The Mitigating 'Hidden' AI Risks Toolkit

UK Government Communications (2025)

Context

Most AI safety attention goes to dramatic failures such as deepfakes or biased algorithms, yet the greater exposure may sit elsewhere. Written by the Behavioural Science Team for the UK Government, this guide draws on the rollout of the GenAI tool Assist across 200+ government organisations. It argues that the costliest risks are low salience and hard to anticipate, seeming mundane while their impact stays unknown until accumulated. Therefore, guardrails & disclaimers cannot catch them, which is why behavioural strategies are needed.

Key Insights

Six types of hidden AI risks through the FORGOOD lens

(Columns 3 and 4 are our interpretive overlay, distinct from the source)

Category	Explanation	Dimension	FORGOOD-applied prompt
Quality assurance	Outputs pass unchecked because reviewers lack subject knowledge to catch errors, or time pressure lets plausible wrong answers through. Worsens when teams are cut on the assumption the tool guarantees quality.	**Goals**	How do you know the tool improves outcomes for users, rather than just speeding up output that no one can meaningfully verify?
Task-tool mismatch	Experimentation pushes a tool toward jobs it was never built for. A single poor public output damages trust in the uses where AI genuinely works.	**Options**	Where would a non-technical route serve users better, and have you compared it before defaulting to the tool?
Perceptions, emotions, signalling	The rollout sends a message. Read as a job threat, or as leadership valuing speed over quality, staff disengage, so measured adoption overstates real value.	**Respect**	Does the rollout respect staff autonomy and dignity, or signal that they are being replaced rather than supported?
Workflow and organisational challenges	Embedding AI adds hidden work (training, QA, prompting) that can lengthen tasks. Removing easy tasks leaves only demanding ones, raising load and weakening the variety that sustains motivation.	**Fairness**	Are the costs and burdens of embedding the tool shared fairly, or pushed onto the teams least able to absorb them?
Ethics	Automation does not create bias, it scales it, widening the speed, frequency and reach of already-biased decisions. Discriminatory patterns may surface only after prolonged use.	**Opinions**	Would the tool's decisions survive the front-page test once their cumulative effect on different groups becomes visible?
Human connection and tech overreliance	The technical fix can quietly displace a non-technical option users accepted more readily, and remove expertise or human contact only people can provide.	**Delegation**	Does your organisation hold the genuine ability, not just the right, to keep meaningful human judgement in the loop?

Marquez, H., & Nicolson, M. (2025). The Mitigating 'Hidden' AI Risks Toolkit. Government Communications, Cabinet Office. https://www.gov.uk/government/publications/mitigating-hidden-ai-risks-toolkit