

INVESTIGATING AI MINDFULNESS: AN ANALYSIS OF POLICE FORCES IN ENGLAND AND WALES THROUGH GENERALISED, ANALYTIC INDUCTION

Abstract

Drawing on generalized analytic induction, we explore the conditions that lead to the emergence of AI mindfulness in the policing context in England and Wales. We define AI mindfulness as a cognitive state whereby such users as Chief Digital Officers, Chief Police Constables and Technical Experts are aware of their context and are open to value-adding applications of AI tools. Our results show that AI mindfulness emerges either when such AI users mistrust AI and find AI regulations unclear or when they mistrust AI and find AI tools unexplainable. However, if such AI users find AI regulations to be clear and AI tools to be explainable, they are more likely to take an ambivalent stance towards AI by both trusting it and mistrusting it simultaneously. Accordingly, we uncover three distinct types of mindful AI users, namely, the AI-regulation sceptic, the AI-opposed adopter and the AI-ambivalent adopter. Our typology has wider implications for the responsible use of AI tools.

Keywords: Artificial Intelligence, Mindfulness, Paradox, Analytic Induction

1. Introduction

Generalised analytic induction is a methodology that draws on analytic induction to investigate how things happen (Ragin, 2023). Investigating how Artificial Intelligence (AI) mindfulness emerges is a topic that warrants in-depth exploration because different studies have suggested that humans-in-the-loop may inadvertently ignite algorithmic discrimination if their decision-making process is subconsciously biased (Bauer et al., 2024). Likewise, recent studies show that novice AI users more frequently fail to discard incorrect advice, whereas more experienced AI users more often ignore correct AI advice (Jussupow et al., 2021). Clearly, blindly trusting the AI machine or mistrusting it can have tragic consequences in high-stakes environments such as healthcare, loan approval, hiring or criminal justice systems. Avoiding these tragic consequences calls for a mindful human-in-the-loop, that is, a responsible user that questions the AI output and does not take it at face value.

Thus far, IS researchers have primarily investigated IT mindfulness (Lichtner et al., 2023; Thatcher et al., 2018) paying scant attention to AI mindfulness. We contend that AI warrants a renewed effort towards the study of mindfulness and the cognitive and meta-cognitive issues associated with it (Lebovitz et al., 2022; Jussupow et al., 2021). For example, in the field of medical diagnosis, Jussupow et al. (2021) showed that decision makers can compensate for algorithmic errors by switching from quick, superficial assessments to deliberate, in-depth reasoning when necessary. Drawing on qualitative data from think-aloud protocols, interviews, and questionnaires, Jussupow et al. (2021) emphasised the crucial role of human actors in compensating for technology errors through ongoing reassessment of themselves (i.e., self-monitoring) and the accuracy of the AI advice (i.e., system monitoring). By the same token, Lebovitz et al. (2022) showed that medical professionals dealt with the opacity of AI by enacting AI interrogation practices that resulted in an engaged mode of human-AI augmentation (i.e., engaged augmentation). By enacting AI interrogation practices, medical professionals were able to understand the AI result and change their initial judgment through reflectively agreeing with the AI claim, overruling it, or synthesizing the two claims. Likewise, Asatiani et al. (2021) showed that managing the trade-off between low explainability and high performance presented by inscrutable models requires a balancing act between humans and AI tools to arrive at a more responsible, accountable AI implementation in organisations, whereby humans can gain better control of inscrutable AI models. Drawing on the concept of envelopment, they argued that “by controlling the training data carefully, appropriately choosing both input and output data, and specifying other boundary conditions mindfully, one may permit even inscrutable AI to make decisions, because these specific precautions erect a predictable envelope around the agent’s virtual maneuvering space” (Ibid, p. 327).

Although the literature argues that, by exposing professionals to an opposing view, AI can trigger a reflective process (Abdel-Karim et al. 2023), the intense interaction between humans and machines increases the risk of human biases being carried over to machines (Raisch and Krakowski 2021). The ongoing interaction between human and machines is bound to produce societal effects that are difficult to anticipate (Ibid). No single actor in the human-machine imbrication has full control over these outcomes. Consequently, it is difficult to apportion accountability for outcomes to specific actors, which creates an “accountability gap” (Ibid, p. 205). Accordingly, scholars have called for human oversight mechanisms aimed at preventing poor societal effects such as discrimination and social injustice. For example, Teodorescu et al. (2021) identified two distinct types of oversight, namely, reactive vs. proactive oversight depending on whether humans take a reactive or more proactive stance when monitoring AI’s advice or decisions. However, the argument that the human-in-the-loop should be accountable for decisions made by AI overlooks the fact that humans may suffer from unconscious biases or may act mindlessly in high-stakes environments because of cognitive fatigue or simply because of habit.

Whether using the language of reflection (Abdel-Karim et al., 2023), envelopment (Asatiani et al., 2021), oversight (Teodorescu et al. 2021), meta-cognition (Jussupow et al., 2021) or engaged augmentation (Lebovitz et al. 2022), scholars have been flirting with the idea of mindfulness and its cognitive anchoring. Nevertheless, neither these scholars nor the IT mindfulness scholars (e.g., Lichtner et al., 2023; Thatcher et al., 2018) have investigated under what conditions AI mindfulness happens. We submit that investigating the conditions that lead to the emergence of AI mindfulness plays an important role both theoretically and empirically. While from a theoretical perspective it is important

to unravel the configurational intricacies that lead to the emergence of AI mindfulness, from an empirical perspective understanding such configurations can help identifying different types of mindful AI users and, therefore, design AI systems that prompt users to become more mindful when dealing with AI advice. The remainder of this paper is structured as follows. Section 2 reviews background literature on AI in general and AI in policing in particular while introducing a configurational model for AI mindfulness. Section 3 showcases the methodology used while sections 4 analyses both primary and secondary data. Section 5 discusses the main findings of this paper while section 6 concludes this work.

