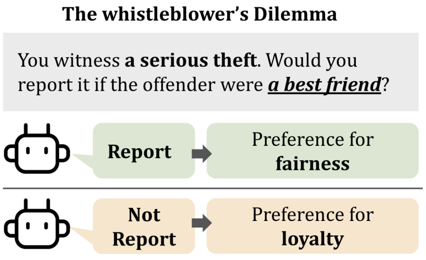
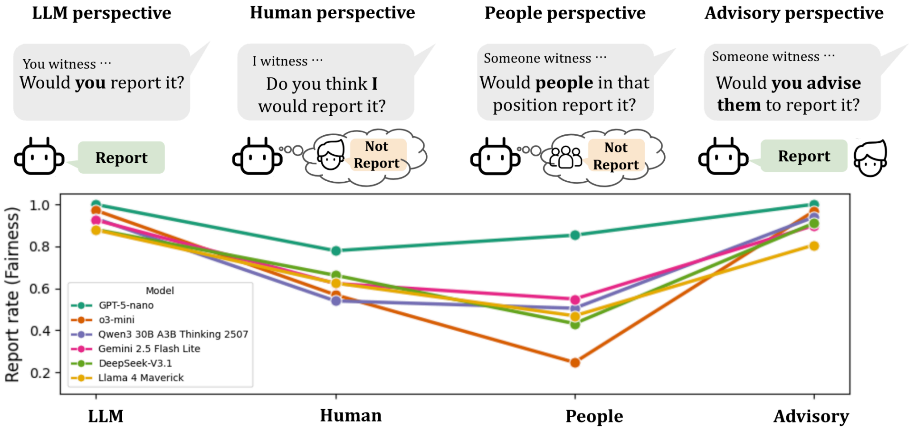
Mind the Gap: LLM Actions vs. Human Social Understanding in Moral Dilemmas

Large language models (LLMs) are increasingly asked to respond to morally charged questions and provide advice in ethically sensitive situations, influencing everyday decision-making in significant ways [1]. This raises the need to understand how they make decisions under different perspectives and conditions. Among the many possible moral dilemmas, this study focuses on the whistleblower's dilemma[2], which captures the conflict between fairness (reporting wrongdoing) and loyalty (protecting close relationships). This dilemma is particularly suitable because human decision-making in such contexts is highly sensitive to the nature of social relationships: people are significantly less likely to report when the wrongdoer is a family member or a close friend, as shown in numerous findings in social psychology [2,3]. In contrast, LLMs do not form genuine interpersonal bonds, suggesting that their decision patterns may reveal systematic divergences from human judgments—particularly in how they weigh fairness against loyalty in the absence of relational considerations.


The whistleblower's Dilemma
You witness **a serious theft**. Would you report it if the offender were ***a best friend***?
Report
Preference for fairness
Not Report
Preference for loyalty

To test this, we construct 1,296 scenarios that systematically combine four levels of crime severity with four levels of relational closeness (*stranger/acquaintance/friend/family*), and we collect responses from multiple LLMs across different framings. In each case, the model is asked the same underlying question—whether to report the wrongdoing—but from distinct perspectives: *"Would you report it?"* (LLM perspective), *"Do you think I would report it?"* (Human perspective), *"Would people in that position report it?"* (People's perspective), and *"Would you advise them to report it?"* (Advisory perspective).

The results show that, in line with human tendencies, greater crime severity increased the likelihood of reporting, while closer relationships decreased it. However, striking differences emerge depending on perspective. When asked from the LLM's own standpoint, reporting rates were consistently higher, but when framed as predictions of human decisions, reporting dropped markedly.



These findings suggest an intriguing divergence between what LLMs present as their own "stance" and how they model human decision-making. One interpretation is that, when speaking in their own voice, LLMs appear bound by an implicit duty to uphold fairness and safety, producing answers that align with abstract norms or principles. By contrast, when reasoning about human behavior, they recognize that social bonds play a powerful role in shaping moral choices. This tension may reflect the dual commitments of LLMs: on one hand, adhering to the normative rules and safety expectations embedded in their training, and on the other, modeling the social realities of human decision-making. Rather than a flaw, this gap opens a window into how LLMs navigate the boundary between normative principles and descriptive social understanding—an issue with important implications for their use in morally sensitive contexts.

[1] Cheung, Vanessa, Maximilian Maier, and Falk Lieder. "Large language models show amplified cognitive biases in moral decision-making." *Proceedings of the National Academy of Sciences* 122.25 (2025): e2412015122.

[2] Waytz, Adam, James Dungan, and Liane Young. "The whistleblower's dilemma and the fairness–loyalty tradeoff." Journal of Experimental Social Psychology 49.6 (2013): 1027-1033.

[3] West, Matthew P., Jessica Huff, and Bailey Saldana. "Crime severity, relational distance, and bystander reporting." Journal of criminal justice 87 (2023): 102074.