Data & Policy (2024), 6: e60
doi:10.1017/dap.2024.72

DATA FOR POLICY PROCEEDINGS PAPER

Human–machine collaboration for enhanced decision-making in governance

Dirk Van Rooy[1]

[1]Centre for Responsible AI, University of Antwerp, Antwerp, Belgium
Corresponding author: Dirk Van Rooy; Email: Dirk.VanRooy@uantwerpen.be

Received: 03 December 2023; **Revised:** 19 April 2024; **Accepted:** 03 May 2024

Keywords: artificial intelligence; cognitive biases; governance decision-making; human–machine collaboration; policy innovation

Abstract

A detailed exploration is presented of the integration of human–machine collaboration in governance and policy decision-making, against the backdrop of increasing reliance on artificial intelligence (AI) and automation. This exploration focuses on the transformative potential of combining human cognitive strengths with machine computational capabilities, particularly emphasizing the varying levels of automation within this collaboration and their interaction with human cognitive biases. Central to the discussion is the concept of dual-process models, namely Type I and II thinking, and how these cognitive processes are influenced by the integration of AI systems in decision-making. An examination of the implications of these biases at different levels of automation is conducted, ranging from systems offering decision support to those operating fully autonomously. Challenges and opportunities presented by human–machine collaboration in governance are reviewed, with a focus on developing strategies to mitigate cognitive biases. Ultimately, a balanced approach to human–machine collaboration in governance is advocated, leveraging the strengths of both humans and machines while consciously addressing their respective limitations. This approach is vital for the development of governance systems that are both technologically advanced and cognitively attuned, leading to more informed and responsible decision-making.

Policy Significance Statement

The importance of human–machine collaborations for developing governance systems that are both technologically advanced and cognitively attuned is examined. A framework is provided and discussed that explores varying levels of automation in such collaboration and how they interact with human cognitive biases, focusing on dual-process models of thinking and their influence on decision-making when integrated with AI systems. This approach emphasizes the transformative potential of combining human cognitive strengths with machine computational capabilities while also addressing their respective limitations. This could lead to more informed, responsible, and efficient decision-making processes, making it highly relevant for policymakers aiming to optimize governance in the age of AI.

1. Introduction

Human–machine collaboration (HMC) is increasingly becoming a pivotal aspect of various sectors, driven by advancements in artificial intelligence (AI) and automation (Geng and Varshney, 2022): In healthcare, AI assists doctors in diagnosing diseases with greater accuracy (e.g. IBM's Watson for

https://doi.org/10.1017/dap.2024.72 Published online by Cambridge University Press

Health). In finance, algorithms perform high-frequency trading, significantly impacting market dynamics. Autonomous vehicles in transportation, such as those developed by Waymo, illustrate the shift toward more complex, safety-critical domains (Ignatious et al., 2023) This trend underscores a broader societal and technological evolution, where human expertise is augmented by machine capabilities, often leading to enhanced efficiency, decision accuracy, and innovation. In the industrial context, this trend has already significantly transformed decision-making structures, leading to the emergence of human–machine collaborative systems (Bhandari, 2021). For instance, Döppner et al. (2019) described a compelling case of HMC within the air cargo industry, providing a vivid illustration of how such partnerships can evolve and function effectively in real-world scenarios. The paper details the integration of a sophisticated decision-support system (DSS) into the workflow of unit load device (ULD) dispatchers, illustrating a key transition from a technology-supporting to a collaborative environment. This transition not only enhanced the efficiency and accuracy of operations but also reshaped the roles and skills of the human workforce, emphasizing adaptability and strategic oversight. Döppner et al.'s (2019) exploration of the mutual learning curve between humans and machines, where each influences and adapts to the other's capabilities, is particularly instructive and suggested potential scenarios in which the integration of AI and automated systems could similarly transform traditional practices in governance and policy decision-making.

2. HMC in the public sector

A number of authors have argued that the integration of HMC in government organizations can lead to efficiency gains, improved decision-making, and enhanced work processes (Callahan and Holzer, 1997; Pi, 2021; Reis et al., 2019). Mikhaylov et al. (2018) envisioned a transformative impact on public service delivery, and identify AI's potential to streamline processes and deliver more personalized and efficient public services. Krafft et al. (2020) described how AI's capability to process vast amounts of data rapidly and accurately enables more informed policy decisions. However, to fully integrate AI in public services and achieve mature HMCs, a number of challenges need to be addressed, such as managerial complexities inherent in cross-sector collaborations essential for AI implementation, including aligning goals among public, private, and non-profit entities and establishing shared knowledge standards (Ahn and Chen, 2020; Pi, 2021). A notable obstacle in this process is the discrepancy in AI understanding between researchers and policymakers, with policy documents often framing AI in human-centric terms, diverging from the technical definitions favored by researchers (Krafft et al., 2020). To address this, a number of authors have underlined the need for a common understanding of AI that is grounded in technical reality, yet also practical for policy development (Mikhaylov et al., 2018; Valle-Cruz et al., 2019). A key argument of this article is that developing a better understanding of the nuances of automation and autonomy within HMC, and how they relate to strengths and weaknesses in human decision-making, can potentially help address this, and as such pave the way for more effective and transparent integrations of AI in governance (see also, Mikhaylov et al., 2018; Valle-Cruz et al., 2019).