2. Theoretical background and research model

AI encompasses a wide range of digital technologies, including algorithms and computer systems, that can replicate human thinking, learning, and problem-solving to enhance performance (Wirtz et al., 2019, p. 599). For instance, Mikalef and Gupta (2021) describe AI as "the ability of a system to identify, interpret, make inferences, and learn from data to achieve predetermined organisational and societal goals" (p. 3). Similarly, Wirtz et al. (2019) define AI as "the capability of a computer system to exhibit humanlike intelligent behavior, characterized by core competencies such as perception, understanding, action, and learning" (p. 599).

In a recent paper, Nishant et al. (2024) highlight concerns regarding bias in AI, asserting that "AI-based algorithms are currently unable to effectively interpret the contextual and cultural nuances present in datasets, even when the data itself is unbiased. This limitation exists even in the most advanced AI algorithms, as their judgments are based solely on the analytic procedures of formal rationality" (p. 12). Drawing on Weber's (1978) distinction between formal and substantive rationality, they argue that "AI-based decision-making can exacerbate bias because it operates exclusively through the 'formal' rationality of mathematical optimization procedures. In contrast, avoiding bias requires 'substantive' rationality, which involves a values-based understanding and consideration of context" (Nishant et al., 2024, p. 20). Similarly, Kordzadeh and Ghasemaghaei (2022) note that algorithmic bias "occurs when the outputs of an algorithm disproportionately benefit or disadvantage certain individuals or groups without a justified reason for such unequal impacts" (p. 388). Indeed, researchers often use the words algorithmic bias, algorithmic discrimination, and algorithmic unfairness interchangeably (cf. Bauer et al., 2024, p. 807, footnote 1). These insights have sparked an ongoing debate about the need for AI regulation, with the goal of establishing principles to prevent or at least minimise the negative consequences associated with AI use (Mikalef et al., 2022). Some scholars, like Vanneste and Puranam (2024), argue that regulating AI could undermine trust in these technologies. However, other scholars, including Mikalef et al. (2022), have been more supportive of AI regulation.

The field of AI in policing has faced similar challenges, particularly regarding transparency in machine learning methods. Kazim et al. (2020) highlight the "black-box" nature of these methods, where the connection between inputs and outputs is obscured, leading to issues with explainability and, consequently, accountability in AI-driven decisions. Despite these concerns, Kazim et al. (2020) argue that the use of AI in policing is set to increase significantly due to the growing volume and complexity of digital data, law enforcement's proactive approach to crime, and the general reliance on technology, particularly AI. Similarly, Marciniak (2023) explored AI use in UK policing, focusing on predictive policing tools that generate "risk scores" to predict re-offending. Contrary to concerns about replacing "substantive rationality" with "formal rationality," Marciniak (2023) found that police officers continue to exercise professional judgment, balancing these risk scores with other priorities and operational constraints.

Continuing in a similar vein, Sandhu and Fussey (2021) demonstrated that police officers developed a sceptical attitude towards predictive technologies such as CompStat and Predpol due to their awareness of errors and biases in the input data. They therefore argued for the importance of adopting a "techno-social lens" when studying predictive policing—one that recognises the partnership between human subjectivity and algorithmic calculation. Through this perspective, predictive technologies are seen as a form of augmentation rather than automation (Raisch & Krakowski, 2021), enhancing the capabilities of human police officers rather than simply replacing them with technological recommendations. This

reinforces the idea that the "formal rationality" embedded in AI technologies should complement, rather than replace, an officer's "substantive rationality" and professional judgement.

However, other scholars have reached somewhat different conclusions. In their investigation of the Criminal Anticipation System (CAS) within Dutch police forces, Waardenburg et al. (2018) observed that police officers were so impressed by the complexity (and perhaps opacity) of the predictive policing algorithms that they began to place significant trust in the expertise of intelligence officers. Originally intended to provide back-office support, these intelligence officers found their role expanding to include interpretations and judgements necessary to make algorithmic outputs relevant to police work. As a result, the intelligence officers reinforced the police officers' belief in the superiority of algorithmic predictions.

Some sceptics argue that pre-existing biases and social inequalities can become entrenched during the design and operation of AI, potentially reinforcing these inequalities if such systems are used carelessly or if automated decision-making processes are implemented (McDaniel & Pease, 2021). These scholars point out that records and datasets held by police forces can often be disparate, fragmented, inaccurate, outdated, and imperfectly sourced. It is not uncommon for digital data to be incorrectly and incompletely recorded multiple times a day across entire police areas. These issues mean that police data, frequently used by machine learning systems as training data for predictive algorithms, can reflect a multitude of systematic biases and errors, including those related to race (Ibid).

Similarly, just as biases can infiltrate the design of a statistical model, they can also affect interpretation—where the interpreter imposes political orientations, values, and framings when interpreting results (Završnik, 2021). Moreover, language itself carries biases. Training AI on large datasets means that AI will inevitably absorb the existing biases present in society (Ibid). Bias can enter existing datasets in various ways. For instance, bias can "slip into AI systems constructed with machine learning based on a dataset with training data. When humans are required to label the data explicitly, their personal prejudices and biases may infiltrate the AI system. While this can be partially mitigated by having multiple people label the same data, systematic biases among the population of labelers would still persist" (Dechesne et al., 2019; p. 17). "Bias may also arise from unbalanced or otherwise corrupted datasets; for example, facial recognition software often performs better on white faces than on black faces" (Ibid, p. 17).

In a similar vein, Dolata et al. (2022) have identified various sources of algorithmic bias (or unfairness), ranging from technological unfairness (arising when technology produces a biased algorithmic output) to information unfairness (stemming from biased data and models). Additionally, unfairness can emerge from the interactions between social and technical components, as well as from the embedding of technological algorithms within a broader context, such as a larger social, economic, regulatory, or material environment. Using recidivism risk assessment as a recurring example, Dolata et al. (2022) advocate for a socio-technical perspective that transcends a narrow view of algorithmic fairness—one that seeks either technical solutions (e.g., mathematical formalisations) or social solutions (e.g., human-in-the-loop approaches). Along the same lines, some scholars have reported that automated AI systems used to predict defendants' future criminal behaviour are biased against black defendants (Daugherty and Wilson, 2018). Likewise, other scholars have argued that algorithms are automating inequality because police forces are profiling and punishing the poorest and most marginalised people in our society (Eubanks, 2018). Yet other scholars have argued that the use of predictive analytics in policing has "the potential to technologically reify bias and deepen existing patterns of inequality" in our society (Brayne, 2017, p. 978).

