvantage.[55]

In line with this, the first example for system comparison is the risk of discriminatory bias, e.g. based on gender and/or race.[56] It is not immediately clear from the algorithm classifications whether it has any risk of discrimination. Algorithm A in System A, while it does not involve any sophisticated techniques, is subject to interviewers' subconscious bias and inconsistency in decision-making, as well as the potential biases embedded in the scorecard design. Algorithm B in System B significantly reduces the risk of human bias at the algorithm application stage because any comparable candidate would receive the same score due to the consistent rules-based technique. In fact, organisations designing algorithmic hiring systems

[54] Arts 24, 25, 32, 35 GDPR.

[55] Recital 75 GDPR.

[56] For example, Art 23 of the Charter of Fundamental Rights of the European Union (OJ/C 326 of 26 October 2012, 391–407).

claim their primary appeal to be reducing the human subconscious bias.[57] However, as past work on algorithmic bias has shown,[58] human biases can be potentially embedded bias in the design of these rules, such as in the scorecard design. The risks may also be affected by geography and jurisdictions, e.g., relevant regulations and cultural stances regarding the gender wage gap, which goes beyond data protection-related concerns. As such, a deeper dive into the process, people, and context is required to assess discrimination risk.

The second example for system comparison is the risk of algorithmic errors. Machine learning algorithms are not necessarily more prone to errors. While System B is rules-based, algorithms used in System B may well be more complex and multi-dimensional and thus more challenging to understand than Algorithm C.1 or Algorithm D.2, increasing the risk of errors going undetected. System D reduces System B's risk of the algorithm error by adding a human review and reversibility of algorithm's predictions but adds the risk associated with using both a third-party data set (Algorithm D.3) and a third-party algorithm (Algorithm D.4), including the difficulty in testing, monitoring, and assessing[59] whether the data and algorithm governance processes are fit for the firm's purpose.

The risks are not solely inherent in whether the system involves AI or not; there are additional risk factors of data processing and of the (in)adequacy of checks and controls, such as discrimination, unlawful data processing, and violation of fundamental human rights. ADM and profiling are associated with the business process and its degree of automation because it refers to the extent to which a decision has automated components. AI is associated with algorithmic techniques and therefore the technology. A broader risk assessment would include considerations beyond these aspects. For example, GDPR's Data Protection Impact Assessment (DPIA)[60] seeks to oblige organisations to assess the full range of impact on the rights of natural persons. Focusing on these aspects of terminologies befuddles the requirements on what governance processes need to be in place by controllers and/or processors for each system or algorithm. In the next section, we propose a different approach that starts with low-level algorithm risks and walk through examples of risk factors encompassing both the technology and the context, highlighting the differences in assessment of the four illustrative examples. To reiterate, the goal is to show that the categorisation as AI and/or ADM gives only partial information on their risks.

[57] Javier Sanchez-Monedero, Lina Dencik, and Lilian Edwards, 'What does it mean to 'solve' the problem of discrimination in hiring? social, technical and legal perspectives from the UK on automated hiring systems' (2020), 458.

[58] Manish Raghavan and others, 'Mitigating bias in algorithmic hiring: Evaluating claims and practices' (2020), 469.

[59] As an example, see: Toader, Adeline, 'Auditability of AI systems–brake or acceleration to innovation?' Available at SSRN 3526222 (2019).

[60] Article 35, Recitals 75, 84, 89–95 GDPR; Article 29 Data Protection Working Party, 'Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is 'likely to result in a high risk' for the purposes of Regulation 2016/679' (4 April 2017) http://ec.europa.eu/newsroom/article29/item-detail.cfm?item id=611236 accessed 26 November 2020. See also: EDPS - Ethics Advisory Group, 'Towards a digital ethics' (2018) https://edps.europa.eu/sites/edp/ files/publication/18-01-25 eag report en.pdf accessed 26 November 2020.

4. BOTTOM-UP GOVERNANCE: CONTEXT, PROCESS, AND TECHNOLOGY

The above shows that designing governance strategies, top-down using overloaded terms such as AI and ADM, can also provide an insufficient coverage of the potential risks. Rather, the specifics of the situation matter, and it is therefore important to contextualise of the system in its domain area and its potential impact. As an illustration, a facial recognition system used by a bank for identity verification at ATMs is likely to raise different considerations to that of the same system used to track VIP customers' movements around a retail store. Understanding the context in which the algorithmic system is used, in particular if it is for ADM, could potentially bring legal/regulatory obligations that will probably need to be addressed as a priority with different sets of guidance that are relevant to the domain area. While guidance may not exist for every obligation, in practice, guidance documents with a misguiding focus could lead to confusion and non-compliance.