Leveraging Simmler and Frischknecht's (2021) influential taxonomy of levels of automation, we will begin by exploring the interplay between automation levels and human decision-making biases. Subsequently, a cognitive framework inspired by dual-process models of thinking is introduced to further explore this interaction within governance systems. The utility of this framework will be demonstrated through a number of policy examples (e.g. urban development, traffic management) that highlight its potential to mitigate a series of cognitive biases. Overall, the argument is made that such a cognitive framework can begin to address the aforementioned discrepancy in AI understanding between researchers and policymakers, and help integrate theoretical insights with practical considerations. The approach outlined suggests a nuanced synergy between human cognitive capabilities and computational power of machines, carefully considering their respective constraints. Although certainly no panacea, the framework could be a useful tool for policymakers striving to refine HMCs in the public sector.

https://doi.org/10.1017/dap.2024.72 Published online by Cambridge University Press

3. Levels of automation

HMC involves the interactive work of human operators and intelligent automation within the same workspace. This collaboration is often facilitated by the design of computer agents that mediate between human-computer interaction and automated plan reasoners (Allen and Ferguson, 2002). This can take various forms, including different levels of assistance and automation (Bao et al., 2023). Simmler and Frischknecht (2021) developed a taxonomy of the levels of automation and technical autonomy in HMC built around two dimensions: automation and autonomy (see Table 1). Automation pertains to the extent to which a machine can operate independently of human input, while autonomy refers to the system's capacity for independent decision-making and learning.

1. Offers decisions (Level 1): Here, the machine acts as a DSS, offering options for the human operator to choose from. For example, a GPS navigation system proposes routes, but the driver decides which one to take.
2. Executes with human approval (Level 2): The machine selects an action and executes it upon receiving human approval. An illustration of this is a thermostat system that suggests temperature adjustments based on weather forecasts but requires homeowner confirmation.
3. Executes if no human vetoes (Level 3): The system acts unless the human operator vetoes it. Automated spam filters in email services, which filter messages unless manually overridden, exemplify this level.
4. Executes and then informs (Level 4): The machine acts independently but informs the human afterward. An example is a security system in a building that locks doors at a set time and notifies security personnel.
5. Executes fully automated (Level 5): The highest level of automation, where the machine operates entirely independently, without human intervention or notification. Autonomous drones used for geographical surveys in remote areas represent this level.

Understanding this taxonomy can provide valuable insights, by offering a lens through which to analyze and formulate strategies around the integration of AI and automation in various contexts, and how it affects decision-making structures and (human) roles. An example of Level 1 automation can be found in medical imaging, where AI systems can outperform human doctors at spotting cancers and other pathologies. However, the final decision still rests with the human doctors who use AI as a diagnostic and research support tool. This setup allows doctors to focus more on treatment regimens and nurturing the doctor-patient relationship (Coppola et al., 2021). The development of semiconductor chips showcases a Level 3 integration of AI and human expertise. In a simulated environment designed to test process engineering, Coppola et al. (2021) revealed that adopting a hybrid approach—initiating with human-directed algorithms and subsequently transitioning to autonomous computer operations—halved the costs of reaching performance targets when compared to exclusive reliance on human designers. As detailed by Kanarik et al. (2023), this

Table 1. *Levels of automation with their main features*[a]

Level	Description	Explanation
1	Offers decision	Technical component suggests options and the human decides
2	Executes with human approval	Technical component acts after human approves
3	Executes if no human vetoes	Technical component acts unless human vetoes
4	Executes and then informs	Technical component acts independently and human is informed about the actions carried out
5	Executes fully automated	Technical component carries out actions independently without informing human

[a] Adapted from Simmler and Frischknecht (2021)

hybrid model embodies a dynamic in which the machine operates autonomously unless human intervention is deemed necessary, with the initial human input being critical for optimizing the system's effectiveness.

No doubt this taxonomy is useful, but it also illustrates an issue already highlighted: AI researchers tend to focus on mostly technical functionality and rational behavior, and often fail to relate it to human cognition, with its strengths and frailties. For instance, AI is often seen as "the ultimate enabler" for automating decision-making tasks in various domains, but it is not a simple substitute for human decision-making. Rapid advances in machine learning have improved statistical prediction, but prediction is only one aspect of decision-making (Goldfarb and Lindsay, 2022). As AI becomes a more integral part of decision-making processes, conflicts will likely arise when human knowledge and experience conflict with the information provided by AI systems, leading to breakdowns in trust and decision-making processes. All this is further exacerbated by the fact that AI notoriously fails in capturing or responding to intangible human factors that go into real-life decision-making, such as ethical, moral, and other human considerations that guide the course of business, life, and society at large (McKendrick and Thurai, 2022). A potential way to address this is by relating such taxonomies to models of cognitive processes involved in the type of human decision-making typical for governance and policy (Van Rooy, 2023).

4. Human decision-making

Governance is a complex and risky decision-making process, where irrational public risk perceptions and overconfident expert predictions can lead to ineffective decision-making (Hardaker et al., 2009). Research has demonstrated how policy decision-making and governance can be significantly affected by cognitive biases, particularly those associated with so-called Type I thinking (Berthet, 2021; Kahneman, 2013). Cognitive biases are described as systematic, universally occurring tendencies in human decision-making that may lead to inaccurate or suboptimal outcomes (see Table 2 for examples). They have been shown to affect the outcome of deliberations and can have substantial effects on society and human wellbeing (Korteling et al., 2023). Understanding and addressing cognitive biases is essential for improving the quality

