Inverse option generation:

Inferences about others' values based on what comes to mind

Jane Acierno[a], Clare Kennedy[b], Fiery Cushman[b], and Jonathan Phillips[a,1]

a. Department of Cognitive Science, Dartmouth College, New Hampshire, USA

b. Department of Psychology, Harvard University, Massachusetts, USA

1. Address correspondence to Jonathan Phillips [jonathan.s.phillips@dartmouth.edu]; 5 Maynard St. Winfred Raven House, Cognitive Science, Hanover, NH 03755

Abstract

Prior work shows that when people try to think of things, such as solutions to a problem, the options that come to mind most often are those that they consider statistically common and valuable. Here, we ask whether ordinary people anticipate this and, therefore, infer that when uncommon solutions come to someone's mind, it is diagnostic of how much those solutions are represented as valuable. To illustrate, imagine your friend is brainstorming what to cook for a vegetarian couple and says aloud, “maybe pizza, burritos, or penne alfredo? No, let’s make a stir-fry.” While some of the options can be explained by both their value and statistical frequency (e.g., pizza or burritos), you might infer that only your friend’s unique love for penne alfredo explains why that option came to mind. Across four studies we demonstrate inferences of this kind, and our results suggest that participants are able to make such inferences by inverting their intuitive understanding of the option-generation process itself. Whereas many current models of our folk theory of mind focus on the core mechanics of deliberative choice – such as the use of beliefs and desires to plan rational action – our results show a much broader folk understanding of pre-deliberative aspects of thought, such as the very process of option generation.

Imagine that you are solving an Escape Room game with some work colleagues, in which you all use various objects and clues to crack codes and solve puzzles in an attempt to exit the room as quickly as possible. At one point, you find a screwdriver and ask your colleague if she has any ideas about what to do with it. Thinking quickly, your colleague says that she is not sure, but she knows that one thing you shouldn't do with the screwdriver is stab somebody.

What is interesting about this case is that the content of the thought that came to your colleague's mind is morally appropriate: she is undoubtedly correct that one should not stab another player with a screwdriver. Yet, at the same time, one gets the sense that there may be something questionable about her moral character for having even had such a thought in the first place; perhaps she is not the kind of person one should spend time locked in a room with.

This kind of inference is interesting in that it seems to be driven not by any judgment or decision your colleague made, but rather by the mere thoughts that happened to come to her mind. In this paper, we explore this kind of phenomenon and ask whether such inferences arise because we have an intuitive theory of how possible actions are generated when solving open-ended decision problems. We first consider evidence for the ability to invert the decision making process to infer other people's latent values that may drive what comes to their minds, and then return to the moral domain to examine whether this ability might also inform moral character judgments.

Prior work on decision making

Prior work on open-ended decision making (*e.g.*, deciding what to have for lunch out of all of the possible things one could eat) illustrates the key role of option generation (Hauser, 2014; Kaiser et al., 2013; Kalis, Kaiser, & Mojzisch, 2013; Phillips, Morris, & Cushman, 2019; Figure 1). In the case of deciding what to have for lunch, a few potential options (*e.g.*, salad, pizza, tacos) may pop into one's mind, and then one goes on to decide which is best among those options. A key focus of contemporary research is to ask how each of these two stages of decision-making works.

Our studies focus on the first stage: option generation. Prior research indicates that several different factors influence which options come to mind, but that statistical frequency and value play an especially important role (Phillips, Morris, & Cushman, 2019; Bear et al., 2020; Srinivasan,

Acierno, & Phillips, 2022; Wang et al., 2024). Statistically frequent options come to mind more often: when thinking of potential dinners, we call to mind dishes that are often eaten; when thinking of zoo animals, we call to mind those often found in zoos. Additionally, valuable options come to mind more often: We are more likely, for instance, to think of dinners and zoo animals that we especially like. Indeed, this role of value includes not just the instrumental value of things, but also their moral value (Acierno, Mischel, & Phillips, 2022; Crockett, 2016; Phillips, Luguri, & Knobe, 2015; Phillips & Cushman, 2017; Phillips & Knobe, 2018).

The influence of value on option generation has been a special target of recent research. This work shows that option generation is especially influenced by value representations that generalize over a wide variety of contexts, and are weighted towards common ones. Put simply, the "valuable foods" that come to mind, for instance, are those that are valuable *in general,* and especially in common past contexts. This means that when a person faces very peculiar present circumstances—say, due to dental surgery they must eat something soft and bland— the options that come to mind may nonetheless include ones that are generally tasty, but poorly suited to these unusual circumstances, such as tacos (Morris et al., 2021; Klein et al., 1995; Johnson & Raab, 2003). This role of general (or "context-free") value is unique to option generation and plays less of a role in the evaluation of options once they have been generated; one is unlikely to actually decide to eat tacos immediately after a dental surgery. At the same time, the options that are generated can also be guided by other context-specific factors that make an option good for the problem faced (Mills & Phillips, 2023). Returning to the previous example, a person will be more likely to call to mind soft, bland foods after dental surgery than they would under normal circumstances.

In summary, prior research has shown that for a given person, the possibilities that are most likely to come to their mind are those that they perceive to be generally valuable, morally good, and common. Our question is, when given information about what came to someone else's mind, can people infer the implicit values and statistical expectations of the person who generated them? That is, based on what comes to mind for any given person, do we infer that they think those options are regarded as high-value (e.g., morally good) and statistically frequent?

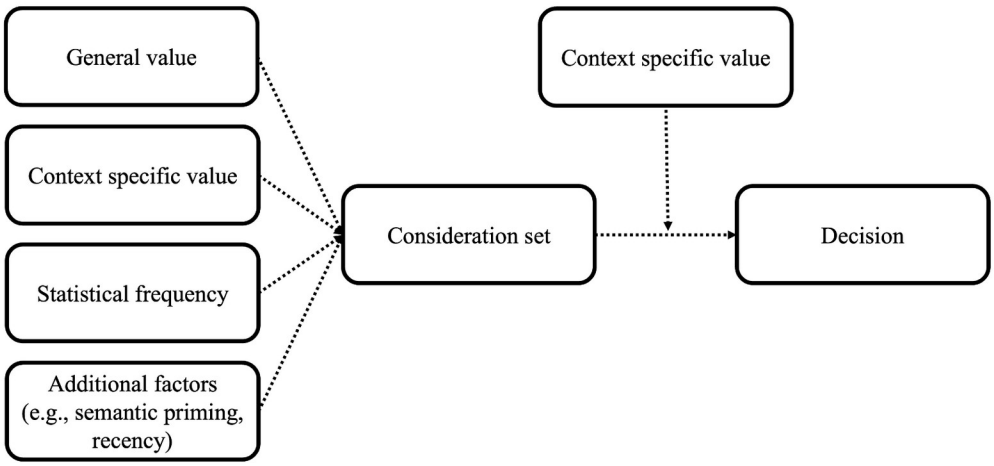

Figure 1. *Schematic model of the option generation process in the context of open-ended decision making.*

Theory of mind and inferred values

Theory of mind allows people to infer other's hidden mental states, such as their beliefs, desires, and preferences, by observing their actions. This is done by inverting a generative model linking mental states to action (Baker, Saxe, & Tenenbaum, 2009; Jara-Ettinger, 2019; Gershman et al., 2016). For instance, if we have a model specifying that *liking* an apple causes people to *reach* for apples then, when we observe somebody reach for an apple, we can invert this model to infer they probably like apples. Recently, a number of researchers have applied this approach to moral inferences specifically (Crockett, Kim, & Shin, 2024; Kleiman-Weiner, Saxe, & Tenenbaum, 2017).

Here, we apply this inverse approach not to physical actions, but to the mental action of calling an option to mind. We investigate whether people have an intuitive theory of how certain options come to mind rather than others, and thus when they find out what comes to another's mind, they can "invert" that cognitive process to infer others' subjective values: realizing that such a thought would be most likely to come to mind if the person sees that thing as either quite valuable or quite likely to occur.

