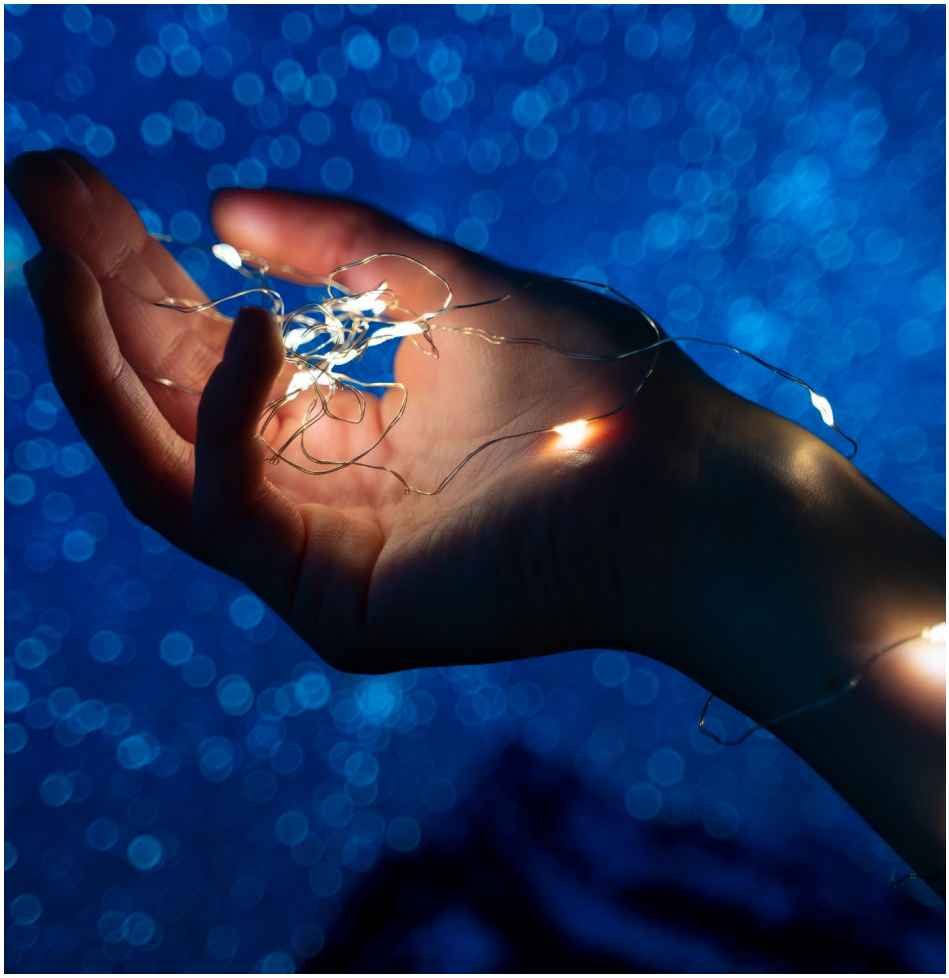
SPAZIO APERTO
OPEN SPACE

AI and Ethical Decisions: Measuring Tendencies for Prosocial Behavior in Large Language Models

Tony Persico, Michael Baggot, Silvia Di Piero



Abstract

The paper investigates AI systems' potential as an ethical agent by systematically testing models on prosocial behaviors such as fairness and altruism. Building on recent advances in AI ethics and machine morality, we will assess large language models through controlled experiments. These tests incorporate experimental games inspired by behavioral economics and a focus on the perception of fairness and justice. The project adopted standardized evaluation metrics to quantify ethical reasoning and benchmark AI results with experimental human behaviors. Future scalability will explore integrating AI into governance and advisory roles, ensuring robust ethical safeguards.

Key words: AI Ethics, Ethical Agency, Prosocial Behavior, Value Alignment, Behavioral Economics, Ethical Decision-Making, Governance and AI.

Introduction

This paper explores the viability of developing artificial intelligence (AI) systems capable of ethical agency. As AI systems increasingly influence social and economic decision-making, the question of their ethical capacities becomes critical. The study is the result of a dedicated session of the conference Communitas 2025, hosted by the Pontifical University of Saint Thomas Aquinas (Angelicum) in Rome. The session investigated whether AI systems can function as ethical agents, entities capable of making decisions that reflect moral principles and support social good. By focusing on prosocial behaviors such as fairness, cooperation, and altruism, we seek to understand the potential for AI to act not only safely but also ethically. Drawing on foundational work in machine ethics, behavioral economics, and value alignment, we design and implement a series of controlled experiments to evaluate ethical reasoning in different widely adopted large language models.

We propose a set of standardized metrics to quantify ethical preferences and alignment with human behaviors. The findings contribute to the ethics-in-AI discourse and offer practical tools for embedding ethical behavior in AI systems deployed in public and policy contexts. Thus, our research aligns with the "ethics in AI" branch of AI ethics, which emphasizes the integration of ethical principles directly into AI design and behavior (Ratti, 2025). This approach is distinct from ethics of AI (focusing on external consequences) and ethics and AI (exploring philosophical implications). Nonetheless, our work also touches on ethics and AI through its engagement with questions about AI moral agency.

