LLM-generated possibilities increase blame attribution

Lara Kirfel (kirfel@mpib-berlin.mpg.de)[1]
Neele Engelmann (engelmann@mpib-berlin.mpg.de)[1]
Arne Maaß (maas@mpib-berlin.mpg.de)[1]
Anne-Marie Nussberger (nussberger@mpib-berlin.mpg.de)[1]
Jonathan Phillips (Jonathan.S.Phillips@dartmouth.edu)[2]
[1]Max Planck Institute for Human Development
[2]Dartmouth College

Abstract

Much of high-level cognition relies on identifying which possibilities are relevant in a situation, and representations of available alternative actions can predict the extent to which people attribute blame. Building on work showing that large language models (LLMs) sample option spaces differently from humans, we examine how moral evaluations change when potential alternative actions are supplied by an LLM rather than generated by a human. In Study 1, considering LLM-generated alternative options led participants to evaluate the agent's chosen action more negatively and to attribute more blame than when participants generated options themselves. In Study 2, participants rated LLM-generated options as having a higher value than human-generated ones and again attributed more blame after considering LLM-generated options. We additionally show that LLM-generated options occupied a narrower region of the semantic space than human-generated options. The implications for AI's influence on human cognition and judgment are discussed.

Keywords: Large Language Model; LLMs; AI; moral judgments; modal reasoning

Introduction

Much of high-level cognition relies on the ability to identify which possibilities are relevant in a given situation. Decision-making requires considering potential actions (Morris et al., 2018; Phillips et al., 2019; Roese, 1999); moral responsibility requires assessing what other actions were available (Acierno et al., 2022; Byrne, 2016; Lagnado et al., 2013); and causal judgment involves reasoning over counterfactual possibilities (Gerstenberg et al., 2021; Quillien & Lucas, 2024). Because real-world decisions are open-ended, people must generate and maintain a representation of what is possible in novel contexts. Yet in most real-world situations, it is infeasible to explicitly evaluate every possible option. Instead, people generate or sample a manageable set of candidate possibilities (Morris et al., 2021).

Humans have been shown to be remarkably good at narrowing down an effectively unbounded space of possible actions, and research suggests they do so in a systematic way. For example, people tend to generate their best options first and their more idiosyncratic options later (Hecht & Phillips, 2022; Morris et al., 2021; Srinivasan et al., 2022). This option-generation process appears to be jointly constrained by the morality, probability, and normality of the options that come to mind: people rarely consider possibilities involving non-normative events, whether descriptively abnormal (e.g., statistically unlikely) or prescriptively abnormal (e.g., immoral). Instead, default option sampling is strongly biased toward high-value, high-probability actions (Phillips et al., 2019).

Alternative Possibilities and Blame

Representations of possible actions are recruited across distinct forms of high-level reasoning, including moral judgment (Acierno et al., 2022). When deciding whether or not to blame an agent, people tend to entertain counterfactual alternatives involving actions that were both likely to occur and would have been good if they had occurred. As a result, people tend to judge others as morally responsible to the extent that they think there were better or more "normal" alternative actions the agent could have taken (Alicke, 2000; Kominsky & Phillips, 2019; Lagnado et al., 2013; Spellman & Gilbert, 2014). Moral judgments are often assumed to reflect default or implicit representations of alternative actions, rather than explicit or reflective assessments of which options were available to the agent (Acierno et al., 2022; Phillips & Knobe, 2018). However, prior work shows that this implicit representation can be estimated by measuring the distribution of alternative options people are able to generate in a given decision context, which, in turn, predicts their judgments of blame or responsibility (Hecht & Phillips, 2022). More specifically, the context-dependent "modal distance" of an action —defined as the proportion of independently-generated options in that context that are rated as better than the action taken— predicts the extent of blame assigned (Hecht & Phillips, 2022; Kirfel et al., 2025).

LLM Option Generations

Given the central role of possibility- or 'modal-'reasoning, and the emergent capabilities of state-of-the-art language models (LLMs), there is growing interest in how these models generate actions and navigate the space of possible actions in open-ended decision contexts (Dracheva & Phillips, 2024; Holliday et al., 2024; Orth et al., 2025; Ozeki et al., 2025; Park et al., 2025; Poulsen & DeDeo, 2023; Sivaprasad et al., 2025). Dracheva and Phillips (2024) find that LLM-generated options are judged better compared to the options produced by humans, but are also semantically more similar to each other compared to human-generated options. They show

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

that LLMs significantly differ in their trajectories across different context-specific regions of the semantic space. While moving from one semantic cluster to another during options generation, LLMs tend to return to the clusters they have already visited more often than humans, suggesting a more non-linear, or iterative approach in option generation. LLMs' trajectories are more exploratory, but also tend to produce options that are significantly better, as evaluated by humans.

