The Role of Digital Literacy in Maintaining Autonomy in AI Decision-Support Balancing the Burdens

Buijsman, Stefan; Carter, Sarah E.; Bermúdez, Juan Pablo

DOI
10.1007/s13347-025-00963-9

Publication date
2025

Document Version
Final published version

Published in
Philosophy and Technology

Citation (APA)
Buijsman, S., Carter, S. E., & Bermúdez, J. P. (2025). The Role of Digital Literacy in Maintaining Autonomy in AI Decision-Support: Balancing the Burdens. *Philosophy and Technology*, *38*(4), Article 136. https://doi.org/10.1007/s13347-025-00963-9

Philosophy & Technology (2025) 38:136
https://doi.org/10.1007/s13347-025-00963-9

COMMENTARY

The Role of Digital Literacy in Maintaining Autonomy in AI Decision-Support: Balancing the Burdens

Stefan Buijsman[1] · **Sarah E. Carter**[1] · **Juan-Pablo Bermúdez**[2]

Received: 5 August 2025 / Accepted: 18 August 2025

Abstract
Integrating AI systems into workflows risks undermining the competence of the people supported by them, specifically due to a loss of meta-cognitive competence. We discuss a recent suggestion to mitigate this through better uncertainty quantification. While this is certainly a step in the right direction, there is a question whether users are sufficiently supported to engage in critical reflection with literacy and tools alone. We therefore suggest that socio-technical system design focused on the role of AI systems is crucial to preserving autonomy, even when supported by uncertainty quantification.

Keywords Autonomy · Uncertainty quantification · AI ethics · Socio-technical systems

1 Introduction

One of the risks of AI integration into workflows is a loss of *meta-cognitive* competence: for the tasks on which AI supports us, we may lose the ability to detect errors, adjust for uncertainty and generally monitor and control our cognitive behavior. The reasons are two-fold: on the one hand, AI systems have low failure transparency which means that it is difficult to spot when a system is unreliable (and needs reflective engagement) and when it is functioning properly (and thus should be given evidential weight). On the other hand, we lose our own task-specific competence over time when AI does the work for us, just as we lose other skills when we outsource them – again reducing the ability to intervene when the system fails to perform as it should. We linked these

✉ Stefan Buijsman
s.n.r.buijsman@tudelft.nl

[1] TU Delft, Jaffalaan 5, 2628 BX Delft, The Netherlands

[2] University of Southampton, University Road, Southampton SO17 1BJ, UK

Published online: 06 October 2025

losses of competence to task-specific autonomy (Buijsman et al., 2025) as losing meta-cognitive competence complicates our ability to accomplish our own goals. The question therefore is: what can we do in the design of human-AI interactions to preserve meta-cognitive competence and autonomy?

1.1 Digital Literacy on Uncertainty

We outlined a number of suggestions in Buijsman et al (2025): redesigning socio-technical systems, improving failure transparency, designing training and adding positive friction. Sass (2025), in a reply to our paper, draws further attention to the options of adding explainable AI and uncertainty quantification (UQ) specifically to help users assess the (different types of) uncertainty of an AI output. Coupled with professional training programs, the idea is that these methods can support meta-cognitive competence by making users more aware when an output requires critical reflection, and when it can be trusted. Sass (2025) thus proposes a way forward to support meta-cognitive competence, by focusing on digital literacy specifically around uncertainty.

The focus on broader AI literacy is one we support, and which fits nicely with our own suggestion to improve failure transparency. UQ is one way to spot model failures, and thus to make it easier to spot when an AI system fails to deliver reliable outputs. We would still suggest to see both explainability and UQ as a part of the puzzle, as ensuring good access to alternative reasons can likewise help to determine when there are model-independent reasons to doubt the suggested decision/output (Buijsman & Veluwenkamp, 2023; Veluwenkamp & Buijsman, 2025). That being said, UQ and explainable AI can certainly be a part of the puzzle to improve/maintain meta-cognitive competence, when validated with the domain-specific empirical evaluations also suggested by Sass (2025).