Table 2. *Cognitive biases affecting system 1 thinking*

Cognitive bias	Description
Anchoring	The tendency to rely too heavily on the first piece of information encountered when making decisions.
Confirmation	The inclination to favor information that confirms existing beliefs and to undervalue information that contradicts them.
Availability	Judgments of likelihood or percentages based on ease of recall (greater "availability" in memory) rather than on actual probabilities.
Overconfidence	The propensity to be more confident in one's own abilities, such as driving, teaching, or spelling, than is objectively reasonable.
Automation	The predisposition to favor suggestions from automated decision-making systems and to ignore contradictory information made without automation, even if it is correct.
Status quo	A preference for the current state of affairs, with alternatives perceived as a change from the baseline.
Omission bias	The tendency to favor an act of omission (not doing something) over one of commission (doing something), due to the perception that harmful consequences are more severe when caused by action rather than inaction.
Hindsight	The inclination, after an event has occurred, to see the event as having been predictable, despite there having been little or no objective basis for predicting it.
Out-of-the-loop	The state of being uninformed about the operations of a system, which can occur when automation is used extensively, and human operators become less involved in the active monitoring of system performance.

https://doi.org/10.1017/dap.2024.72 Published online by Cambridge University Press

Table 3. *Characteristics of system 1 and system 2 and application in AI design*

Type	Characteristics	Key aspects	Collaboration focus on:
Type 1	Does not require working memory. Automatic Instinctive and rapid, operates subconsciously	Uses heuristics (mental shortcuts), susceptible to biases.	Aligning with users' mental models and cognitive maps for intuitive interaction.
Type 2	Controlled, slow, limited capacity and conscious	Involves careful processing of information, evaluating options, and considering consequences.	Facilitate analytical mindset with information and tools for data analysis and decision-making, motivate users to pay attention.

of policy decision-making and governance. The emergence of HMC and AI-driven decision support systems has added an additional layer of complexity, and indeed urgency, to this issue (Van Rooy, 2023).

A good way to understand cognitive biases and their impact on human decision-making is by framing them in Kahneman's influential classification of fast (Type I) and slow (Type II) thinking (Kahneman, 2013) (see Table 3). It might be useful to clarify these concepts in some detail, for those unfamiliar with the behavioral science literature: Type I thinking is instinctive and rapid, often operating at a subconscious level. It is the kind of thinking we employ for quick decisions, such as catching a ball or answering straightforward questions (e.g. how much is 2 + 2?). Central to Type I thinking are heuristics, which are mental shortcuts or rules of thumb that our brains use to accelerate decision-making. While this mode of thinking is efficient, it is also prone to biases and systematic errors in judgment. For instance, if you have recently read about a robbery in your neighborhood, you might overestimate the likelihood of crime occurring nearby. This is known as the availability heuristic or bias, where the probability of events is judged based on how readily examples come to mind.

Type II thinking, in contrast to the rapid and instinctive nature of Type I thinking, is characterized by its slower, more deliberate, and conscious process. It is the type of thinking we engage in when faced with complex problems or significant decisions. For example, when solving a challenging math problem or deciding between multiple job offers, Type II thinking comes into play. This analytical form of thinking involves a thorough process of information evaluation. It requires one to carefully process the information at hand, consider various options, and deliberate on the potential consequences of different choices. Ideally, our interactions with AI-driven systems should be structured in a way that supports and even encourages this analytical mindset at the appropriate times. This means providing users with ample information and the necessary tools to conduct thorough data analysis, evaluate different options, and ultimately arrive at well-informed decisions. This is particularly vital in scenarios that are either high-pressure or inherently complex, where the decision-making process requires more than just an instinctive response.

When conceptualizing HMCs, understanding the interplay between these two types of thinking can be very useful in creating a balance between user friendliness and ease of use on the one hand, and support for complex decision-making processes on the other (Van Rooy, 2023). This balance is crucial for effective collaborations with AI systems, especially under pressure or in situations that demand a high level of accuracy and thoughtfulness (AlKhars et al., 2019; Nurse et al., 2022). Essentially, this means creating collaborations that (1) allow users to make efficient, instinctive decisions when appropriate, minimizing the cognitive load for routine tasks; but also (2) provide the necessary tools and information for more considered and analytical decision-making when required (see Table 3). A good example of an application that emphasizes the balance between Type I (intuitive) and Type II (deliberative) thinking is found in autonomous vehicle interfaces: These systems are designed to allow drivers to make quick, intuitive decisions based on dashboard displays and alerts (Type I thinking), while also offering the ability to engage more deeply with navigation settings and system checks for complex scenarios (Type II thinking). This dual

https://doi.org/10.1017/dap.2024.72 Published online by Cambridge University Press

approach helps ensure safety by allowing for rapid responses to immediate driving conditions, while also supporting thorough decision-making when navigating more complicated driving tasks or settings.

5. Enhancing decision-making through human–machine integration

The exploration of cognitive biases within the frameworks of Type I and Type II thinking offers valuable insights into the complexities of human decision-making. Understanding these biases is crucial, not just for identifying the limitations and strengths of human cognition, but also for shaping the way we interact with technology. It is important to build HMCs in which both parties are "aware", or at least reminded, of each other's strengths and weaknesses: Human decision-makers are vulnerable to cognitive biases, while machines have difficulty handling new and dynamic contexts with incomplete information (Xiong et al., 2022). Weser and Zhang (2010) also highlight the challenge of incomplete information in dynamic environments and the difficulties in predicting future changes and maintaining a complete world model. These challenges are further compounded by the limitations of information processing, as discussed by Walton (2018), which can impact the effectiveness of AI and machine learning in these contexts. The consequences of all of this are well documented: Biased or unrepresentative AI models and poor data quality have led to erroneous or unfair outcomes in a variety of domains (Aldoseri et al., 2023). In collaborations, humans would have to be motivated at appropriate times to engage in Type II thinking and become aware of limitations on the algorithmic side, rather than relying on its input as "the ultimate enabler".