Some prior work offers circumstantial evidence in favor of this possibility. Morwedge and colleagues (2014) find that, regardless of the content of a thought, the more the thought simply "pops into one's mind", the more we believe it reveals meaningful self-insight. For example, when instructed to randomly (vs. deliberatively) think of an attractive person

other than their significant other, the spontaneous thoughts are believed to provide greater self-insight into their level of attraction to the person (Morewedge, Giblin, & Norton, 2014). Relatedly, Critcher and colleagues (2012) find that immoral decisions that were made with less reflection are thought to be more revealing of underlying character traits, due to the assumption that quick decisions reflect decision certainty and thus unambiguous behavioral motivation. Additional work outside of the moral domain by Gates and colleagues (2021) demonstrates that people draw inferences from decision times and incorporate this information when inferring how careless a decision-maker was. However, this picture may be complicated by findings in Pizarro et al. (2003) suggesting that impulsive immoral behavior is discounted, perhaps because it is interpreted as a reaction to an external situational factor rather than revealing some genuine internal desire of the agent. Here, we build on this line of work but look specifically at how the options that come to mind, holding fixed the actual actions, may influence moral character judgements.

In summary, if the ordinary folk theory of mind includes an accurate representation of the option-generation process, then people may conclude that the options that spontaneously come to a person's mind are diagnostic of their values, including their moral values. We test this possibility in four studies. Our studies draw on two core principles of causal inference: diagnosticity and screening off. First, our studies test the principle that information about what comes to mind is diagnostic about a person's subjective perceptions of that option's frequency, general value, and situation-specific value, among other properties. Second, our studies test the principle that when an option that comes to mind satisfies one of these criteria, this screens off its diagnosticity regarding the other properties. This is a core principle of causal inference. Just as knowledge of a rainstorm "screens off" the diagnostic value of wet grass for inferring that the sprinkler was on, if a person is asked to think of a pub food and says "hamburger", its commonness screens off any strong inference about the person's values. If, instead, the person says "Moules frites", the fact that it is relatively less common leads to greater diagnostic value in terms of the person's own preferences.

Study 1: Inferred values of ordinary category members based on what comes to mind

In Study 1, we tell participants which zoo animals come to a person's mind when planning a zoo trip, and then ask whether our participants use

this information to draw inferences about their preferences for those animals. More specifically, we built on a paradigm used in Mills and Phillips (2023) in which participants were asked to decide which zoo animal to take a group of young children to see with limited time at the zoo. In this study, participants were asked to list all of the zoo animals that came to mind even if those animals were ultimately not selected when making this decision. Prior results indicate that people consider cute and harmless animals the best choices in these specific circumstances. In line with this, cute and harmless animals, e.g., flamingos or penguins, were often among those that came to participants' minds. This is not particularly surprising because they are good *situation-specific* options. At the same time, Mills and Phillips also noticed that other zoo animals that were neither cute nor harmless, e.g., lions or rhinos, also frequently came to mind, even though they are poor options for the situation. Those that were not good options for the specific situation faced, were generally common or especially notable (Mills & Phillips, 2023).

These prior findings provide a test case for our predictions of diagnosticity and screening-off. Specifically, we predict that participants should infer that the previous person must especially like a particular zoo animal when this animal came to mind despite not being able to be explained away either by (1) commonality or (2) situation-specific value. To illustrate, if a prior participant thought of taking the small children to see a lion, a hyena, and a flamingo, then participants should specifically infer that this person must especially like hyenas because this animal is neither a good option for the problem being solved (*cf*. flamingo) and is not a zoo animal that is common (*cf*. lion), and so the only explanation for why it would come to mind is that the person must particularly like hyenas.

Methods

In Study 1 we utilized existing data from work by Mills and Phillips (2023) in which participants were told to imagine that they are taking a group of children to the zoo but there is only time to see one animal. Mills and Phillips' (2023) participants reported both the animal they selected, and also up to eight additional animals that they considered while making the decision. For the present study, we asked a new group of participants to rate how much the previous participant likes each animal in general (as a measure of context-independent subjective value), as well as how good the animal would be for the particular problem faced involving young children. To measure whether the inferred values were influenced by having come to

mind during decision making, we also collected data from a separate group of participants who were simply asked to rate how much each zoo animal is liked in general with no additional information about a specific decision context. These ratings provide a baseline against which we can compare participants inferred values for animals that came to mind during decision making.

Participants

Initial choice sets were taken from Mills and Phillips' (2023) participants ($N = 100$). Participants with choice sets consisting of fewer than 2 animals were dropped from analyses, leaving 97 remaining participants.

English speaking participants ($N = 155$) located in the United States were recruited via Prolific. Our sample consisted of 66 female participants, 87 male, and 2 non-binary. General value ratings were collected by asking "Independent of their decision at the zoo, how much do you think the previous participant generally likes this animal?" using a 100-point sliding scale with exact numbers hidden, ranging from "this animal is likely one of their least favorite" to "this animal is likely one of their most favorite". To determine context-specific value ratings, we asked "Specifically considering the decision they had to make at the zoo, how good would it be to go see each of these animals?" using a 100-point sliding scale with exact numbers hidden, ranging from "this animal would be among the worst animals they could choose" to "this animal would be among the best animals they could choose".

Finally, to capture baseline differences in how much different animals are like, an additional sample was collected asking how much people generally like each of the 66 animals considered by Mills and Phillips' (2023) original participants ($N = 100$).

Results

As a first pass, we ask whether participants overall made the inference that the options that come to mind are generally well-liked. Indeed, collapsed across each choice set, average ratings for how much the previous participant generally likes each animal were significantly above the scale midpoint, $t(154) = 13.39$, $p < .001$. This confirms that indeed people infer that others like the animals that come to mind. This should not be surprising, given evidence that people may simply think of zoo animals that are generally well-liked (Mills & Phillips, 2023).

Building on this, we next ask the more specific question of whether participants infer that others like a given zoo animal when that animal coming to mind cannot be explained by either the general frequency of thinking of that zoo animal or by the fact that the external problem faced may have prompted that zoo animal to come to mind. To do this, we operationalized how common it was to think of a particular zoo animal by calculating (based on Mills & Phillips, 2023) the number of times a given animal was generated, divided by the total number of animals generated across all choice sets. Additionally, we calculated the mean rating of how good participants thought each specific zoo animal would be for the specific problem faced, *i.e.*, taking children to the zoo. We can use these two to predict inferred general value. Specifically, we expect people will infer the prior participant liked the animal more than people generally do when that animal is generally unlikely to come to mind *and* when the animal is not actually a good solution for the problem at hand.

To ask this statistically, we conducted a generalized linear mixed effects regression using the "lmer" package in R predicting the difference between inferred value and general value using the commonality of the option and the rating of how good the option is in the given context. The analysis revealed a main effect of commonality, $\chi^2(1) = 81.051$, $p < .001$, such that more common options were more likely to be selected. Additionally, we found a main effect of context-specific value, $\chi^2(1) = 16.457$, $p < .001$, such that options with high context-specific value were more likely to come to mind. Critically, we also found an interaction between commonality and context-specific value, $\chi^2(1) = 33.728$, $p < .001$, such that the higher the commonality and context specific value was, the smaller the difference between inferred and general value for that animal. Specifically, options were inferred to be especially well-liked when it was not common to think of them, and they were also not a good option for the problem being solved (Figure 2).