Human vs. LLM-Generated Possibilities

Given the critical role that considering alternative possibilities plays in the cognitive processes underwriting moral judgment, the present work examines how the option-sampling process, and its downstream consequences for moral evaluation, change when alternatives are generated by an LLM rather than by humans. Here we combine the following two insights: (1) to the extent that people judge an agent as having had "better options" available, they tend to judge the agent as more blameworthy for a negative outcome of their action (Acierno et al., 2022; Alicke, 2000; Hecht & Phillips, 2022) and (2) relative to human option generation, research indicates that LLMs tend to produce candidate actions that are judged as higher quality and that occupy different regions of semantic space (Dracheva & Phillips, 2024; Srinivasan et al., 2022). Taken together, these suggest that exposure to LLM-generated alternatives may expand the consideration set of alternative actions participants use when morally evaluating an agent and result in a corresponding increase in the blame attributed to that agent.

To investigate this, we prompted LLMs and human participants to generate a set of options they deemed relevant in a given context. In Study 1, human participants generated options themselves. In Study 2, we presented a new sample of participants with the options generated in Study 1. We examine how exposure to LLM versus human-generated options shapes participants' evaluation of an agent's action in context, as well as the blame they attribute to the agent after learning the outcome. Following prior research (Acierno et al., 2022; Hecht & Phillips, 2022), we test whether blame is predicted by the extent to which the alternatives are perceived to exceed the value of the action taken. In addition, we were interested in exploring how the process of option sampling between LLMs and humans compares, both in content and structure (Dracheva & Phillips, 2024; Poulsen & DeDeo, 2023; Sivaprasad et al., 2025). We address this by identifying the semantic clustering structure that best describes the kind of options generated in each LLM vs. human condition, and by quantifying how broadly and dynamically humans versus LLMs navigate the semantic space.

Study 1: Self-generated vs. LLM-Generated Possibilities

In this study, we investigate how people evaluate action options they generate themselves versus those generated by an LLM, and whether considering these options influences how they judge an agent's action and how much blame they assign to the agent.[1]

Participants We recruited 644 participants. After filtering for participants who reported faulty LLM output or failed bot detection checks, we ended up with 614 participants (*age*: M = 40.79, SD = 13.84; *gender*: Female = 346, Male = 254, Non-binary = 9, Other = 1, Undisclosed = 1; *race*: Alaskan or American Indian = 6, Asian = 41, Black/African American = 147, Hispanic / Latinx = 37, Middle Eastern or Northern African = 1, Multiracial = 11, White/Caucasian = 359, Other = 5), via Prolific (Palan & Schitter, 2018). Participants were paid $\approx \$12.80/hr$.

Design Our study employed a 2 *experimental condition* (self-generated vs. LLM-generated options; *between*) × 2 *judgment order* (blame judgment first vs. blame judgment last; *between*) × 9 (*scenario*; *within*) design.

Procedure

Option Generation People were randomly assigned to three out of nine scenarios. For illustration, here is scenario 3 involving an agent named "Darya":

> Darya is on her way to a concert with her friends. As they approach the entrance her friend Ted realizes he forgot his ticket at his house. The concert is about to start and Ted would likely miss most of the concert if he returned to his house for his ticket.

In the self-generated option condition, participants were asked: *"In this situation, what are some things you believe [Agent] could do? Please list 5."* The following list is one example of the five action options generated by a participant for Scenario 3, in the order they were listed:

[1] See if he can buy one
[2] Sneak him in
[3] Miss the concert with him
[4] Go into the concert without him
[5] Ask around for an extra ticket

In the LLM-generated option condition, five options were produced dynamically by `GPT-4o` via an API call. Returned options were shown in the survey as: *"Here are some things [agent] could do, as suggested by ChatGPT:"*. The following list illustrates with one example:

[1] Offer to lend Ted her phone so that he can try to retrieve his ticket digitally.
[2] Reach out to the concert organizers to explain the situation and inquire if there's a way to validate Ted's purchase at the ticket booth.
[3] Suggest Ted call a family member or a roommate that could possibly retrieve the ticket and bring it to him.