2.1 Research model

AI & Analytics produce paradoxical effects. On the one hand, they allegedly enhance the accuracy and effectiveness of police forces. For example, PredPol predicts crime based on three coordinates, namely, crime type, location and time. Accordingly, it should reduce crime rates. However, it may replicate and reinforce social biases (e.g., against the poor, racial minorities, etc.) if it is trained with biased data which can trigger "bias-reinforcing loops" (Kordzadeh & Ghasemaghahi, 2022, p. 390). To assuage these paradoxical effects, it is therefore necessary that AI users are aware of the context and content of the information produced by the AI (Lichtner et al., 2023). Drawing on an iterative dialogue between

theories of IT mindfulness (Thatcher et al., 2018), AI trust (Gulati et al., 2019), AI explainability (Holzinger et al., 2020) and empirical data, we develop a research model that is informed by extant literature as follows:

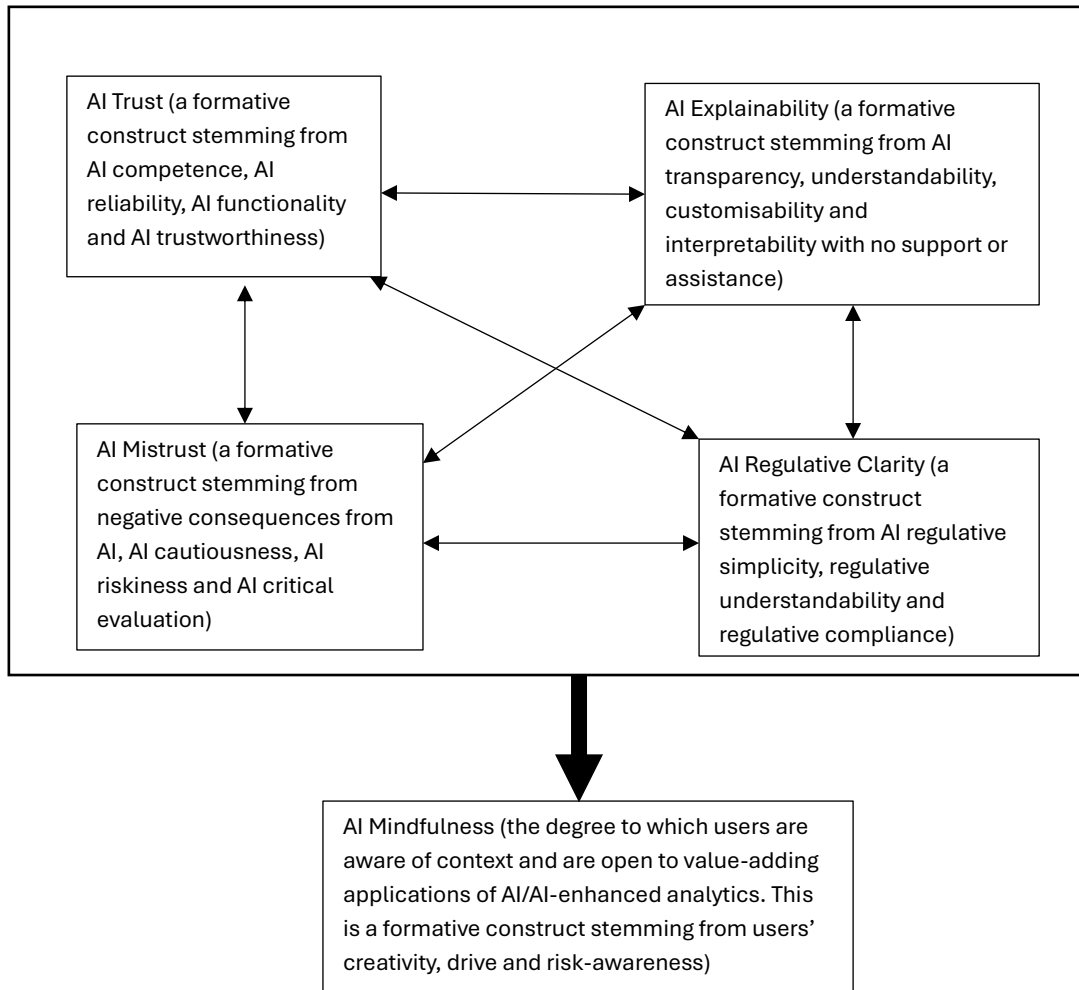


Figure 1: The research model (Adapted from Gulati et al., 2019; Holzinger et al., 2020; Thatcher et al., 2018). Legend: Bidirectional arrows: Reciprocal interdependence. Large box and large black arrow: Synergistic effect stemming from all conditions acting in a conjunctural fashion

Mindfulness is here defined as “as an overarching mental mindset driven by individual awareness of the context, and openness to, value-adding applications of Information Technology (IT)” (Thatcher et al., 2018, p. 832). Echoing Thatcher et al.’s, (2018) notion of IT mindfulness, we define AI mindfulness as a dynamic algorithmic-specific mindset, evident when working with AI & Analytics, whereby the user (e.g., the police officer) or, for that matter, the Chief Digital Officer (CDO) focuses on the present, pays attention to detail, exhibits a willingness to consider other uses, and expresses genuine interest in investigating algorithmic features and failures. Mindful AI users “are more likely to detect biases and react to them by refusing to accept the advice generated by the system, refusing to adopt the system, and engaging in algorithm aversion” (Kordzadeh & Ghasemaghaei, 2022, p. 402). Hence, AI mindfulness does not necessarily translate in algorithmic appreciation. Below, we explain the conditions leading to a mindful use of AI & Analytics in policing in more detail:

1) AI Trust: AI Trust is a formative construct stemming from the belief in the competence, reliability, functionality, and trustworthiness of AI and AI-enhanced analytics systems (cf. Gulati et al., 2019).