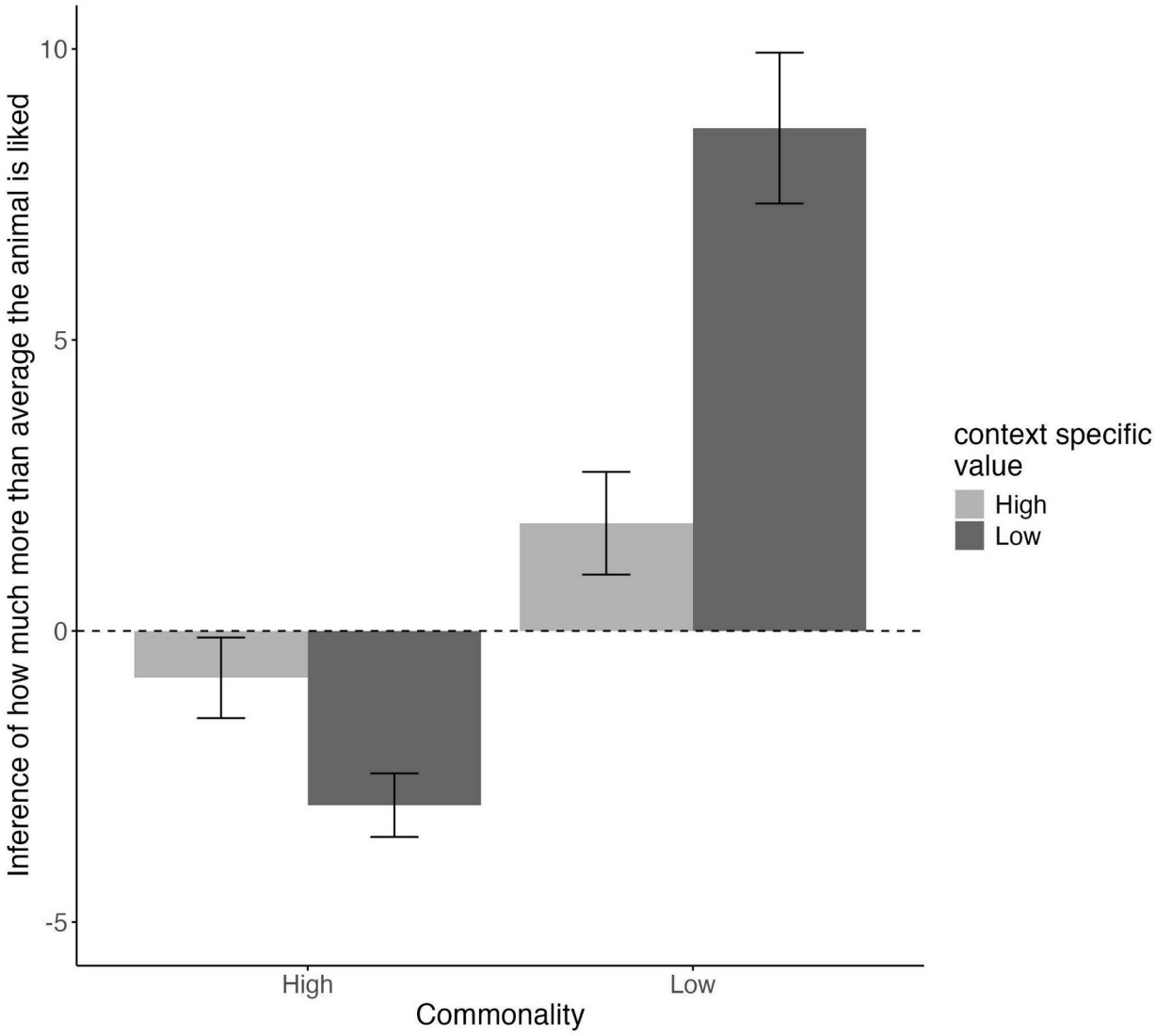

Figure 2. *Barplot of participants' inferences of how much animals that came to mind were liked in general. Left bars depict inferences about animals that were more common to think of; right bars depict inferences about animals that were more uncommon to think of. Lighter bars depict inferences about animals that were better options for the specific decision problem faced; Darker bars depict inferences about animals that were worse options for the specific decision problem faced. Error bars depict +/- 1 SEM.*

Discussion

This first study demonstrated the core principles of diagnosticity and screening off in participants' inferences about others' subjective values based on what comes to mind: their inferences were sensitive both to the

commonality of the animal and to how good that animal is for the specific problem being faced, such that thinking of a particular zoo animal was only seen as diagnostic when its coming to mind was not screened off by either of these two features. This specific pattern suggests that participants may have a nuanced intuitive model of the option generation process that they are inverting to make inferences about others' preferences. If this is right, we would expect that participants' theory of mind inferences based on what comes to mind may also be sensitive to other idiosyncratic features of the option generation process. To explore this possibility, we next consider whether participants' value inferences are sensitive to the order in which an option comes to mind, another signature feature of option generation during open-ended decision making.

Study 2: Option generation order guides value inferences

Prior work shows not only that higher-value options are more likely to come to mind when problem-solving, but also that the highest-valued options are most likely to come first (Phillips et al., 2019; Srinivasan, Acierno, & Phillips, 2022; Johnson & Raab, 2003; Klein et al., 1995). If participants are inverting their intuitive understanding of the option generation process, and that process has this signature feature, then we may expect to find the same signature in participants' inferences of others' subjective values. In other words, we should find that participants are most likely to infer that others particularly like a given option when they generate that option earlier, rather than later, in the decision-making process.

To test this, we utilized the data collected by Morris and colleagues (2021) in which participants are asked to decide what to make for a meal after having dental surgery, and listed all of the foods that came to mind. They then rated both how much they liked each food in general, and how good a solution it was in the specific situation; the majority of the variance for what came to mind was predicted by general liking. We present a new set of participants with the different choice sets generated by participants in Morris et al. (2021), and ask them to infer how much the previous participants liked each of the foods that came to mind in general and also how good each food was for the specific dinner problem faced. We hypothesized that participants would, on average, infer that the options generated earlier during the option-generation process were regarded as having a higher general value.

Methods

Participants

English speakers located in the United States were recruited via Prolific (N = 155, 79 male, 69 female, and 2 other).

Procedure

The present study used choice sets generated by previous participants as part of a study by Morris and colleagues (2021). From Morris and colleagues' raw data, 6 participants' choice sets were not included because they included options that were not words.

Participants were told that previous study participants were given the following hypothetical situation, and asked "what would you cook yourself for dinner tonight?".

> *Imagine that you just got dental surgery, and your doctor gives you food restrictions for the night. You're supposed to avoid food with seeds, foods that require too much chewing, and foods that are moist.*

Participants were informed that previous study participants wrote lists of foods in the order in which the foods came to mind, and that for the present study they were going to see 12 previous participants' lists of foods and would be asked to estimate how much the person liked each of the foods. An example choice set was provided for clarity.

> *For example, if you see the following list:*
> *1. Pasta*
> *2. Rice*
> *3. Pizza*
> *... that means that pasta came to mind first, rice came to mind second, and pizza came to mind last.*

After completing a comprehension check, participants were then randomly presented with 12 choice sets, out of 196 possible choice sets, generated by previous participants in a study by Morris and colleagues (2021). On 7-point Likert scales, participants were asked to infer how much the previous participant liked each food ("*this food is likely one of their [**least favorite** (1) / **favorite** (7)] dishes*") as well as the context-specific value of each dish ("*this food would be among the [**worst** (1) / **best** (7)] dishes they could make in this situation*").

We ask whether subjects' inferences are sensitive to the order in which an option was generated such that they inferred that foods generated earlier in the decision making process were more liked (in general) by the person.

Results

Before tackling our primary question of whether generation-order guides value inferences, we first wanted to confirm that earlier options had higher value relative to later options for the original option-generation data from Morris et al. (2021). While this has been found in other work, it was not analyzed in the original paper. Our secondary analysis of the Morris et al. (2021) data confirms this pattern of results. Specifically, employing the "lmerTest" function in lme4 package (Bates, et al., 205) over a maximal linear mixed-effects model, we observed that the order in which an option came to mind was predictive of how much that option was "generally liked" by the person who came up with it, such that earlier options were rated more highly than later options, $t(112.45) = -2.192$, $p < .05$ (Figure 3).