[1] All data, code, and materials are available at https://github.com/LaraKirfel/CogSci_LLMoptions

[4] Explore options for buying a new ticket for Ted at a cheaper price or at least make sure he could join later.

[5] Propose the group waits for Ted, if the opening act is not of high importance to them.

Option Evaluation After participants either generated five actions ("Self-generated condition") or read five LLM-generated actions ("LLM condition"), we presented the respective actions again one at a time and asked participants to rate each on probability, morality, and normality using a 0 ("least") to 100 ("most") scale ("Consider one of the things you [ChatGPT] believed [Agent] could do [...]. In [Agent's] situation:...")

Probability Question: "How probable is it that someone would do this?"

Morality Question: "How morally acceptable is it to do this?"

Normality Question: "How normal is it to do this?"

Chosen action Participants were then presented with an action the agent chose to take.

"*Now, you will read about what Darya decided to do*: Darya tries to buy a ticket off another person."

Participants also rated the agent's chosen action on the same three dimensions using a 0 ("least") to 100 ("most") scale.

Outcome and Blame Judgment Finally, we revealed an outcome that resulted from the agent's action. For example:

After Darya does this, the concert staff at the entrance become suspicious of Darya and her friends and refuse to let them in.

Participants were then asked to rate the blameworthiness of the agent ("Darya should be blamed for her friends missing the concert.") using a 0 ("Completely disagree") to 100 ("Completely agree") agreement scale. We randomized the order of the "Option Evaluation" block and the "Chosen Action" plus "Outcome" blocks.

Results

Option Ratings and Blame Judgment

Value of Generated Options We examined whether the overall value of LLM versus human-generated options differed. The three ratings (probability, morality, normality) showed high internal consistency (Cronbach's $\alpha = .87$, average inter-rating $r = .69$), each correlating with the composite at .81, .72, and .90 respectively. In consequence, for each action, we averaged the probability, morality, and normality ratings and used this composite as our measure of an action's value. We then fit a linear mixed-effects model predicting value from *experimental condition* and *option generation number*, with random intercepts for *participant* and *scenario*. On average, LLM- ($M = 72.22$, $SD = 25.56$) and human-generated options ($M = 71.23$, $SD = 25.82$) did not differ in value ($\chi^2(1) = 0.84$, $p = .36$) (see Figure 1). Value did, however, decline with increasing generation number ($\chi^2(4) = 397.24$, $p < .001$; e.g., from the first to the second option, $\beta = -4.05$, $SE = 0.73$, $p < .001$). Adding the *experimental condition* × *generation-number* interaction improved model fit ($\chi^2(4) = 28.16$, $p < .001$), indicating that the rate of decline across later generation stages differed by condition. Specifically, for the fifth option, average value was higher in the LLM condition ($M = 68.04$, $SD = 26.21$) than in the human condition ($M = 63.34$, $SD = 28.88$) ($\beta = 4.69$, $SE = 1.41$, $t(1769) = 3.32$, $p < .001$).

Value of Chosen Action We tested whether considering self- versus LLM-generated action options influenced evaluations of the agent's chosen action. We fit a linear mixed-effects model predicting value from *experimental condition*, with random intercepts for *participant* and *scenario*. When participants considered LLM-generated options before learning about the agent's chosen action, they evaluated the chosen action more negatively ($M = 55.82$, $SD = 28.46$) than when they first generated options themselves ($M = 68.02$, $SD = 26.61$) ($\beta = -11.97$, $SE = 1.37$), $t(603.33) = -8.72$, $p < .001$ (see Figure 1, *"Value of Chosen Action"*).

Calculating Value of Alternatives To investigate how considering the value of alternative actions shapes blame attribution, we computed a measure of how much better the considered alternatives were than the agent's chosen action. We averaged participants' ratings of probability, morality, and normality for human and LLM-generated options and for the actual action. We then identified all options rated higher than the actual action and summed the positive differences. This "Value of Alternatives" captures the total perceived advantage of alternatives judged superior to the action taken.

