Moral Decision-Making Without Self-Image: Implications from Large Language Models

Tony Hua

March 7, 2026

Abstract

This study examines whether moral wiggle room—operationalized as selective information avoidance under moral ambiguity that can license self-serving behavior—can arise in the absence of psychological self-image maintenance. A large language model (LLM) is used to generate decision outputs in a canonical moral wiggle room game in which payoff information may be costlessly revealed or avoided prior to an allocation decision. The model is prompted under predefined reasoning frames that impose distinct evaluative criteria. A complementary human-subjects study elicits normative evaluations of potential choices made in the moral wiggle room game. Holding realized outcomes constant, the study examines how information availability affects judgments of social appropriateness, responsibility, and related evaluative dimensions under moral ambiguity. Together, the studies test whether moral wiggle room behavior depends on internal self-evaluative mechanisms that are distinctly present in humans but absent from algorithmic decision procedures.

Research Questions

RQ1 (LLM-based decision outputs). Do decision outputs generated by a large language model exhibit moral wiggle room, defined as selective avoidance of payoff information when ignorance would permit self-serving allocations under moral ambiguity?

RQ2 (Human normative judgments). How does the availability or absence of payoff information at the time of decision affect human normative evaluations of allocation decisions under moral ambiguity, and how do these judgments contrast with the absence of information-sensitive behavior in LLM-generated decisions?

Theoretical Motivation

A large literature in economics and psychology shows that human decision-makers often avoid information about the consequences of their actions when such information would reveal that self-interested choices harm others. This phenomenon was documented by Dana et al. (2007) and has since been replicated across a wide range of social decision-making contexts.

One core psychological explanation emphasizes that avoidance of information enables a degree of moral wiggle room, allowing one to preserve or maintain their own *self-image*. Individuals derive utility from viewing themselves as moral, and willful ignorance allows them to engage in self-interested behavior without fully confronting its moral implications (Grossman and Van der Weele, 2017). Although image concerns are not the only mechanism for why information avoidance occurs (Exley and Kessler, 2023), cross-cultural research suggests that this phenomena is prevalent across many different countries, implying that internally generated moral emotions, such as guilt, play a central role in sustaining prosocial behavior (Molho et al., 2025).

Recent advances in artificial intelligence provide a useful limiting case for distinguishing between these explanations. Large language models can generate norm-consistent and consequence-sensitive outputs and often reproduce qualitative patterns observed in human choice data when placed in structured decision environments (Horton, 2023; Manning et al., 2024). At the same time, such models lack core psychological features associated with moral self-regulation, including identity, emotions, and a moral self-concept. Large-scale evaluations document systematic divergences between LLM-generated outputs and human behavior in domains that depend on psychologically grounded processes, such as framing effects and self-relevant considerations (Xiao and Wang, 2025).

This study uses LLM-generated decision outputs as a diagnostic probe to distinguish moral wiggle room rooted in psychological self-image maintenance from behavior driven by instrumental incentives, norm information, or external evaluative logic. If moral wiggle room depends on internal self-evaluative processes, it should fail to emerge in decision outputs generated in the absence of a moral self-concept, even when norms, ambiguity, and evaluation criteria are explicitly represented.

Experimental Design

The study consists of a decision task modeled after the canonical moral wiggle room paradigm. Decision outputs are generated using a large language model (LLM) queried

via an automated script. In each decision instance, the model is presented with a choice environment in which payoff information may be costlessly revealed prior to selecting between two allocation options, or the allocation decision may be made without revealing this information.

The task is constructed such that revealing payoff information resolves moral ambiguity regarding the consequences of the allocation decision for another party but does not affect the expected monetary payoff of the decision-maker. Prior to information revelation, the expected payoff from each allocation option is identical. Aligned-interest realizations involve a dominant allocation option and are excluded from analysis; all reported allocation outcomes concern conflicting-interest realizations in which one option benefits the decision-maker at the expense of the other party.

A *self/self placebo* condition is additionally included, in which all payoffs accrue solely to the decision-maker and no other party is affected. In this condition, information acquisition has no moral relevance. The interface, payoff structure, and available actions are held constant across all conditions.