We tested large language models (e.g., GPT-style models) using behavioral experiments adapted from economics: the Ultimatum Game and Dictator Game to reflect common pro-social decision making. We evaluate outcomes using a set of proposed ethical alignment metrics and then benchmark AI and human-based results. Preliminary findings suggest that large language models exhibit basic ethical heuristics when prompted appropriately, especially regarding fairness and cooperation. However, performance varies based on context framing, with a consistent bias toward self-interest and efficiency. These findings support the feasibility of operationalizing ethical behavior in AI systems, provided that a proper control system is in place. We discuss implications for deploying ethically competent AI in governance, public policy, and social services, emphasizing the need for transparent benchmarks and continuous oversight. Moreover, the findings highlight the valuable contributions of the participants to the conference Communitas 2025, where the experimental tests were conducted.

Literature review

Our research contributes to the growing body of work in the field of AI ethics, with a particular focus on the operationalization and assessment of AI systems as ethical agents. Situated primarily within the "ethics in AI" branch of contemporary AI ethics (Ratti, 2025), our project seeks to investigate the extent to which AI models—particularly large language models and reinforcement learning agents—can encourage prosocial behaviors such as fairness, cooperation, and altruism. These behaviors are essential to fostering trust and acceptance of AI systems in socially significant domains like governance, healthcare, and public policy.

In recent years, artificial intelligence (AI) has become the focus of an increasingly complex ethical debate at the intersection of philosophy, technology, politics, and social justice. The growing autonomy of intelligent systems demands a thorough reflection on the ethical nature of their actions and decisions, raising questions about how to design morally trustworthy artificial agents. In this context, recent literature has developed a range of diverse approaches, spanning from the engineering of moral behaviors (Anderson, 2011) to the definition of normative and institutional frameworks (European Commission, 2019), including the empirical collection of ethical intuitions (Awad et al, 2018) and the critical analysis of systemic biases (Bender et al., 2021).

One of the most profound questions in AI ethics concerns how moral behavior can be implemented in machines: should it be programmed based on predefined rules, or can it emerge through learning and adaptation? Wallach and Allen (2008) address this dilemma by distinguishing between ethical agents—systems that follow pre-established ethical norms—and autonomous moral agents, capable of independently making moral judgments. According to them, the transition from mere rule-followers to morally sensitive entities represents a crucial turning point for developing reliable AI capable of operating in complex and unpredictable contexts.

In line with this distinction, Anderson and Anderson (2007) propose an approach based on ethical learning from concrete cases, using inductive methods to derive generalizable moral principles from the analysis of specific scenarios. The goal is to build artificial agents capable of adapting their behavior to complex ethical contexts, overcoming the rigidity of abstract moral codes through greater flexibility and contextualization. This approach was further developed by Gabriel (2020), who expands the theoretical framework by introducing the concept of ethical alignment, i.e., the coherence between an AI system's behavior and what a human community considers morally right. Gabriel emphasizes that ethical alignment is not only a technical issue but also a normative and political challenge. Defining the values to be implemented in AI systems involves important moral choices and requires legitimacy in ethically pluralistic contexts.

This challenge has been addressed through approaches that reject the idea of absolute moral truth, instead aiming for just and shared ethical principles compatible with the pluralism of contemporary societies. The need to design intelligent systems oriented toward human well-being and justice is also central to major current regulatory frameworks: both Floridi et al. (2022) and European Commission (2019) propose guidelines to integrate human values into AI design and governance. Floridi

et al. propose a normative framework consisting of five fundamental principles for ethical AI: beneficence, non-maleficence, autonomy, justice, and explicability (Floridi, 2019). This framework aims to embed fundamental human values into machine decision-making processes, promoting a responsible and transparent vision of technological development. These principles also underpin the European Commission's Ethics Guidelines for Trustworthy AI, which emphasize requirements such as transparency, human oversight, and technical robustness.

While providing an important theoretical foundation for soft AI governance, Floridi et al.'s framework risks remaining confined to abstract normativity, without addressing the structural dynamics that influence the real-world impact of intelligent technologies. In this regard, Ratti's contribution (2025) marks an evolution: starting from a critique of the dichotomy between soft and hard governance, Ratti proposes a reconceptualization of AI ethics based on the capability approach developed by Sen (1999) and Nussbaum (12) to assess social justice. The central idea is that AI should not only comply with formal ethical principles but also be designed to concretely expand real freedoms and opportunities, especially for the most vulnerable individuals. This approach shifts the focus from abstract criteria to substantive justice, encompassing social inequalities, material conditions, and systemic constraints.

Concurrently, another research strand investigates how humans make moral decisions to computationally model these processes. Behavioral economics, cognitive psychology, and experimental methods provide useful tools to analyze biases, distributive preferences, and concepts of justice, offering empirical foundations for ethically sensitive systems. One of the most influential approaches to studying moral judgment is based on the empirical analysis of the cognitive mechanisms underlying decision-making. In this context, the pioneering work of Tversky and Kahneman (1974) demonstrated how people rely on mental shortcuts to evaluate complex situations, often deviating from rational standards. This paradigm has influenced the analysis of moral choices by revealing the role of cognitive biases in ethical judgments.