1.2 Acting on Uncertainty in Socio-technical Systems

More importantly, though, we believe that the framing by Sass (2025) suggests fewer options than we actually have (namely either minimize AI use, or focus on literacy to better work with existing systems), and puts the onus too much on the users of AI systems, as opposed to on system designers. The conclusion then is that UQ methods and literacy are needed. Yet even under the assumption that the methods are calibrated (an ongoing challenge) and that uncertainty is typically low enough to prevent alert fatigue, there is a question whether users are sufficiently supported to engage in critical reflection with literacy and tools alone. To illustrate this worry, consider the setup of self-driving cars' AI systems. The system is supposed to be monitored constantly by the driver, who will intervene when needed. However, because mistakes are few and far between, drivers tend to stop paying attention and reaction times tend to increase (Payre et al., 2016). While a range of warning systems has been developed to give warnings when drivers' attention slips, and to hand over control, it is still difficult for drivers to intervene when needed, precisely because it is hard to keep paying attention when the AI system is mostly reliable.

It is an open question whether this will happen with warnings of AI uncertainty. It may well be that with the proper training and room for reflection users manage to engage in critical reflection when triggered. But we nevertheless worry that, as e.g. when Kosmyna et al (2025) observed a drastic reduction in cognitive effort in students writing with LLM support versus those unaided, many of the AI implementations will make users insensitive to such warnings. If attention fades because the task is partially automated and users have little to do other than wait for a case with high uncertainty, or because the system's design rewards quick decisions and punishes those who take the time to critically assess AI outputs, then UQ may offer few benefits. In short, UQ may be a useful element, but the larger issue is finding ways of designing the sociotechnical system as a whole in ways that sustain human metacognitive engagement.

These concerns do not lead to a minimization of AI use or a rejection of UQ. Instead, we hope that they can help us design better roles for AI systems, aided by UQ. For instance, if an AI system has mostly low uncertainty, that may be a reason to let it make automated decisions in those cases, only letting people look at the cases with moderate to high uncertainty (thus using the uncertainty as a filter). Sass's (2025) suggestion of ensemble modeling may be used to position AI systems as exploring possibilities and ramifications, as opposed to getting single suggested decisions. Options such as these refer to the issue back to the designers of the socio-technical system, who need to ensure that users have enough meaningful and engaging work for them to remain alert and in a position to critically reflect on outputs. Literacy alone cannot ensure that.

Funding Not applicable.

Data Availability Not applicable.

Declarations

Ethics Approval Not applicable.

Consent to Publication The authors each consent to publication of the submitted manuscript.

Competing Interests The authors declare that there are no conflicts of interest.

References

Buijsman, S., Carter, S. E., & Bermúdez, J. P. (2025). Autonomy by Design: Preserving Human Autonomy in AI Decision-Support: S. Buijsman et al. *Philosophy & Technology*, *38*(3), 97.

Buijsman, S., & Veluwenkamp, H. (2023). Spotting when algorithms are wrong. *Minds and Machines, 33*(4), 541–562.

Kosmyna, N. et al. (2025) Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task. Preprint retrieved from https://arxiv.org/abs/2506.08872

Payre, W., Cestac, J., & Delhomme, P. (2016). Fully automated driving: Impact of trust and practice on manual control recovery. *Human Factors, 58*(2), 229–241.

Sass, R. (2025). Meta-cognitive competence and AI-assisted decision-making: Revisiting the role of explainable AI and uncertainty quantification. *Philosophy and Technology, 38*(3), 1–5. (Forthcoming).

Veluwenkamp, H., & Buijsman, S. (2025). Design for operator contestability: Control over autonomous systems by introducing defeaters. *AI and Ethics, 5*, 3699–3711. https://doi.org/10.1007/s43681-025-00657-0