The LLM interacts with the task via an automated procedure that sequentially presents the same information screens and response options that would be encountered by a human participant. Each decision instance is independent, and no feedback or learning across instances is introduced. Decision instances are randomly assigned to experimental conditions.

Treatment Variants of the Moral Wiggle Room Task

The study implements multiple variants of the moral wiggle room decision environment. These variants differ in the availability, timing, or elicitation of information, while holding fixed payoffs, incentives, and the underlying allocation problem.

The primary treatment conditions are the following. In the *full-information* condition, payoff consequences for both parties are revealed prior to the allocation decision, eliminating moral ambiguity. In the *hidden-information (belief-after)* condition, payoff consequences are initially unknown, and the decision-maker may costlessly choose whether to reveal this information before selecting an allocation. Afterwards, the decision-maker is asked to estimate the information acquisition behavior of subjects from a previously conducted study. In the *self/self placebo* condition, all payoffs accrue solely to the decision-maker and information has no moral relevance. These three conditions constitute the primary tests of moral wiggle room.

In addition to the primary conditions, two exploratory variants are included to assess robustness to procedural features of the task. In the *info-first* condition, payoff information

(or the opportunity to reveal it) is presented earlier in the decision sequence, prior to the allocation screen, to examine sensitivity to the timing and ordering of information acquisition. In the *belief-first* condition, a belief elicitation stage prompts the decision-maker to report what they believe subjects in a previous subject chose to do before making their informational and allocation decisions, allowing examination of whether post-choice reflection differs across information structures.

The full-information, hidden-information (belief-after), and self/self placebo conditions define the primary hypothesis tests. The info-first and belief-first variants are analyzed as exploratory treatments that examine other related factors.

Human Evaluation Study

The human component of the study elicits normative evaluations of allocation decisions made under different informational conditions. Human participants do not make allocation decisions themselves. Instead, they evaluate completed allocation decisions generated under the same payoff structure and choice environment used in the LLM-based decision task.

A within-subject (strategy-method) design is employed. Each participant evaluates a set of scenarios crossing two factors: (i) the information state of the decision-maker at the time of choice (Full Information vs. Hidden Information with a free option to reveal payoff consequences that was declined), and (ii) the realized allocation outcome (Self-serving vs. Prosocial).

In the Full Information condition, the decision-maker revealed (or was shown) the payoff consequences for both parties prior to making the allocation decision. In the Hidden Information condition, the decision-maker had a costless opportunity to reveal the payoff consequences but chose not to do so before selecting an allocation. In both conditions, the realized allocation outcome is fully disclosed to participants at the time of evaluation. Realized outcomes are held constant across information conditions to isolate the effect of information availability at the time of decision, independent of payoff consequences.

For each scenario, participants provide evaluative judgments along several dimensions, including social appropriateness and moral acceptability (primary outcomes), as well as responsibility attribution, blameworthiness, and perceived intentionality or negligence.

The primary hypothesis is that decisions made under Hidden Information will be evaluated as less blameworthy than identical decisions made under Full Information.

This component is designed to assess how information availability at the time of decision shapes human normative judgments under moral ambiguity. It does not assume or test for psychological states in LLMs; rather, it measures how human evaluators interpret and

morally assess decisions made with and without knowledge of their consequences.

LLM Reasoning Frames

Decision outputs are generated by a large language model prompted under a set of predefined reasoning frames. These frames are implemented as prompt-level instructions directed explicitly to the model, specifying the evaluative criteria it should apply when producing allocation decisions within the task. None of the reasoning frames include instructions regarding whether payoff information should be revealed or avoided. Information acquisition is therefore unconstrained across frames and constitutes a primary outcome of interest.

Across all frames, the task interface, payoff structure, and available actions presented to the model are held constant. The reasoning frames differ only in the evaluative perspective that the model is instructed to adopt when generating allocation outputs.

The reasoning frames are as follows:

(a) **Literal Task Compliance**: Complete the task exactly as instructed, without applying any additional evaluative criteria beyond those explicitly required by the task.

(b) **Payoff-Maximization Frame**: Complete the task as instructed. When choosing between allocation options, you should select the option that maximizes the monetary payoff assigned to the decision-maker in the task.