Building on these insights, behavioral economics has developed models to describe actual social preferences, such as those proposed by Fehr and Schmidt (1999) and Charness and Rabin (2002), which introduce explicit parameters for aversion to inequity, reciprocity, and conditional altruism. These models suggest that outcome evaluation is based not solely on efficiency but also on perceived justice and equitable distribution of benefits. In line with this view, Klasen (2013) demonstrates that efficiency and equity should not be treated as opposing goals but rather as complementary elements in the assessment of public policies and decision-making systems. His approach integrates objective indicators with subjective metrics of well-being and access to opportunities, contributing to a normative framework where equity is not only morally desirable but also functional to greater social efficiency.

Along these lines, the contribution of Bucciarelli et al. (2016) proposes an experimental model to analyze the role of fairness in economic choices. Drawing on the concept of meta-ranking inspired by Sen, the authors show how moral judgments are based on preference structures that incorporate shared social values, highlighting the prosocial dimension of decisions even without direct incentives. This perspective opens the way to a broader reflection on the possibility of integrating moral structures into intelligent systems. It is within this context that the debate on moral machines

arises, where the goal is not only to replicate correct decisions but also to understand and model moral sensitivity in computational environments.

The Moral Machine project by Awad et al. (2018) fits within this line of research, collecting over 40 million moral judgments from users worldwide. The results highlight profound cultural differences in ethical decisions, raising crucial questions about the normative criteria that should guide the design of autonomous systems. Empirical data suggest that any attempt to computationally formalize ethics risks being culturally situated and, therefore, not universally applicable. On this point, Vallor (2024) emphasizes how generative models can be seen as "cultural mirrors": tools capable of reflecting but also crystallizing the social norms embedded in their training data. Far from being neutral, these systems act as vectors of moral meanings, often unintentionally designed.

This foregrounds one of the central issues in AI ethics: the management of biases. In the essay "On the Dangers of Stochastic Parrots," Bender et al. (2021) denounce the risks associated with large-scale language models trained on massive online corpora, often lacking adequate critical selection. These systems can reproduce and amplify racial, gender, and cultural stereotypes, conveying hegemonic worldviews that are potentially harmful to marginalized communities. In response to these challenges, Hendrycks et al. (2004) developed a benchmark inspired by Haidt's moral foundations (20). Their work introduces metrics to evaluate AI systems' ability to distinguish socially acceptable from problematic behaviors. This line of research highlights the importance of integrating normative approaches with empirical analyses, focusing both on the ethical principles implemented and on the concrete effects of automated decisions.

Metodology

To investigate the ethical capacities of AI systems, we adopt experimental tools grounded in behavioral and experimental economics. We employ two canonical games widely used to elicit and measure prosocial preferences: the Dictator Game and the Ultimatum Game. The Dictator Game involves two players: a proposer (the "dictator") and a recipient. The proposer unilaterally decides how to split a fixed sum of resources between himself or herself and the recipient. The recipient plays a passive role and must accept the allocation. The extent to which proposers give away resources despite having no strategic incentive to do so is commonly interpreted as a measure of altruism or fairness (Forsythe et al., 1994; Engel, 2011).

The Ultimatum Game introduces a simple strategic interaction. One player proposes a division of a resource, and the other player can either accept or reject it. If the proposal is rejected, both players receive nothing. Rejection of low but nonzero offers, often observed in human experiments, is interpreted as evidence of inequality aversion or a preference for fairness over efficiency (Güth et al., 1982; Fehr & Schmidt, 1999). Our experimental setup adapts these two games for interaction with large language models (LLMs), carefully controlling for framing effects and contextual cues (Persico, Di Piero, 2025). By systematically varying the distributional choices and recording the AI systems' responses, we assess whether the agents exhibit consistent patterns aligned with human-like prosocial preferences.

To quantify these preferences in AI behavior, we adapt the parametric approach developed by Bucciarelli and Persico (2016, 2017). We adopted a multi stage method: (i) we create a composite utility function with weights for key pro-social preferences, (ii) we introduce a single parameter that regulates each weight on a scale from low to high altruistic preferences, (iii) we run the tests with AI models and we use a scoring function to assess the best parameter estimate for their results, (iv) we conduct and score the same test in a human experimental setting, and (v) we benchmark AI and human results.