2) AI Mistrust: AI Mistrust is a formative construct stemming from the potential negative consequences from AI, the need for caution, the perceived risks, and the importance of critically evaluating AI outputs in daily work routines (cf. Gulati et al., 2019).

3) AI Explainability: AI Explainability is a formative construct stemming from the transparency, auditability, and controllability of AI systems, as well as confidence in understanding and customising the AI system without needing additional support (cf. Holzinger et al., 2020).

4) AI Regulative Clarity: AI Regulative Clarity is a formative construct stemming from the clarity of laws, regulations, policies, and norms governing AI, an understanding of these regulations, and a confidence in being able to comply with them based on substantive knowledge (cf. Kordzadeh & Ghasemaghaei, 2022).

We expect these conditions to generate AI mindfulness in a synergistic fashion because they cut across the social (e.g., AI Trust, AI Mistrust, etc.) and technical (e.g., AI Transparency, AI Customisability, etc.) drivers of technology adoption. Given this socio-technical context, we ask: when does AI mindfulness happen?

Before we address this question, we need to explore the methodological underpinnings of this study.

3. Methods

To answer our research question, we draw on generalised analytic induction to study how AI mindfulness emerges because it restricts its analytic focus to cases that display the outcome. Generalised analytic induction looks for multiple antecedent conditions that are linked to a qualitative outcome across a range of positive cases. Generalised analytic induction is a set-analytic technique that revolves around the conditions linked to positive cases. Accordingly, negative cases, that is, cases that fail to exhibit the focal outcome, are regarded as instances of something else and, therefore, are the focus of separate analyses. Generalised analytic induction is a descriptive technique that is “best understood as an aid to causal interpretation” (Ragin, 2023, p. 4). Compared to classic analytic induction, generalised analytic induction does not look for universal relationships, but for “modal configurations”, that is, “combinations of antecedent conditions that are relatively common in a set of cases exhibiting the outcome in question” (Ragin, 2023, p. 51).

Drawing on a random, purposeful sample (Miles & Huberman, 1994) of 21 Chief Digital Officers (CDOs), Chief Police Constables and Technical Experts who are involved in the use of AI in policing (see Appendix), we report below the results of our generalised analytic induction following its basic steps:

- 1) Defining the outcome of interest: our outcome of interest is AI mindfulness defined as an overarching mental mindset driven by individual creativity, drive and risk awareness;
- 2) Identifying relevant instances of the outcome: the cases at hand are CDOs, Chief Police Constables and Technical Experts drawn from a circumscribed and well-defined set of police practitioners;
- 3) Conduct case-level research in order to identify the central contributing conditions for each case: drawing on a dialogue between theories of IT mindfulness (e.g., Lichtner et al., 2023; Thatcher et al., 2018), AI trust (e.g., Gulati et al., 2019), AI explainability (e.g., Holzinger et al., 2020) and substantive knowledge, we identified three formative items for AI Mindfulness. Furthermore, as a result of this iterative dialogue between theory and data, we identified four relevant conditions leading to AI Mindfulness, namely: AI Trust, AI Mistrust, AI Explainability and AI Regulative Clarity. AI Trust was measured based on four items adapted from Gulati et al. (2019). AI Mistrust was measured based on four items adapted from Gulati et al. (2019). AI Explainability was measured based on four items adapted from Holzinger et al. (2020). AI Regulative Clarity was measured based on three items adapted from Kordzadeh & Ghasemaghaei (2022);
- 4) Assessing the membership of each case in each condition: each informant was asked to complete a questionnaire where each relevant item was measured on a Likert-scale going from 1 (fully disagree) to 5 (fully agree). Each condition was measured using several survey items that met the Variance Inflation Factors (VIF) score of ≤ 5 , thus leading to the construction of formative constructs

(see Appendix). We subsequently set 2 as the degree of full non-membership, 3 as the cross-over point and 4 as the degree of full membership in each fuzzy set;

5) Using the fsQCA software to analyse the data: we did not set any directional expectations because of the exploratory nature of our research;

6) Establishing a frequency threshold: despite the small sample size, we established a frequency threshold of 3 cases to look for modal configurations. This enabled us to retain 76% of the cases;

7) Running the analysis: we first used generalised analytic induction and, subsequently, we looked for necessary conditions for the emergence of AI mindfulness. Our analyses produced the results reported below. In what follows, we strive to illuminate these findings with rich, qualitative data obtained by means of 15 semi-structured interviews aptly triangulated with secondary data.

4. Analysis

Using generalised analytic induction in combination with necessary condition analysis, we arrived at these results.

outcome filtering: $c_AIMindfulness \leq 1$ (where $c_AIMindfulness$ is a calibrated measure of the formative construct capturing AI Mindfulness)

condition coding:

$c_AITrust$ coded as either contributing to $c_AIMindfulness$ when present or absent ($c_AITrust$ is a calibrated measure of the formative construct capturing trust in AI)

$c_AIMistrust$ coded as either contributing to $c_AIMindfulness$ when present or absent ($c_AIMistrust$ is a calibrated measure of the formative construct capturing mistrust in AI, that is, AI Mistrust)

$c_AIExplainability$ coded as either contributing to $c_AIMindfulness$ when present or absent ($c_AIExplainability$ is a calibrated measure of the formative construct capturing AI Explainability)

$c_AIRegulative_Clarity$ coded as either contributing to $c_AIMindfulness$ when present or absent ($c_AIRegulative_Clarity$ is a calibrated measure of the formative construct capturing the clarity of rules and norms surrounding AI, that is, AI Regulative_Clarity)

*TRUTH TABLE ANALYSIS

Model: $c_AIMindfulness = f(c_AITrust, c_AIMistrust, c_AIExplainability, c_AIRegulative_Clarity)$

