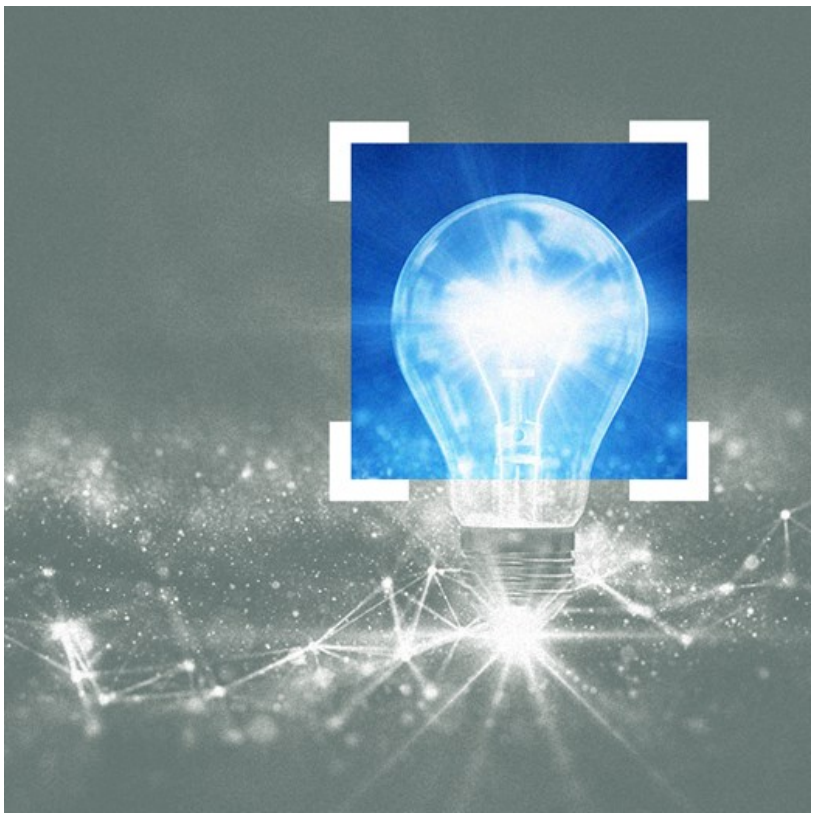
AI bias

Poama, A.; Fosch Villaronga, E.; De Silva, S.; Froggatt, A.; George, D.; Goldberg, S.; ... ; Tanna, M.



Expert Essentials

(In Progress)

Sam De Silva (ed.) et al.

https://doi.org/10.1093/9780198972877.001.0001

Published: 14 May 2025 - **Online**

ISBN: 9780198972877

ARTICLE

AI bias

Andrei Poama, Eduard Fosch-Villaronga

https://doi.org/10.1093/9780198972877.003.0045

Published: 02 June 2025

Abstract

This chapter examines how artificial intelligence (AI) bias can undermine legal (or legally relevant) norms and standards. It does so by introducing a conceptual distinction between *bias in AI* (arising from flawed data, programming choices, or emergent algorithmic behaviour) and *bias towards AI* (where human decision-makers either overtrust or unjustifiably dismiss AI outputs). This distinction can equip legal practitioners with a deeper, yet straightforward understanding of various AI biases and the risks they raise. To mitigate these risks, the chapter explores preventive and corrective strategies, including regulatory sandboxes, fairness-aware AI design, auditing laws, and legal oversight mechanisms. Addressing AI bias is not merely a technical challenge—it is a professional responsibility for legal practitioners who seek to properly navigate the relationship between law and AI.

Keywords: artificial intelligence (AI), AI bias, algorithmic bias, AI governance, AI regulation, automation bias, algorithmic decision-making, fairness in AI, EU AI Act, bias in/towards AI

Subject: Artificial Intelligence, Law

Section: Technology

Section editors: Joanna Bryson, Jennifer Davidson, Katherine Evans, Charles Kerrigan, Tharin Pillay, Scott Robbins, Jacqui Taylor

Introduction

To understand the relevance and risks of artificial intelligence (AI) bias in legal practice, we must first define what we mean by *AI*. In 2024, the EU became the first jurisdiction to establish a comprehensive legal framework for AI.[1] The AI Act (AIA) establishes a risk-based regulatory approach, differentiating AI applications based on their potential harm to individuals and society. The AIA legally defines what qualifies as an AI system. Under the AIA, an *AI system* is defined in Art. 3.1 AIA as:

> 'AI system' means a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments.