5.1. A cognitive framework for HMC

Although research into the impact of incorporating intelligent machines on decision-making and decision-makers is relatively new (Bhandari et al., 2021), several authors have made (similar) suggestions: Xiong et al. (2022) highlighted the potential for superior performance by leveraging human and machine capabilities, while Geng and Varshney (2022) emphasized the importance of integrating human cognitive strengths with machine computational capabilities. More recent research collectively underscores the enhanced decision-making outcomes achieved through strategic human–AI system integration that leverages the complementary strengths of both, with frameworks suggesting that this synergy can surpass individual human or AI capabilities (Lai et al,. 2022; Rastogi et al., 2022; Steyvers et al., 2022). Similarly, Van Rooy (2023) described how HMC needs to consider how the interaction between humans and machines can either exacerbate or mitigate the impact of cognitive biases associated with Type I thinking. If we take the Simmler and Frischknecht (2021) taxonomy as a starting point, we can examine how different cognitive biases could impact different levels of HMC. Each level of automation presents unique challenges where human cognitive biases can impact the effectiveness and safety of the system. Recognizing and mitigating these biases is crucial for designing and managing HMCs. Table 4 summarizes strategies that can be employed, from diverse option generation to designing algorithms for debiasing, and that can collectively ensure more unbiased and informed decision-making. In the complex area of policy decision-making, these strategies could play out in different ways. Here are illustrative examples for each level:

1. Offers decision-anchoring, confirmation bias, availability heuristic: Imagine a policy decision on urban development. The technical system suggests several options, including building a new park, a commercial center, or residential apartments. At this level, decision-makers are highly involved in the selection process, making certain biases more likely: A policy maker might gravitate toward the first option presented (anchoring bias) or choose an option that aligns with their preexisting beliefs about urban planning (confirmation bias). They might also pick an option based on a recent event, like a successful park inauguration they attended (availability heuristic). To counteract such biases, the system could offer a diverse range of options and provide transparent algorithms that explain how each option aligns with urban development goals (Khediri et al., 2021). Diverse option generation combats anchoring by providing a broad spectrum of choices, reducing the likelihood that the first option disproportionately influences the decision (George et al., 2000;

Table 4. *Level automation, most likely bias, and mitigation techniques*

Level of automation	Most likely bias	Mitigation techniques
1. Offers decision	Anchoring Confirmation Availability	Diverse option generation: Ensure the machine presents a broad spectrum of choices. Algorithmic transparency: Make the decision process of the machine transparent to highlight how options are generated. Critical review sessions: Implement regular sessions where decisions are reviewed and biases are identified and discussed.
2. Executes with human approval	Overconfidence Automation	Risk communication: Clearly communicate the risks associated with each option provided by the machine. Regular training: Educate humans on the potential pitfalls of overconfidence and automation biases. Independent verification: Introduce a process where a third party or an independent system verifies the decisions before execution.
3. Executes if no human vetoes	Status quo Omission	Default option rotation: Rotate default actions to prevent status quo bias. Active confirmation requirement: Require active confirmation from humans for critical decisions, even in non-veto scenarios. Feedback Mechanisms: Provide feedback on the consequences of non-intervention to highlight the impact of omission bias.
4. Executes and then informs	Hindsight Automation	Post-decision analysis: Conduct analyses after decisions to identify and learn from any hindsight biases. Regular updates and summaries: Keep humans informed regularly about automated actions to reduce complacency. Critical questioning prompts: Implement prompts that encourage humans to critically assess and question automated actions.
5. Executes fully automated	Out-of-the-loop unfamiliarity	Periodic human engagement: Involve humans periodically in the decision process to maintain familiarity. Simulation training: Use simulations to train humans in intervening and understanding automated systems. System transparency reports: Provide regular reports on system performance and decision logic to maintain human understanding of the system.

Gong et al., 2017). Alternatively, Rastogi et al. (2022) demonstrated how applying time constraints based on an AI system's confidence level can also effectively reduce the impact of anchoring bias. Algorithmic transparency can mitigate confirmation bias by allowing decision-makers to understand how options are generated, encouraging them to consider alternatives they might otherwise dismiss. Critical review sessions provide a structured environment to address availability bias by ensuring decisions are not overly influenced by information that is more easily recalled but may not be the most relevant.

https://doi.org/10.1017/dap.2024.72 Published online by Cambridge University Press

