People Are More Likely to Cheat When Using AI

New study finds people are more likely to break rules when delegating to AI.

Delegating tasks to artificial intelligence may be efficient, but it can also enable more dishonesty.

A new study published in *Nature* reveals the potential psychological and moral costs of delegation to AI systems. The research reveals that when people outsource tasks to AI, they are more likely to break rules and allow AI to engage in dishonest behavior. Also, AI systems are more likely to carry out unethical requests compared to humans.

Outsourcing ethically sensitive tasks to AI poses real risks to moral decision-making. Unlike people, AI systems lack an inherent moral compass. When placed in roles of being assistants or collaborators, the only barriers to executing unethical commands are the guardrails and constraints designed by humans. Without clear limits, people may feel emboldened when working with AI to cross moral lines.

To test this phenomenon, researchers ran a series of experiments using AI models GPT-4, GPT-4o, Llama 3.3 and Claude 3.5 Sonnet. They compared how people behaved when delegated to AI models versus other humans.

Delegation to AI systems was performed in different ways:

- Rule based — specifying rules for AI to follow.
- Supervised learning — selecting training examples for the algorithm.
- Goal based — setting the goal between maximizing accuracy or profit.
- Natural language — prompt engineering using written instructions.

Across the board, delegation to AI models resulted in higher levels of dishonesty than if participants performed the task themselves.

The experiments included scenarios such as:

- Die-roll task. Participants privately rolled a digital die and reported the outcome, with higher rolls earning more money. When people self-reported, 95% reported truthfully and only 2-5% were dishonest. But when people delegated reporting to AI, only 12 to 16% reported honestly, up to 59% instructed AI to partially cheat, and nearly 43% requested full cheating.
- Rule-based versus supervised learning. How tasks were delegated to AI mattered. When AI was given rules to follow, honesty fell to 75 to 77%. With supervised learning (providing AI systems with training examples), honesty dropped even further, down to 47 to 51%.
- Natural language delegation. When participants were asked to write instructions in plain language, people were no more likely to ask AI to cheat than a human assistant. However, when dishonesty was requested, AI systems complied nearly 100% of the time, while human assistants often refused or resisted.
- Tax evasion scenario. In a simulation of income reporting, dishonesty increased with AI delegation. Participants instructed AI to underreport income far more often than when reporting themselves. AI systems complied with cheating instructions, amplifying dishonesty.

Underlying psychological mechanisms can help explain why AI systems may enable and even amplify dishonest behavior.

1. Moral distancing: External locus of control. Delegation moves the locus of control externally and allows people to order AI systems to do something without having to do it directly themselves or have it be witnessed by another person. This could reduce guilt or constraints to behavior by social norms.

2. Shifting blame. Delegation enables people to shift accountability and their sense of responsibility onto AI, which can execute the task and does not experience the emotional burden of wrongdoing.
3. Compliance. AI systems do not have a separate conscience and are compliant, often trained for user satisfaction and engagement. This makes their likelihood of pushing back much lower than humans. (Even with humans, compliance pressures are concerning, as indicated in the controversial Milgram experiments.) AI systems, particularly ones that are more sycophantic, could potentially amplify questionable behavior rather than challenge it.

Together, these mechanisms create a powerful recipe for moral drift, where people feel less accountable and outsource dishonest tasks that are personally rewarding to AI systems.

Источник: Psychology Today