While this definition is legally authoritative within the EU context, it is not uncontroversial. Critics argue that it risks conflating different types of AI systems and obscure differences between rule-based and learning-based models. At its core, an AI system or agent comprises a mechanism for generating outputs—such as predictions, classifications, or recommendations—based on training data and a set of inferential rules. In machine-learning systems, these rules are typically learned from training data; in rule-based systems, they are programmed by human developers. Here, it is important to note that the inferential rules can be entirely determined by human agents, but deployed by artificial ones (this is often called *rule-based AI* or *AI expert systems*) or they can be formed on the basis of the input data by artificial agents themselves (this is called *machine-learning AI*). An orthogonal distinction between different types of AI agents is that between predictive and generative AI: predictive AI agents forecast future events based on past data, while generative AI agents create new data that resembles old data (in particular, text, images, or sound).[2]

Because AI systems learn from correlations and infer relations between data points, they inevitably classify information. In some cases, classifications introduce deviations from norms or standards, which can be understood in terms of bias. While not all deviations are problematic, some create systematic distortions that lead to unfair or discriminatory outcomes. Given that many current AI systems, particularly those based on machine learning, do not follow preset rules like traditional software,[3] their ability to infer, adapt, and operate with varying degrees of autonomy makes them particularly vulnerable to bias, which can come from the human decisions taken on those classifications, be embedded in the algorithms' datasets, or get introduced through model design.[4] Defining *bias*, therefore, is essential for understanding and assessing *AI* bias.

We take *bias* to generally refer to 'a systematic departure from a [genuine] norm or standard of correctness.'[5] This definition of bias is tied to the norm-theoretic account of bias, which takes biases to be, *qua* departures from genuine norms, considered invariably problematic. The contending *functional* account takes biases to be functionally useful 'best predictions' for a decision or event in the context of incomplete information and uncertainty. While many of the norms and standards that legal professionals (should) care about are specifically legal—such as the principles of proportionality in criminal law or the prohibition of non-discrimination in labour law—other norms and standards that matter within the legal domain can be epistemic (e.g., accuracy, reliability, coherence, evidence sensitivity), moral (e.g., respect, trust), economic (e.g., utility, efficiency), aesthetic (e.g., symmetry), and so on. Biases, as defined here, can deviate from multiple kinds of norms and standards at once. For example, racist or sexist biases can simultaneously violate epistemic norms of accuracy *and* moral norms of respect. Biases can also affect different mental processes, such as when someone is systematically more likely to *perceive* an individual from a certain group as dangerous or more likely to *believe* that members of that group pose a risk.[6] That being said, biases do not necessarily depend on

specific mental states. They may operate without conscious intention or awareness (e.g., implicit bias[7]), *and* biases can emerge through structural processes where *no* individual's mental state is at play (e.g., datasets might be biased because they are incomplete). Some of these structural biases without mental states can form a distinctive case of AI bias. Note that, in this chapter, we sometimes use 'input data' to refer to training data used during development. This differs from the common technical usage, where 'input' refers to data processed by the model during deployment. We discuss this (sub)type of bias in section 'Biases towards AI'.

AI biases

With these shorthand definitions in hand, we can now analyse different AI biases through examples drawn from the legal domain or closely relevant to it. Here, we propose to distinguish between two types of AI bias, namely: bias *in* AI and bias *towards* AI. Biases in AI arise from flaws or limitations in the training data or the algorithm design or from human influence or emerge autonomously from the algorithm, absent direct human agency. Bias *towards* AI occurs when human decision-makers develop skewed perceptions or inadequately and systematically rely on specific AI tools. Put differently, bias in AI occurs at the AI design and development stages, while bias towards AI materializes during deployment and use.

Biases in AI

Bias in AI can occur because of AI input concerns, specifically due to biases in the data on which the AI is trained. For instance, decision-making processes within law enforcement increasingly rely on intelligence derived from large and complex datasets that predict where and when crimes are most likely to occur.[8] Called *predictive policing*, that software type calculates the probability of crimes being committed in certain areas or by certain categories of individuals based on data about arrests done by the police in the past. The problem here is that past arrest data can reflect various (e.g., sexist, classist, racist) human biases held by the police towards members of specific groups—a phenomenon often referred to as historical bias. This can lead to systematically inaccurate predictions about the likelihood of individuals from those groups committing offenses or crimes occurring in areas where they disproportionately reside.[9]