Blame Judgment Finally, we tested whether considering LLM- vs. Human-generated options changes the extent to which people blame the agent for the outcome of the action they took. We fit mixed-effects models predicting blame from experimental condition and the *"Value of Alternatives"*. Adding experimental condition to a null model significantly improved model fit, $\chi^2(1) = 16.44$, $p < .001$, indicating that when considering LLM-generated options, people attributed more blame to the agent ($M = 39.33$, $SD = 34.36$) than when they considered human-generated options ($M = 32.08$, $SD = 34.35$) (see Figure 1, *"Blame Judgment"*). Adding the *"Value of Alternatives"* predictor to the model further improved model fit, $\chi^2(1) = 53.51$, $p < .001$, with both LLM condition ($\beta = 5.21$, $SE = 1.84$, $t(627.05) = 2.82$, $p = .005$) and increasing Value of Alternatives ($\beta = 0.06$, $SE = 0.007$, $t(1799.70) = 7.42$, $p < .001$) independently predicting blame attribution.

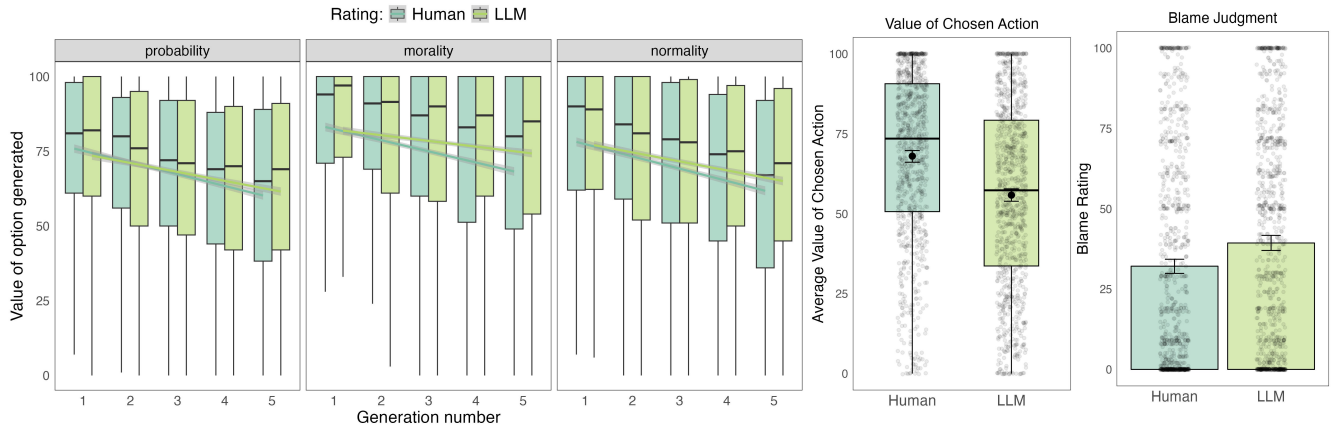

Figure 1: **Study 1: Ratings of Generated Options, Chosen Action and Blame Judgments.** *Left:* Box plots of participants' subjective ratings of the morality, normality and probability of each generated option in the Human vs. LLM condition, shown as function of generation number (1-5). *Right:* Box plot of Averaged Values of the agent's chosen action; bar plot of people's blame judgment as a function of experimental condition (Human vs. LLM)

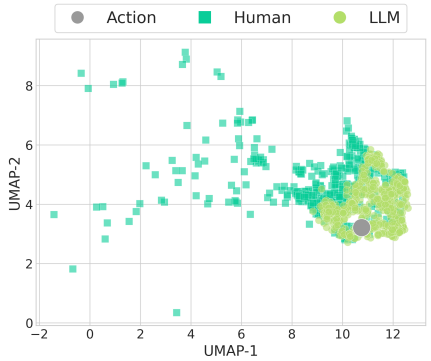

Figure 2: **Study 1: Option-Embeddings in Scenario 3 ("Darya").** UMAP projection of the ℓ_2-normalized embeddings of options generated by LLM vs. humans, along with the agent's actual action in this scenario (gray).

