Routledge
Taylor & Francis Group

Judicial Review

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/rjdr20

Automation Bias and the Principles of Judicial Review

Tatiana Kazim & Joe Tomlinson

To cite this article: Tatiana Kazim & Joe Tomlinson (2023): Automation Bias and the Principles of Judicial Review, Judicial Review, DOI: 10.1080/10854681.2023.2189405

To link to this article: https://doi.org/10.1080/10854681.2023.2189405

Published online: 19 Apr 2023.

JUDICIAL REVIEW
https://doi.org/10.1080/10854681.2023.2189405

Routledge
Taylor & Francis Group

Automation Bias and the Principles of Judicial Review

Tatiana Kazim[a] and Joe Tomlinson[b]

[a]Public Law Project; [b]York Law School, University of York

1. Though much of it remains out of sight, it is no secret that the use of automated decision-making systems in government is now widespread and growing. In practice, many such systems are partially, rather than fully, automated; there is a 'human in the loop'. In such a system, there are a range of downstream risks in respect of the quality of administrative decision-making. One such potential risk is automation bias – a well-documented psychological phenomenon whereby decision-makers put too much, or inappropriate, trust in computers and thus effectively abdicate to some extent their own discretionary judgement. The question of how judicial review may respond to this risk is on the horizon.

2. This article reviews what the existing evidence base tells us about where the risk of automation bias arises, as well as its nature and extent. It further explores the principles of administrative law which – though not yet applied by the courts in practice – may be called upon to regulate this risk.

What is automation bias and when does it arise?

3. Automation bias is a form of cognitive bias. It is about the way a human decision-maker interacts with an automated system – the research literature tends to refer to the 'human-automation team' – and how decision effectiveness may be undermined by automation bias. It was first identified as a phenomenon in cockpit crews and seen as a 'decision short-cut'.[1] In 1996, it was defined by Mosier and Skitka as the human 'tendency to use automated cues as a heuristic replacement for vigilant information seeking and processing'.[2] This definition is still used widely.

4. In administrative decision-making systems, automation bias is most relevant in the context of partial automation, where there is an 'official in the loop' working alongside an automated process. By way of example, an increasingly common scenario is where

[1]Earl L Weiner, 'Beyond the Sterile Cockpit' (1985) 27 *Human Factors* 75.
[2]Kathleen L Mosier and Linda J Skitka, 'Human Decision Makers and Automated Decision Aids: Made for Each Other?' in Raja Parasuraman and Mustapha Mouloua (eds), *Automation and Human Performance: Theory and Applications* (CRC Press 1996) 205.

an automated system directs a streaming or triage system, which determines the type or quality of human judgement required in a particular case.

5. Until recently, the evidence base on automation bias in decision-making had not emerged in the context of public administration. Instead, the existing studies spanned multiple fields where decision-making is salient. Aviation was the original site of research into automation bias, and many of the more recent studies also took place in an aviation context. Other prominent research fields where studies have been undertaken include bail decision-making, healthcare and the military. An extensive review of 74 papers conducted by Goddard, Roudsari and Wyatt in 2012 found that automation bias 'appears to be a fairly robust and generic effect across research fields'.[3] However, human interactions with automated systems are not always marked by automation bias. In 2003, Dzindolet and others observed under-utilisation, or 'disuse', of automated systems alongside over-reliance, or 'misuse'.[4]

6. More recently, Alon-Barkat and Busuioc published the first paper on automation bias and selective adherence in a public administration context.[5] They conducted three studies between February 2020 and February 2021, all taking place in the Netherlands, with an aggregated sample of 2,854 participants. In study 1, participants were asked to act as hypothetical school board members making decisions about whether to renew teacher contracts. Each participant was given two inputs: a qualitative evaluation by the human resources person of the educational association, and a numeric prediction of the teachers' potential to perform well in the future. Participants were divided into two groups. One group was told that the numeric prediction was generated by an algorithm; the other group was told that it was generated by a human being. Study 2 was identical, except for one modification designed to test 'selective adherence', that is, a decision-maker's tendency to defer to the algorithm when its predictions match pre-existing stereotypes. To test for this, the teachers' names were manipulated as a cue for their ethnic identity. In one group, all three teachers were given traditionally Dutch names. In the other group, the teacher with the lowest numeric prediction was given a traditionally Moroccan name. Study 3 replicated the previous two, but with a sample of civil servants.