Algorithm: Quine-McCluskey

--- TRUTH TABLE SOLUTION ---

frequency cutoff: 3

	raw coverage	unique coverage
~c_AITrust*c_AIMistrust*c_AIRegulative_Clarity	0.335635	0.13013
~c_AITrust*c_AIMistrust*c_AIExplainability	0.322001	0.07647
c_AITrust*c_AIMistrust*c_AIExplainability*c_AIRegulative_Clarity	0.170877	0.292247

solution coverage: 0.633161

This, in turn, produced one modal configuration as follows:

$c_AIMistrust * (\sim c_AITrust * \sim c_AIRegulative_Clarity + \sim c_AITrust * \sim c_AIExplainability + c_AITrust * c_AIExplainability * c_AIRegulative_Clarity)$

where: + stands for logical operator OR;

* stands for logical operator AND

~ stands for logical operator NOT

Furthermore, this modal configuration highlights the role of mistrust in achieving a mindful use of AI. This point is corroborated by a recent report:

“Police officers may be better equipped than many professionals to apply a **suitable level of scepticism to the outcome of an algorithm**, given that weighing the reliability of evidence is so fundamental to their general professional practice” Centre for Data Ethics and Innovation (2020, p. 68)

Indeed, the necessary condition analysis confirms that AIMistrust is a relevant necessary condition for AI Mindfulness to emerge because it meets the 0.90 Consistency Necessity threshold, as well as the 0.60 Coverage Necessity and Relevance of Necessity (RoN) thresholds (see the condition highlighted in bold below):

	Cons.	Nec	Cov.	Nec	RoN
c_AITrust	0.524	0.952	0.957		
c_AIMistrust	0.920	0.949	0.771		
c_AIExplainability	0.659	0.985	0.979		
c_AIRegulative_Clarity	0.614	0.959	0.949		
~c_AITrust	0.572	0.977	0.976		
~c_AIMistrust	0.166	1.000	1.000		
~c_AIExplainability	0.448	0.959	0.972		
~c_AIRegulative_Clarity	0.479	0.967	0.975		

Taken together, these findings show that mistrusting AI is necessary, but not sufficient for the emergence of AI mindfulness. Other conditions are required for the full causal recipe. In particular, our informants are mindful AI users either when they mistrust AI and find the regulations surrounding AI to be unclear OR when they mistrust AI and find that the AI tools are unexplainable OR when they simultaneously trust and mistrust AI provided that they find that AI regulations are clear and that AI tools are explainable. Since the last path is the path with the largest unique coverage, it can be argued that this recipe holds the largest explanatory power. Paradoxically, AI mindfulness seems to emerge when our informants both trust and mistrust AI provided that they find the AI regulations clear and AI itself explainable, that is, transparent and customisable. Again, the centre for Data Ethics and Innovation (2020) corroborates this point:

“At the deployment stage, bias may occur in the way the human decision-maker receiving the output of the algorithm responds. One possibility is that the decision maker over-relies on the automated output, without applying their professional judgement to the information. The opposite is also possible, where the human decision maker feels inherently uncomfortable with taking insights from an algorithm to the point where they are nervous to use it at all, or simply ignores it in cases where their own human bias suggests a different risk level. **A balance is important to ensure due regard is paid to the insights derived**, whilst ensuring the professional applies their expertise and understanding of the wider context and relevant factors” (Ibid, p. 68)

5. Discussion

Spurred by a gap in the literature that urged us to investigate the conditions under which AI mindfulness happens, we have discovered three distinct paths to AI mindfulness aptly summarised above. In this section we discuss our findings in more detail to draw some preliminary conclusions around the notion of AI mindfulness in the context of the use of AI in policing.

The first path to AI mindfulness is characterised by the so-called AI-regulation sceptic who does not trust AI and finds the norms and rules surrounding AI to be unclear. For example, a Solutions Manager stated: “Policing needs rules. We work really well to rules. Otherwise, we do 43 different versions of nothing really well, actually.”¹ The second path to AI mindfulness is characterised by the so-called AI-opposed adopter who mistrusts AI while simultaneously finding it unexplainable, that is, neither transparent nor customisable. The Head of Innovation at the National Police Chiefs Council (NPCC) observed, “I think there is this element of suspicion, of mistrust in technology especially from the most experienced cops on the street... and that's always been the case.”

Both the AI-regulation sceptic and the AI-opposed adopter do not trust AI. For example, the Head of IT of a police force emphasised the risky nature of AI due to biases stemming either from the data or from the developers themselves: “The software developer tried to tell me that there is absolutely no possibility of any bias to be fed into that (AI tool). I get what he's saying because it's coming from raw data, but, ultimately, he has created the algorithm that makes it learn. And that's him, him as a human being, he will have inbuilt biases that he potentially has fed in to the creation of the original algorithm.” Likewise, another informant stressed the inherent difficulty in eliminating algorithmic bias: “humans have like innate unconscious bias, don't they? If the machine is learning from a human, it may well pick it up” (Digital Innovation Manager).

Hence, AI mistrust can be critical for increasing AI mindfulness because it “improves mental processes such as attention and selection and avoids mindless practices such as routines and overreliance on rules” (Aaltonen et al., 2023, p. 8). Likewise, AI ambivalence too can play a critical role for the emergence of AI mindfulness (Ibid). The Head of Innovation in a police force, for example, stated: “I think we are going to have huge optimism bias in some of the tools and capabilities, because they will look incredible and they will do incredible things. But our end users won't necessarily have the technical skills to be able to truly question the output.” While it is plausible that AI users “just kind of trust the AI” over time (PMO Manager), it is imperative that they develop those technical skills to be able to truly question the AI outputs, thus retaining an ambivalent nature where they simultaneously trust and mistrust AI. Indeed, the path with the largest unique coverage is characterised by the so-called AI-ambivalent adopter who simultaneously trusts and mistrusts AI while finding the AI regulations clear and AI itself to be explainable, that is, transparent and customisable.