As a second preliminary analysis, we next ask whether participants overall could predict how much prior participants liked particular options that came to mind. That is, regardless of the order of options, were participants able to infer which options were more liked than others? We also found that this was the case: participants' inferred preference ratings linearly predicted prior participants' actual preference rating, $t(749.73) = 7.867$, $p < .001$.

Turning to our primary analysis, we now ask whether participants' inferences of others' preferences (*i.e.*, how much the person likes each food) were sensitive to the order in which an option was generated. In other words, independent of how much the food was actually liked, do participants infer that options that were generated later in the decision process were probably liked less than those that were generated earlier? To ask this question statistically, we used a linear mixed-effects model to ask whether option order predicted inferred preference ratings controlling for how much the food was actually liked by the person who generated that option. Naturally, controlling for the original participants' actual food preferences will artificially deflate the observed effect of order, as prior work has demonstrated that participants' preferences exert a causal effect on the order in which the food options come to mind (Morris et al., 2021); however, this step is necessary to ensure that any observed effect of order

on inferences is due simply to information about order, rather than information derived from specifics of the food items. This analysis revealed that even when controlling for how much original participants liked the foods, option order was indeed predictive of inferred preference ratings, $t(89.59) = -6.767$, $p < .001$, such that participants are more likely to infer a higher preference for options generated earlier in the choice set versus later (Figure 3).

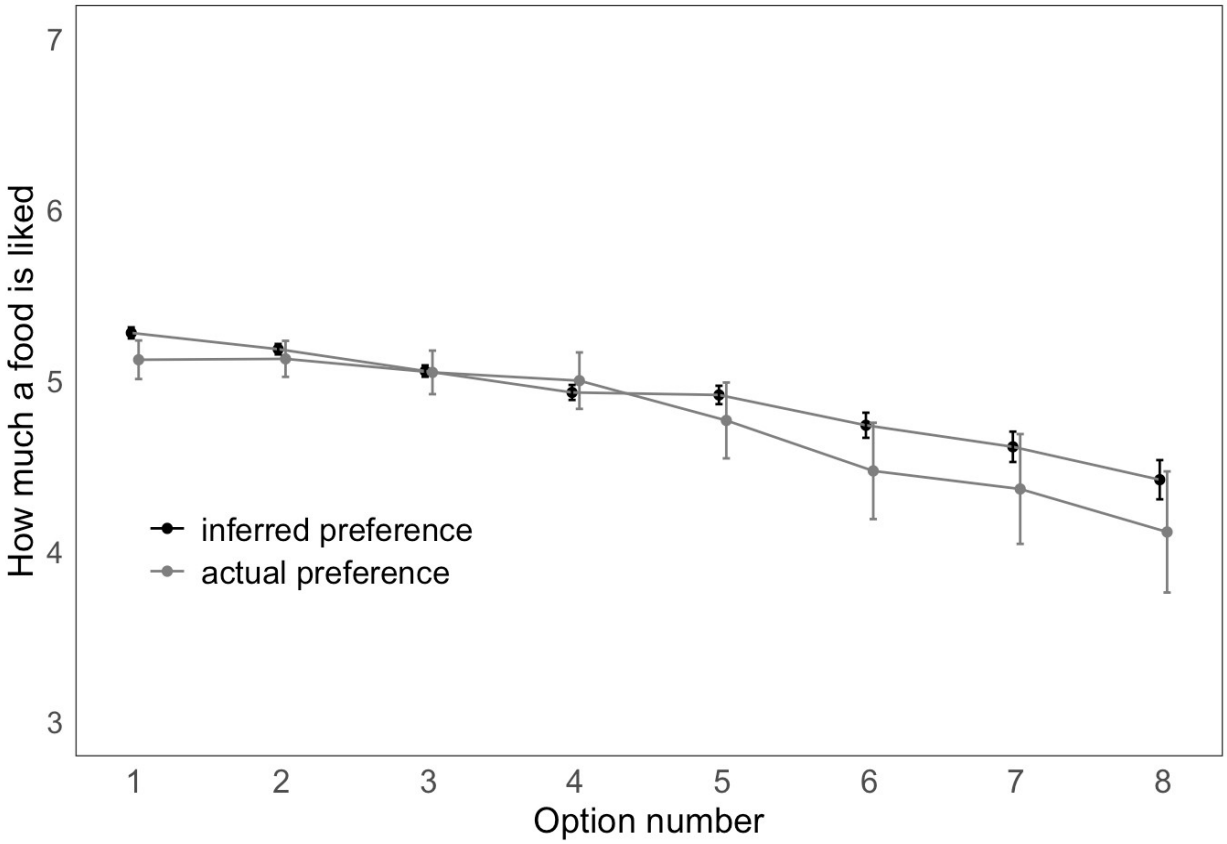

Figure 3. *Plot of Morris and colleagues' (2021) participants' ratings of how much they liked each food option they generated (light-gray points), and our participants' inferred ratings of how much the previous participants liked those food options (dark-gray points). X-axis represents the order in which each food option was generated. Error bars depict +/- 1 SEM.*

Discussion

This second study asked whether participants' value inferences based on what comes to mind are sensitive to the order in which an option came to mind---a signature feature of option generation during open-ended decision making. Using a paradigm from Morris et al. (2021), we found additional evidence that participants not only seem to have an intuitive model for the process of calling options to mind, but also found that their inferences are sensitive to how this process unfolds over time as people generate additional options. This ability to predict others' preferences based on the order in which they generated the option provides further evidence that participants may be inverting a model of the option-generation process in inferring the latent values that drive what comes to others' minds. Study 1 and 2 together provide clear evidence for a sophisticated capacity for making theory of mind inferences based on the options that come to mind during open-ended decision making. In the next two sets of studies we ask whether this capacity can also guide more consequential social inferences, such as inferring others' *moral* values.

Study 3: Moral character inferences based on what comes to mind

In Study 3 we hold fixed the actual action that is done, and ask whether people treat information about what comes to mind as diagnostic of a person's moral character. Specifically, across three closely related studies, we ask whether merely thinking of (and immediately rejecting) an action that would cause harm to another person leads to an inference that the agent in question must place a low value on others' wellbeing. Additionally, we manipulate the extent to which inferences about a person's moral character would be screened off by alternative explanations of why a harmful action comes to mind. In the first study (Study 3a), we compare harmful actions to morally neutral actions, and in a second study (Study 3b), we compare neutral actions to actions that are bizarre and unhelpful but not harmful. A third study (Study 3c) offers a combined replication of the Studies 3a and 3b.

General Procedure in Study 3

Participants were asked to read a short description of people playing an "escape room" game. In an escape room game, participants use various objects and clues to crack codes and solve puzzles in an attempt to exit the room as quickly as possible. In the scenario presented, one of the players named Kat finds an object and asks the other player named Max what they should do with it. The object in question is either a "safe" object that is not

usually associated with actions that would harm others (i.e., screwdriver, pencil, string of lights) or a “dangerous” object that is more commonly associated with actions that would harm others (i.e., knife, gun, whip). Max has an immediate first thought about what to do with the object but immediately rejects the idea and tells Kat that he cannot think of anything. The thought is either morally neutral (in Studies 3a-c), immoral (in Studies 3a and 3c), or bizarre (Studies 3b and 3c).

For example, participants might read the following vignette containing an immoral thought with a safe object.