2. Executes with human approval—overconfidence, automation bias: Consider a policy on traffic management. The automated system proposes to implement a new AI-driven traffic light control system and awaits human approval. The policy maker, confident in their understanding of traffic systems, might overlook potential risks or issues with the AI system due to overconfidence bias, which stems from an excessive belief in their own insights, and dismisses external data. In contrast, they might also show automation bias, where they rely too heavily on the system's recommendations, assuming the automated decisions are more reliable than human judgment without critical evaluation. To mitigate, clear communication of risks and regular training sessions on AI systems' limitations could be implemented (Schemmer et al., 2022). Risk communication is essential here, as it clearly delineates potential pitfalls associated with each option, addressing overconfidence by ensuring decision-makers are aware of what could go wrong. Regular training helps make the operators more cognizant of the limitations of both their judgment and the automated system, thus reducing both overconfidence and automation bias. Independent verification by a third party or system introduces an additional layer of scrutiny, which can help catch errors or biases that the primary decision-maker might miss. Although not the focus of the current paper, it should be obvious that explainable AI (xAI) will play a crucial role here. However, it is worth mentioning that while xAI clarifies AI processes, it does not inherently correct for, and can in fact exacerbate decision bias (Bertrand et al., 2022).
3. Executes if no human vetoes—status quo, omission bias: In a policy decision about environmental regulations, the system automatically updates regulations based on new data unless vetoed. This level's default action approach makes it susceptible to status quo and omission biases. Policy makers might prefer not to intervene, maintaining the status quo (status quo bias) or avoid making an active decision due to fear of being responsible for potential negative outcomes (omission bias). This can be mitigated by rotating the default options, forcing individuals to actively confirm their choices, and providing clear alternatives to the default. By doing so, individuals are prompted to engage in more deliberate decision-making processes, reducing the influence of default bias and promoting a more thoughtful evaluation of options, disrupting the status quo bias (Sunstein, 2014; Vargas and Lauwereyns, 2021). The requirement for active confirmation for critical decisions ensures that omission bias is addressed by making inaction (not vetoing) a conscious choice rather than a passive default. Feedback on the consequences of non-action can further highlight the costs of omission bias, encouraging more active engagement.
4. Executes and then informs—hindsight, automation bias: In a health policy scenario, an automated system implements a new vaccination strategy and informs policy makers postimplementation. They might believe, in hindsight, that they anticipated the success or failure of the strategy (hindsight bias), or they may not scrutinize the strategy assuming the system's infallibility (automation bias). Post-decision analysis and critical questioning prompts can be employed to encourage reflective thinking (Goddard et al., 2012; Lyell and Coiera, 2017). Post-decision analysis allows for reflection and learning from past actions, addressing hindsight bias by highlighting discrepancies between expected and actual outcomes. Regular updates and summaries keep decision-makers in the loop, reducing the out-of-sight, out-of-mind mentality that feeds automation bias. Critical questioning prompts encourage a proactive mindset, urging decision-makers to critically evaluate actions taken by the automated system.
5. Executes fully automated—out-of-the-loop unfamiliarity: In a fully automated financial policy, the system adjusts interest rates based on economic indicators without human input or information. With no human intervention, the main risk is losing touch with the system's operations. This could lead to policy makers losing touch with the decision-making process (out-of-the-loop unfamiliarity), impairing their ability to intervene during economic crises. Regular system performance reports and simulation training for policymakers, to familiarize them with the system's logic and functionality, can maintain engagement and understanding of the system's logic and functionality, ensuring that human operators maintain the capability and confidence to intervene when necessary.

https://doi.org/10.1017/dap.2024.72 Published online by Cambridge University Press

(Green, 2022). Transparency reports on system performance and decision-making logic help mitigate out-of-the loop unfamiliarity by keeping decision-makers informed about how decisions are made, even if they are not directly involved in day-to-day operations.

5.2. Biases across automation levels: understanding susceptibility and mitigation

Table 4 thus highlights the most likely biases emerging from the characteristics of each automation level. As a simple rule of thumb, the characteristics of the different levels of automation that make them more or less susceptible to specific cognitive biases are closely related to the degree of human involvement and oversight in the decision-making process. As automation increases, the types of biases shift from those influenced by direct human decision-making toward those related to overreliance on or underengagement with automated systems (Cummings, 2004; Mosier and Skitka, 1999). More specifically, varying levels of automation necessitate different degrees of human supervision, which in turn influences the likelihood of certain cognitive biases while potentially mitigating others. In the initial stages of automation, where machines provide decisions (Level 1) and execute actions subject to human approval (Level 2), human participation plays a critical role in mitigating automation bias. However, this involvement can introduce other biases: confirmation bias at Level 1, where operators may favor options that confirm preexisting beliefs, and anchoring bias at Level 2, due to the cognitive load of approving from a set of complex options, with the initial choice often unduly influential (DeKay, 2015). At Level 3, the dynamic shifts slightly as the system acts unless explicitly vetoed by a human, which can mitigate automation bias. However, this setup may inadvertently nurture overconfidence bias, where the presumption of the system's reliability could deter operators from actively questioning or intervening. This level of automation risks engendering a false sense of security in the system's capabilities, potentially leading to missed opportunities for critical evaluation or necessary intervention. Overall, automation bias tends to increase as the level of automation rises, particularly in scenarios where systems execute actions independently (Level 4) and without human oversight (Level 5). Hindsight bias is more likely at higher levels of automation because the lack of their immediate involvement may lead individuals to believe, after outcomes are known, that they would have predicted or made different decisions. This retrospective certainty is less common at lower levels of automation, where decisions require active human input and the consequences of those decisions are more directly observable. The diminished role of humans in monitoring and intervening in the system's operations also increase the risk of out-of-the-loop unfamiliarity, where humans lose touch with how decisions are made and are not prepared to intervene effectively in unusual or critical situations. This unfamiliarity is less of an issue at lower automation levels, where human interaction with and oversight of automated processes help maintain an understanding of the system's functioning and decision-making logic.