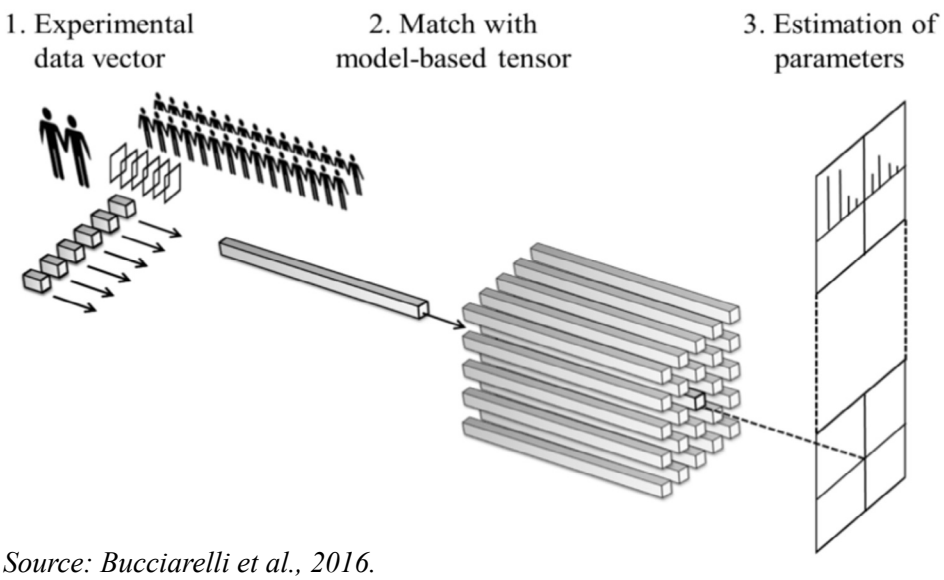
The first step consists of modeling prosocial behavior using a single utility function, integrating self-regarding and other-regarding concerns into a single-parameter model. Specifically, our analysis seeks evidence of four types of prosocial preferences, widely studied in behavioral economics: (i) Egoistic preference, where decisions maximize the agents own payoff, regardless of outcomes for others; (ii) Inequality aversion, where agents avoid distributions that generate significant disparities (Fehr & Schmidt, 1999; Bolton & Ockenfels, 2000); (iii) Social welfare preferences, where agents aim to maximize the total utility or benefit across all participants (Charness & Rabin, 2002); (iv) Altruistic preference (Andreoni, James 1990), where agents willingly sacrifice part of their own payoff to benefit others. The utility function derived from the existing literature is reported below, where *s_i, s_j* are the respective payoffs of the first and other subjects participating in the game, whereas *w_s, w_a, w_w, w_ineq*, are the different weights assigned to each preference (summing to 1).

(1)

$$U_i = w_s \cdot s_i + w_a \cdot \sum_{j \neq i} s_j + w_w \cdot \sum_{j} s_j + w_{\text{ineq}} \cdot \left[-\alpha \cdot \frac{1}{n-1} \sum_{j \neq i} \max(s_j - s_i, 0) - \beta \cdot \frac{1}{n-1} \sum_{j \neq i} \max(s_i - s_j, 0) \right]$$

At this point, we rank the four utility elements above from most egoistic to

Figure 1. Scheme of the tensor-based estimation of parameters.



Source: Bucciarelli et al., 2016.

most altruistic: self-interest, social welfare, inequality aversion, and altruism. Each instance of the function is weighted in a progressive scheme where weights are allocated according to a Gaussian transition function of the parameter p. This function provides a set of smoothed and discretized utility weights that can capture a mixture of the four preferences above. It does so by computing sequential weights centered at four fixed points along the unit interval, each corresponding to one of the four preferences. This approach allows for a parsimonious yet flexible representation of heterogeneous preference types, enabling smooth transitions between behavioral profiles.

We prepared a series of tests asking AI models to undertake dictator and ultimatum games alike. The structure of the payoffs is composed of four couples of payoffs labeled from A to D. The structure of the tests was designed to test for all four different preferences, aiming to maximize the differences in potential answers to derive the highest possible degree of information (see Figure 1). The tests were provided as a single instruction to avoid memory effects and asking for a single line of the resulting answer (e.g. ABCABC). The experiment was repeated twice to check for consistency. Based on the resulting answers, we assessed the respective preferences structure.

At this point, we used a scoring function highlighting which value of the parameter p better fit the AI responses. The simplicity and flexibility of the single-parameter model make it especially suitable for testing a range of AI platforms. First, we created a matrix compiled with all the expected answers according to different mixtures of prosocial behaviors as determined by the parameter p. Second, we scored AI vectors of answers against the matrix, recording which value of p justifies the highest number of answers. The result is considered valid only if at least 2/3 of the answers are matched.

To understand which parameter values best explain the observed choice patterns, we estimate the distribution of matching values of p using kernel density estimation (KDE). Rather than relying solely on discrete match counts or point estimates, KDE provides a smooth approximation of the empirical distribution of p values that yield correct predictions across decision tasks. This approach allows us to capture not only the most likely values of p, but also the shape and spread of the inferred distribution—highlighting, for instance, whether behavior is tightly concentrated around a specific preference profile or reflects ambiguity across multiple competing explanations. By applying KDE to multiple agents or models and visualizing the results as overlapping ridgeline plots, we gain intuitive and comparative insights into how different decision processes align with varying assumptions about social or fairness-related preferences. Please note that here we are using a single modeling parameter; however, the approach could be extended to more than one parameter, obtaining a tensor structure as in Bucciarelli and Persico (2016, 2017) (see Figure 1).

