Latent Neural Coupling of Risk and Time Preferences in LLMs Mirrors Human Biases

YAN LENG*, University of Texas at Austin, USA
TRUNG NGUYEN, University of Texas at Austin, USA

Large language models (LLMs) now stand in for human decision makers in many settings, but it is unclear whether their choices arise from structured internal representations or surface-level pattern matching—and whether the economic preferences they express are correlated as in humans. We address these questions by analyzing three canonical preference domains – risk, time, and social – across several current LLMs.

Our study proceeds in three stages. (1) Behavioral elicitation: Using standard prospect theory, intertemporal-choice, and dictator-game vignettes, we confirm that the models reproduce familiar biases – risk aversion and temporal discounting. (2) Latent probing: Contrastive prompts that elicit risk-averse versus risk-seeking responses allow us to train a sparse linear probe and identify a one-dimensional "risk axis" in activation space that reliably orders the models' lottery choices. (3) Causal steering: Small perturbations along this axis predictably shift behavior: nudging the model toward risk seeking lowers its implied discount rate, whereas nudging toward risk aversion raises it—reversing the positive risk-aversion/patience correlation observed in humans. The same manipulation produces only a weak, model-specific increase in self-interested dictator allocations, indicating that social motives are encoded in largely separate subspaces.

We contribute three findings. First, LLMs encode risk attitudes along a structured linear continuum from aversion to seeking. Second, this dimension simultaneously governs time preferences, showing that the models internalize a probabilistic relationship between risk and patience rather than hard-coding each bias independently. Third, social preferences remain largely decoupled, highlighting domain-specific representation. Taken together, these results show that LLMs, even without real-world experience, acquire interpretable latent coordinates for abstract economic traits. The probe-and-steer method we introduce offers a practical tool for mapping and manipulating such traits, positioning LLMs as scalable testbeds for behavioral theory and as synthetic participants in social-science research.

For the full paper, please visit https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5284661.

CCS Concepts: • **Computing methodologies** → **Natural language generation**; **Learning latent representations**; **Neural networks**; **Cognitive science**; • **Information systems** → **Language models**; • **Applied computing** → **Sociology**; **Economics**.

Additional Key Words and Phrases: Generative AI; LLMs; Mechanistic Interpretability; Linear Representation; Risk preference; Time preference; Social Preference

ACM Reference Format:
Yan Leng and Trung Nguyen. 2025. Latent Neural Coupling of Risk and Time Preferences in LLMs Mirrors Human Biases. In *The 26th ACM Conference on Economics and Computation (EC '25), July 7–10, 2025, Stanford, CA, USA*. ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3736252.3742588

*Both authors contributed equally.

Authors' Contact Information: Yan Leng, yan.leng@mccombs.utexas.edu, University of Texas at Austin, Austin, TX, USA; Trung Nguyen, trungnguyen@utexas.edu, University of Texas at Austin, Austin, TX, USA.

EC '25, July 7–10, 2025, Stanford, CA, USA

ACM ISBN 979-8-4007-1943-1/2025/07
https://doi.org/10.1145/3736252.3742588