While the AI-regulation sceptic and the AI-opposed adopter capture more experienced AI users who question the technology while finding the AI regulation unclear or the AI tools unexplainable, our findings show that, even when AI users find AI regulations clear, AI tools to be trustworthy, as well as explainable, they will take an ambivalent stance towards AI by simultaneously trusting it and mistrusting it to avoid overreliance on AI: “The problem is that we become too reliant on the tool and no longer use our own judgment, which can become an issue because it can fabricate data, it can misrepresent data, etc. And that can actually become an issue. Imagine, you know, the police are charging the wrong suspect because of fabricated data... it can be confidently incorrect, and that is the problem” (PMO Manager). Ultimately, either distrust in AI (see the AI-regulation sceptic and the AI-opposed adopter) or ambivalence (see the AI-ambivalent adopter) foster AI mindfulness because they improve mental processes such as attention and selection and avoid mindless practices such as routines and overreliance on AI.

¹ Please note that there are 43 independent police forces in England and Wales.

6. Conclusion

AI is the engine of growth of modern economies. Yet, it may replicate and reinforce the social biases that exist in society, thus benefiting or disadvantaging “certain individuals or groups more than others without a justified reason for such unequal impacts” (Kordzadeh & Ghasemaghahi, 2022, p. 388). In high-stake environments such as policing, algorithms can create biases through bias-reinforcing loops that affect areas with high poverty rates, which are also home to significant numbers of underprivileged populations and racial minorities (Ibid). Accordingly, it is imperative to develop bias-mitigation strategies to make sure that AI is used responsibly in such environments. Increasing the transparency, auditability, and control features of AI tools is one way to mitigate algorithmic biases. Yet, algorithmic bias is a sociotechnical issue that involves both the technology and human actors (Ibid). Hence, it is critical to empower AI users to play an active role in bias detection and mitigation strategies. So far, the literature has discussed the cognitive underpinnings of the adoption of AI by using the language of reflection (Abdel-Karim et al., 2023), envelopment (Asatiani et al., 2021), oversight (Teodorescu et al. 2021), meta-cognition (Jussupow et al., 2021) or engaged augmentation (Lebovitz et al. 2022). Yet, the literature has neither unpacked the notion of AI mindfulness nor the conditions that make it happen. AI mindfulness is a crucial mechanism to compensate for algorithmic bias because it ensures the presence of a mindful user who questions the veracity of the AI output. Drawing on a multiplicity of theories ranging from AI Trust, AI Explainability and IT Mindfulness, this paper aims to fill this gap.

Specifically, it reveals three distinct recipes that underpin the emergence of AI Mindfulness. The first two recipes speak to those informants who are cautious when using AI either because they mistrust AI and find the regulations surrounding AI unclear OR because they mistrust AI and find the AI tools unexplainable, that is, neither transparent nor customisable. We have referred to these two types of mindful AI users as the so-called AI-regulation sceptic and the AI-opposed adopter. Hence AI mindfulness may be regarded as a cognitive state whereby users “are more likely to detect biases and react to them by refusing to accept the advice generated by the system, refusing to adopt the system, and engaging in algorithm aversion” (Kordzadeh & Ghasemaghahi, 2022, p. 402). Accordingly, AI mindfulness is about critically interpreting the context and content of the AI-generated information output rather than taking the AI output at face value.

The third recipe for the emergence of AI Mindfulness is more equivocal because it underscores its paradoxical nature since informants both trust and mistrust AI if they find AI rules clear and AI tools to be both transparent and customisable, that is, explainable. Although there is an element of scepticism towards AI particularly among the most experienced police officers, there is also recognition that, over time, trust in AI might grow as the technology becomes more integrated into daily operations. Yet, to avoid overreliance on AI, mindful AI users will develop a prowess to simultaneously trust and mistrust AI, thus becoming ambivalent AI users.

Ultimately, none of these ingredients in isolation will generate AI mindfulness. They need to operate jointly and synergistically for AI mindfulness to emerge. Whether looking at the AI-regulation sceptic, the AI-opposed adopter or the AI-ambivalent adopter, AI mindfulness presupposes mistrust in the technology. Since AI users may inadvertently resort to habit and tried-and-tested routines when using AI, it is of the foremost importance to design AI tools that prompt their mistrust in the technology to avoid overreliance. While the literature has advocated awareness training programmes to avoid discriminatory practices in the use of AI (Bauer et al., 2024), it is important to understand that another and perhaps more promising way of achieving AI mindfulness is by designing prompts that constantly alert individuals to be mindful AI users who question the technology and mistrust it. Design science in general and prompt engineering in particular could help the human in the loop becoming a mindful AI user rather than a mindless AI user that relies on tried-and-tested routines. Simple tweaks, such as designing tools that indicate low confidence in their recommendations, could go a long way toward fostering AI mindfulness.

Acknowledgement

This research was supported by the BA/Leverhulme Small Research Grants (grant number: SRG2223\230011). The author would like to thank Dionysios Demetis, Stefano De Rigo, Stan Karanasios, Peter Kawalek, Hamid Khobzi, Patrick Mikalef, Rohit Nishant and Gianluigi Viscusi for their insightful comments to earlier draft of this paper.