4.1 Risk Factors: An Illustration

In this section, an illustrative set of relevant risk considerations is presented, across a range of dimensions. The purpose of this section is to show how practitioners undertaking more holistic risk-based analysis beyond the technology helps understand the types of mitigation strategies as appropriate for their system, in a manner more useful than one driven by a solely by a broad classification as to whether their employs 'AI.'

Risks in our illustrative examples of four hiring systems include organisation-specific risks, i.e., operational and security risks in relation to the business, but also the potential impact on society and people. The organisational risks are also framed in the context of its societal impact, e.g., the risk of algorithmic complexity is framed around the impact of the resulting likelihood of errors on who is hired. An enterprise risk framework would typically include ethical assessments and encapsulate legal implications as a part of its associated reputational, regulatory, and behavioural/conduct risks. This is further discussed in section 3.3. The context is especially crucial. For example, if a system involves ADM, it may be subject to any relevant regulatory and legal obligations, e.g., under GDPR, and a system used for hiring is subject to national and EU legislation, e.g., non-discrimination laws. Some risks may be mitigated and accepted, e.g., it is difficult for an online system to be completely free of cybersecurity risks. Others that have associated legal and compliance requirements may lead the key stakeholders to scrap the model development entirely in an early risk assessment.[61]

Figure 16.6 outlines six general risk dimensions that, regardless of the process or technique employed, are important to consider in any risk governance: Context (Domain, Potential Impact), Process (Technical, Business), and Technology (Technique, Data). These are aligned to those proposed in technology risk management literature, which provide similar permutations of these six risk categories. A review of relevant literature on technology risks summarises the key dimensions as: project, relationships, solution, and environment risks.[62]

[61] For an example of a standard enterprise risk process, see: Alex Dali and Christopher Lajtha, 'ISO 31000 risk management—"The gold standard"' (2012) 45(5) EDPACS 1.

[62] Hazel Taylor, Edward Artman, and Jill Palzkill Woelfer, 'Information technology project risk management: bridging the gap between research and practice' (2012) 27(1) *Journal of Information Technology* 17.

Context		Process		Technology	
Domain	**Potential Impact**	**Technical**	**Business**	**Technique**	**Data**
• Regulated industry • Market-level considerations set by policymakers and regulators • Oversight mechanism in place	• Scale, materiality (e.g. number of people impacted) • Likelihood of potential impact • Audience / user • Internal vs. external • Vulnerability of users • Human rights • (Ir)reversibility • Duration of impact	• Technical process and workflow • Complexity of system • Interaction with other systems • Process in live environment	• Non-technical business process and workflow • Definition of risk appetite, value prioritisation • Decision vs. insight • Meaningful human involvement / handover • Governance fit-for-purpose • Consistency • Failsafe mechanisms / contingency plan	• AI / ML • Rules-based • Third party • Complexity / interpretability of algorithm • Speed and scale of retraining / learning • Opacity/ control (third party) • Margins of error / variations in accuracy • Consistency in prediction	• Personal / sensitive information • Identifiability if anonymised • Volume, velocity, variety of data • Third party data sources

Figure 16.6 Dimensions of risk factors for algorithmic systems

This section uses more generic terms (technical process vs. project, business process vs. relationships, technology vs. solution, context vs. environment) to apply to implementations beyond a typical information technology project, to reflect the reality that algorithmic systems are increasingly embedded in a wider array of business functions. *Process* encompasses both technical and business actions taken in decision-making, e.g., any automated validation checks or stakeholder approvals, while *technology* is focused on the design, build, and testing of the algorithm (technique) and the data sets used. *Context* includes the domain area, including the relevant regulations beyond data protection, and the potential impact on people, market, and society. The risks mentioned in section 2 can be grouped into these categories. Figure 16.6 connects each dimension to the potential risk factors.

These risk dimensions and factors are representative of those referenced in technology risk management literature,[63] but we do not prescribe these dimensions and factors as 'a' or 'the' comprehensive and complete framework; instead, their role is to demonstrate the nuances and details of potential risks of a system. Note that we use these examples of risk dimensions and factors showing how guidance documents could define a system and assess the risk, to move beyond broad classifications. We show these risk-based considerations capture a more holistic view of a system risk than described through its categorisation as AI or non-AI.

The risk factors in Figure 16.6 are applicable to all types of algorithms and should be factored in any assessment of system risk. These are our illustrative examples of what guidance documents can more broadly incorporate into their risk management process; while we are not prescribing a definitive framework, such a systematic and standardised approach to governance intends to facilitate the detection of potential unknown risks.