Although certain biases are more commonly associated with specific levels of automation, it is not impossible for a specific bias to manifest itself across different levels in slightly different ways. Confirmation bias, for example, can influence decisions across the board, from actively selecting among machine-provided options at lower levels to uncritically accepting automated decisions at higher levels, always favoring information that aligns with pre-existing beliefs. Overconfidence in one's judgments or the system's accuracy might lead to overlooking errors at initial levels, where human input is significant and persist as automation increases, assuming infallibility of machine operations. Automation bias can start as a subtle preference for automated suggestions, even when human control is prevalent, and evolve into an overreliance on fully automated systems, underestimating the need for oversight. This underscores the complex interplay between human cognitive tendencies and automation levels, while also highlighting the importance of tailored mitigation strategies. Although certain mitigation strategies are more optimally aligned with specific levels due to the nature of human–machine interaction at those stages, some of these strategies are adaptable and can be applied across various levels. For instance, at the first level where machines offer decisions (Level 1), the use of diverse option generation combats anchoring by providing a broad array of choices, preventing the undue influence of the first option. As we move to higher levels, such as when machines execute and then inform humans of their actions (Level 4), the

https://doi.org/10.1017/dap.2024.72 Published online by Cambridge University Press

strategy shifts to post-decision analysis, encouraging reflection and learning from past actions, addressing the same anchoring bias in a context where initial impressions could still unduly influence retrospective judgment. Similarly, to counteract automation bias at a level where the machine executes actions with human approval (Level 2), regular training about the system's capabilities and limitations keeps human operators critically engaged, reducing reliance on automated suggestions. In contrast, at the highest level of automation (Level 5), where machines operate fully independently, periodic human engagement through simulation training is essential to maintain familiarity with the system's operations, ensuring a critical stance toward automation persists.

6. Discussion

Some of the illustrative examples above dealing with higher automation levels, seem for the moment to be solidly situated in the future. However, while not exactly matching the fictional scenarios, the use of AI in banking mirrors the concept of "Executes Fully Automated" in financial policy, where decisions are made based on data analysis without human intervention, potentially leading to "out-of-the-loop unfamiliarity" if not monitored properly (Agarwal et al., 2021). Already, we have seen policy failures that can be placed at the "Executes Fully Automated" level, where the technical component carries out actions independently without informing humans: In both the Dutch childcare allowance scandal (Van de Vijver and De Raedt, 2023) and the Robo Debt scandal in Australia (Rinta-Kahila et al., 2023), automated systems were used to detect and act without proper oversight on suspected benefit fraud, leading to serious consequences for the affected individuals. The use of algorithms in the detection of fraud led to issues such as racial profiling, discrimination against certain groups, and the imposition of exorbitant debts on individuals, often with lower incomes or belonging to ethnic minorities (D'Rosario and D'Rosario, 2020; Monarcha-Matlak, 2021). These scandals highlight the risks associated with the "Executes Fully Automated" level of automation, where the technical component acts independently without informing humans, leading to out-of-the-loop unfamiliarity and the potential for serious harm when human intervention is required in unexpected situations or system failures. This evolution calls for an urgent need to establish clear frameworks to identify where the final decision-making authority lies, be it with humans or machines, and to implement corresponding mitigation strategies. Such frameworks should not only pinpoint the decision-makers, but also include strategies to mitigate the risks of biases inherent in HMCs.

Our analysis of the complexities observed in HMC within governance led us to focus predominantly on enhancing Type II thinking. The necessity for enhanced Type II thinking became apparent through our analysis, as it addresses the kind of oversight failures exemplified by, for instance, the Robodebt and childcare scandals, which were attributed to insufficient scrutiny. However, future research could also explore scenarios where Type I thinking is equally crucial, such as in dynamic environments like emergency response systems, where rapid and instinctive decision-making can be vital. Examining such contexts could provide a more comprehensive understanding of how to balance intuitive and analytical processes to optimize decision-making in various automated systems.

HMC thus emerges as a key factor in the transformation of decision-making paradigms in the landscape of governance and policy making. The benefits of such HMC are extensive, from offering improved efficiency, a decrease in cognitive workload for routine tasks, to a strengthened approach to intricate and high-stakes decisions. However, the integration of HMC in governance is not without its challenges, and careful consideration must be given to the potential biases in human and machine decision-making. Insights from behavioral science can be useful in optimizing these collaborations, suggesting that aligning AI systems with human cognitive processes can reduce errors, enhance user experience, and promote more effective and harmonious HMCs. The impact on policy and governance from such collaborations is anticipated to be significant, necessitating a concerted effort from policymakers and AI technologists alike. Together, they must strive to develop HMCs that transcend the mere sum of their individual parts, achieving a synergy where the combined effect is greater than the contributions of each component.

Data availability statement. This review paper does not contain primary data. All data referred to in this paper are available from the original sources cited in the references. No new data were created or analyzed in this study.

https://doi.org/10.1017/dap.2024.72 Published online by Cambridge University Press

Author contribution. Conceptualization: DVR; Methodology: DVR; Investigation: DVR; Data Curation: DVR; Writing – Original Draft: DVR; Writing – Review & Editing: DVR; Visualization: DVR; Supervision: DVR; Project Administration: DVR.

Provenance. This article is part of the Data for Policy 2024 Proceedings and was accepted in Data & Policy on the strength of the Conference's review process.

Funding statement. Part of this research was supported by grants from the BELSPO, the Belgian Defense research agency. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interest. The author declares none.

References

Agarwal A, **Singhal C, and Thomas R** (2021) *AI-Powered Decision Making for the Bank of the Future*. McKinsey & Company.

Ahn MJ and Chen Y-C (2020) *Artificial Intelligence in Government:: Potentials, Challenges, and the Future*. Proceedings of the 21st Annual International Conference on Digital Government Research.