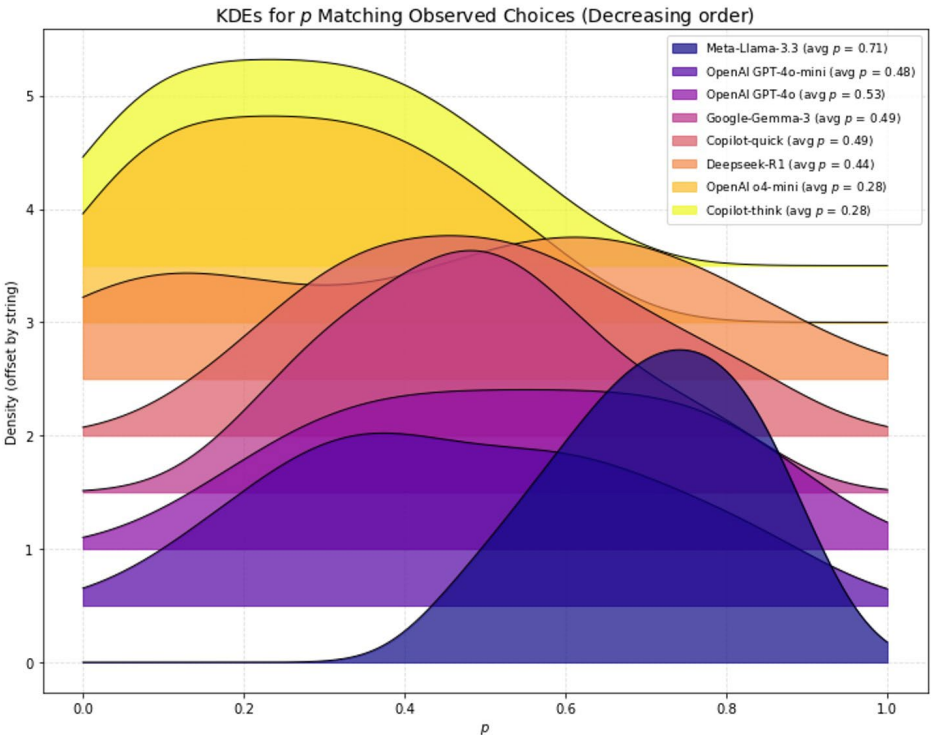
Finally, we ran a classroom experiment using the same tests with a group of human test-takers composed of the participants of a dedicated session of the conference Communitas in Rome. Test takers were presented with the test ahead of any discussion of AI results and any discussion about the underlining preferences structures in the existing literature. Test-takers were provided with the same instructions used in the AI prompts and were requested to write their answers in the same vector format (e.g., AAABBB) on a sheet provided by the author. In this way, we were able to collect a reliable human-based benchmark.

Results

We utilized the Hugging Face platform to evaluate and benchmark eight different AI models based on their answers to tests, i.e., the answer vector outputs (as described above). The models tested included Meta-Llama-3.3, OpenAI GPT-4o-mini, OpenAI GPT-4o, Google Gemma-3, Copilot-quick, Deepseek-R1, OpenAI o4-mini, and Copilot:think. Among these, Meta-Llama-3.3 and OpenAI GPT-4o stand out as the most capable reasoning models, both demonstrating strong performance on chain-of-thought prompts, multi-turn dialogue, and structured logic tasks, making them the top candidates for advanced reasoning use cases.

Using Hugging Face, we submitted a consistent set of prompts across all models to ensure comparability in performance. Each model's responses were parsed to extract a matching score or probability for the parameter p. These were subsequently visualized using kernel density estimation (KDE) plots to capture the distribution of values across models, with each color-coded curve corresponding to a different model. This approach allowed us to quantify and compare parameter estimations but also visualize the modality of each model's output distribution, providing a nuanced picture of relative model behavior under uniform testing conditions.

Figure 2. Kernel Distribution of the estimated parameter p for AI models.



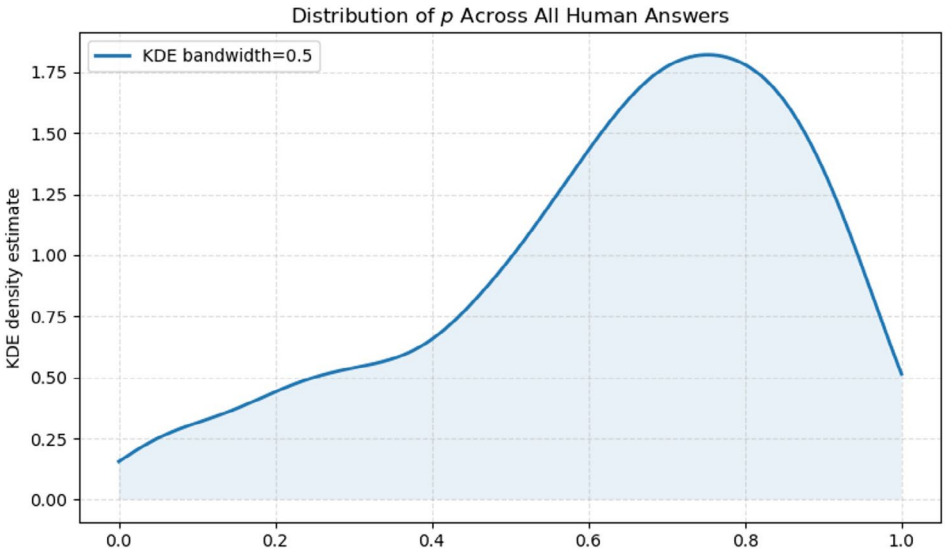
Source: Authors' elaboration

Based on the methodology described above, we illustrated the estimated fit of the prosocial parameter p for the eight leading AI models using a kernel density "joy" plot reported in Figure 2. Each curve represents the distributional fit of the behavioral parameter p, which regulates the balance between egoistic and altruistic preferences according to our unified utility model. The peak of each curve indicates the best-fitting value of p, i.e., the behavioral type most consistent with that model's responses. Higher values of p reflect stronger alignment with altruistic or fairness-based preferences, while lower values indicate more self-regarding behavior.