4.2 Case Study: Risk Assessment

These risk factors just described (Figure 16.6) will be explored for each of the example hiring algorithms described above. This exercise demonstrates how a system may have variable risk

[63] Ibid.

Table 16.2 Summary of four hiring system classifications in case study

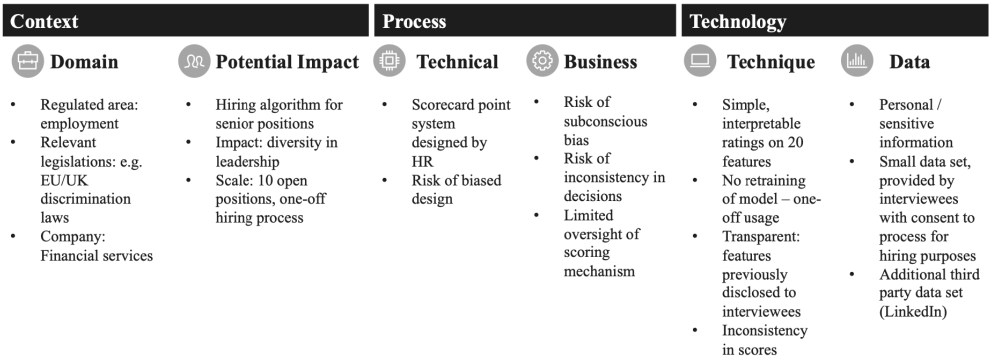
System	Algorithms	Classification
A	A: scorecard	Neither AI nor ADM
B	B: rules-based	Not AI, solely-ADM profiling with legally significant effect
C	C.1: ML, C.2: rules-based	AI, not solely ADM
D	D.1, D.2: ML, D.3: rules-based, D.4: third-party ML	AI, not solely ADM

assessments in each dimension, which provides more valuable insights into the relative risks of each of the systems than their classifications as AI or ADM.

To assist recall in referring back to the illustrative examples of section 3.1, Table 16.2 summarises the key properties of Systems A–D.

4.2.1 System A risks

System A risks are discussed along the three dimensions and six sub-dimensions in Figure 16.7. The hiring process is governed by relevant employment and anti-discrimination laws for the local jurisdiction, which forms a part of the context to frame the risk assessment. While this is a UK-based company, GDPR remains law in the UK as in the EU member states.[64] In addition, it would be subject to the Equality Act of 2010, prohibiting discrimination based on a set of protected characteristics. Even if the applicant features such as age, race, sex, and disability status are not directly considered in the scorecard, indirect discrimination, in which the process or procedure is less favourable in practice based on these protected characteristics, is also illegal. While the domain area (hiring) is the same in the four case studies, the use case and industry would be important in determining the risk level.



Context		Process		Technology	
Domain	Potential Impact	Technical	Business	Technique	Data
• Regulated area: employment • Relevant legislations: e.g. EU/UK discrimination laws • Company: Financial services	• Hiring algorithm for senior positions • Impact: diversity in leadership • Scale: 10 open positions, one-off hiring process	• Scorecard point system designed by HR • Risk of biased design	• Risk of subconscious bias • Risk of inconsistency in decisions • Limited oversight of scoring mechanism	• Simple, interpretable ratings on 20 features • No retraining of model – one-off usage • Transparent: features previously disclosed to interviewees • Inconsistency in scores	• Personal / sensitive information • Small data set, provided by interviewees with consent to process for hiring purposes • Additional third party data set (LinkedIn)

Figure 16.7 System A risks

The scale of the potential impact appears low from a societal perspective, given only ten people are hired, and this is a one-off process. The potential impact from an internal company perspective depends on the level of influence of senior leaders and the importance of diversity

[64] European Union (Withdrawal) Act 2018, s 3; Data Protection Act 2018.

in leadership to the organisation. The technical process risk appears low, given the system only has one algorithm component (scorecard) with no interactions with any complex systems. It is also used in a one-off analysis rather than being called constantly in a live environment.

By contrast, the business process risk may be considered significant due to the potential for subconscious interviewer bias and the risk of inconsistency in decisions with varying score distribution for each interviewer. The presence of hiring discrimination in human-driven processes are well-documented through field experiments and paired audit studies.[65] Without appropriate testing and oversight for implicit or subconscious biases, it is difficult to ascertain that the scoring algorithm is not resulting in indirect discrimination, especially when the top ten scorers are automatically hired without further review.