7. Alon-Barkat and Busuioc found that in none of the three studies were participants more likely to follow the numeric prediction when generated by an algorithm compared to a human expert. In studies 1 and 2, the differences were small and statistically insignificant. In study 3, participants were actually less likely to follow the algorithmic advice compared to human advice. However, it is important to note that the results of

[3]Kate Goddard, Abdul Roudsari and Jeremy C Wyatt, 'Automation Bias: A Systematic Review of Frequency, Effect Mediators, and Mitigators' (2012) 19 *Journal of the American Medical Informatics Association* 121, 123.
[4]Mary T Dzindolet and others, 'The Role of Trust in Automation Reliance' (2003) 58 *International Journal of Human–Computer Studies* 697.
[5]Saar Alon-Barkat and Madalina Busuioc, 'Human–AI Interactions in Public Sector Decision Making: "Automation Bias" and "Selective Adherence" to Algorithmic Advice' (2023) 33 *Journal of Public Administration Research and Theory* 153.

study 3 are attributed to the very high-profile childcare benefits scandal, involving reliance by the Dutch tax authorities on an AI algorithm that used nationality (amongst other criteria) to flag high-risk benefits applicants for further scrutiny. This scandal received widespread news coverage shortly before Alon-Barkat and Busuioc's third study was carried out and they take the results to be indicative of a growing reluctance to trust algorithms in the scandal's aftermath. As for studies 1 and 2, Alon-Barkat and Busuioc say:

> One possible explanation for this discrepancy is a relative skepticism about the performative capacity of AI algorithms, with many participants, based on their self-reporting, still under-exposed to their performative capacities (in studies 1 and 2), or exposed to their negative consequences (in study 3, following the benefits scandal). This is an important difference to earlier studies on automation applied in areas well-accustomed to such devices (aviation, healthcare), characterized by routine use of reliable automation, resulting in high levels of trust in their performance.[6]

8. Given that automation bias has, as mentioned above, been found to be a 'robust and generic effect' in a wide range of contexts, including bail decision-making, this suggests that it could also arise in the context of administrative decision-making. Although automation bias did not seem to be significant in Alon-Barkat and Busuioc's study, it may be significant in situations where officials are accustomed to using AI and have overcome any initial scepticism, and have not been exposed to any high-profile AI-related scandal.

9. Overall, the evidence base suggests that automation bias could arise in the context of administrative decision-making as well as in other decision-making contexts. However, it cannot be presumed that the presence of an automated system working alongside a human official necessarily leads to actual automation bias. A material question, therefore, is what conditions affect the prevalence and extent of automation bias.

What factors affect the prevalence and extent of automation bias?

10. 'Effect mediators' is a term used by Goddard, Roudsari and Wyatt to describe factors affecting the presence and extent of automation bias.[7] There are several strands of research on effect mediators. Four prominent effect mediators are noteworthy as regards administrative decision-making.

11. First, the matter of trust appears to be highly relevant. Definitions of trust vary in this context. Trust can be understood in behavioural terms (e.g. a tendency on the part of the 'truster' to be vulnerable to the 'trustee'). Perhaps more commonly, however, it is understood as belief or attitude (e.g. that the trustee will help the truster to achieve their goals, or belief in the ability, integrity and benevolence of the trustee). Dzindolet

[6]ibid 165.
[7]Goddard, Roudsari and Wyatt (n 3).

and others found that the default position is trust in automation. Before any experience with the automated support system, participants in their study initially considered it to be trustworthy and reliable. This initial position changed depending on their interactions with the system. After observing that the automated system make errors, participants distrusted even reliable aids, unless an explanation was provided regarding why the aid might err. Knowing why the automated system might err increased trust in the aid and increased automation reliance – even when the trust was unwarranted.

12. In 2015, Hoff and Bashir carried out a review of the empirical research on factors influencing trust in automation.[8] They devised a three-layer model of trust in automation, comprising:
    (1) dispositional trust, i.e. an individual's overall tendency to trust automation, independent of context or specific system. Variables affecting dispositional trust include culture, age, gender and personality;
    (2) situational trust, i.e. an individual's tendency to trust automation in a given situation. Variables fall into two categories: the external environment; and the internal, context-dependent characteristics of the operator; and
    (3) learned trust, i.e. an individual's tendency to trust automation depending on their experience with it.

13. A second apparently salient factor is cognitive load. Parasuraman and Manzey propose that automation bias occurs when multiple tasks compete for the system-user's attention.[9] However, Lyell and Coiera found that automation bias can also arise in single-task situations.[10] This was more likely to be the case where there was a high degree of 'verification complexity' – complexity in verifying that automation is performing correctly. They also suggest that the complexity of the task itself can increase the likelihood of automation bias. They suggest that cognitive load theory may be able to explain both their findings and the findings of Parasuraman and Manzey. Cognitive load can be increased by the complexity of a task or the number of tasks. The higher the cognitive load, the more likely automation bias is to arise. So, automation bias may arise when there are multiple tasks competing for a user's attention or when the user is carrying out a single task, and the task itself is complex and/or verifying the information provided by the automated system is complex.[11] The influence of cognitive load appears to be independent of the influence of trust: in high cognitive load situations, the user may over-rely on the automated system, even if their trust in it