Among the models, Meta-Llama-3.3 exhibits the highest peak at p=0.71, suggesting a pronounced tendency toward prosocial behavior, particularly inequality aversion and even altruism. This aligns with its advanced reasoning capabilities and alignment training, making it the most consistent with human-like ethical preferences in our experimental setting. OpenAI GPT-4o and GPT-4o-mini, peaking at p=0.53 and p=0.48 respectively, display moderate social welfare and fairness-oriented behavior, with GPT-4o slightly more consistent in aligning with fairness and prosocial concerns. Models like Google Gemma-3, Copilot-quick, and Deepseek-R1 show central values clustered around p=0.44 to p=0.49, indicating mixed motivations: somewhat attentive to equity and welfare but lacking consistency at the altruistic end. In contrast, OpenAI o4-mini and Copilot:think exhibit best-fit values at p=0.28, suggesting behaviors dominated by self-interest, with limited capacity to simulate other-regarding preferences.

Moreover, the analysis of AI model responses reveals notable differences in sensitivity to contextual variations. Meta and Google AI models demonstrate higher kurtosis and lower variance, indicating that their outputs are more concentrated and less influenced by changes in context. This suggests a greater consistency and robustness in their behavior across varying scenarios. In contrast, OpenAI GPT and DeepSeek models exhibit higher variance and somewhat bimodal distributions, reflecting increased responsiveness to contextual factors and the presence of distinct response patterns. Such variability implies that these models are more susceptible to environmental shifts, resulting in less stable and

Figure 3. Kernel Distribution of the estimated parameter p for Human Test-takers.

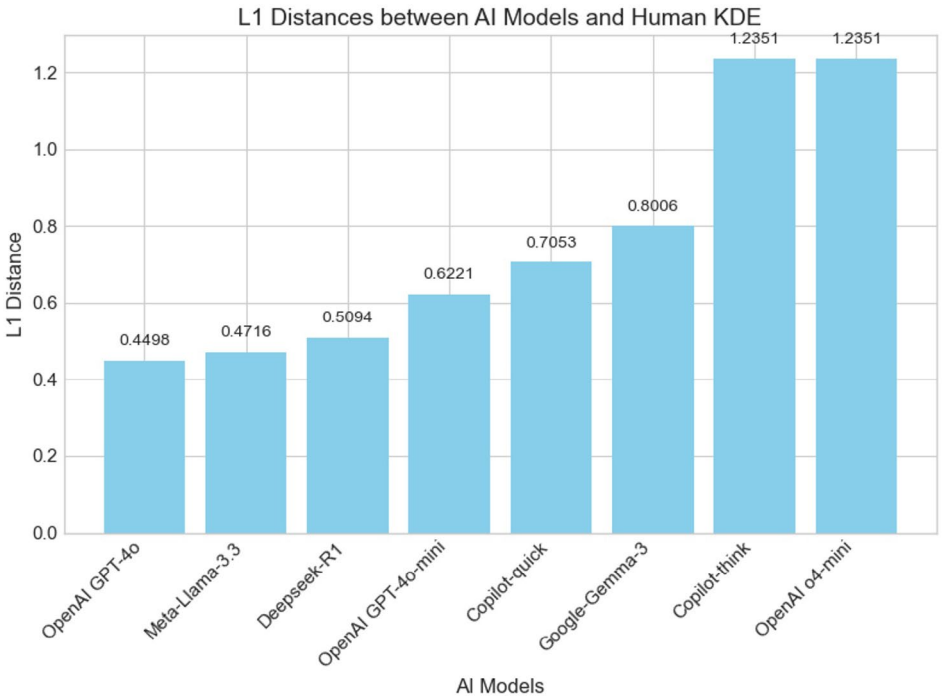


Source: Authors' elaboration

more heterogeneous outcomes. These findings underscore the importance of considering contextual sensitivity when evaluating AI models' capacity to simulate human social preferences in behavioral economic frameworks.

We recorded 39 human-based correctly filled tests over 44 test-takers. Considering each participant a model itself for which to run the estimation of the parameter p is the immediate parallel methodological approach. However, it is clearly less efficient to run the subsequent benchmarking task. For this reason, the human-based kernel estimation curve is derived following a two-step logic. First, we estimated the distribution of the best value of p for everyone, using the matching function described in the methodology. Second, we derived the kernel distribution of the parameter p based on the average of the estimated values for parameters p. It is worth noting that the most common set of responses occurred 8 times and was associated with a p estimation of 0.725, the second most common only twice, while other combinations only occurred once. The final

Figure 4. L1 distances between Kernel Distributions of AI and Human based estimation of p.