> *During the game, one of the players, Kat, found a screwdriver, but she wasn’t sure what they could do with it, so she asked another player named Max if he had any ideas. The first thing that Max thought was, "I better make sure not to stab my teammate with this screwdriver." Max immediately knew he wasn't going to do this and told Kat that he couldn't think of anything.*
>
> *In the end, they used the screwdriver to release a nail to gain access to a special box.*

Other participants viewed vignettes where the object presented to Max was inherently dangerous (e.g., a knife) and Max again had the same kind of thought (about not stabbing someone with the knife). In other conditions, the thought that came to Max’s mind was instead morally-neutral (not cutting anything important with the knife) and in other conditions, the thought was odd but not a morally wrong thing to do (not cutting his own hair with this knife). We used a total of three different safe objects (pencil, lights, screwdriver) and three different dangerous objects (gun, whip, knife), each of which prompted object-specific thoughts that were either immoral, morally neutral, or bizarre.

We then asked participants to rate Max’s moral character on a sliding scale from 0 (morally bad) to 100 (morally good). Participants were asked to explain their answer in a textbox. Before being debriefed, participants in Study 3a and 3c completed a short demographic questionnaire, asking them about self-reported age, gender, race, religiosity, and political affiliation.

Study 3a: Immoral vs. morally neutral actions

Study 3a tests whether people consider it diagnostic of a person’s moral character that an immoral candidate action came to their mind

spontaneously, even if it was immediately rejected. We predicted that immoral candidate actions would be judged diagnostic of poor moral character when there was not another plausible explanation of why they came to mind, but such an inference should be screened off when another plausible explanation is available. Thus, if Max thinks of a harmful action he could perform with a screwdriver (a safe object uncommonly used for stabbing), his moral character would be judged more negatively than if he thinks of a harmful action he could perform with a knife (a dangerous object more commonly used for stabbing). This design allows us to hold the thought constant while manipulating the context, thereby getting around the possibility that people may judge Max to have a poor moral character simply because he considered the possibility of harming someone. In this first study, we compare cases of an immoral thought coming to mind to cases where a more morally neutral thought comes to mind for both safe and dangerous objects.

Participants

English speaking participants ($N = 401$) located in the United States were recruited via Amazon Mechanical Turk. Our sample consisted of 223 female participants, 176 male, and 1 non-binary, with a mean age of 36.26 ($SD = 12.58$).

Results and interim discussion

Our central prediction concerns how participants' moral character evaluations of Max vary as a function of both the kind of thought that came to mind (neutral vs. immoral) and whether the object that prompted the thought changes how common it would be to think of an action that could harm someone (*i.e.*, a safe vs. dangerous item). Statistically, we subjected participants' moral character assessments to an ANOVA that included the kind of thought Max had, the kind of item that prompted the thought, and the interaction between them. This analysis revealed a marginally-significant interaction effect between thought-type (neutral vs. immoral) and item (safe vs. dangerous), $F(1,393) = 3.063$, $p = .081$. In the context of this interaction, we also found a main effect of thought-type (neutral vs. immoral) such that agents with neutral thoughts were judged to have better moral character than agents with immoral thoughts, $F(1, 393) = 6.950$, $p = .009$, as well as a main effect of item (safe vs. dangerous) such that agents interacting with dangerous objects were judged to have better moral

character than agents interacting with safe objects, $F(1, 393) = 34.876$, $p < .001$.

A series of t-tests further examined the effects of thought-type for each type of object. Critically, when comparing neutral and immoral thoughts with safe objects, we found that agents with neutral thoughts are judged to be more moral than agents with immoral thoughts, $t(195) = -3.07$, $p = .002$. However, we did not observe a significant difference between the immoral and neutral thought when the object was a dangerous one, $t(198) = -.630$, $p = .592$. Taken together, these findings suggest that in cases in which harmful actions have a low probability of coming to mind (e.g.., when asked what to do with a typically safe object), character inferences were made based on the morality of the thoughts, whereas in cases in which harmful actions have a high probability of coming to mind (*e.g..*, when asked what to do with a typically dangerous object), no such moral character inference is made.

An additional but unpredicted pattern was that in the neutral thought conditions, we found a significant effect of object kind, such that agents who have neutral thoughts about dangerous objects ($M = 69.583$, $SD = 20.759$) were considered considerably more moral than agents who had neutral thoughts about safe objects ($M = 60.766$, $SD = 19.192$), $t(188) = 3.038$, $p = .003$.

In summary, we found evidence that negative moral character inferences only occurred when a harmful action came to mind that would be typically uncommon to consider (*i.e.*, the action was prompted by a “safe” object); we did not see similar negative character inferences when harmful actions came to mind but would be common to consider given the object that prompted the thought. In other words, we found evidence for the same combination of diagnosticity and screening off in inferences about others’ *moral* values.

We also found that participants saw agents as particularly moral when they had only neutral thoughts despite their thoughts being prompted by dangerous objects. One way to make sense of this finding is that when prompted by a dangerous object, the probability of thinking of an action that harms other is relatively high, and so when no such thought arises, the *absence* of such a thought may be attributed to the agent placing especially low value on actions that harm others---they are the kind of person who would never even consider such an idea. We continue to explore this effect in the following studies.

Study 3b: Bizarre vs. neutral actions

While the results of the prior study provides some initial evidence that people are making moral character inferences based on the options that come to others' minds, a reasonable alternative explanation for this finding is that the observed difference in assessments of Max's character were not based on the morality of the actions that Max considered, but were instead based on perceptions of whether Max was actually trying to help his teammates by thinking of useful options for the specific problem they faced (the escape room game). It is particularly clear that Max's thoughts are unhelpful when there is no external reason to consider doing an immoral action, but Max considers this anyway.

We further investigate this alternative explanation by asking a new group of participants to make moral character assessments of Max after Max has considered and rejected actions that are obviously not useful in the context, but which are not clearly morally wrong. We again compare moral character evaluations of Max in these cases to those in which Max considers non-bizarre and morally neutral actions. If the prior findings were due to the morality of Max's thoughts, then we would not expect to find a difference in moral character evaluations between bizarre and neutral thoughts; if they are instead due to the perceived differences in Max's cooperativeness, we should see a similar differences in character evaluations between bizarre and neutral thoughts.

Participants

English speakers located in the United States were recruited via Amazon Mechanical Turk ($N = 401$).

Results and interim discussion

We analyzed participants' moral character judgements with an ANOVA containing the type of thought Max had (neutral vs bizarre) and the type of item that prompted the thought (safe vs dangerous). This analysis revealed a significant main effect of thought-type, such that an agent with neutral thoughts was considered slightly more moral than an agent with bizarre thoughts, $F(1, 393) = 3.888$, $p = .049$, but no main effect of item-type, $F(1, 393) = .008$, $p = .930$, and no significant interaction between thought-type and item-type, $F(1, 393) = 2.442$, $p = .119$.

Importantly, the observed main effect of thought-type was again driven by positive character judgments for agents with neutral thoughts when confronted with a dangerous object. Similar to the pattern previously

observed, participants viewed agents as particularly moral when they had a neutral thought when presented with an object that is frequently associated with harm (neutral vs. bizarre, $t(189) = -2.513$, $p = .013$). However, more critically for our primary question in this study, when examining the difference between neutral and bizarre thoughts in the safe object conditions, we did not find an effect of thought-type, $t(204) = -0.335$, $p = .738$.

This set of analyses investigated whether the previously observed effect could instead be explained by perceived differences in whether the agent was attempting to be helpful in the cooperative game he was involved in. Our results revealed that, when prompted by non-dangerous items, agents who thought of particularly useless actions were *not* judged to be morally worse, suggesting that the previously observed effect cannot simply be attributed to participants judging Max as immoral for having abnormal and unhelpful thoughts.

Study 3c: Combined replication

In this third study, we combined aspects of the prior two studies, now including immoral, bizarre, and neutral thoughts, but otherwise directly replicated the prior procedures. In addition to providing a direct replication of the effects observed in Study 3a and 3b, this design also allows us to more directly compare moral character assessments based on having immoral vs. simply bizarre actions come to mind. If participants were making moral character inferences about Max based on what the thoughts that come to his mind reveal about his underlying values, then we should expect them to infer that Max has a poor moral character specifically when an immoral action comes to Max's mind and this inference is not screened off.