[8]Kevin Anthony Hoff and Masooda Bashir, 'Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust' (2015) 57 *Human Factors* 407.
[9]Raja Parasuraman and Dietrich Manzey, 'Complacency and Bias in Human Use of Automation: An Attentional Integration' (2010) 52 *Human Factors* 381.
[10]David Lyell and Enrico Coiera, 'Automation Bias and Verification Complexity' (2017) 24 *Journal of the American Medical Informatics Association* 423.
[11]ibid.

is low. It is important to note, however, that the more recent work of Lyell, Magrabi and Coiera suggests that 'underload' can also increase the effects of automation bias: if cognitive load is too low, this may lead to allocation of insufficient cognitive effort to the task, thereby increasing reliance on automation.[12] This may suggest that there is a cognitive load 'sweet spot' that will optimise the effectiveness of the human-automation team.

14. A third factor is the consequences of the decision. Studies suggest that 'criticality' is an important predictor of omission errors.[13] If an omission error is more critical, a decision-maker may be more likely to examine the output of an automated system more closely.

15. A fourth factor is the influence of pre-existing stereotypes. Cowgill's research and Albright's research, both in the context of judicial bail decisions aided by algorithmic risk scores, suggest that the degree of reliance placed in automation can vary depending on racial bias.[14] Albright found that the algorithm's recommendation was more likely to be overridden in favour of harsher bond conditions for black defendants than similar white defendants. Albright gives two reasons for this overall effect.[15] First, it can be explained at the county level: judges in 'whiter' counties were more receptive to the new practice than judges in 'blacker' counties. Second, it can be explained at an individual level: individual judges were more likely to deviate from the recommended bond condition for moderate-risk black defendants than for similar moderate-risk white defendants. This suggests that automation bias may interact with pre-existing stereotypes in the following way: where algorithmic advice departs from pre-existing stereotypes, decision-makers may be more likely to override it; but where algorithmic advice conforms to pre-existing stereotypes, they may be more likely to conform to it. The significance of pre-existing stereotypes is supported by Alon-Barkat and Busuioc's research.[16] While the authors did not find evidence of automation bias, they did find evidence of 'selective adherence'. In study 2, they found that when a low prediction score was assigned to a teacher from a negatively stereotyped ethnic minority, participants were significantly more likely to rely on it in their decisions and less likely to override it.

[12]David Lyell, Farah Magrabi and Enrico Coiera 'The Effect of Cognitive Load and Task Complexity on Automation Bias in Electronic Prescribing' (2018) 60 *Human Factors* 1008.
[13]Kathleen L Mosier and Dietrich Manzey, 'Humans and Automated Decision Aids: A Match Made in Heaven?' in Mustapha Mouloua and Peter A Hancock (eds), *Human Performance in Automated and Autonomous Systems* (CRC Press 2019).
[14]Bo Cowgill, 'The Impact of Algorithms on Judicial Discretion: Evidence from Regression Discontinuities' (2018) Working Paper <http://www.columbia.edu/~bc2656/papers/RecidAlgo.pdf> accessed 8 February 2022; Alex Albright, 'If You Give a Judge a Risk Score: Evidence from Kentucky Bail Decisions' (2019) Discussion Paper <http://www.law.harvard.edu/programs/olin_center/Prizes/2019-1.pdf> accessed 24 August 2022.
[15]Albright (n 14).
[16]Alon-Barkat and Busuioc (n 5).

How can automation bias have a negative impact on the quality of decisions?

16. It is important to note that automated support can improve decision-making overall and often does.[17] However, our focus is on how automation bias has the capacity to decrease the quality of decision-making and potentially give rise to legal errors. In this context, the existing research provides some helpful distinctions.

17. Mosier and Skitka distinguished between two negative effects of automation bias:[18] first, 'omission errors', which occur 'when decision makers fail to notice problems because an automated aid fails to detect them'; second, 'commission errors', which occur 'when people inappropriately follow an automated decision aid directive or announcement'. In 2010, Parasuraman and Manzey linked the distinction between omission and commission errors to two different functions of automated support systems.[19] Some automated support systems have an alert function, which 'makes the user aware of a situational change that might require action'.[20] Omission errors are related to the alert function – the decision-maker does not respond to a critical feature of their situation because the support system does not alert them as necessary. Other automated support systems have a recommendation function, which 'involves advice on choice and action'.[21] Commission errors are related to the recommendation function – the decision-maker follows the advice of the automated system even though it is incorrect. In a similar vein, others have drawn the line between 'reliance' and 'compliance'.[22]