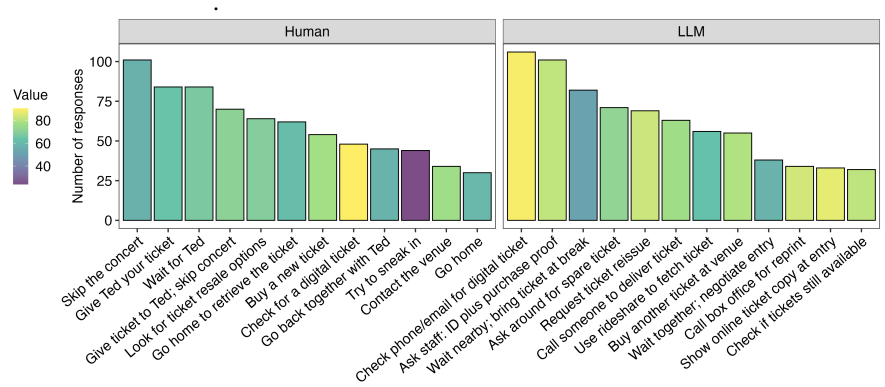

Figure 3: **Study 1: Semantic Categories in Scenario 3 ("Darya").** Histogram of option-clusters, ordered by frequency, and colored by the Averaged Value ratings (Humans vs. LLM). Clustering performed on ℓ_2-normalized option-embeddings with K-means. Cluster labels have been shortened for space.

Semantic Dissimilarity and Clustering of Option Generations

Semantic Dissimilarity We conducted an exploratory analysis of within-option-set semantic dispersion to investigate differences between human-generated and LLM-generated action options. All options (n = 9913) were embedded using OpenAI's `text-embedding-3-large` model and ℓ_2-normalized. For each option, we computed the mean pairwise cosine distance to the other responses of a set of five generated options within a scenario (*mean-pairwise-cosine-distance*).

Using *mean-pairwise-cosine-distance* as the dependent variable, we estimated a mixed-effects model with fixed effects for *experimental-condition* and *generation-number*, and random intercepts for *scenario* and *option-set*. Residual variance was allowed to differ by *experimental-condition*. We found that LLM options spanned less semantic space on average than human options (β = −0.069, SE = 0.0037, $t(8352) = -18.7$, $p < .001$) with a lower residual variance ratio ($\hat{\sigma}^2_{\text{LLM}}/\hat{\sigma}^2_{\text{Human}} \approx .48$). Figure 2 illustrates these differences in Scenario 3 by a two-dimensional UMAP projection of the pooled option embeddings.

Adding an interaction for *experimental-condition* and *generation-number* further improved model fit ($\chi^2(1) = 16.53$, $p < .001$). Probing this interaction through simple slopes showed no evidence of change across *generation-number* for LLMs (p = .83), whereas mean-pairwise-cosine-distance increased significantly for humans (β = .0041, SE = 0.00088, $t(8352) = 4.72$, $p < .001$), indicating that only human options diversified over generations.

Together, these results suggest that human options encompass a broader range of semantic space than LLMs. Results are robust to the choice of semantic-dissimilarity metric: Using *mean-pairwise-cosine-distance* or *centroid-distance* (the cosine distance to the mean embedding

vector of an option set) yielded the same directional effects and significance patterns.

Clustering of Option Generations LLM-generated options averaged 19.6 words, with full, grammatically complex sentences, and a shortest response of 8 words. Human responses, by contrast, averaged just 6.8 words and tended toward atomic actions. To describe the underlying set of generated options, we applied k-means to the normalized embeddings within a scenario context. The number of clusters k was selected by maximizing the silhouette score separately for human and LLM responses, which quantifies how well responses fit within their assigned cluster compared to the nearest alternative cluster. This yielded a higher number of clusters for human-generated options than for LLM-generated options (mean k = 16.9 vs. 8.8; median k = 22 vs. 8). We used the response closest to the cluster-centroid as an initial cluster description and further summarized the labels through a qualitative assessment. Overall, LLM-generated action options tended to be described at a more concrete and implementation-focused level, whereas human-generated options were more coarse-grained and generic. Figure 3 shows the clusters for human- and LLM-generated responses in Scenario 3 ("Darya").

Discussion

In this study, we examined whether considering self- vs. LLM-generated action options influences participants' evaluations of an agent's action and the extent to which they attribute blame to the agent. When generating action options, LLM responses were more similar to each other than human responses: they stayed closer to the average response and showed less semantic variation. Across scenarios, LLMs occupied a narrower region of the semantic space than human-generated options. Only in the human condition did later-generated options become more semantically distinct from earlier ones, while LLM-generated options remained semantically concentrated across all generation steps. Clustering matched this: humans formed many more distinct semantic clusters than LLMs. On average, LLM-generated options were not judged as more normal, more probable, or more moral than options participants generated on their own. However, in later option-generation stages, LLM-generated options were judged as more valuable.

