The Hazards of Putting Ethics on Autopilot

Research shows that employees who are steered by digital nudges may lose some ethical competency. That has implications for how we use the new generation of AI assistants.

By Julian Friedland, David B. Balkin, Kristian Ove R. Myrseth

[Accepted version, *MIT Sloan Management Review*, June 2024]

The generative AI boom is unleashing its minions. Enterprise software vendors have rolled out legions of automated assistants that use large-language model (LLM) technology such as ChatGPT to offer users helpful suggestions or execute simple tasks. These so-called copilots and chatbots can increase productivity and automate tedious manual work. But if they are not thoughtfully implemented, they risk diminishing employees' decision-making competency, especially when ethics are at stake.

Our examination of the consequences of "nudging" techniques, used by companies to influence employees or customers to take certain actions, has implications for organizations adopting the new generation of chatbots and automated assistants. Companies implementing generative AI agents are encouraged to tailor them to increase managerial control. Microsoft, which has made copilots available across its suite of productivity software, offers a tool that allows enterprises to customize copilots, allowing them to more precisely steer employee behavior. Such tools will make it much easier for companies to essentially put nudging on steroids – and based on our research into the effects of nudging, that may over time diminish individuals' own willingness and capacity to reflect on the ethical dimension of their decisions.

AI-based nudges may be particularly persuasive, considering the emerging inclination among individuals to discount their own judgments in favor of what the technology suggests. At its most pronounced, this can become a kind of techno-chauvinistic hubris, which discounts human cognition in favor of AI's more powerful computational capacities. That's why it will be particularly important that employees be encouraged to maintain a constructively critical perspective on AI output and that managers pay attention to opportunities for what we call *ethical boosting* – behavioral interventions that utilize mindful reflection, as opposed to mindless reaction. As we'll discuss, this is a way to help individuals grow in ethical competence, rather than allow those cognitive skills to calcify.

Digital nudges, especially in the form of salient incentives and targets, can lead to subtle motivational displacement by obfuscating the ultimate aims of the team or organization and shifting proximal goals. When a performance measure becomes the main objective, it ceases to function as an effective measure – a phenomenon known as Goodhart's law. For example, copilots might be designed to nudge customer-facing workers to maintain five-star ratings by offering bonus points or financial rewards. But if workers focus entirely on increasing their ratings, rather than delivering great customer service in the hopes that they will receive a high rating, they may be tempted to game the system by misleading customers. That is, the ratings themselves may become goals in their own right, potentially neglecting important qualities that are difficult to measure, such as honesty and trustworthy behavior.

The implications of nudging are particularly pernicious in ethically nuanced contexts which require self-awareness of the values we care most deeply about. By uncritically accepting AI copilot guidance, managers may neglect to consider the 'why' underlying their decisions. In this

article, we'll explain how that leads to the risk that their ethical competence may degrade over time – and what to do about it.

From Reactive Nudging to Reflective Boosting

Nudges tend to exploit what Daniel Kahneman dubbed "thinking fast," a reactive mode that contrasts with "thinking slow," that is, reflective thinking, as described in our recent paper "Beyond the brave new nudge: Activating ethical reflection over behavioral reaction" (Academy of Management Perspectives). Such interventions can leverage mild financial incentives or emotional triggers, including joy, fear, empathy, social pressure, or reputational rewards, to induce individuals to act as they arguably should upon ethical reflection. Heavy reliance on these incentives can reactively shift attention toward the extrinsic reward, thereby supplanting and weakening the ethical motives they are intended to encourage. This is because moral maturity and autonomy are ultimately achieved through instilling good habits aimed at *intrinsic* – as opposed to *extrinsic* – rewards. While nudging interventions can be effective when used carefully and sparingly – for leading agents to increased self-awareness and autonomy – the power and pervasiveness of Gen-AI technology is ripe for overuse, which could instigate a nudge riot of motivational displacement and dependency, crowding out good habits of ethical reflection. It could also backfire in other ways by causing some employees to recoil from what they perceive to be excessive paternalism or surveillance. Managers should take care to avoid setting up a virtual Brave New World in which ethical behavior is perpetually conditioned, via automatic cognitive responses, to do what is lauded by the AI and its designers.

Though reliance on behavioral nudges cannot be entirely avoided, especially in processes where risk management or regulatory compliance are highly salient, the good news is that checking mechanisms can be introduced to keep humans mindfully engaged, and to trigger

ethical reflection before action. This can guard against the tendency of cognitive skills to atrophy from disuse. Given the many current limitations of LLMs, including tendencies to produce biased and inaccurate information, lack of comprehension and logical coherence, managers should prioritize such engagement triggers to keep people thinking critically about AI copilot output even in the absence of ethical choices or nudges.

How can individuals develop their abilities to think reflectively about ethical choices and resist the easy default options that nudges present, not only in the workplace but in their many interactions as consumers and citizens? We see promise in ethical boosting, which is rooted in a positive view of human potential to learn and grow. Where nudging promotes reactivity and seeks to steer subjects to choose specific behaviors without much thought of their own, boosting is a long-term developmental exercise to encourage habits of mindfulness and reflection. Boosts could take the form of mental rules of thumb, or *heuristics* – such as the Golden rule, the best for all concerned, and one's virtuous self-image – that help identify and think through ethical dilemmas.

Boosting principles could also target negative contingencies by correcting unhealthy workplace patterns via reminders at key inflection points. Here, even AI copilots can play a role, if they nudge us to think instead of just clicking the box that's easiest to click. We found that Microsoft's copilot was already fairly adept at warning of subtle, potentially offensive language in emails. But we can also choose to exercise our brains by rewriting accordingly in our own words, rather than accepting the bland system recommendation. To boost such a mindset, messaging apps might invite users to take time before responding to a potentially rude or hostile message chain, thereby allowing tempers to cool and the more reflective mind to engage. An image of a person rage-typing might serve as an effective speed bump helping users to build

virtuous self-awareness. Likewise, training such as the ARPA's Sirius program aims to instill cognitive skills such as gaining competency at recognizing one's own biases and assumptions.

In conclusion, managers should be heedful of the rhetorical Siren song underlying the generative AI branding as personal "copilots," in contrast with "decision support" or "assistants." While the latter terms acknowledge the technology is subservient to the user, copilot connotes a more capable, autonomous, and even responsible role for the technology. A copilot is fully qualified to fly the plane in a pilot's absence; the cachet of competence implied by the term subtly invites employees to trust in and abide by AI-driven nudges. If AI copilots enable greater managerial control and efficiency at the cost of declining ethical competence in the workforce, managers may want to consider installing some reflective speed bumps.