18. This existing evidence and literature therefore suggest that automation bias could lead to poor-quality administrative decisions in a number of ways. Sometimes, a poor decision could be rooted in bad information provided to an official by an automated system, for example where, because of an error in its design or operation, an automated system fails to take account of a relevant factor when generating a risk score. Relying on the risk score, the human decision-maker also fails to take account of that factor. Or, because of an error in its design or operation, the system takes account of an irrelevant or non-existent factor when generating the risk score. The human decision-maker then relies on the risk score and does not correct for the presence of the irrelevant or non-existent factor. Another scenario may arise where the automated system is providing accurate information, but automation bias still leads to poor decisions, for example where a system was designed to provide one piece of information relevant to a decision and the human decision-maker should take account of other pieces of information, but instead solely or mostly relies on the output of the automated system and overlooks the other relevant

[17]Goddard, Roudsari and Wyatt (n 3).
[18]Mosier and Skitka (n 2) 205.
[19]Parasuraman and Manzey (n 9).
[20]ibid 391.
[21]ibid.
[22]Hoff and Bashir (n 8), relying on Stephen R Dixon and Christopher D Wickens, 'Automation Reliability in Unmanned Aerial Vehicle Control: A Reliance-Compliance Model of Automation Dependence in High Workload' (2006) 48 *Human Factors* 474.

information. Another example is where an automated system is designed to provide a recommendation for one type of decision then the human decision-maker uses this information out of context, to make a different type of decision.

Can administrative law respond to automation bias?

19. Given the potential of automation bias to lower the quality of administrative decisions, an important question is how administrative law may regulate it. We are concerned with judicial review in this article but, as a preliminary matter, it is important to note that there are a range of potential mitigations public authorities may be able to put in place where automation bias could risk negatively affecting decision-making. Broadly, they can try to improve the design and operation of automated support systems (for example through the use of better data) or improve officials' capacity to respond appropriately to them (for example through training). A 'right first time' approach to decision-making remains the best way to comply with principles of administrative law.

20. In terms of administrative law issues that may arise, a number of existing grounds of review could be relevant. By way of example:
    (1) It could be argued that to take into account an automated system's prompt to the exclusion of other factors that ought to affect the decision might see a decision-maker effectively fail to take account of relevant considerations, even though the scope of that doctrine is limited. Alternatively, in cases where a decision-maker uses the output of an automated system out of context, to make a decision that the output was not intended to support, they may be taking account of an irrelevant consideration or making an error of fact).[23]
    (2) In cases where a decision-maker effectively unreflectively follows what an automated system prompts that could be to fetter their discretion.[24]
    (3) Overreliance on an automated prompt – and the information underpinning it – may amount to a breach of the duty of inquiry if it means the decision-maker has not sufficiently aquatinted themselves with the facts relevant to the decisions.[25]

21. This is not necessarily a comprehensive survey: there could well be further ways in which existing grounds relate to automation bias in particular circumstances. The wider point, however, is that administrative law already provides a set of legal principles ostensibly capable of responding to the type of harms automation bias could generate in public administrative systems.

[23] *E v Secretary of State for the Home Department* [2004] EWCA Civ 49, [2004] QB 1044.
[24] *Lavender v Minister of Housing and Local Government* [1970] 1 WLR 1231.
[25] *Secretary of State for Education and Science v Tameside MBC* [1977] AC 1014.

22. The more complex question as regards the grounds of judicial review in this context is whether their application in a particular case requires a court to engage in a detailed review of the alleged existence of automation bias in an administrative process or whether it can resolve the case without interrogating this issue.

23. Some cases may be simple if the latter applies. For instance, if the product of automation bias is that a decision-maker does not take into account a relevant consideration, this will be visible to the subject of the decision, and the argument can be made about the failure to consider an arguably relevant consideration. This could well be done without getting into the details of whether or not the apparent failure was caused by automation bias.

24. A more complex case, however, is where the claim turns on the existence of automation bias in a system, for instance if there was an automated triage system that ordered cases based on apparent risk and all cases flagged as high risk were receiving adverse decisions. Such circumstances may well involve the court interrogating whether decision-makers had become too dependent on the automated signals. This is very complex territory for a court, not only due to the difficulty in accessing evidence but also in how the assessment of that evidence might become highly technical in place. In practice, harmful automation bias is likely to be a subtle practice of a decision-maker or group of decision-makers that evolves over time. While there may be occasions where public officials may openly admit to the infamous line of 'the computer says no', those occasions are likely to be rare. It is much more likely that automation bias is more apparent in administrative data on decision-making patterns, as such data is more likely to reveal where decision-makers are not departing from the prompts of automated systems. This, of course, requires such data to be collected by government, for it to be of sufficient granularity to interrogate for automation bias, and for it to be in the public domain. As public law litigators will know, this is all far from a given in contemporary public administration.