When Should Algorithms Resign?

Umang Bhatt[1,2*] Holli Sargeant[3,4*]

[1]Center for Data Science, New York University; New York City, NY, USA

[2]The Alan Turing Institute; London, UK

[3]Faculty of Law, University of Cambridge; Cambridge, UK

[4]Berkman Klein Center for Internet & Society, Harvard University; Cambridge, MA, USA

This paper discusses *algorithmic resignation*, a strategic approach for managing the use of AI systems within organizations. Algorithmic resignation involves the deliberate and informed disengagement from AI assistance in certain scenarios, by embedding governance mechanisms directly into AI systems. Our proposal is not merely about disuse of AI but includes guiding when and how these systems should be used or avoided. We discuss the multifaceted benefits of algorithmic resignation, spanning economic efficiency, reputational gains, and legal compliance. Further, we outline the operationalization of resignation through various methods such as positive and negative nudges, stakeholder incentive alignment, and careful consideration of the level of AI engagement. Using techniques like barring access to AI outputs selectively or providing explicit disclaimers on system performance, algorithmic resignation not only mitigates risks associated with AI but also leverages its benefits, ensuring the responsible and effective use of AI systems.

[*] Both authors contributed equally. Names listed alphabetically.

Introduction

Tesla recently won a pivotal lawsuit related to its Autopilot system. Jurors held that a fatal accident was not the fault of Tesla's driver-assistance system because all drivers were informed they must maintain full control over the vehicle (*1*). Given the driver was ultimately held responsible, the case underscores the legal challenges of humans and AI systems working together.

When organizations grant members access to AI systems, they are responsible for members' use of such tools. Although AI may be used judiciously by some members, human oversight does not inherently guarantee proper use. Even partial automation may lull some members into complacency (*2*). Just as most drivers are not drunk, many members may use AI appropriately; however, the stakes of AI misuse are high.

Unlike ex-post enforcement against drunk drivers, we advocate building governance mechanisms into AI systems directly. We argue for **algorithmic resignation**, a phenomena where organizations resign algorithmic assistance in favor of (unaided) human decision-making. Resignation places governance into system design and provides concrete guidance for appropriate use, distinguishing permissible levels of automation from inappropriate or unlawful ones. Such an approach would allow organizations to control when and how AI systems are used.

What is Algorithmic Resignation?

Algorithmic resignation is the strategic disengagement or limitation of AI assistance in specific scenarios. This concept is not merely about the disuse of AI but is about embedding governance mechanisms within AI systems, guiding when and how these systems should be used or abstained from. Organizations can embed algorithmic resignation directly into the AI systems they build, procure, or deploy. Deciding when to mandate the disuse of AI systems can depend on a variety of factors (*3*). One set of factors may depend on the AI system's performance. For instance, when an AI system encounters scenarios it has not been trained on (known as "out-of-distribution" instances), or when the system is highly uncertain in its predictions, resignation can act as a safeguard against potentially erroneous predictions.

Other factors depend on the user of the system. As preferences and expertise may differ across members, organizations can enable each member to have access to AI systems when most appropriate for them (*4*). Junior doctors may benefit from assistance more than senior physicians; however, even for the latter, nudging experienced doctors to view predictions from AI systems may counteract overconfidence (*5*). One practical application could be selectively enabling non-native English speakers to use large language models, like ChatGPT, for composing clearer communications, while prohibiting its use for critical decision-making tasks (*6*).

Beyond individual personalization, organizations can elect to create broad policies for algorithmic resignation. Perhaps a law firm allows the use of ChatGPT for internal research or communications but not for external client interactions. Likely organizations would craft a strategy to decide not only how much AI assistance to provide but also to which members.

Operationally, resignation can be implemented in various ways. It can range from completely barring access to AI outputs in certain scenarios to softer measures like explicit disclaimers or guidance. For instance, a large language model might refuse to provide responses to queries that border on legal advice, instead offering general legal information that is unregulated. By promoting disuse of AI systems via resignation, organizations can prevent misuse or, worse, abuse of AI-assisted decision-making.