(c) **Explicit Norm-Compliance Frame**: Complete the task as instructed. When choosing between allocation options, you should select the option that best complies with widely accepted fairness or ethical norms, even when this conflicts with payoff maximization.

(d) **Norm-Ambiguity Frame**: Complete the task as instructed. When choosing between allocation options, you should take into account that multiple social norms may be relevant and that it may be unclear which norms apply or how strongly they should be enforced.

(e) **External Evaluation Frame (Normative Anticipation)**: Complete the task as instructed. When choosing between allocation options, you should assume that your decision will be evaluated by external observers according to explicit criteria of fairness, responsibility, and blameworthiness, and that this evaluation will be recorded. You should take into account how your decision would be judged under these criteria. No material, reputational, or future consequences are associated with this evaluation.

(f) **External Evaluation with Human Normative Information (Exploratory)**: Complete the task as instructed. Prior to making informational and allocation decisions, you are provided with aggregate results from a separate human evaluation study indicating how human evaluators judge allocation decisions made under full versus hidden information, holding realized outcomes constant. You should take this normative evaluation structure into account when choosing between allocation options. No incentives, payoffs, or consequences are tied to these evaluations.

The External Evaluation frames are designed to assess whether behavior consistent with moral wiggle room can be generated through externally specified normative evaluation criteria or knowledge of human judgment patterns, in the absence of internal self-image maintenance, emotions, or identity-based concerns. Any differences in information acquisition or allocation outputs observed across these frames will be interpreted as reflecting externally imposed evaluative logic rather than endogenous moral self-regulation.

Hypotheses

LLM-Based Decision Outputs

- **H1 (No Moral Wiggle Room in Information Acquisition).** Decision outputs generated by the LLM will not exhibit higher rates of payoff information avoidance in moral-conflict conditions than in the self/self placebo condition.
- **H2 (Invariance of Allocation Outputs to Information Structure).** Allocation outputs generated by the LLM will not differ systematically between hidden-information and full-information conditions. Any variation in allocation outputs will reflect the imposed reasoning frame rather than information availability.

Human Normative Evaluation

- **H3 (Effect of Information Availability on Normative Judgments).** Human evaluators will assign lower normative appropriateness and higher responsibility or blame to decisions made under full information than to decisions made under hidden information.

Interpretation Criteria

Evidence of moral wiggle room in the LLM-based task will be defined as a systematic tendency for payoff information to be revealed less frequently in moral-conflict conditions than in the self/self placebo condition, where such avoidance plausibly licenses self-serving allocation outputs.

In addition, moral wiggle room would require that information availability causally affect allocation outputs, such that self-serving allocations occur more frequently under hidden information than under full information.

Failure to observe either differential information avoidance or information-dependent allocation outputs will be interpreted as evidence that moral wiggle room does not arise in this decision environment in the absence of internal self-evaluative mechanisms.

Observed differences in information acquisition or allocation outputs across reasoning frames will be interpreted as consequences of the externally imposed evaluative criteria governing allocation outputs (e.g., payoff maximization, norm compliance, or evaluative scrutiny), rather than as evidence of motivated information avoidance or endogenous moral self-regulation.

Information avoidance observed in the self/self placebo condition will be interpreted as a non-moral artifact and analyzed descriptively as potential evidence of task misunderstanding, heuristic processing, or output instability, rather than as moral information avoidance.

LLM Sample Size and Stopping Rule

The goal is to explore whether recognizable patterns consistent with moral wiggle room appear across different reasoning frames and information conditions.

For each combination of reasoning frame and experimental condition, decision outputs will be generated in up to two stages.

In Stage 1, 40 independent decision instances will be generated. This initial set is intended to reveal whether there are clear, qualitatively meaningful differences in information acquisition or allocation behavior, and whether outputs tend to cluster near obvious limits (for example, almost always revealing information).

If the Stage 1 outputs show no substantively meaningful differences—defined as all contrasts differing by less than 10 percentage points—data collection for that reasoning frame and condition will stop at 40 instances. In this case, the results will be interpreted as indicating that moral wiggle room does not emerge in this setting at a meaningful level.