The inconsistency in scoring is also a challenge for the scorecard. The same inputs (CV, LinkedIn profiles, and answers to interview questions) may lead interviewers to give different scores, given the potential subjectivity in interpreting the hiring scorecard guidelines and in assessing the candidate. However, the scorecard as an algorithm is simple, interpretable, with a limited number of features. There is limited risk on its fit for purpose given it was designed as a one-off system for this hiring cycle and will not be re-trained or re-used in the future. Given the 20 features are previously disclosed to the interviewees, the scorecard is transparent.

The personal data collected through candidate-submitted CVs and interview reports should be processed in accordance with relevant data protection regulations (e.g., GDPR) and stored securely. There is limited regulatory risk around consent associated with the usage of data provided voluntarily by the interviewee; however, greater scrutiny should be required for the use of social media data, as it may be outdated or inaccurate, and that no protected feature nor its proxy is being considered or collected from the data set, e.g., information about their marital status or religious / political group memberships. A valid legal basis would be required for all personal data involved, including social media data.

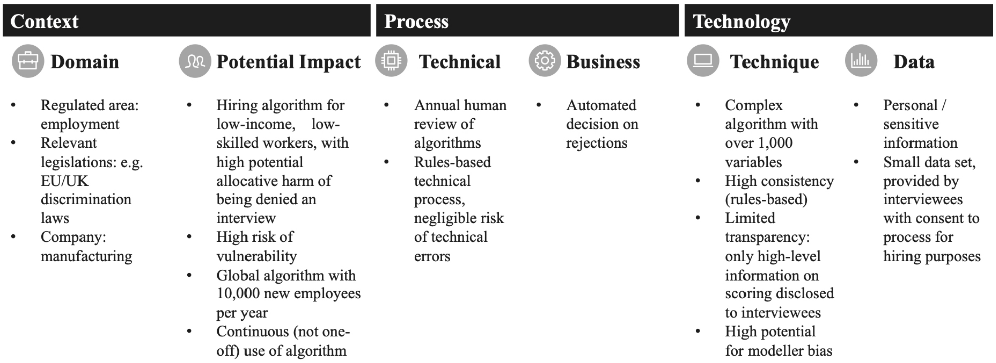
While System A is human-driven and non-ADM with a simple, non-AI algorithm, there are still risks, especially associated with the use of third-party social media data, inconsistencies in scoring, and the potential for unintended discrimination. These exact factors are mentioned in the ICO guidelines[66] as common features of AI, but they are also present in non-AI, non-ADM systems. Guidelines could include the potential risks of non-AI algorithms and be targeted at specific risk factors (e.g., third-party data usage), regardless of whether AI is used, to provide a more holistic recommendation on mitigation strategies.

4.2.2 System B risks

The main risk difference between System B and System A is the systematic, rules-based nature of the scorecard. Rather than being based on independent human judgement of each interviewer, the importance of each feature is pre-weighted in a scorecard. This results in much greater consistency in results of the technique with low risk of calculation errors.

[65] Stijn Baert, 'Hiring discrimination: an overview of (almost) all correspondence experiments since 2005' in *Audit Studies: Behind the Scenes with Theory, Method, and Nuance* (Springer 2018).

[66] Information Commissioner's Office, 'Big data, artificial intelligence, machine learning and data protection' [2017] Data Protection Act and General Data Protection Regulation https://ico.org.uk/media/fororganisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf accessed 15 February 2021.



Context		Process		Technology	
Domain	Potential Impact	Technical	Business	Technique	Data
- Regulated area: employment - Relevant legislations: e.g. EU/UK discrimination laws - Company: manufacturing	- Hiring algorithm for low-income, low-skilled workers, with high potential allocative harm of being denied an interview - High risk of vulnerability - Global algorithm with 10,000 new employees per year - Continuous (not one-off) use of algorithm	- Annual human review of algorithms - Rules-based technical process, negligible risk of technical errors	- Automated decision on rejections	- Complex algorithm with over 1,000 variables - High consistency (rules-based) - Limited transparency: only high-level information on scoring disclosed to interviewees - High potential for modeller bias	- Personal / sensitive information - Small data set, provided by interviewees with consent to process for hiring purposes