A different input problem concerns missing data rather than bias in data collection. For instance, face recognition algorithms have been found to be systematically less accurate for darker skin tones than for lighter ones. One salient explanation for this biased output is the underrepresentation of darker skin tones in the algorithm's training data—a form of representation bias: when used for criminal suspect identification purposes, these tools disproportionately result in higher false positives for darker skinned individuals compared to their lighter skinned counterparts.[10]

In both examples, AI outputs (crime predictions, suspect face identification scores) are biased against certain categories of the population because of data input concerns. The norm that these outputs systematically depart from is an epistemic one (i.e., accuracy) and, possibly, a moral one (i.e., the equal respect due to individual members of the relevant demographic groups). Both examples illustrate how machine-learning AI tools can reproduce human biases embedded in the input or generate new biases due to data gaps. We call this *data bias*.

However, biased AI output can also stem from the AI's human-programmed rules of inference rather than its data. This can happen with (fully or partly) rule-based tools where human algorithm developers decide which variables to compute and what weight to assign to them (i.e., how much a variable matters for deciding an outcome). One (infamous) example here is the *Correctional Offender Management Profiling for Alternative Sanctions* (COMPAS) software, which predicts criminal defendants' offense and flight risks from legal proceedings. Initially developed to assist bail decisions, COMPAS is now increasingly used for sentencing purposes across various US jurisdictions. Despite their purported success, studies have found that COMPAS is

biased, with a false positive rate twice as high for black defendants, compared to white defendants.[11] Like in the previous examples, one explanation could be biased or incomplete data. However, another equally potent explanation pertains to the human choices regarding the variables included in the survey that generates the data for the algorithm *and* the weightings assigned to these variables to predict recidivism scores. We call this *programming bias*, which may encompass subtypes such as measurement bias (bias in the features selected or how they are collected), aggregation bias (using overly general models that ignore subgroup differences), and learning bias (biases introduced by the model choice or optimization procedure).

Finally, biased AI output might not be substantially driven by data input problems or human programmers' biases. Instead, they might be inherent to the structure of the algorithm—a phenomenon we refer to as *emergent algorithmic bias*, building on ideas of outcomes arising from complex model interactions. Emergent algorithmic biases are unintended consequences that arise from the complex interactions within machine-learning systems, even when trained on robust training data. These biases manifest as algorithms autonomously process and adapt to information, leading to unforeseen and often undesirable outcomes.[12] For instance, machine-learning models often fail to distinguish between causally relevant and incidental patterns, leading them to rely on spurious correlations as shortcuts for performing tasks. This issue is particularly concerning in sensitive machine-learning AI in the medical domain, where models trained for diagnostic or procedural assessment may inadvertently focus on irrelevant features rather than clinically meaningful cues. A recent study on video-based AI for surgical skill assessment finds that an AI system designed to evaluate procedural performance could rely on unreliable temporal features within video frames rather than actual markers of surgical proficiency.[13] Specifically, the model may be latching onto superficial visual cues, such as lighting changes, camera angles, or background motion, to generate skill evaluations. A related study further highlights this issue in medical AI applications.[14] It shows how AI systems trained for clinical decision-making can become overly reliant on artefacts in medical images, such as hospital-specific scanning techniques or non-diagnostic visual markers, rather than on actual pathological features. Such spurious correlations reflect a broader challenge in AI design: machine-learning models often optimize for predictive accuracy based on patterns in the training data without distinguishing between causation and mere correlation. This can significantly compromise an AI system's accuracy, reliability, and generalizability.

While we focus here on biases originating during the design and development stages, it is important to note that some forms of bias may emerge during deployment as AI systems interact with users and environments. For instance, feedback bias can arise when models learn from user-generated input that reflects pre-existing or contextual biases. Similarly, hardware limitations (e.g., in sensors or imaging devices) can introduce bias into AI outputs, particularly in embodied systems like robotics or autonomous vehicles. While these cases blur the line between development and deployment, they do not compromise the cogency of our distinction between bias in AI and bias towards AI, which is meant analytically, not dichotomously.[15]

A related concern worth noting is bias that arises when AI systems are deployed in contexts that differ significantly from those they were designed or trained for. For example, applying an AI model developed in one country's legal or economic context to a different jurisdiction can lead to outcomes that are inappropriate or harmful. While our typology distinguishes bias in AI (development) from bias towards AI (use), we recognize that mismatches between development and deployment contexts introduce distinctive risks that deserve separate attention. We also recognize that problematic deployment dynamics might problematically affect earlier stages in the life cycle of an algorithm (such as data collection, data processing, and algorithmic design).