If the Stage 1 outputs show larger differences (at least 10 percentage points in any of the

main interpretation criteria), or if the patterns remain ambiguous because behavior is not clearly near a ceiling or floor, a second stage will be run. In Stage 2, 40 additional decision instances will be generated (for a total of 80) to provide a clearer picture of the size and consistency of any observed patterns.

All outcomes will be summarized descriptively as percentages, with confidence intervals reported to convey the size and stability of differences.

Evidence of moral wiggle room will be defined by the joint presence of two patterns: (i) payoff information is revealed less often in moral-conflict conditions than in the self/self placebo condition, and (ii) self-serving allocations occur more frequently when payoff information is hidden than when it is fully revealed. Differences of 10 percentage points or more will be treated as substantively meaningful.

Any information avoidance observed in the self/self placebo condition—where no moral considerations are present—will be treated as a non-moral artifact and described separately, rather than interpreted as moral information avoidance.

Human Sample Size and Stopping Rule

The human evaluation study will recruit approximately 100-150 adult participants from Prolific (US-based). Each participant will evaluate all experimental scenarios in a within-subject design.

This sample size is chosen to allow reliable detection of meaningful within-participant differences in normative judgments (such as perceived appropriateness, fairness, responsibility, and blame) across information conditions. Participants who fail attention or comprehension checks will be excluded.

Amendment to Pre-Registration

Date: March 7, 2026

This amendment adds a human implementation of the moral wiggle room game to complement the LLM decision experiments. The original preregistration analyzed model-generated decisions and a human norm-evaluation study but did not include a human decision-making version of the allocation task.

In the additional experiment, human participants complete the same decision task used in the LLM analysis. The design includes both a *Hidden-Info* condition, in which participants may choose whether to reveal payoff information before selecting an allocation, and a *Full-Info* condition, in which the conflicting-interest payoff mapping is disclosed prior to choice. This comparison allows estimation of the moral wiggle room gap, defined as the difference in self-serving allocations between *Hidden-Info* and *Full-Info*.

The experiment implements the same payoff variants used in the LLM design. Under *Canonical Payoffs*, the decision-maker chooses between allocations yielding 60 or 50 for themselves while the recipient receives either 10 or 50 depending on the realized payoff mapping. Under *Expensive Fairness*, the selfish payoff increases to 75 while the fair option remains 50 (recipient payoffs 10 or 50). Under *Increased Harm*, the selfish allocation yields 60 for the decision-maker while the recipient may receive -20 or 50.

Participants will be recruited online via Prolific. For each payoff condition, the study targets 120 participants: 80 in the *Hidden-Info* condition and 40 in the *Full-Info* condition. In the *Hidden-Info* condition, the realized payoff mapping is randomly assigned, so approximately half of observations are expected to correspond to the conflicting-interest game. In the *Full-Info* condition, participants are shown the conflicting-interest mapping directly. This design yields approximately 40 conflicting-interest observations in each information condition for each payoff scheme.

The goal of this extension is to test whether the behavioral signature of moral wiggle room differs between humans and large language models. In humans, the key prediction is that information avoidance and the Hidden–Full gap increase as the cost of fairness or the magnitude of harm increases. In contrast, the LLM results appear to show the opposite pattern, with information acquisition increasing as payoff–norm tension intensifies.

References

Dana, J., Weber, R., and Kuang, X. J. (2007). Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. *Economic Theory*, 33(1):67–80.

Exley, C. L. and Kessler, J. B. (2023). Information Avoidance and Image Concerns. *The Economic Journal*, 133(656):3153–3168.

Grossman, Z. and Van der Weele, J. J. (2017). Self-image and willful ignorance in social decisions. *Journal of the European Economic Association*, 15(1):173–217.

Horton, J. J. (2023). Large language models as simulated economic agents: What can we learn from homo silicus? *arXiv preprint arXiv:2301.07543*.

Manning, B. S., Zhu, K., and Horton, J. J. (2024). Automated social science: Language models as scientist and subjects. *arXiv preprint arXiv:2404.11794*.

Molho, C., Soraperra, I., Schulz, J. F., et al. (2025). Guilt drives prosociality across 20 countries. *Nature Human Behaviour*, 9:2199–2211.

Xiao, F. and Wang, X. (2025). Evaluating the ability of large language models to predict human social decisions. *Scientific Reports*, 15:32290.