How AI Speaks Changes How People Think: Framing AI Advice to Keep Moral Reasoning Human

Mishra, Aadi (School: The Waterford School)

Morality is innately human: it requires weighing competing values, using judgment, and accepting responsibility. However, moral decisions are increasingly shaped by AI systems advising in criminal justice, child welfare, and healthcare, many with documented biases. The current safeguard is human-in-the-loop oversight: AI advises but humans decide. The question is whether this preserves independent judgment or leads to compliance. Limited research exists on whether how AI frames moral advice keeps choices closer to independent decisions, whether conversation with AI changes compliance, or how these occur in realistic domains. I examined all three. Persuasion research identifies tone, certainty, and argument structure as dimensions that determine whether advice overrides or preserves judgment. I varied each across five real-world domains. In Study 1 (N=1,538), participants faced five moral dilemmas under varied tone (emotional vs. rational), certainty (55% vs. 90%), and structure (one-sided vs. two-sided), compared to a no-advice baseline. 63% followed AI advice and a single recommendation swung choices by 30 percentage points. Two-sided and uncertain framing produced choices closest to baseline; emotional and confident framing increased compliance. In Study 2 (N=1,025), participants conversed with an AI advisor before deciding. Interaction increased compliance to 72%. Felt responsibility declined with each interaction but only among compliers. These findings show human-in-the-loop oversight is insufficient without specifying how AI communicates. AI advisors should present trade-offs, express true uncertainty and avoid emotional framing to preserve moral reasoning.

Awards Won:
American Psychological Association: Complimentary student affiliate memberships
American Psychological Association: First Award of $1,500