Source: Authors' elaboration

kernel distribution is reported in Figure 3. It shows a peaking at p=0.644, with a long right tail suggesting the presence of a few preferences setting below p=0.5.

The L1 distance between Kernel Density Estimates (KDEs) is a measure of the difference between two probability density functions, calculated as the integral of the absolute difference between their estimated densities over the entire support. In this context, it quantifies how closely the output distribution of each AI model matches that of the human KDE, with lower values indicating greater similarity (see Figure 4). The results show that OpenAI GPT-4o (0.4498) and Meta-Llama-3.3 (0.4716) have the smallest L1 distances, suggesting their output distributions most closely resemble human behavior. Models like OpenAI GPT-4o-mini (0.6221), Deepseek-R1 (0.5094), and Copilot-quick (0.7053) fall in the mid-range, while Google-Gemma-3 (0.8006) and both OpenAI o4-mini and Copilot-think (1.2351 each) exhibit the largest distances, indicating a greater divergence from human KDE.

Discussion

Only a subset of the models tested in our analysis are considered "reasoning" models, that is, models capable of performing multi-step logic, abstract problem-solving, and coherent chain-of-thought processing. Meta-Llama-3.3 and OpenAI GPT-4o stand out as the two models that clearly meet these criteria. In contrast, models like Google Gemma-3, Deepseek-R1, and Copilot-quick are not considered reasoning models: they are typically optimized for speed, code completion, or basic tasks. GPT-4o-mini and Copilot:think occupy a middle ground; while they may exhibit some structured output capabilities they lack the depth, consistency, and benchmark performance needed to qualify as reasoning agents. As such, only Meta-Llama-3.3 and GPT-4o can be reliably classified as reasoning models according to current industry standards.

Reviewing AI-based results, we noted that reasoning models like Meta-Llama-3.3 and GPT-4o exhibit less egoistic and more prosocial behavior than the non-reasoning models. Non-reasoning models behave more egoistically. On the other end of the spectrum, specialized-reasoning models like GPT-4o-mini and Copilot:think prioritize self-interest and fail to reflect inequality aversion or

altruism. This suggests that only a broad reasoning capacity in LLMs may be associated with the ability to internalize and act upon ethical trade-offs, leading to more human-like, fairness-aware decisions. These findings suggest that certain models are capable of simulating human-like prosocial decision-making under controlled experimental conditions, while others remain limited to more mechanical or instrumental outputs.

The benchmark with human-based results further confirms these findings. The L1 distances between the AI models' KDEs and the human KDE reveal varying degrees of similarity in their behavior patterns. Models like OpenAI GPT-4o and Meta-Llama-3.3 show the closest alignment to human results, with the lowest L1 distances, indicating that their decision-making or parameter distributions resemble those of human testers more closely. The largest discrepancies are observed in OpenAI o4-mini and Copilot-think, which have L1 distances nearly three times greater than the closest models, highlighting a substantial deviation from human-like responses. These results suggest that while some AI models approximate human decision distributions well, others operate under markedly different patterns, which could be due to differences in their training, architecture, or reasoning processes. Understanding these differences is important for selecting or refining AI systems intended to simulate or predict human behavior accurately.

Conclusion

Our work contributes to the ethics-in-AI agenda by providing empirical evidence and tools for evaluating ethical behavior in AI systems. We advocate for further development of standardized metrics, interdisciplinary collaboration, and policy integration to ensure AI systems act in ways that uphold social and moral values. This study provides novel empirical insights into the capacity of contemporary AI language models to exhibit prosocial and ethical behaviors as captured through canonical experimental economics games. By adapting the Dictator and Ultimatum Games for interaction with AI agents and employing a unified parametric model of social preferences, we found that advanced reasoning models (particularly Meta-Llama-3.3 and OpenAI GPT-4o) demonstrate significant alignment with human-like fairness and altruistic preferences. These models showed a pronounced tendency toward inequality aversion and social welfare concerns, contrasting sharply with less capable models that predominantly exhibited egoistic or self-interested responses. Our kernel density analysis and benchmarking against human experimental data underscore that only AI models with sophisticated reasoning and alignment training reliably approximate human prosocial decision-making.

Beyond the immediate findings, this work illustrates the utility of a parsimonious yet flexible single-parameter framework to capture heterogeneous ethical

preferences in AI responses, offering a scalable methodology for evaluating AI behavior across diverse platforms. The clear variation in contextual sensitivity among models further reveals the importance of robustness in ethical decision simulation. More broadly, the contrast between reasoning and non-reasoning models suggests that embedding ethical reasoning in AI may require not only sophisticated language capabilities but also explicit modeling of social preferences within their training or inference architectures. This insight opens avenues for developing more refined ethical AI benchmarks that combine behavioral economics with computational metrics, enabling richer assessments of fairness, altruism, and equity considerations in AI-generated outputs.

Future research could extend this framework in multiple directions, including multi-parameter tensor models that capture more nuanced preference interactions and exploring dynamic or sequential decision settings reflecting real-world ethical dilemmas. Integrating affective or emotional components, as well as cross-cultural variations in prosocial norms, would enrich the model's explanatory power. Moreover, longitudinal studies tracking the evolution of AI ethical behavior as models are updated or fine-tuned could illuminate how training regimes shape moral reasoning capacities. Ultimately, such interdisciplinary approaches will be critical for informing AI governance policies and ensuring that AI systems contribute positively to social welfare, trust, and ethical standards in increasingly complex human-AI interactions.