Figure 16.8 System B risks

However, the risk profile of System B is broader than the scorecard itself, as outlined in Figure 16.8. The multi-national jurisdiction requires ensuring compliance with local legislation. The scale of the potential impact is extremely high, not only due to the number of people impacted, but also due to the relative vulnerability of the applicants, for whom being denied the job interview may cause more harm than for those in System A due to their low-skill, low-income positions. Given the algorithm is used continuously in a live environment, the risk is also dynamic; in contrast to the one-off use of an algorithm in System A, the algorithm may be inaccurate or outdated. While there is an annual review of algorithms, this may not be sufficiently frequent to identify any errors. Because the algorithm is rules-based, there is a low likelihood of errors in the technical process, but there may be a high business process risk due to the automated nature of the job rejections that occur with only an annual review of the decisions. It also risks being caught by the Article 22 GDPR prohibition on solely automated decision-making with legal or similarly significant effects,[67] depending on whether one of the exemptions specified in Article 22 applied[68] (given the power imbalance between the applicants and the company, it is unlikely that informed consent would be an option[69]).

While it is more consistent in outcomes than System A due to its dependency on pre-programmed rules, there is also limited transparency. Only high-level information on the scoring mechanism would be disclosed to the applicants to prevent gaming, and even if the full algorithm were disclosed, it would be difficult to interpret. Compared to System A, the data risks are low given it is based on only the small data set provided by the applicant for the purpose of the hiring process.

System B, while non-AI ADM, has varying risk levels across its context, process, and technology. It shares similar features to AI, e.g., complexity and limited interpretability, despite its rules-based algorithm. Guidance documents could help practitioners understand their compliance responsibilities for non-AI algorithms by tailoring it to the risk factors involved.

[67] Art 22 GDPR.

[68] Art 22(2) GDPR.

[69] Recital 43, GDPR; Article 29 Data Protection Working Party ‘Guidelines on Consent under Regulation 2016/679’ (2018) *wp259rev.01* https://ec.europa.eu/newsroom/article29/items/623051/en accessed 28 November 2021.

Context		Process		Technology	
Domain	**Potential Impact**	**Technical**	**Business**	**Technique**	**Data**
• Regulated area: employment • Relevant legislations: e.g. EU/UK discrimination laws • Company: manufacturing	• Hiring algorithm for low-income, low-skilled workers, with high potential allocative harm of being denied an interview • High risk of vulnerability • Global algorithm with 10,000 new employees per year • Continuous (not one-off) use of algorithm	• Automated controls and checks • Human review of algorithms	• Human review of predicted performance from the algorithm but non-automated decision	• Simple ML algorithm, interpretation reviewed • Model with consistency (non-random) • Potential bias: past candidates may not be reflective of new candidates	• Personal / sensitive information • Small data set, provided by interviewees with consent to process for hiring purposes

Figure 16.9 System C risks

4.2.3 System C risks

System C is similar to System B except that instead of a rules-based scorecard, past employee data are used to predict new candidate performance based on similar features. In this illustrative example (Figure 16.9), given it is an ML algorithm, any outliers or exceptions are flagged by a rules-based algorithm (e.g., if any features are outside of an expected range, or a candidate is from a non-traditional background).

The risks are arguably comparable or lower across the board compared to System B. The primary new risk is that the current employee demographics or skill sets are not representative of those of new job applicants. This skewed sample could result in looking for 'similar' candidates to those already holding the position, replicating any past biases in hiring practices. For example, if in the past, the company hired mostly men in highly ranked universities, the algorithm may overlook women with high performance from foreign regional universities. There would be additional bias deriving from the fact that the algorithm is only trained on past candidates who were both accepted into their roles and accepted their offers of employment. Without information on candidates who rejected the roles but were a good fit and on candidates who were turned away, the algorithm affected by resulting representation bias in the employee data could be a poor fit to the new candidate data. For example, a technology company's AI algorithm to review job applicants' resumes and score them from 1-star to 5-star was scrapped after it was found to penalise resumes that included the word 'women's,' as in 'women's chess club captain' and downgrade graduates of two all-women's colleges.[70]

The risks of biases in System C are different but not necessarily worse than human bias embedded in the scorecard design in System B. Scorecard designers, for example, may rate degrees in technical fields higher than degrees in humanities, which is not necessarily a prerequisite to a low-skill position but may cause significant gender bias.

While System C risks are mitigated through automated controls, System B risks are only annually reviewed with no specific testing for biases. It reduces the potential for errors compared to an automated scorecard.

[70] Jeffrey Dastin, 'Amazon scraps secret AI recruiting tool that showed bias against women' (2018) 9 San Francisco, CA: Reuters 2018.