Taking stock, there are three types (or sources) of bias in AI: (1) *data bias*, which includes both historical biases (where prejudiced patterns are encoded into training datasets) and representation biases (where certain groups are systematically underrepresented or excluded); (2) *programming bias* (when human choices about inference rules, variables, and weightings introduce disparities in decision outcomes), and (3) *emergent algorithmic bias* (when AI systems develop biases due to the structure of their adaptive learning processes in interactions with users). Following Friedman and Nissenbaum, emergent bias arises *ex post* 'only in a context of use' – meaning

when idiosyncratic or highly contextual features of the (otherwise unbiased or not significantly biased) training dataset *become* a source of bias when the algorithm generates algorithmic outputs that inform decisions about actual decisions.[16] The key takeaway here is that only one form of bias in AI—emergent algorithmic bias—can be attributed to artificial agents as such. The other forms of bias arise from either flawed training data or the broader socio-psychological dynamics of human involvement in AI design.

Biases towards AI

Biases in AI can become legally objectionable (and, sometimes, legally actionable) if decision processes within the legal domain are entirely delegated to AI agents. However, this is rarely the case: in most instances, AI systems are designed and implemented with a 'human-in-the-loop' safeguard, ensuring that no artificial agents possess exclusive legitimate authority over decisions, especially those that affect individuals. More commonly, biases in AI influence legally relevant decisions because *human* decision-makers trust or rely on AI tools. Put differently, biases in AI can become particularly problematic when humans make biased decisions *as a result of* AI recommendations—namely, when human individuals adopt beliefs, attitudes, or behaviours towards AI systems that systematically deviate from genuine norms, a process that can happen without the awareness of the human decision-makers.

A salient type of bias towards AI is *automation bias*, broadly defined as the human 'tendency to use automated cues as a heuristic replacement for vigilant information seeking and processing.'[17] More recently, it has been described as 'an attitude in which the operator of an autonomous system will defer to its outputs to the point where they overlook or ignore evidence that the system is failing.'[18] Under this definition, automation bias directly connects to epistemic norms—particularly, evidence responsiveness. However, automation bias can also be interpreted as a violation of a legal norm, namely statutory discretion. According to this norm, only certain human decision-makers have decisional discretion in the legal domain, meaning that excessive deference to AI systems risks either restricting discretion (fettering discretion) or improperly transferring decision-making power to AI (illegitimate authority delegation).[19]

An example of automation bias in safety-critical scenarios comes from studies on human reliance on robots during emergencies. Research from Georgia Tech found that during simulated fire evacuations, participants followed a robot's instructions—even when the robot had previously demonstrated unreliable behaviour, such as leading them in the wrong direction or displaying error messages.[20] This suggests that in high-stress environments, people may default to trusting automated agents over their own judgement, even when clear evidence of the system's fallibility exists. Similarly, automation bias can affect legal professionals who rely on AI-driven decision aids. Studies on automation in search-and-rescue contexts show that users' trust and reliance on AI increase with the system's perceived reliability, even when errors are embedded into its recommendations.[21] This has significant implications for legal AI tools, where deference to algorithmic outputs could lead to systematic errors in risk assessment, sentencing, or evidence evaluation.

Bias towards AI can also cut in a direction opposite to automation bias, for instance, when people distrust or avoid algorithms even when they are demonstrably superior to human judgement. This is *algorithmic aversion*, a tendency to resist using algorithms or integrating their output into decisions. There is no theoretically salient account addressing the legal implications of algorithmic aversion, but its effects can be inferred from research in human–computer interaction, behavioural economics, and law. People are prone to avoiding algorithms after witnessing or learning that they make errors—even when the overall accuracy of the AI remains superior to human judgement.[22] Depending on the context of implementation, this tendency could be extremely problematic. For example, in healthcare, doctors might resist AI diagnostic tools despite their proven accuracy, potentially leading to suboptimal patient care or malpractice risks if human oversight fails to catch what the AI would have identified.

On one possible explanation, algorithmic aversion merely reinstantiates other (widely spread) human biases. For instance, status quo bias may lead decision-makers to favour familiar, human-led processes over AI-driven alternatives, even when the latter is demonstrably more effective. Likewise, confirmation bias may cause individuals to reject algorithmic outputs that contradict their pre-existing beliefs—or, conversely, to accept outputs that align with their views without sufficient scrutiny. Tendencies to resist changing entrenched decisional procedures or habits can prompt systematic departures from accuracy or reliability norms.[23] In employment, for instance, hiring managers might override algorithmic resume screening, believing that they are better at recruiting, and thus unintentionally lead to discriminatory hiring patterns. If these decisions result in a workforce that is demonstrably less diverse, legal liability could arise—even if the initial motivation was algorithmic aversion rather than explicit bias.