Participants

English speaking participants ($N = 810$) located in the United States were recruited via Amazon Mechanical Turk. Our sample consisted of 357 female participants, 445 male, and 8 non-binary or 'other', with a mean age of 38.77 ($SD = 12.04$).

Results

As in Studies 3a and 3b, we conducted an ANOVA on participants' moral character assessments using the type of item that prompted the thought (safe or dangerous) and the type of thought Max had (immoral,

neutral, or bizarre). This analysis revealed a significant interaction between thought-type and item-type, $F(2, 804) = 8.744$, $p < .001$ (Figure 4). In the context of this interaction, we also observed a main effect of thought-type, $F(2, 804) = 4.963$, $p = .007$, such that neutral thoughts ($M = 66.05$, $SD = 20.644$) were judged to be more moral than bizarre thoughts ($M = 63.25$, $SD = 20.448$), which in turn were considered more moral than immoral thoughts ($M = 60.22$, $SD = 25.017$). Additionally, we observed a main effect of item-type, $F(1, 804) = 28.364$, $p < .001$, such that agents whose thoughts were prompted by dangerous items were judged to be more moral overall ($M = 67.080$, , $SD = 20.476$) than agents whose thoughts were prompted by safe items ($M = 58.970$, $SD = 23.413$).

Finally, we again observed that participants who had neutral thoughts in response to dangerous objects were judged to have a better moral character than agents who had neutral thoughts in response to safe objects, $t(260) = -3.055$, $p = .002$.

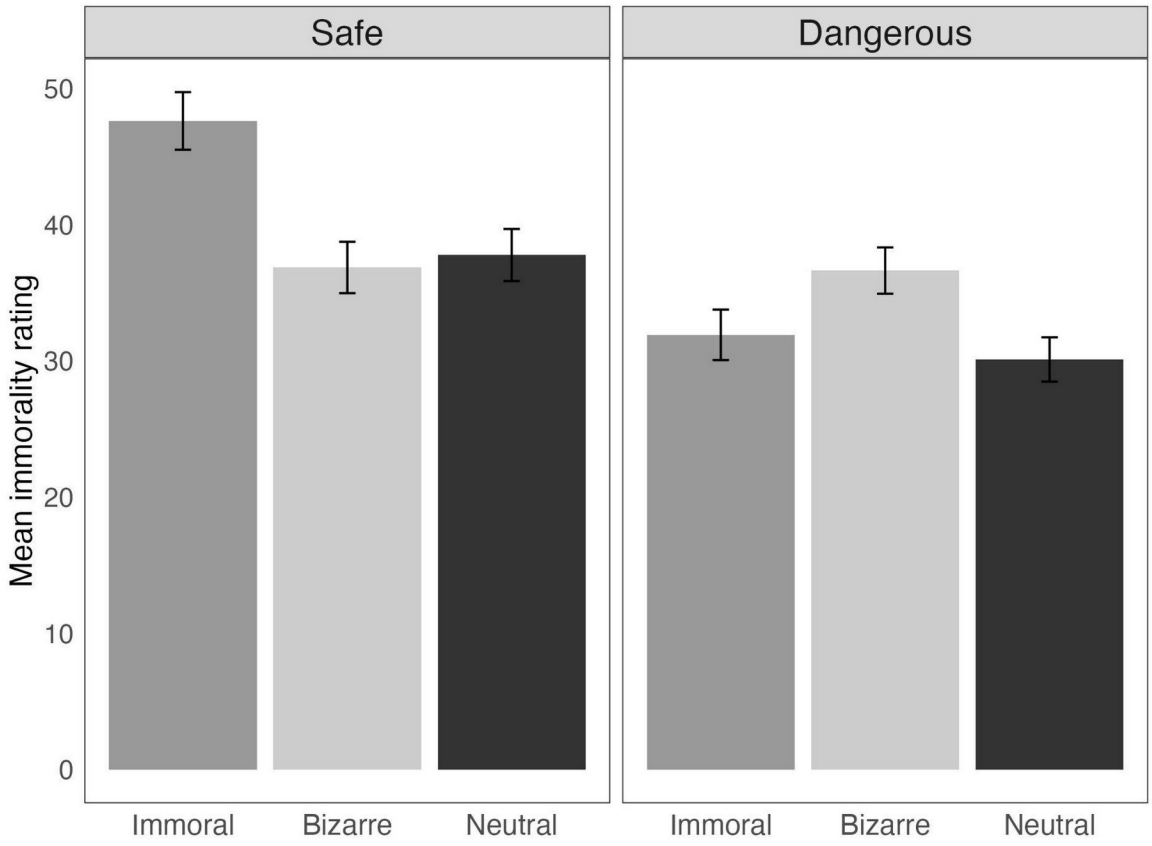

Figure 4. *Barplot of participants' mean moral character judgements of agents based on the presence of an immoral (medium-gray), bizarre (light-gray), or neutral (dark-gray) thought upon encountering an object*

infrequently associated with harm (a "safe" object) or an object frequently associated with harm (a "dangerous" object). Error bars depict +/- 1 SEM.

Study 3 Discussion

Building on the idea that people have an intuitive model of the option generation process, we hypothesized that moral character inferences based on what comes to mind should also combine information about how common it would be to consider a given option, the value of that action in the specific context, and the general value that an agent may place on that action. More specifically, we predicted that when immoral thoughts come to mind, people will infer that the person who had this thought may have a negative moral character when the thought is neither a good option nor a common thing to consider in the decision context.

We found evidence for the same combination of diagnosticity and screening off observed in Studies 1 and 2: people specifically do not infer negative moral character based on immoral thoughts when such thoughts are common in the decision context. Critically however, when the context cannot explain the presence of an immoral thought, people infer that the agent must have an immoral character. Taken together, we find evidence that merely thinking of an action that causes harm, despite immediately and explicitly rejecting the action, leads to an inference of underlying immoral values that may be guiding others' option-generation process.

Study 4: Inverse Values Moral Judgment Study

In this final study, we explore whether the previously observed inferences about others' moral character based on what comes to their mind are sensitive not only to the presence of harmful thoughts, but also the *ease* with which such thoughts come to mind. Similar to the logic of Study 2, the question in this study concerns whether participants' value inferences are sensitive not only to the output of the option-generation process, but also to the way in which that process operates. More specifically, we consider a comparison between moral judgments of two people who both end up considering and doing similarly harmful actions, but for one of them, thinking of such actions comes easily, while for the other, thinking of them proves to be more difficult. Given the role of implicit values in calling options to mind, we hypothesized that participants would infer morally worse values for the person who can more easily think of harmful actions.

As a particularly stringent test, we explore this effect in the context of human rights workers conducting a training session that teaches targeted

populations how to effectively handle hate speech. The key idea is that in such a context, it is actually ideal for those leading the training to be able to generate harmful speech easily and quickly. However, based on the findings in the prior studies, it also seems that being able to do so should also lead participants to infer that the person may have morally suspect values. Put together, we predict that in such cases participants may exhibit a kind of paradoxical pattern in their moral judgments, whereby the better they judge the human rights worker to be at training targeted populations, the worse they judge their underlying moral character to be. If found, such a pattern would provide strong evidence that the kind of value inferences participants are making concerns latent moral values (those that drive the option-generation process) and not the explicit moral values (those that guide which actions are actually chosen).

Methods

Participants

English speaking participants ($N = 398$) located in the United States were recruited via Amazon Mechanical Turk. Our sample consisted of 195 female participants, 200 male, 2 non-binary, and 1 ‘other’, with a mean age of 40.98 ($SD = 28.84$).