Aldoseri A, **Al-Khalifa KN and Hamouda AM** (2023) Re-thinking data strategy and integration for artificial intelligence: concepts, opportunities, and challenges. *NATO Advanced Science Institutes Series E: Applied Sciences 13*(12), 7082. https://doi.org/10.3390/app13127082

Allen JF and Ferguson G (2002) *Human-Machine Collaborative Planning.* Proceedings of the Third International NASA Workshop on Planning and Scheduling for Space, Houston, TX, October 27–29.

Bao Y, **Gong W and Yang K** (2023) A literature review of human–AI synergy in decision making: from the perspective of affordance actualization theory. *Systems 11*(9), 442. https://doi.org/10.3390/systems11090442

Berthet V (2021) The impact of cognitive biases on professionals' decision-making: a review of four occupational areas. *Frontiers in Psychology 12*, 802439. https://doi.org/10.3389/fpsyg.2021.802439

Bertrand A, **Belloum R**, **Eagan JR, and Maxwell W** (2022) How cognitive biases affect XAI-Assisted decision-making: a systematic review (AAAI; ACM SIGAI, trans.). In *5th AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society, AIES 2022.* Association for Computing Machinery, Inc. https://doi.org/10.1145/3514094.3534164

Coppola F, **Faggioni L**, **Gabelloni M**, **De Vietro F**, **Mendola V**, **Cattabriga A**, **Cocozza MA**, **Vara G**, **Piccinino A**, **Lo Monaco S**, **Pastore LV**, **Mottola M**, **Malavasi S**, **Bevilacqua A**, **Neri E, and Golfieri R** (2021) Human, All Too Human? An all-around appraisal of the "Artificial Intelligence Revolution" in medical imaging. *Frontiers in Psychology 12*, 710982. https://doi.org/10.3389/fpsyg.2021.710982.

Cummings M (2004) *Automation bias in intelligent time critical decision support systems. In AIAA 1st Intelligent Systems Technical Conference.* American Institute of Aeronautics and Astronautics. https://doi.org/10.2514/6.2004-6313.

DeKay ML (2015) Predecisional information distortion and the self-fulfilling prophecy of early preferences in choice. *Current Directions in Psychological Science 24*(5), 405–411. https://doi.org/10.1177/0963721415587876

Döppner DA, **Derckx P and Schoder D** (2019) Symbiotic co-evolution in collaborative human-machine decision making: exploration of a multi-year design science research project in the air cargo industry. In *Proceedings of the 52nd Hawaii International Conference on System Sciences. Hawaii International Conference on System Sciences.* https://doi.org/10.24251/hicss.2019.033

D'Rosario M and D'Rosario C (2020) Beyond RoboDebt. *International Journal of Strategic Decision Sciences 11*(2), 1–24. https://doi.org/10.4018/ijsds.2020040101

Geng B and Varshney PK (2022) Human-machine collaboration for smart decision making: current trends and future opportunities. In *2022 IEEE 8th International Conference on Collaboration and Internet Computing* (*CIC*), 61–67. https://doi.org/10.1109/CIC56439.2022.00019

George JF, **Duffy K, and Ahuja M** (2000) Countering the anchoring and adjustment bias with decision support systems. *Decision Support Systems 29*(2), 195–206. https://doi.org/10.1016/s0167-9236(00)00074-9

Goddard K, **Roudsari A and Wyatt JC** (2012) Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association: JAMIA 19*(1), 121–127. https://doi.org/10.1136/amiajnl-2011-000089

Goldfarb A and Lindsay JR (2022) Prediction and judgment: why artificial intelligence increases the importance of humans in war. *International Security 46*(3), 7–50. https://doi.org/10.1162/isec_a_00425

Gong M, **Lempert R**, **Parker A**, **Mayer LA**, **Fischbach J**, **Sisco M**, **Mao Z**, **Krantz DH and Kunreuther H** (2017). Testing the scenario hypothesis: an experimental comparison of scenarios and forecasts for decision support in a complex decision environment. *Environmental Modelling and Software 91*, 135–155. https://doi.org/10.1016/j.envsoft.2017.02.002

Green B (2022) The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Review 45*, 105681. https://doi.org/10.1016/j.clsr.2022.105681

Hardaker JB, **Fleming E and Lien G** (2009) How should governments make risky policy decisions? *Australian Journal of Public Administration 68*(3), 256–271. https://doi.org/10.1111/j.1467-8500.2009.00638.x

Ignatious HA, **El-Sayed H**, **Khan MA and Mokhtar BM** (2023) Analyzing factors influencing situation awareness in autonomous vehicles-a survey. *Sensors 23*(8). https://doi.org/10.3390/s23084075

Kahneman D (2013) *Thinking, Fast and Slow*, 1st Edn. Farrar, Straus and Giroux. https://www.amazon.com/Thinking-Fast-Slow-Daniel-Kahneman/dp/0374533555

https://doi.org/10.1017/dap.2024.72 Published online by Cambridge University Press

Kanarik KJ, **Osowiecki WT**, **Lu YJ**, **Talukder D**, **Roschewsky N**, **Park SN**, **Kamon M**, **Fried DM and Gottscho RA** (2023) Human-machine collaboration for improving semiconductor process development. *Nature 616*(7958), 707–711. https://doi.org/10.1038/s41586-023-05773-7

Callahan K and Holzer M (1997) *Government at Work: Best Practices and Model Programs.* SAGE Publications.