Benefits of Resignation

Algorithmic resignation offers a range of benefits, spanning economic, reputational, and legal domains. These advantages not only improve the operational efficiency of organizations but also align with broader ethical and regulatory expectations.

Financial

From a financial perspective, algorithmic resignation can lead to significant cost savings and increased efficiency. By enabling selective use of AI, organizations can optimize decision-making processes, ensuring that AI systems are used where they add the most value and avoiding costly errors associated with over-reliance on AI. By providing structured choices on when to use AI, a firm helps employees exercise better control and judgment when they otherwise may not act in the best interests of the firm (*7*). The selective approach approach can be seen as a strategy to align the interests of various stakeholders, thereby optimizing overall performance and minimizing the friction and costs that arise when different parties interests diverge.

Reputational

Implementing algorithmic resignation will demonstrate a commitment to responsible AI, setting a precedent for meaningful commitment to the conscious application of technology. Such commitment is a display of trustworthiness, as stakeholders and the public more broadly observe the organization behaving in *good faith* (*8*). With resignation powering the conscious use of AI systems, stakeholders are assured that the organization's use of AI enhances, not replaces, human judgment and expertise. The reputational backlash from AI misuse is well-documented and litigated in the courts (*9*).

Legal

In the face of increasing calls for external regulation of AI, there is a strong incentive for organizations to develop internal governance mechanisms like algorithmic resignation. Algorithmic resignation allows organizations to self-regulate meaningfully, by embedding governance mechanisms directly within AI systems. This concept then offers a proactive, self-regulatory approach that aligns with, and even anticipates, external regulatory demands.

Emerging legislation, such as the EU's Artificial Intelligence Act and the US Executive Order on AI, is concerned with responsible use and categorising the risk of using different types of AI systems (*10,11*). By embedding principles of algorithmic resignation into their operations, organizations can proactively address these emerging legal requirements. This not only aids in compliance but also reduces the risk of legal challenges stemming from over-reliance on AI, such as those related to errors, privacy breaches, or unethical decision-making. The advent of legislation marks a critical juncture for organizations to reevaluate their AI strategies and adopt internal mechanisms like algorithmic resignation to ensure responsible and compliant AI use.

By strategically implementing resignation, organizations can harness the power of AI while maintaining control, ensuring ethical compliance, and enhancing their reputation in the market.

Considerations of Resignation

The successful deployment of algorithmic resignation requires careful consideration of several factors. These considerations ensure that resignation is not only effective but also aligns with organizational goals and regulatory requirements.

Directionality of Selectivity

Directionality refers to the guidance provided for AI use or disuse. Organizations will decide how to nudge members towards intended use. Encouraging the use or disuse of AI can be achieved through various methods. Nudges can have positive or negative affect, depending on their intended effect and the context in which they are applied (*12*). Positive nudges are crafted to coax individuals gently towards choices that are beneficial both for their personal well-being and the greater societal good. Conversely, negative nudges, while still preserving the essence of choice, subtly discourage actions or decisions that are deemed harmful or less desirable. Such nudges work on the premise of dissuasion rather than prohibition, leveraging human psychology to steer behavior away from less favorable outcomes.

Nudging in respect of AI system can be implemented through restraints, disclosures, and guidance to members using AI systems. A positive nudge could remind members to use AI systems by potentially forcing members to view a AI system recommendation before proceeding or subtly indicating with visual cues that AI system prediction may help. For example, hospital administrators could add a prompt in a clinical decision support system encouraging doctors to review AI suggestions for patient treatment. On the other hand, a negative nudge would deter members from AI assistance, ranging from disclosures about AI shortcomings to restraints on access. Social media companies may embed a warning in an automated content moderation tool indicating high uncertainty in AI-generated predictions, suggesting manual review. This strategic approach ensures that AI is utilized where it adds the most value, avoiding over-reliance while capitalizing on its benefits.