Context		Process		Technology	
Domain	**Potential Impact**	**Technical**	**Business**	**Technique**	**Data**
• Regulated area: employment • Relevant legislations: e.g. EU/UK discrimination laws • Company: professional services	• Hiring algorithm for recent university graduates • Multi-national European offices with 100 new employees per year • Continuous (not one-off) use of algorithm	• High risk of technical errors, with high level of interactions between data and algorithms	• Human review of all algorithms and outputs • Final decisions made through human deliberations	• Complex ML algorithms with varying interpretability • Potential inconsistency in results, e.g. due to errors • Limited transparency: third party algorithm inaccessible and non-auditable, limited disclosure to applicants	• Personal / sensitive information • Third party data sets with potential errors, biases • Social media data set with questionable consent and quality

Figure 16.10 System D risks

4.2.4 System D risks

System D (Figure16.10) involves AI but not fully automated ADM. Its hiring algorithm targets recent university graduates with a scale of impact between Systems A and C. While the system is used continuously (vs. one-off), it is used once a year for the graduate application cycle, rather than a rolling process used in System B.

The technical process risk is high due to the complex interactions between multiple different algorithms, which exacerbates the potential for error. The business process risk is relatively low due to the human review of algorithms and outputs, with final decisions made through human deliberations. The manual validation of a random sample of the final output data set mitigates the risks of errors. The technique-related risk consideration is relatively high, not only due to the complexity of the algorithms and the potential inconsistency in the results, but also due to the third-party algorithm (AI-as-a-Service) used that is inaccessible and non-auditable for the organisation. The proprietary nature of the algorithm would also limit the organisation's ability to disclose its explanations to the applicants. The data risk is high, as in addition to the social media data set used in System A, it uses additional data purchased from a third party, which may have its own set of biases through human encoding (e.g., what factors are important in ranking universities?) and through choices in the data collection process (e.g., were some universities, e.g., foreign, not included in the ranking?).

4.2.5 Illustrative examples takeaways

The risk factors reveal important and nuanced comparisons across the four systems. For example, the rules-based technique in System B is arguably more complex than ML-based techniques in System C and System D because it has over 1,000 variables, despite the technique of algorithms in C and D being more advanced. Due to the high number of variables, it could be harder to identify errors / discrepancies or explain an outcome in System B than in Systems C and D.

A risk assessment that is both more detailed and holistic, beyond the categorisation of AI and/or ADM, requires consideration of factors beyond whether it uses AI. This indicative framework shows that a bottom-up approach, starting with a set of risk factors rather than top-down starting with the system categorisation, can holistically reveal and account for the context-specific complexities of such systems. This enables a more targeted risk and impact assessment and mitigation strategy.

4.3 Risk Factors' Role in Algorithmic Impact Assessments

Various approaches have been proposed and are under development to assist organisations in a more holistic assessment of an algorithmic system's potential impact. Identifying the risks factors as demonstrated in the case study could facilitate an algorithmic impact assessment through its evaluation of the process and technology risks within the specific context of the system's deployment. The potential impact of a system can be assessed (ex ante) through various review instruments, e.g., with GDPR's Data Protection Impact Assessment (DPIA) for data protection. Having a more holistic consideration of the impact in the context of the risk factors in guidance could potentially give a more effective assessment of likelihood and scale of harm by providing more nuanced distinctions between the different risk factors.[71] Amongst others,[72] the Belgian Privacy Commission published a 'white list' and 'black list' in relation to GDPR, setting out when a Data Protection Impact Assessment (DPIA) is always required ('black list') and not required ('white list').[73] These are associated with the type of data and processing rather than focusing on AI or other technical methods employed.

Some academics have proposed approaches to undertake impact assessments broader than data protection.[74] In addition, the Canadian regulator appears, at this moment, fairly unique with its obligatory Algorithmic Impact Assessment (AIA), which applies to public entities that seek to develop or to procure any ADM system.[75] More guidance may be underway—in their White Paper on AI, the European Commission identified several AI risk factors to consider in future regulatory framework, whereas Council of Europe's Commissioner for Human Rights explicitly called for the development of fundamental rights impacts review of AI systems.[76] Several scholars have proposed approaches for impact assessments. For instance, Mantelero

[71] Broader approaches to impact assessments more generally are not entirely new, see e.g., Paul De Hert, 'A human rights perspective on privacy and data protection impact assessments' in *Privacy Impact Assessment* (Springer 2012); more critical about broadening GDPR's DPIA to other rights are Niels van Dijk, Raphaël Gellert, and Kjetil Rommetveit, 'A risk to a right? Beyond data protection risk assessments' (2016) 32(2) *Computer Law & Security Review* 286.