Procedure

Participants read one of four vignettes about two people named John and Sarah. John and Sarah work for a human rights organization that trains volunteers for difficult situations involving hate speech. The four scenarios involve John and Sarah training (1) Jewish volunteers facing a group of neo-Nazi protesters, (2) volunteers protesting an organization of white supremacists, (3) spies learning to withstand enemy torture, and (4) Pride March participants facing an anti-LGBTQ group. In all four situations John and Sarah insult the trainees in order to prepare them for the upcoming encounter with the extreme group. In each condition, either John or Sarah easily generates hateful slurs and insults to yell at the trainees, while the other must rely on a reference sheet of suggested insults.

After reading the scenario about the training, participants were asked to make a forced-choice judgment of whether John or Sarah (1) did a better job training the trainees and (2) is a stronger advocate of human rights. They were additionally asked to explain their decisions in an open-ended textbox. Finally, participants completed a short demographic questionnaire,

asking them about self-reported age, gender, race, religiosity, and political affiliation, and then were debriefed.

Results and discussion

We recoded each response as selecting the trainer who more easily thought of harmful actions or the agent who had more difficulty thinking of harmful actions, and used a generalized linear mixed effects model (with a random intercept and slope by scenario) to ask whether the kind of trainer that was selected for differed significantly for the two questions: (1) who was a better trainer and (2) who was a stronger advocate for human rights. We found that it did, z-ratio = -7.927, $p < .001$. Specifically, for the trainer who was easily able to generate harmful actions, participants tended to think this trainer did a better job at training (chosen on 83.41% of trials) but also tended to think this agent was also a worse advocate for human rights (chosen on only 27.64% of trials). Correspondingly, the trainer who had difficulty generating harmful actions was judged to be a better advocate for human rights (chosen on 72.36% of trials), but was also judged to be worse at the training (chosen on only 16.58% of trials). This finding, along with the patterns observed in Studies 3a-c, demonstrates that not only can people infer others' moral values based on the options that come to their mind, but also based on *how* those options are generated.

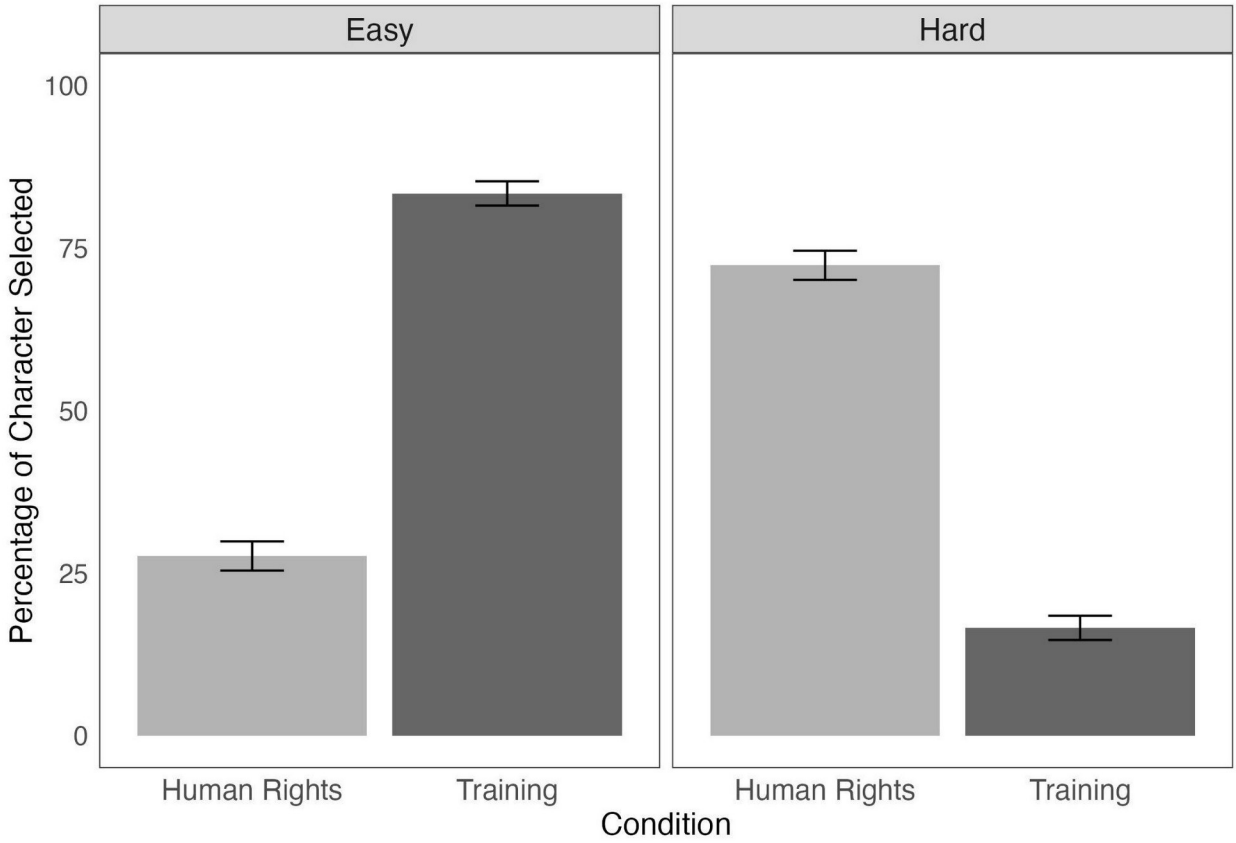

Figure 5. *Barplot of the percentage of times each agent was selected as the stronger advocate of human rights (light-gray) or better trainer (dark-gray) when harmful actions came to mind with ease ("Easy") or with difficulty ("Hard"). Error bars depict +/- 1 SEM.*

General Discussion

Building on work demonstrating that options generated in the two-stage decision making model are informed by subjective value and commonality, we ask whether people have an intuitive theory of the decision-making process that allows them to infer others' subjective values from the thoughts that come to mind. We find that people infer others' subjective values based on the options generated in open-ended decision making contexts (Study 1) and the order in which options came to mind (Study 2). This also informs their subsequent moral character evaluations (Study 3), even when the immoral options that come to mind are optimal given task-specific demands (Study 4). Critically, across these studies, we also find that participants' inferences about others' values based on what comes to mind follow the core principles of diagnosticity and screening off:

participants only make strong inferences about others' value when the consideration of these options cannot be otherwise explained away.

Our work sheds light on the underlying cognitive processes involved in person-perception by demonstrating that subjective value and character inferences are informed by the options that come to mind, perhaps through an inversion of an intuitive understanding of the option-generation process itself. Of course, our study provides only indirect evidence regarding the precise mechanism by which people infer a person's values from the options that come to mind. Consistent with this interpretation, people not only treat what comes to mind as diagnostic of values, but they also screen off its diagnostic value when other alternative explanations are plausible. An alternative way one may be tempted to explain these inferences is by proposing that people have learned a statistical regularity between the options that come to a person's mind (in particular kinds of contexts) and that person's values. While this sort of general statistical learning occurs in many domains (Frost et al., 2015), there is reason to doubt it offers a complete explanation in this case because we rarely get direct, observable information about the *thoughts* that come to others' minds. Without such data, learning the nuanced set of inferences we observed in our data would prove quite difficult. Rather, it seems more plausible that participants have an intuitive theory of how options are generated from their first-person experience with option generation, and are able to infer how such a process would likely need to be altered to generate particular options. In this way, our proposal is more somewhat aligned with simulation theory within theory of mind (for more on the historical debate between simulation theory and theory-theory, see Gordon, 1986; Goldman, 1989; Gopnik & Wellman, 1992; Gopnik & Meltzoff, 1997; Nichols & Stitch, 2003; Saxe, 2005). It will be important for future studies to target the developmental processes that give rise to the pattern of inferences we demonstrate here.