Importantly, unlike automation bias, not all forms of algorithmic aversion are problematic. In some cases, scepticism towards AI may be justified—for instance, when decision-makers have legitimate concerns and evidence about an algorithm's poor performance. Given this distinction, it might be fitting to propose a new term—such as *anti-automation bias*—to distinguish between cases where algorithmic aversion leads to irrational rejection of AI and those where scepticism is warranted. For example, studies have shown that even when algorithms outperform human decision-makers on average, individuals may abandon them after witnessing a single error. In one experiment, participants preferred human forecasts over algorithmic ones—even after being shown that the algorithm was more accurate overall. The term 'anti-automation bias' would highlight the symmetrical contrast with automation bias, capturing the valence between excessive trust in AI and unwarranted scepticism towards it.[24]

Summary

Taking stock, we identified two types of AI bias: bias in AI and bias towards AI, with each type covering different specific subtypes of bias (*see* Table 1).

Table 1

Typology of AI biases

Types of AI bias			**Attributable to**
Bias in AI (design and development)	Data bias (historical bias, representation bias)	Incorrect data	Humans
		Missing	Humans
	Algorithmic bias	Programming bias (human-based choices)	Humans
		Emerging bias (algorithm inner structure)	AI and human-in-the-loop
Bias towards AI (deployment and use)	Automation bias	Objectionable human biases (overreliance)	Humans
	Algorithm aversion	Objectionable human biases (avoidance)	Humans
		Anti-automation biases (scepticism)	Humans

Based on this distinction, we can venture some legally significant observations. First, there is only one subtype of AI bias—the emergent algorithmic bias we discuss in section 'Biases in AI'—that does not flow from or instantiate an already-existing *human* bias. This has implications for how we determine and allocate (moral or legal) responsibility for AI bias. Specifically, it gives us reasons to be sceptical about claims that attribute most or all of the responsibility for biased decisions to artificial agents themselves.[25] Moreover, some biases may be inherent to the machine-learning paradigm itself, given that choices about data, variables, and modelling procedures are unavoidable and may introduce structural imperfections, even absent explicit human bias. Second, and insofar as most biases in AI are distinctly human, there is an argument for using AI systems more for *monitoring* purposes—namely, to assess the scope, frequency, and magnitude of our biases—rather than for decision-making or decision-taking purposes.[26] Third, and finally, the distinction between biases in AI and biases towards AI suggests that the former are particularly problematic *if* the latter occur. For instance, absent automation bias, bias in AI would arguably have less impact on the decision taken by legal professionals (or people in general). If this observation is correct, strategies for countering AI bias should prioritize tackling bias towards AI—provided that robust human oversight is in place. Moreover, attention to the deployment context is crucial, as mismatches between an AI system's design assumptions and its actual use environment can significantly amplify risks.

Countering AI bias: preventive and corrective strategies

AI bias presents serious challenges across legal, social, and professional domains.[27] Countering AI biases is important if we care about the norms from which they depart. Given our human fallibility, we cannot eliminate AI bias entirely, but we can improve our efforts to reduce its risks and/or shrink its scope.[28] To mitigate these risks, two complementary approaches should be pursued: preventive strategies, which aim to minimize bias before AI systems are deployed, and corrective strategies, which focus on auditing and adjusting AI systems after biases have been identified to mitigate or contain their impact.

Preventive strategies

Preventing AI bias at the design stage is the most effective way to reduce discriminatory or distorted outcomes. Regulatory frameworks, technical interventions, and structured evaluation processes can help ensure that AI systems are fair and reliable before they are deployed in high-stakes settings.

One promising approach is regulatory sandboxes, controlled environments where AI models are tested under realistic conditions before full deployment. The European Union's AIA mandates in Art. 57 that all member states establish at least one AI regulatory sandbox by 2026, allowing policymakers and external auditors to assess AI risks in law enforcement, hiring, and healthcare before implementing these systems.[29] However, sandboxes do not operate in a vacuum; their effectiveness depends on the robustness of the training data, the inclusiveness of the testing environments, and the extent to which assessment standards account for diversity. This threefold attention is crucial because legal and technical frameworks themselves may be incomplete or outdated, especially given the novelty of AI practices. Standards often fail to account for diversity, as seen in examples like ISO 13482:2014 for service robotics, or medical devices that inadequately consider gender differences, leading to disproportionate harms for women.[30] If datasets, testing zones, and regulatory framework lack diversity, biases may persist despite pre-deployment testing.[31]