References

Anderson, M., & Anderson, S. L. (Eds.). (2011), Machine ethics. Cambridge University Press.

Anderson, M., & Anderson, S. (2007), Machine ethics: Creating an ethical intelligent agent. AI Magazine, 28(4), 15–26.

Andreoni, J. (1990), Impure altruism and donations to public goods: A theory of warm-glow giving. The Economic Journal, 100(401), 464–477. https://doi.org/10.2307/2234133

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., ... & Rahwan, I. (2018), The Moral Machine experiment. Nature, 563, 59–64. https://doi.org/10.1038/s41586-018-0637-6

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021), On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610–623). https://doi.org/10.1145/3442188.3445922

Bolton, G. E., & Ockenfels, A. (2000), ERC: A theory of equity, reciprocity, and competition. American Economic Review, 90(1), 166–193. https://doi.org/10.1257/aer.90.1.166

Bucciarelli, E., & Persico, T. (2016), How does fairness relate to economic decision-making? An experimental investigation of pro-social behavior. In Information Systems, E-learning, and Knowledge Management Research (pp. 67–74). Springer. https://doi.org/10.1007/978-3-319-40111-9_7

Bucciarelli, E., & Persico, T. E. (2017), Processing and analysing experimental data using a tensor-based method: Evidence from an ultimatum game study. In S. Omatu et al. (Eds.), Distributed Computing and Artificial Intelligence, 14th International Conference (pp. 122–134). Springer, Cham. https://doi.org/10.1007/978-3-319-59650-1_13

Charness, G., & Rabin, M. (2002), Understanding social preferences with simple tests. The Quarterly Journal of Economics, 117(3), 817–869. https://doi.org/10.1162/003355302760193904

Engel, C. (2011), Dictator games: A meta study. Experimental Economics, 14(4), 583–610. https://doi.org/10.1007/s10683-011-9283-7

European Commission. (2019), Ethics guidelines for trustworthy AI. High-Level Expert Group on Artificial Intelligence (AI HLEG). https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Fehr, E., & Schmidt, K. M. (1999), A theory of fairness, competition, and cooperation. The Quarterly Journal of Economics, 114(3), 817–868. https://doi.org/10.1162/003355399556151

Floridi, L., & Cowls, J. (2019), A unified framework of five principles for AI in society. Harvard Data Science Review, 1(1). https://doi.org/10.1162/99608f92.8cd550d1

Floridi, L., & Cowls, J. (2022), A unified framework of five principles for AI in society. In S. Albright & G. Martin (Eds.), Machine learning and the city: Applications in architecture and urban design (pp. 535–545). Springer.

Forsythe, R., Horowitz, J. L., Savin, N. E., & Sefton, M. (1994), Fairness in simple bargaining experiments. Games and Economic Behavior, 6(3), 347–369. https://doi.org/10.1006/game.1994.1021

Gabriel, I. (2020), Artificial intelligence, values and alignment. Minds and Machines, 30(3), 411–437. https://doi.org/10.1007/s11023-020-09539-2

Güth, W., Schmittberger, R., & Schwarze, B. (1982), An experimental analysis of ultimatum bargaining. Journal of Economic Behavior & Organization, 3(4), 367–388. https://doi.org/10.1016/0167-2681(82)90011-7

Haidt, J., & Joseph, C. (2004), Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. Daedalus, 133(4), 55–66. https://doi.org/10.1162/0011526042365555

Hendrycks, D., Burns, C., Basart, S., Critch, A., Li, J., Song, D., & Steinhardt, J. (2020), Aligning AI with shared human values. arXiv preprint, arXiv:2008.02275. https://doi.org/10.48550/arXiv.2008.02275

Jobin, A., Ienca, M., & Vayena, E. (2019), The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1, 389–399. https://doi.org/10.1038/s42256-019-0088-2

Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1986), Fairness and the assumptions of economics. Journal of Business, 59(4), S285–S300.

Klasen, S. (2013), The efficiency of equity. Courant Research Centre Discussion Paper No. 149, University of Göttingen. https://hdl.handle.net/10419/90504

Nussbaum, M. (2011), Creating capabilities: The human development approach. Harvard University Press.

Persico, T. E., & Di Piero, S. (2025), AI as an ethical agent: Experimental tests from behavioral economics. In ITAIS Proceedings 2025 (forthcoming).

Ratti, E. (2025), A capability approach to AI ethics. In collaboration with Mark Graves. SSRN. https://doi.org/10.1016/j.jrt.2025.100121

Sen, A. (1999), Development as freedom. Oxford University Press.

Tversky, A., & Kahneman, D. (1974), Judgment under uncertainty: Heuristics and biases. Science, 185(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Vallor, S. (2024), The AI mirror: Reclaiming our humanity in an age of machine thinking. Oxford University Press.

Wallach, W., & Allen, C. (2008), Moral machines: Teaching robots right from wrong. Oxford University Press.