Khediri A, **Laouar MR and Eom SB** (2021) Improving intelligent decision making in urban planning: using machine learning algorithms. *International Journal of Business Analytics (IJBAN) 8*(3), 40–58. https://doi.org/10.4018/IJBAN.2021070104

Korteling JEH, **Paradies GL, and Sassen-van Meer JP** (2023) Cognitive bias and how to improve sustainable decision making. *Frontiers in Psychology 14*, 1129835. https://doi.org/10.3389/fpsyg.2023.1129835

Krafft PM, **Young M**, **Katell M**, **Huang K and Bugingo G** (2020) Defining AI in policy versus practice. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society.* 72–78. https://doi.org/10.1145/3375627.3375835.

Lai V, **Carton S**, **Bhatnagar R**, **Liao QV**, **Zhang Y and Tan C** (2022) Human-AI collaboration via conditional delegation: a case study of content moderation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, Article Article 54. https://doi.org/10.1145/3491102.3501999

Lyell D and Coiera E (2017) Automation bias and verification complexity: a systematic review. *Journal of the American Medical Informatics Association: JAMIA 24*(2), 423–431. https://doi.org/10.1093/jamia/ocw105

McKendrick J, and Thurai A (2022, September 15). AI Isn't ready to make unsupervised decisions. *Harvard Business Review.* https://hbr.org/2022/09/ai-isnt-ready-to-make-unsupervised-decisions

Mikhaylov SJ, **Esteve M and Campion A** (2018) Artificial intelligence for the public sector: opportunities and challenges of cross-sector collaboration. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 376* (2128) 20170357.

Monarcha-Matlak A (2021) Automated decision-making in public administration. *Procedia Computer Science 192*, 2077–2084. https://doi.org/10.1016/j.procs.2021.08.215

Mosier KL and Skitka LJ (1999) automation use and automation bias. proceedings of the human factors and ergonomics society… *Annual Meeting Human Factors and Ergonomics Society. Meeting 43*(3), 344–348. https://doi.org/10.1177/154193129904300346

Pi Y (2021) Machine learning in Governments: benefits, challenges and future directions. *JeDEM—eJournal of eDemocracy and Open Government 13*(1), 203–219.

Rastogi C, **Zhang Y**, **Wei D**, **Varshney KR**, **Dhurandhar A, and Tomsett R** (2022) Deciding fast and slow: the role of cognitive biases in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction 6*(CSCW1), 1–22. https://doi.org/10.1145/3512930

Reis J, **Santo PEI and Melão N** (2019) Artificial intelligence in government services: a systematic literature review. In *Advances in Intelligent Systems and Computing* (pp. 241–252). Springer International Publishing.

Rinta-Kahila T, **Someh I**, **Gillespie N**, **Indulska M, and Gregor S** (2023) Managing unintended consequences of algorithmic decision-making: the case of Robodebt. *Journal of Information Technology Teaching Cases* 204388692311655. https://doi.org/10.1177/20438869231165538.

Schemmer M, **Kühl N**, **Benz C and Satzger G** (2022) On the influence of explainable AI on automation bias. In *arXiv* [cs.HC]. arXiv. http://arxiv.org/abs/2204.08859

Simmler M. and Frischknecht R (2021) A taxonomy of human–machine collaboration: capturing automation and technical autonomy. *AI & Society 36*(1), 239–250. https://doi.org/10.1007/s00146-020-01004-z.

Steyvers M Tejeda H, **Kerrigan G and Smyth P** (2022) Bayesian modeling of human-AI complementarity. *Proceedings of the National Academy of Sciences of the United States of America 119*(11), e2111547119. https://doi.org/10.1073/pnas.2111547119

Sunstein CR (2014) *Active Choosing or Default Rules? The Policymaker's Dilemma.* https://doi.org/10.2139/ssrn.2437421

Valle-Cruz D, **Alejandro Ruvalcaba-Gomez E**, **Sandoval-Almazan R and Ignacio Criado J** (2019) A review of artificial intelligence in government and its potential from a public policy perspective. In *Proceedings of the 20th Annual International Conference on Digital Government Research*, 91–99. https://doi.org/10.1145/3325112.3325242

Van Rooy D (2023) Designing the interaction between humans and autonomous systems: the role of behavioral science. In De Sainz Molestina D, Galluzzo L, Rizzo F, Spallazzo D (eds.), *IASDR 2023: Life-Changing Design*, 9–13 October, Milan, Italy. Conference Series. https://doi.org/10.21606/iasdr.2023.457

Vargas DV and Lauwereyns J (2021) Setting the space for deliberation in decision-making. *Cognitive Neurodynamics*, *15*(5), 743–755. https://doi.org/10.1007/s11571-021-09681-2

Walton P (2018). Artificial intelligence and the limitations of information. *Information. An International Interdisciplinary Journal 9*(12), 332. https://doi.org/10.3390/info9120332

Weser M, **Off D and Zhang J** (2010) HTN robot planning in partially observable dynamic environments. In *2010 IEEE International Conference on Robotics and Automation*, 1505–1510. https://doi.org/10.1109/ROBOT.2010.5509770

Xiong W, **Fan H**, **Ma L and Wang C** (2022) Challenges of human—machine collaboration in risky decision-making. *Frontiers of Engineering Management 9*(1), 89–103. https://doi.org/10.1007/s42524-021-0182-0

Cite this article: Van Rooy D (2024). Human–machine collaboration for enhanced decision-making in governance. *Data & Policy*, 6: e60. doi:10.1017/dap.2024.72

https://doi.org/10.1017/dap.2024.72 Published online by Cambridge University Press