[72] IAPP, 'EU Member State DPIA Whitelists, Blacklists and Guidance' [2021] International Association of Privacy Professionals https://iapp.org/resources/article/eu-member-state-dpia-whitelists-and-blacklists/ accessed 20 February 2021.

[73] Wim Nauwelaerts and Paul Greaves, 'Belgian Privacy Commission Issues Guidance on Data Protection Impact Assessments Under the GDPR' [2018] Data Matters: Cybersecurity, privacy, data protection, internet law, and policy, Sidley https://datamatters.sidley.com/belgian-privacy-commission-issues-guidance-on-dataprotection-impact-assessments-under-the-gdpr accessed 15 February 2021.

[74] For example: Margot E Kaminski and Gianclaudio Malgieri, 'Multi-layered explanations from algorithmic impact assessments in the GDPR' (2020); Kaminski, Margot E, and Gianclaudio Malgieri. 'Algorithmic impact assessments under the GDPR: producing multi-layered explanations.' U of Colorado Law Legal Studies Research Paper 19–28 (2019).

[75] Government of Canada, 'Algorithmic Impact Assessment' (28 July 2020) https://www.canada.ca/en/ government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmicimpact-assessment.html accessed 25 November 2020.

[76] Council of Europe Commissioner for Human Rights, 'Unboxing Artificial Intelligence: 10 steps to protect Human Rights' (1 May 2019) https://rm.coe.int/ unboxing-artificial-intelligence-10-steps-to-protect-human-rights-reco/1680946e64 accessed 25 November 2020.

developed an approach for an ethical impact assessment on AI for ethical boards, either in-company, or by independent actors.[77] AINOW developed a practical framework for public agency accountability for ADM.[78] Another approach aiming at a practically applicable fundamental rights impact assessment to ADM for private organisations aimed at incorporating fundamental rights into GDPR's DPIA.[79]

4.4 Mitigation Measures

The identification of holistic risk factors, both organisation-specific and those concerning broader societal impact, is essential for both AI and non-AI methods in order to identify the appropriate and proportional controls. Organisations should implement recommended control strategies targeted to a risk, regardless of whether the guidance is framed around AI.

Mitigation measures seek to be proportionate to the risk's potential impact and its likelihood of occurrence, including the business impact of non-compliance with any legal or regulatory obligation and the societal impact in accordance with the organisational ethics and values. Regarding our case study, if the hiring system is used for recruiting a few specialist roles, it may be subject to less regulatory scrutiny than if it is used to hire all administrative staff, due to the relative *impact* of each model, thus incentivising the organisation to focus on the latter model. System B used in a large-scale setting has a greater material societal impact on the financial opportunities of those more likely to be from a disadvantaged socioeconomic background. A system used by a large global company also has greater potential impact in scale than one used by a small firm. Mitigation strategies can reduce the system residual risk, as a part of a standard governance process.

The identification of specific risk factors helps facilitate more effective governance. In System D, which concerned hiring using multiple algorithms and data sources, the primary risk is the complexity of the technical process and workflow, with high levels of interaction between algorithms and data sources. Automated controls can be put in place to mitigate the risk of unidentified errors, such as checks that the input data fits within an expected distribution. Any outliers and anomalies could be flagged for human review. Appropriate governance processes should be identified for AI systems throughout its development lifecycle, from design to deployment.[80]

System A and System D both use social media data sets. The concerns around consent and quality could be mitigated by disclosing to the applicants any unfavourable findings in the social media data and allow them to challenge them. Guidelines could be given to the interviewees in System A, and constraints to Algorithm D.2, to prevent the use of potentially sensitive or legally protected information that is not relevant to the individual's suitability to perform the job. To reduce human subconscious bias in all four systems, training could be

[77] Alessandro Mantelero, 'AI and Big Data: A blueprint for a human rights, social and ethical impact assessment' (2018) 34(4) *Computer Law & Security Review* 754.

[78] Dillon Reisman and others, 'Algorithmic impact assessments: A practical framework for public agency accountability' [2018] AI Now Institute 1.

[79] Heleen L Janssen, 'An approach for a fundamental rights impact assessment to automated decision-making' (2020) 10(1) *International Data Privacy Law* 76.

[80] Michelle Lee, Luciano Floridi, and Alexander Denev, 'Innovating with confidence: Embedding AI governance and fairness in a financial services risk management framework' (2020) 34(2) *Berkeley Technology Law Journal*.

provided to interviewers and algorithm developers to increase their awareness of biases, and the algorithms' output should be tested for potential unfair bias. There has been a significant amount of academic interest in the formalisation and testing of fairness in algorithms,[81] which could be applied to systematically identify discriminatory biases in algorithms. While these are just a few examples of mitigation techniques, they address specific risks that are not necessarily unique to AI systems.