Another preventive measure involves adopting structured models that help operationalize the consideration of potential biases throughout the AI development life cycle. The High-Level Expert Group on AI, established by the European Commission in 2019, developed the Assessment List for Trustworthy AI.[32] This self-assessment checklist guides AI developers and deployers in implementing the seven key requirements for trustworthy AI: human agency and oversight, technical robustness and safety, privacy and data governance, transparency,

diversity, non-discrimination and fairness, societal and environmental well-being, and accountability. Although not explicitly targeting bias, Assessment List for Trustworthy AI helps organizations identify and mitigate potential biases in AI systems before deployment. Systematically and holistically addressing these aspects could help companies get closer to compliance with other pieces of legislation, mainly the AIA.[33] Additionally, a key preventive measure involves early scoping of impacted stakeholders—identifying who may be directly or indirectly affected by an AI system, and ensuring that development and testing processes reflect the diversity and norms of the deployment context.

Proactively addressing bias in natural language processing models should be part of a preventive AI governance strategy, ensuring that AI-driven decision-making tools do not systematically disadvantage individuals based on language, gender, or ethnicity. In that sense, given that language-based AI systems can reinforce and perpetuate societal biases, particularly when trained on historical data that reflects pre-existing stereotypes. Research has shown, however, that most bias detection and mitigation techniques focus on English-language models, leaving bias in non-English AI applications largely underexplored.[34] Puttick and others emphasize that 'biases in non-English AI applications remain critically underexamined, leading to significant fairness and accountability concerns when these models are deployed across diverse linguistic and cultural contexts.' This raises linguistic and cultural fairness concerns, as AI systems deployed in multilingual and diverse environments may operate with hidden biases that disadvantage certain demographic groups.

Corrective strategies

Concerning corrective strategies, one important instrument is *monitoring*, which can be done through the adoption of auditing laws that require public or private organizations to constantly check for and publicize data about AI-assisted decisions. Art. 72 EU AIA mandates that providers of high-risk AI systems implement and maintain a post-market monitoring system proportionate to the AI technology's nature and the specific risks associated with its deployment. Providers are required to document their monitoring activities to ensure ongoing compliance with regulatory standards. Here, the thought is that the harms that stem from AI bias—for instance, errors in identifying suspects or in allocating criminal sanctions or jobs—can only be corrected if they are adequately documented.[35] Absent audit laws, another important strategy is *litigation*: lawyers could be more active in bringing (individual or class-action) cases before the court for situations where they can prove either bias in AI (in particular, biases in the algorithm rules decided by humans) or bias towards AI (in particular, automation bias, which can trigger legal liability in cases of fettering discretion). A third corrective strategy concerns the *suspension* of AI applications that have been found or are reasonably thought to have a discriminatory (because biased) impact. Specifically, legally authorized agents (be they courts, parliaments, or, sometimes, heads of the executive) could impose moratoria that temporarily prohibit using AI systems until the source of bias is appropriately addressed—for instance, by improving the datasets of algorithms that underrepresent certain skin colours or tones for artificial face recognition systems. Note that suspension (moratorium) strategies are controversial. This is because they can lead to exploitative dataset improvement processes. One infamous example here is that of Google paying homeless people an exploitative sum of money to have their facial data collected to improve its data gap in black faces.[36]

Conclusions

AI bias is a multifaceted phenomenon that requires careful analysis to understand its implications for law, ethics, and decision-making. A key distinction must be drawn between biases in AI systems and biases towards AI systems. Importantly, only one form of bias—bias inherent to an AI system's algorithm structure—can be attributed to AI as such. Other biases stem either from the data used to train AI models or from human interactions with AI outputs, making AI more of a mirror of human biases rather than their independent source. This does not mean that emergent AI biases should concern us less, but it does mean that we should be sceptical when confronted with claims or theories that agonize about the difficulty of construing or closing alleged 'responsibility gaps' for AI-related biases. At an extreme, we should be sceptical about 'agency laundering' attempts to attach responsibility for AI biases to AI systems instead of the humans who design or use them.[37] Finally, the fact that most types of AI biases are human rather than artificial does not imply that AI systems are infallible when it comes to satisfying relevant norms; it only implies that their fallibility might often be *ours* and that, whenever that is the case, we should more readily (and responsibly) own up to it. Nor should our analysis be understood as a claim about the *frequency* of some AI biases as compared to others. Rather, it is a claim about the scope (or types) of AI biases that might occur in the design and deployment of AI systems.