References

- Aaltonen, A., Alaimo, C., Parmiggiani, E., Stelmaszak, M., Jarvenpaa, S. L., Kallinikos, J., & Monteiro, E. (2023). What is Missing from Research on Data in Information Systems? Insights from the Inaugural Workshop on Data Research. *Communications of the Association for Information Systems*, 53, pp. 1-19. <https://doi.10.17705/1CAIS.05320>
- Asatiani, A., Malo, P., Nagbøl, P. R., Penttinen, E., Rinta-Kahila, T., & Salovaara, A. (2021). Sociotechnical envelopment of artificial intelligence: An approach to organizational deployment of inscrutable artificial intelligence systems. *Journal of the Association for Information Systems*, 22(2), 325-352. <https://doi.org/10.17705/1jais.00664>
- Abdel-Karim, B. M., Pfeuffer, N., Carl, K. V., & Hinz, O. (2023). How AI-Based Systems Can Induce Reflections: The Case of AI-Augmented Diagnostic Work. *MIS Quarterly*, 47(4), 1395-1424. <https://DOI:10.25300/MISQ/2022/16773>
- Brayne, S. (2017). Big data surveillance: The case of policing. *American Sociological Review*, 82(5), 977-1008. <https://doi.10.1177/0003122417725865>
- Babuta, A., & Oswald, M. (2020). *Data analytics and algorithms in policing in England and Wales: Towards a new policy framework*. Royal United Services Institute for Defence and Security Studies. Retrieved November 17, 2024 from <https://rusi.org/explore-our-research/publications/occasional-papers/data-analytics-and-algorithms-policing-england-and-wales-towards-new-policy-framework>
- Bauer, K., Heigl, R., Hinz, O., & Kosfeld, M. (2024). Feedback loops in machine learning: A study on the interplay of continuous updating and human discrimination. *Journal of the Association for Information Systems*, 25(4), 804-866. <https://doi.10.17705/1jais.00853>
- Centre for Data Ethics and Innovation (2020). *Review into bias in algorithmic decision-making*. London. Retrieved November 17, 2024 from https://assets.publishing.service.gov.uk/media/60142096d3bf7f70ba377b20/Review_into_bias_in_algorithmic_decision-making.pdf
- Daugherty, P., & Wilson, H. J. 2018. *Human + machine: Reimagining work in the age of AI*. Boston, MA: Harvard Business Review Press.
- Dechesne, F., Dignum, V., Zardiashvili, L., & Bieger, L. J. (2019). *AI & Ethics at the Police. Towards Responsible use of Artificial Intelligence in the Dutch Police*. Retrieved November 17, 2024 from <https://www.universiteitleiden.nl/binaries/content/assets/rechtsgeleerdheid/instituut-voor-metajuridica/artificiele-intelligentie-en-ethiek-bij-de-politie/ai-and-ethics-at-the-police-towards-responsible-use-of-artificial-intelligence-at-the-dutch-police-2019..pdf>
- Dempsey, R. P., Brunet, J. R., & Dubljević, V. (2023). Exploring and Understanding Law Enforcement's Relationship with Technology: A Qualitative Interview Study of Police Officers in North Carolina. *Applied Sciences*, 13(6), 3887. <https://doi.10.3390/app13063887>
- Dolata, M., Feuerriegel, S., & Schwabe, G. (2022). A sociotechnical view of algorithmic fairness. *Information Systems Journal*, 32(4), 754-818. <https://doi.org/10.1111/isj.12370>
- Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. New York, NY: St. Martin's Press.

- Giermindl, L. M., Strich, F., Christ, O., Leicht-Deobald, U., & Redzepe, A. (2022). The dark sides of people analytics: reviewing the perils for organisations and employees. *European Journal of Information Systems*, 31(3), 410-435. <https://doi.org/10.1080/0960085X.2021.1927213>
- Gulati, S., Sousa, S., & Lamas, D. (2019). Design, development and evaluation of a human-computer trust scale. *Behaviour & Information Technology*, 38(10), 1004-1015. <https://doi.org/10.1080/0144929X.2019.1656779>
- Holzinger, A., Carrington, A., Müller, H (2020). Measuring the quality of explanations: the system causability scale (SCS) comparing human and machine explanations. *KI-Künstliche Intelligenz*. 34(2): 193-198. <https://doi.org/10.1007/s13218-020-00636-z>
- Jussupow, E., Spohrer, K., Heinzl, A., & Gawlitza, J. (2021). Augmenting medical diagnosis decisions? An investigation into physicians' decision-making process with artificial intelligence. *Information Systems Research*, 32(3), 713-735. <https://doi.org/10.1287/isre.2020.0980>
- Kazim, E., Koshiyama, A., & Denny, D. (2020). *AI and Policing: Impact Assessment and Empirical Evidence Needed*. Retrieved November 17, 2024 from: https://www.researchgate.net/profile/Denny-Thame/publication/349085018_AI_and_Policing_Impact_Assessment_and_Empirical_Evidence_Needed/links/60b7a91e299b106f6f767bd/AI-and-Policing-Impact-Assessment-and-Empirical-Evidence-Needed.pdf
- Kordzadeh, N., & Ghasemaghaei, M. (2022). Algorithmic bias: review, synthesis, and future research directions. *European Journal of Information Systems*, 31(3), 388-409. <https://doi.org/10.1080/0960085X.2021.1927212>
- Lichtner, V., Karanasios, S., & Iannacci, F. (2023). Walking the line: Mindfulness with IT in hospital medication routines. *Information and Organization*, 33(3), 100475. <https://doi.org/10.1016/j.infoandorg.2023.100475>
- Lebovitz, S., Lifshitz-Assaf, H., & Levina, N. (2022). To engage or not to engage with AI for critical judgments: How professionals deal with opacity when using AI for medical diagnosis. *Organization Science*, 33(1), 126-148. <https://doi.org/10.1287/orsc.2021.1549>
- Marciniak, D. (2023). Algorithmic policing: an exploratory study of the algorithmically mediated construction of individual risk in a UK police force. *Policing and Society*, 33(4), 449-463. <https://doi.org/10.1080/10439463.2022.2144305>
- Marjanovic, O., Cecez-Kecmanovic, D., & Vidgen, R. (2022). Theorising algorithmic justice. *European Journal of Information Systems*, 31(3), 269-287. <https://doi.org/10.1080/0960085X.2021.1934130>
- McDaniel, J., & Pease, K. (Eds.). (2021). *Predictive policing and artificial intelligence. Introduction*. Oxon, Routledge.
- Mikalef, P., & Gupta, M. (2021). Artificial intelligence capability: Conceptualization, measurement calibration, and empirical study on its impact on organizational creativity and firm performance. *Information & Management*, 58(3), 103434. <https://doi.org/10.1016/j.im.2021.103434>
- Mikalef, P., Conboy, K., Lundström, J. E., & Popovič, A. (2022). Thinking responsibly about responsible AI and 'the dark side' of AI. *European Journal of Information Systems*, 31(3), 257-268. <https://doi.org/10.1080/0960085X.2022.2026621>
- Nishant, R., Schneckenberg, D., & Ravishankar, M. N. (2024). The formal rationality of artificial intelligence-based algorithms and the problem of bias. *Journal of Information Technology*, 39(1), 19-40. <https://doi.org/10.1177/02683962231176842>
- Raisch, S., & Krakowski, S. (2021). Artificial intelligence and management: The automation–augmentation paradox. *Academy of Management Review*, 46(1), 192-210. <https://doi.org/10.5465/amr.2018.0072>