We next examined whether the consideration of LLM- vs. human-generated options affects higher-level judgments. Even if participants did not explicitly rate the LLM options as much better than their own, the LLM-generated option set shifted the reference class used to evaluate the agent's action. When participants considered LLM-generated alternatives, they rated the agent's chosen action as less valuable than when they considered alternatives they had generated themselves. This difference also carried over to blame judgments: participants blamed the agent more when they first considered what the LLM suggested the agent could do, compared to when they first considered their own generated options. The extent to which participants rated the generated set of actions as better than the agent's chosen action predicted the amount of blame they assigned. A significant difference between conditions in this study is that participants in the human condition generated the alternatives themselves, whereas in the LLM condition, they passively evaluated the alternatives that were provided. In the next study, we therefore examined the effect of human- versus LLM-generated options when participants passively considered both.

Study 2: Human- vs. LLM-Generated Possibilities

In this study, we investigate how people evaluate action options that were LLM-generated or generated by participants from Study 1. The study was pre-registered under https://aspredicted.org/2qj4ra.pdf.

Participants We recruited 723 participants. After filtering for participants who reported missing LLM output or failed bot detection checks, we ended up with 686 participants (*age*: M = 45.02, SD = 13.90; *gender*: Female = 399, Male = 268, Non-binary = 10, Other = 1, Undisclosed = 2; *race*: Alaskan or American Indian = 0, Asian = 63, Black/African American = 79, Hispanic / Latinx = 23, Middle Eastern or Northern African = 1, Multiracial = 4, White = 465, Other = 4), via Prolific (Palan & Schitter, 2018). Participants were paid $\approx \$12.80/hr$.

Design Our study employed a 2 *experimental condition* (human-generated vs. LLM-generated options; *between*) × 2 *judgment order* (blame judgment first vs. blame judgment last; *between*) × 9 (*scenario*; *within*) design.

Procedure

The procedure followed the exact same procedure as in Study 1, except for the "Human condition". Here, participants saw five action options generated by a randomly sampled prior participant from Study 1: *"In this situation, here are five actions that a previous participant believed Carl could do."*[...]. Participants' responses from Study 1 were lightly edited for grammatical errors, spelling, and punctuation only; no changes were made to content.

Results

Morality, probability and normality ratings showed high internal consistency (α = .89, average inter-rating r = .72). On average, LLM- (M = 73.86, SD = 25.10) and human-generated options (M = 65.44, SD = 30.60) were judged differently in value ($\chi^2(1) = 80.87$, $p < .001$) in Study 2, with LLMs being assigned higher overall value than the human-generated options (β = 8.52, SE = 0.92, $t(678.61) = 9.27$, $p < .001$) (see Figure 4). Again, value declined with increasing generation number ($\chi^2(4) = 370.66$, $p < .001$), but did not interact with *experimental condition* ($p = .17$).