Incentives and Stakeholder Trade-offs

The successful implementation of algorithmic resignation requires an understanding of the various incentives driving different stakeholders. Within organizations, individuals (like managers or employees) may have personal motivations that differ from those of the organization directors or owners. There are economic costs that arise from the inefficiencies or losses incurred due to misaligned objectives (*7*). These differing incentives mean that each stakeholder may desire a different level of use. To manage individuals acting in their best interest, there is often a need for oversight and incentives to encourage behavior that is in the best interest of the organization. For instance, in complex healthcare settings, the decision to use tools like EKGs can be influenced by various stakeholders' interests. Hospitals might favor the use of AI for efficiency and improved diagnostics, while insurers might be cautious about over-use of AI due to their hefty costs. The challenge lies in balancing these incentives for use of AI systems.

Level of Engagement

The *level of engagement* considers the extent to which members use AI systems in their decision-making processes. A high level of engagement (e.g, AI significantly influences an entire outcome) may conflict with regulatory oversight. Algorithmic resignation could have prevented recent misuse of large language models in the legal profession. An attorney in the US was sanctioned for using ChatGPT to draft a brief that generated fake judicial opinions and legal citations (*13*), and a federal judge in Brazil is facing investigations for a judgement that contained excerpts from ChatGPT, citing non-existent and incorrect details. Though level of engagement with AI was high in both instances, it is worth considering if legal professional rules would have been breached had the prose from ChatGPT been edited (i.e., if AI were selectively used).

Organizations operating in the EU must ensure their members are not making solely automated decisions with legal or similarly significant effects on individuals. Unless they do so within specific conditions of GDPR data processing (i.e., for business necessity, or with express consent) (*14*). Even if a human participates in the decision-making process, it is considered a solely automated decision if they are unable to "influence the causal link between the automated processing and the final decision," the "automated processing remains the only element justifying the approach" (*15*). For decisions subject to regulation, resignation would need to provide restraint and thus bar access to AI to ensure the appropriate level of engagement.

In conclusion, the implementation of algorithmic resignation necessitates a nuanced understanding of the directionality of AI system interaction, the varied incentives of stakeholders, and the level of engagement with AI systems. By carefully considering these aspects, organizations can develop robust strategies for the responsible and effective use of AI, aligning technology use with broader organizational objectives and ethical standards.

References and Notes

1. Molander v. Tesla Inc., No. RIC2002469, (Cal. Super. Ct. Oct. 31, 2023).

2. C. K. Morewedge, *et al.*, Human bias in algorithm design, *Nature Human Behaviour*, **7**, 1822 (2023).

3. R. Parasuraman, V. Riley, Humans and automation: Use, misuse, disuse, abuse, *Human factors*, **39**, 230 (1997).

4. U. Bhatt, Trustworthy Machine Learning: From Algorithmic Transparency to Decision Support, Ph.D. thesis, University of Cambridge (2023).

5. C. R. Sunstein, *Decisions about Decisions: Practical Reason in Ordinary Life* (Cambridge University Press, 2023).

6. A. D. Giglio, M. U. P. d. Costa, The use of artificial intelligence to improve the scientific writing of non-native english speakers, *Revista da Associação Médica Brasileira*, **69**, e20230560 (2023).

7. M. Jensen, W. Meckling, Theory of the firm: Managerial behavior, agency costs and ownership structure, *Journal of Financial Economics*, **3**, 305 (1976).

8. O. O'Neill, Linking trust to trustworthiness, *International Journal of Philosophical Studies*, **26**, 293 (2018).

9. M. Holweg, R. Younger, Y. Wen, "the reputational risks of ai", *California Management Review Insights* (2022).

10. European Parliament, "Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying

down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts" (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD), European Parliament, 2023).

11. J. R. Biden, "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence" (Executive Order 14085, The White House, 2023).

12. R. H. Thaler, C. R. Sunstein, *Nudge: Improving decisions about health, wealth, and happiness* (Penguin, 2009).

13. Opinion and Order on Sanctions, Mata v Avainca, Inc No. 1:22-cv-01461 (SDNY 22 June 2023).

14. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJ L 119/1.

15. Opinions of AG Pikamäe, SCHUFA Holding and Others II (C-634/21) [2023] ECR.