Further reading

Custers, B, and E Fosch-Villaronga (eds), *Law and Artificial Intelligence: Regulating AI and Applying AI in Legal Practice*, vol 35 (Springer Nature 2022).
Google Scholar Google Preview WorldCat COPAC

Forrest, KB, *When Machines Can Be Judge, Jury, and Executioner: Justice in the Age of Artificial Intelligence* (World Scientific Publishing 2021).
Google Scholar Google Preview WorldCat COPAC

Fosch-Villaronga, E, and A Poulsen, 'Diversity and Inclusion in Artificial Intelligence' in B Custers and E Fosch-Villaronga (eds), *Law and Artificial Intelligence: Regulating AI and Applying AI in Legal Practice* (2022).
Google Scholar Google Preview WorldCat COPAC

Rigotti, C, and E Fosch-Villaronga, 'Fairness, AI & Recruitment' (2024) 53 *Computer Law & Security Review* 105966.
Google Scholar WorldCat

Vydra, S, A Poama, S Giest, A Ingrams, and B Klievink, 'Big Data Ethics: A Life Cycle Perspective' (2021) 14 *Erasmus Law Review* 24.
Google Scholar WorldCat

1 *See* Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonized rules on AI and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139, and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797, and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance), <https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng> accessed 20 May 2025.

2 For an analysis of the distinction between predictive and generative AI, see A Narayanan and S Kapoor, *AI Snake Oil: What Artificial Intelligence Can Do, What It Can't, and How to Tell the Difference* (Princeton University Press 2024).

3 See the distinction between rule-based and machine-learning AI at the end of the paragraph above.

4 A Caliskan, JJ Bryson, and A Narayanan, 'Semantics Derived Automatically from Language Corpora Contain Human-Like Biases' (2017) 356(6334) Science 183.

5 T Kelly, *Bias: A Philosophical Study* (Oxford University Press 2022) 63. For a defence of the functional account, see GM Johnson, 'Varieties of Bias' (2024) 19(7) Philosophy Compass e13011).

6 See Kelly (n 5) 63.

7 For a philosophical analysis of implicit bias, see J Holroyd, R Scaife, and T Stafford, 'What Is Implicit Bias?' (2017) 12(10) Philosophy Compass e12437.

8 *See* Europol Innovation Lab, 'Observatory Report' (2024) <www.europol.europa.eu/cms/sites/default/files/documents/AI-and-policing.pdf> accessed 15 January 2025.

9 For a theoretical discussion of the biases at play in AI-based police predictions and an ethnographic study of their actual use by police forces, see S Brayne, *Predict and Surveil: Data, Discretion, and the Future of Policing* (Oxford University Press 2021).

10 On this, see especially, J Buolamwini and T Gebru, 'Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification' (Conference on Fairness, Accountability and Transparency, PMLR, New York, January 2018).

11 For an empirical analysis of the biases at play in the design of COMPAS, see J Angwin, J Larson, S Mattu, and L Kirchner, 'Machine Bias' in K Martin (ed), *Ethics of Data and Analytics* (Auerbach Publications 2022). For a normative analysis of the wrong-making features of these biases, see C Castro, 'What's Wrong with Machine Bias' (2019) 6(15) Ergo: An Open Access Journal of Philosophy 405.

12 RV Yampolskiy, *AI: Unexplainable, Unpredictable, Uncontrollable* (CRC Press 2024).

13 D Kiyasseh, J Laca, TF Haque, M Otiato, BJ Miles, C Wagner, and others, 'Human Visual Explanations Mitigate Bias in AI-Based Assessment of Surgeon Skills' (2023) 6(1) NPJ Digital Medicine 54.

14 AJ DeGrave, JD Janizek, and SI Lee, 'AI for Radiographic COVID-19 Detection Selects Shortcuts over Signal' (2021) 3(7) Nature Machine Intelligence 610.

15 For an analysis of the distinction between analytical distinctions and dichotomies, see H Putnam, *The Collapse of the Fact/Value Dichotomy and Other Essays* (Harvard University Press 2004).

16 See B Friedman and H Nissenbaum, 'Bias in Computer Systems' (1996) 14(3) ACM Transactions on Information Systems (TOIS) 336.

17 KL Mosier and LJ Skitka, 'Human Decision Makers and Automated Decision Aids: Made for Each Other?' in R Parasuraman and M Mouloua (eds), *Automation and Human Performance* (CRC Press 2018).