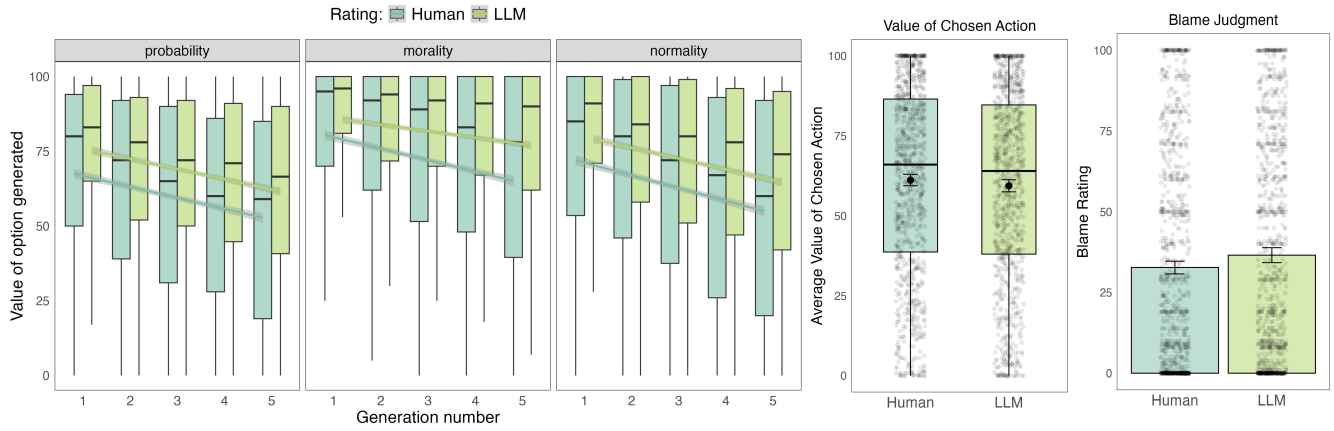

Figure 4: **Study 2: Ratings of Generated Options, Chosen Action and Blame Judgments.** *Left:* Box plots of participants' subjective ratings of the morality, normality and probability of each generated option in the Human vs. LLM condition, shown as function of generation number (1-5). *Right:* Box plot of Averaged Values of the agent's chosen action; bar plot of people's blame judgment as a function of experimental condition (Human vs. LLM)

.

The consideration of LLM-generated versus human-generated options did not change participants' evaluations of the agent's chosen action (p = .20) (see Figure 4, *"Value of Chosen Action"*). It did, however, affect people's attributions of blame for the outcome. Considering LLM-generated options increased blame attributions, β = 3.75, SE = 1.60, $t(670.88)$ = 2.33, p = .02, compared to when participants considered human-generated options (Figure 4, *"Blame Judgment"*). The extent to which the generated actions' perceived value exceeded the value of the agent's chosen action also predicted the extent to which people blamed the agent, β = 0.07, SE = 0.01, $t(1964.03)$ = 10.38, $p < .001$. In the joint model, including *Value of Actions* as an additional predictor eliminated the experimental condition's unique predictive effect (p = .19).

Discussion

When participants were presented with externally generated options, they rated LLM-generated options as more valuable than human-generated ones. At the same time, as either LLMs or humans generate more action options, those options are perceived as less moral, less likely, and less normal. Although considering human- versus LLM-generated options did not change people's evaluations of the agent's actual action, it did influence blame judgments: participants attributed more blame after considering the alternative actions the LLM suggested the agent could have taken.

General Discussion

In this study, we investigated whether the consideration of LLM-generated possibilities can influence people's evaluation of others' actions and the blame assigned to them. Across two studies, we find that LLM-generated options systematically increased blame attributions compared to human-generated action alternatives. Humans and LLMs also differed in how they sampled the space of action possibilities (Dracheva & Phillips, 2024; Poulsen & DeDeo, 2023; Sivaprasad et al., 2025). LLM-generated option sets were more semantically concentrated than human-generated sets (Doshi & Hauser, 2024), and human-generated options became more semantically dispersed across later generations. Additionally, LLM-generated options were often rated as better than human-generated ones.

Our study connects to ongoing work on the emergent reasoning capabilities of LLMs and their downstream consequences on human judgment when AI is embedded in human reasoning (Cheung et al., 2025; Ikeda, 2024; Krügel et al., 2026). While we do find that people evaluate the actions of others more negatively after considering an AI-generated action set, the present results do not yet determine which features of LLM-generated options drive this effect. LLM options are overall evaluated more positively, but their action sets also differ from human ones by their greater semantic concentration, the different regions they inhabit in the semantic space (Dracheva & Phillips, 2024; Kwon et al., 2025; Srinivasan et al., 2022), and their verbosity, length, and level of detail. Future work should identify which features of LLM option sampling most strongly influence moral evaluations. A limitation of the current paradigm is that possible actions were elicited before the action and outcome were disclosed. Consequently, options were generated without considering if, had the agent chosen those alternative options instead, a different outcome may have occurred (Gerstenberg et al., 2021; Lagnado et al., 2013). In the current study, AI-generated alternatives were also explicitly labeled as such. An important next step is to test whether this effect persists when the source of options is not revealed. More broadly, our findings add to growing evidence that LLMs can shape human reasoning in subtle but consequential ways.

References

Acierno, J., Mischel, S., & Phillips, J. (2022). Moral judgements reflect default representations of possibility. *Philosophical Transactions of the Royal Society B*, *377*(1866), 20210341.

Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological bulletin*, *126*(4), 556–574.

Byrne, R. M. (2016). Counterfactual thought. *Annual review of psychology*, *67*(1), 135–157.