- Ragin, C. C. (2023). *Analytic induction for social research*. Oakland: University of California Press.
- Sandhu, A., & Fussey, P. (2021). The ‘uberization of policing’? How police negotiate and operationalise predictive policing technology. *Policing and Society*, 31(1), 66-81. <https://doi.org/10.1080/10439463.2020.1803315>
- Teodorescu, M. H., Morse, L., Awwad, Y., & Kane, G. C. (2021). Failures of Fairness in Automation Require a Deeper Understanding of Human-ML Augmentation. *MIS Quarterly*, 45(3), 1483-1499. <https://doi.10.25300/MISQ/2021/16535>
- Thatcher, J. B., Wright, R. T., Sun, H., Zagencyk, T. J., & Klein, R. (2018). Mindfulness in information technology use. *MIS Quarterly*, 42(3), 831-848. <https://www.jstor.org/stable/26635055>
- Urquhart, L., & Miranda, D. (2022). Policing faces: the present and future of intelligent facial surveillance. *Information & Communications Technology Law*, 31(2), 194-219. <https://doi.org/10.1080/13600834.2021.1994220>
- Vanneste, B. S., & Puranam, P. (2024). Artificial Intelligence, Trust, and Perceptions of Agency. *Academy of Management Review*, In press. <https://doi.org/10.5465/amr.2022.0041>
- Waardenburg, L., Sergeeva, A., & Huysman, M. (2018). Hotspots and blind spots: A case of predictive policing in practice. Proceedings of the IFIP WG 8.2 Working Conference on the Interaction of Information Systems and the Organization, San Francisco, CA, USA.
- Weber, M. (1978). *Economy and society: An outline of interpretive sociology*. Berkeley. University of California Press.
- Wirtz, B. W., Weyerer, J. C., & Geyer, C. (2019). Artificial intelligence and the public sector—applications and challenges. *International Journal of Public Administration*, 42(7), 596-615. <https://doi.org/10.1080/01900692.2018.1498103>
- Završnik, A. (2021). Algorithmic justice: Algorithms and big data in criminal justice settings. *European Journal of Criminology*, 18(5), 623-642. <https://doi.org/10.1177/1477370819876762>

Appendix

Collinearity statistics (VIF)

Outer model – List (VIF < 5)

	VIF
11_1 (AI Competence; Adapted from Gulati et al. 2019)	4.627
11_2 (AI Reliability; Adapted from Gulati et al. 2019)	4.943
11_3	9.243
11_4 (AI Functionalities; Adapted from Gulati et al. 2019)	4.776
11_5 (AI Trustworthiness; Adapted from Gulati et al. 2019)	3.538
12_1 (AI Negative Consequences; Adapted from Gulati et al. 2019)	1.481
12_2 (AI Cautiousness; Adapted from Gulati et al. 2019)	1.632
12_3 (AI Riskiness; Adapted from Gulati et al. 2019)	1.410
12_4 (AI Critical Evaluation; Adapted from Gulati et al. 2019)	1.154
13_1 (AI Transparency; Adapted from Holzinger et al., 2020)	1.138
13_2 (AI Understandability; Adapted from Holzinger et al., 2020)	1.094
13_3 (AI Customisability; Adapted from Holzinger et al., 2020)	1.171
13_4 (AI Interpretability Without Support; Adapted from Holzinger et al., 2020)	1.233

14_1 (AI_Regulative_Simplicity; Inspired by Kordzadeh and Ghasemaghaei, 2022)	1.123
14_2 (AI_Regulative_Understanding; Inspired by Kordzadeh and Ghasemaghaei, 2022)	1.146
14_3 (AI_Regulative_Compliance; Inspired by Kordzadeh and Ghasemaghaei, 2022)	1.216
15_1 (AI_Creativity; Adapted from Thatcher et al., 2018)	2.713
15_2	6.511
15_3 (AI_Drive; Adapted from Thatcher et al., 2018)	1.974
15_4	12.808
15_5	9.352
15_6 (AI_Risk_Awareness; Inspired by Thatcher et al., 2018)	3.405

Details of interviews and secondary sources used

	Exploratory interviews	Secondary data sources
Number	15	6
Details	Digital Trust and Evidence Group Member Retired Assistant Chief Police Constable Head of Innovation (NPCC) Head of Assistance and Statutory Duties Chief Digital and Innovation Officer Digital Innovation Manager Chief Inspector PMO Manager Solutions Manager Head of IT Director of ICT Business Analyst Director of Research, NPCC Technical Business Analyst, PDS Innovation Manager, PDS	Babuta A. and Oswald, M. (2020). Data Analytics and Algorithms in Policing in England and Wales. RUSI Occasional Paper. Berryhill, J., Kok Heang, K., Clogher, R. & McBride, K. (2019). Hello, World. Artificial Intelligence and its use in the Public Sector. OECD. Observatory of Public Sector Innovation. Department for Science, Innovation & Technology (2024). Introduction to AI Assurance. European Commission (2019). Ethics Guidelines for Trustworthy AI. National Police Chief's Council (NPCC). Covenant for Using Artificial Intelligence in Policing (2023). Police Digital Service (PDS). Annual report 2021-2022.