18 J Zerilli, I Goñi, and MM Placci, *Automation Bias and Procedural Fairness: A Short Guide for the Public Sector* (Zenodo 2024) <https://zenodo.org/records/13132781> accessed 20 May 2025.

19 Zerilli, Goñi, and Placci (n 20) 5.

20 *See* GaTech, *Emergencies: Should You Trust a Robot?* (Georgia Tech News Center 2016) <https://news.gatech.edu/news/2016/02/29/emergencies-should-you-trust-robot> accessed 20 May 2025.

21 *See* JM Ross, JL Szalma, PA Hancock, JS Barnett, and G Taylor, 'The Effect of Automation Reliability on User Automation Trust and Reliance in a Search-and-Rescue Scenario' (2008) 52(19) Proceedings of the Human Factors and Ergonomics Society Annual Meeting 1340.

22 *See* BJ Dietvorst, JP Simmons, and C Massey, 'Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err' (2015) 144(1) Journal of Experimental Psychology: General 114.

23 See CR Sunstein and JH Gaffe, 'An Anatomy of Algorithm Aversion' (2024) 26 Columbia Science & Technology Law Review 290.

24 For an empirical analysis of these biases in the public sector, S Alon-Barkat and M Busuioc, 'Human–AI Interactions in Public Sector Decision Making: "Automation Bias" and "Selective Adherence" to Algorithmic Advice' (2023) 33(1) Journal of Public Administration Research and Theory 153. Alon-Barkat and Busuioc do not find evidence of automation bias but find some evidence of 'selective adherence' (what we here call anti-automation bias).

25 For a normative discussion of such (psychologically plausible) phenomena, see A Rubel, C Castro, and A Pham, 'Agency Laundering and Information Technologies' (2019) 22(4) Ethical Theory and Moral Practice 1017.

26 This is briefly suggested by Sunstein in relation to non-cognitive biases (for instance, sexist, racist or classist biases) in CR Sunstein, 'Algorithms, Correcting Biases' (2019) 86(2) Social Research: An International Quarterly 499.

27 E Fosch-Villaronga and A Poulsen, 'Diversity and Inclusion in Artificial Intelligence' in B Custers and E Fosch-Villaronga (eds), *Law and Artificial Intelligence: Regulating AI and Applying AI in Legal Practice* (Springer 2022).

28 S Townson, 'Manage AI Bias Instead of Trying to Eliminate It' (2023) 64(2) MIT Sloan Management Review 1.

29 T Buocz, S Pfotenhauer, and I Eisenberger, 'Regulatory Sandboxes in the AI Act: Reconciling Innovation and Safety?' (2023) 15(2) Law, Innovation and Technology 357.

30 E Fosch-Villaronga and H Drukarch, 'Accounting for Diversity in Robot Design, Testbeds, and Safety Standardization' (2023) 15(11) International Journal of Social Robotics 1871.

31 K Prifti and E Fosch-Villaronga, 'Towards Experimental Standardization for AI Governance in the EU' (2024) 52 Computer Law & Security Review 105959.

32 *See* P Ala-Pietilä, Y Bonnet, U Bergmann, M Bielikova, C Bonefeld-Dahl, W Bauer, and others, *The Assessment List for Trustworthy Artificial Intelligence (ALTAI)* (European Commission 2020) <https://altai.insight-centre.org/> accessed 27 May 2025.

33 A Fedele, C Punzi, and S Tramacere, 'The ALTAI Checklist as a Tool to Assess Ethical and Legal Implications for a Trustworthy AI Development in Education' (2024) 53 Computer Law & Security Review 105986.

34 A Puttick, L Rankwiler, C Ikae, and M Kurpicz-Briki, 'The BIAS Detection Framework: Bias Detection in Word Embeddings and Language Models for European Languages' (*arXiv* 2024).

35 For a discussion of the effectiveness of AI audit laws, see A Hilliard, A Gulley, A Koshiyama, and E Kazim, 'Bias Audit Laws: How Effective Are They at Preventing Bias in Automated Employment Decision Tools?' [2024] International Review of Law, Computers & Technology 1.

36 For a critical discussion of this case, see S Lazar, 'Legitimacy, Authority, and the Political Value of Explanations' (*arXiv* 2022).

37 See M Da Silva, 'Responsibility Gaps' (2024) 19(9–10) Philosophy Compass e70002; A Rubel, C Castro, and A Pham, 'Agency Laundering and Information Technologies' (2019) 22(4) Ethical Theory and Moral Practice 1017.