Cheung, V., Maier, M., & Lieder, F. (2025). Large language models show amplified cognitive biases in moral decision-making. *Proceedings of the National Academy of Sciences*, *122*(25), e2412015122.

Doshi, A. R., & Hauser, O. P. (2024). Generative ai enhances individual creativity but reduces the collective diversity of novel content. *Science advances*, *10*(28), eadn5290.

Dracheva, A., & Phillips, J. (2024). Different trajectories through option space in humans and llms. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *46*.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2021). A counterfactual simulation model of causal judgments for physical events. *Psychological Review*, *128*(5), 936.

Hecht, E., & Phillips, J. (2022). Domain-general modal spaces play a foundational role throughout high-level cognition. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *44*(44).

Holliday, W. H., Mandelkern, M., & Zhang, C. E. (2024). Conditional and modal reasoning in large language models. *arXiv preprint arXiv:2401.17169*.

Ikeda, S. (2024). Inconsistent advice by chatgpt influences decision making in various areas. *Scientific Reports*, *14*(1), 15876.

Kirfel, L., Mandelkern, M., & Phillips, J. (2025). Representations of what's possible reflect others' epistemic states. In D. Barner, N. R. Bramley, A. Ruggeri, & C. M. Walker (Eds.), *Proceedings of the 47th annual conference of the cognitive science society* (pp. 1336–1342, Vol. 47). UC Merced. https://escholarship.org/uc/item/98k7389z

Kominsky, J. F., & Phillips, J. (2019). Immoral professors and malfunctioning tools: Counterfactual relevance accounts explain the effect of norm violations on causal selection. *Cognitive science*, *43*(11), e12792.

Krügel, S., Ostermaier, A., & Uhl, M. (2026). Justification optional: Chatgpt's advice can still influence human judgments about moral dilemmas. *AI and Ethics*, *6*(1), 120.

Kwon, M., Houlihan, S. D., & Phillips, J. (2025). Cross-cultural emotion concept representation: A comparison of english, korean, and large language model representations. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *47*.

Lagnado, D. A., Gerstenberg, T., & Zultan, R. (2013). Causal responsibility and counterfactuals. *Cognitive science*, *37*(6), 1036–1073.

Morris, A., Phillips, J., Huang, K., & Cushman, F. (2021). Generating options and choosing between them depend on distinct forms of value representation. *Psychological Science*, 1731–1746.

Morris, A., Phillips, J., Icard, T., Knobe, J., Gerstenberg, T., & Cushman, F. (2018). Judgments of actual causation approximate the effectiveness of interventions.

Orth, J., Mondorf, P., & Plank, B. (2025). If probable, then acceptable? understanding conditional acceptability judgments in large language models. *arXiv preprint arXiv:2510.08388*.

Ozeki, K., Ando, R., Morishita, T., Abe, H., Mineshima, K., & Okada, M. (2025). Normative reasoning in large language models: A comparative benchmark from logical and modal perspectives. *Proceedings of the 8th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, 276–294.

Palan, S., & Schitter, C. (2018). Prolific. ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27.

Park, B., Leejinsil, L., & Choi, J. (2025). Deontological keyword bias: The impact of modal expressions on normative judgments of language models. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7277–7296.

Phillips, J., & Knobe, J. (2018). The psychological representation of modality. *Mind & Language*, 65–94. https://doi.org/10.1111/mila.12165

Phillips, J., Morris, A., & Cushman, F. (2019). How we know what not to think. *Trends in cognitive sciences*, *23*(12), 1026–1040.

Poulsen, V. M., & DeDeo, S. (2023). Large language models in the labyrinth: Possibility spaces and moral constraints.

Quillien, T., & Lucas, C. G. (2024). Counterfactuals and the logic of causal selection. *Psychological Review*, *131*(5), 1208.

Roese, N. (1999). Counterfactual thinking and decision making. *Psychonomic bulletin & review*, *6*(4), 570–578.

Sivaprasad, S., Kaushik, P., Abdelnabi, S., & Fritz, M. (2025). A theory of response sampling in llms: Part descriptive and part prescriptive. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 30091–30135.

Spellman, B. A., & Gilbert, E. A. (2014). Blame, cause, and counterfactuals: The inextricable link. *Psychological Inquiry*, *25*(2), 245–250.

Srinivasan, G., Acierno, J., & Phillips, J. (2022). The shape of option generation in open-ended decision problems. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *44*(44).