5. DISCUSSION

The example Systems A-D show it is their specific properties that exhibit different risk profiles, which are not necessarily aligned with each of their classification as AI or non-AI. The complexity and resulting lack of transparency of non-AI System B raises its technical risk compared to Systems C and D, despite Systems C and D using AI techniques. The continuous usage of System B (rolling application) exacerbates the risk of any errors going undetected, coupled with a high potential societal impact of any error on 10,000 low-income employees globally.

Despite the fact that systems deploying AI are not necessarily higher risk than those not using AI, as concerns over AI have risen with rapid adoption across industries of increasingly complex algorithmic techniques, much guidance has been framed around 'AI'. However, to align to the technology-agnostic nature of concerns in data protection and risk management, and for practitioners to effectively comply with its requirements, it is important to for organisations to apply the recommendations from guidance in a bottom-up approach, starting with the risk factors rather than to AI algorithms. The focus of guidance on AI may give the false impression that the risk of any AI system is greater than those of non-AI systems, potentially fostering complacency among organisations in the governance of complex, high-risk non-AI algorithms while encouraging disproportionate scrutiny on low-risk, simply structured AI algorithms.

Targeting organisational risk governance on ambiguous terminology is ultimately counterproductive to its practical implementation. The conflation and overloading of definitions of AI makes it problematic if organisations rely solely on these terminologies to target their governance strategies. These categories of AI vs. non-AI fail to convey the full range and depth of risks, thus limiting an organisation's understanding of what governance measures should be employed. Due to the focus of guidance documents on big data, ML, and AI, an additional step is needed for organisations to unpack to what extent they apply to non-AI algorithms with similar risks. It is important for organisations to interpret the guidance for GDPR and other relevant compliance requirements—not only for ADM models that use AI techniques—but for all algorithmic systems, with the full understanding and consideration of the risk levels of each individual system.

We illustrated the usefulness of a bottom-up *risk-oriented* approach in a more holistic understanding of system risk, with examples across different dimensions in order to better understand a system's risks. The factors we introduce and consider, though consistent with established organisational risk management practices, may not be comprehensive, nor is the

[81] Ben Hutchinson and Margaret Mitchell, '50 years of test (un) fairness: Lessons for machine learning' [2019] 49.

framework the best or the only way of categorising a taxonomy of types of risks. Rather, the objective was to demonstrate that a risk-driven approach may support more actionable insights compared to an approach that identifies risks typical of AI-based systems. Further guidance focused on educating on specific risks and their contexts, processes, and technologies, would help organisations implement their recommendations in practice.

Of course, the confusion around terms may still exist due to their inherent ambiguity, and they may still be used to reference the range of techniques and processes being used. However, rather than applying the AI-specific recommendations to 'AI' systems only, organisations should investigate the risks a system may pose regardless of how the technology is defined or classified.

Organisations will continue looking for a way to navigate the increasingly complex risk landscape, given the availability of data and advancements in AI. In light of these trends, further work is needed to help practitioners apply the recommendations from guidance documents to their specific use cases.

6. CONCLUSION

A full understanding of system risk requires a holistic and fine-grained view of not only of the technology, but the systems purposes, aims and broader context. This chapter highlights the gaps and ambiguity in the definitions of key terminologies, disentangling the intersections between different possible overlap among AI technique, ADM process, and profiling purpose. It is not only that the scope of terms should be clarified—an inherently difficult task due to the nuances and ambiguity of these definitions. An organisation should not take the focus of guidance materials being framed around AI as an indication of that AI should be subject to exceptional and different governance strategies. Rather than a top-down approach of applying blanket governance changes to all systems using AI, organisational risk management should entail more a holistic approach considering technology-agnostic risk factors. A case study of four hiring systems showed the limitations of these terminologies, i.e., AI, non-AI distinctions, in explaining the system risk. We suggested three dimensions (context, process, technology) and six sub-dimensions of risk factors (technique, data, technical process, business process, domain, and potential impact) as exemplar considerations for driving a more complete risk assessment, accounting for the relevant regulations and guidance materials. This would support more appropriate and effective organisational governance regimes across all algorithmic systems.

ACKNOWLEDGEMENTS

The Compliant & Accountable Systems Group acknowledges the financial support of the Aviva, UK Engineering & Physical Sciences Research Council (EP/R033501/1, EP/P024394/1), and Microsoft through the Microsoft Cloud Computing Research Centre.