Our findings also raise new questions about the degree of correspondence between actual option generation and the folk theory of it. For example, prior work demonstrates that the initial options people generate are sampled from a relatively local region of semantic space (Srinivasan, Acierno, & Phillips, 2022; He, Richie, & Bhatia, 2024). Do people draw inferences about each other's semantic representations from information about what comes to mind? For instance, suppose that two people are trying to think of US presidents and one generates the sequence "Washington, Regan, Lincoln" while the other generates the sequence

“Nixon, Regan, Hoover”; would people infer different semantic representations of Regan in these two cases?

In the current work, we present the options considered during decision making, including the action ultimately selected by the agent. By ensuring that the selected action is not morally objectionable, we are able to isolate how merely considering certain options impacts character inference. However, future research could explore how the morality of both considered and selected options influences character judgments. For example, informing participants that an agent considered both moral and immoral options and then chose one or the other may help us explore these factors’ independent impact on character judgments. Given the well-documented negativity bias in moral character evaluations, such that immoral behavior disproportionately influences character judgments (Reeder & Coovert, 1986; Skowronski & Carlston, 1987; Riskey & Birnbaum, 1974; Mende-Siedlecki, Baron, & Todorov, 2013), the presence of immoral options will likely have a particularly negative impact on character evaluations. Finally, the present work also holds potential significance for legal proceedings as they suggest that jurors’ beliefs that a defendant even contemplated immoral activities could lead jurors to make character judgments that influence further inferences, e.g., about intent.

Conclusion

Building on work in decision making and option generation, we show that ordinary people use information about what options come to a person’s mind in order to infer their values. Consistent with conceptualizations of theory of mind as a form of inverse reinforcement learning (Jara-Ettinger, 2019), we show that people treat “what comes to mind” as especially diagnostic of a person's values when other candidate explanations are absent, draw inferences based not only on what does come to mind but also what does not, and accurately assume that more highly valued items are considered first. These findings suggest that people possess a sophisticated folk theory of option generation, opening new avenues for research in the structure and development of theory of mind.

Data availability statement: All data is publicly available in the following anonymized repository: https://osf.io/3dj2u/?view_only=be5a92d3f9754777948201c2cf4e4005.

References

Acierno, J., Mischel, S., & Phillips, J. (2022). Moral judgments rely on default representations of possibility, *Philosophical Transactions B.*

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition, 113*(3), 329-349. https://doi.org/10.1016/j.cognition.2009.07.005

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., Dai, B., Grothendieck, G., Green, P. & Bolker, M. B. (2015). Package 'lme4'. *Convergence* 12, 2.

Bear, A., Bensinger, S., Jara-Ettinger, J., Knobe, J., & Cushman, F. (2020). What comes to mind?. *Cognition, 194*, 104057. https://doi.org/10.1016/j.cognition.2019.104057

Critcher, C. R., Inbar, Y., & Pizarro, D. A. (2013). How Quick Decisions Illuminate Moral Character. *Social Psychological and Personality Science, 4*(3), 308-315. https://doi.org/10.1177/1948550612457688

Crockett, M. J. (2016). How formal models can illuminate mechanisms of moral judgment and decision making. *Current Directions in Psychological Science*, *25*(2), 85-90. Chicago

Crockett, M. J., Kim, J. S., & Shin, Y. S. (2024). Intuitive Theories and the Cultural Evolution of Morality. *Current Directions in Psychological Science*, 09637214241245412.

Frost, R., Armstrong, B. C., Siegelman, N., & Christiansen, M. H. (2015). Domain generality versus modality specificity: The paradox of statistical learning. *Trends in Cognitive Sciences*, 19(3), 117-125.

Gates, V., Callaway, F., Ho, M. K., & Griffiths, T. L. (2021). A rational model of people's inferences about others' preferences based on response times. *Cognition*, *217*, 104885. https://doi.org/10.1016/j.cognition.2021.104885

Gershman S. J., Gerstenberg T., Baker C. L., Cushman F. (2016). *Plans, Habits, and Theory of Mind. PLOS ONE 11*(9): e0162246. https://doi.org/10.1371/journal.pone.0162246

Goldman, A.I. (1989). Interpretation psychologized. *Mind & Language, 4, 161-185. https://doi.org/10.1111/j.1468-0017.1989.tb00249.x*

Gopnik, A., & Wellman, H. M. (1992). Why the child's theory of mind really *is* a theory. *Mind & Language, 7*(1-2), 145–171. https://doi.org/10.1111/j.1468-0017.1992.tb00202.x

Gopnik, A., & Meltzoff, A. N. (1997). *Words, Thoughts, and Theories*. The MIT Press. https://doi.org/10.7551/mitpress/7289.001.0001

Gordon, R. M. (1986). Folk Psychology as Simulation. *Mind & Language, 1*(2), 158-171. https://doi.org/10.1111/j.1468-0017.1986.tb00324.x

Hauser, J. R. (2014). Consideration-set heuristics. *Journal of Business Research, 67*(8), 1688–1699.

He, L., Richie, R., & Bhatia, S. (2024). Limitations to optimal search in naturalistic active learning. *Journal of experimental psychology. General*, *153*(5), 1165–1188. https://doi.org/10.1037/xge0001558

Jara-Ettinger, J. (2019). Theory of mind as inverse reinforcement learning. *Current Opinion in Behavioral Sciences, 29*, 105-110. https://doi.org/10.1016/j.cobeha.2019.04.010

Johnson, J. G., & Raab, M. (2003). Take The First: Option-generation and resulting choices. *Organizational Behavior and Human Decision Processes, 91*(2), 215-229. https://doi.org/10.1016/S0749-5978(03)00027-X

Kaiser, S., Simon, J. J., Kalis, A., Schweizer, S., Tobler, P. N., & Mojzisch, A. (2013).The cognitive and neural basis of option generation and subsequent choice. *Cognitive, Affective, & Behavioral Neuroscience,13*(4), 814–829.

Kalis, A., Kaiser, S., & Mojzisch, A. (2013). Why we should talk about option generation in decision-making research. *Frontiers in psychology*, *4*, 555.

Kleiman-Weiner, M., Saxe, R., & Tenenbaum, J. B. (2017). Learning a commonsense moral theory. Cognition, 167, 107-123.

Klein, G., Wolf, S., Militello, L., & Zsambok, C. (1995). Characteristics of skilledoption generation in chess. Organizational Behavior and Human DecisionProcesses, 62 (1), 63–69.

Mende-Siedlecki, P., Baron, S. G., & Todorov, A. (2013). Diagnostic value underlies asymmetric updating of impressions in the morality and ability domains. The Journal of neuroscience : the official journal of the Society for Neuroscience, 33(50), 19406–19415. https://doi.org/10.1523/JNEUROSCI.2334-13.2013

Mills, T., & Phillips, J. (2023). Locating what comes to mind in empirically derived representational spaces. *Cognition, 240*, 105549. https://doi.org/10.1016/j.cognition.2023.105549

Morewedge, C. K., Giblin, C. E., & Norton, M. I. (2014). The (perceived) meaning of spontaneous thoughts. *Journal of Experimental Psychology: General, 143*(4), 1742–1754. https://doi.org/10.1037/a0036775

Morris, A., Phillips, J., Huang, K., & Cushman, F. (2021). Generating options and choosing between them depend on distinct forms of value representation. *Psychological science, 32*(11), 1731-1746.

Nichols, S., & Stich, S. P. (2003). *Mindreading: An integrated account of pretence, self-awareness, and understanding other minds.* Clarendon Press/Oxford University Press. https://doi.org/10.1093/0198236107.001.0001

Phillips, J., & Cushman, F. (2017). Morality constrains the default representation of what is possible. *Proceedings of the National Academy of Sciences, 114*(18), 4649-4654. https://doi.org/10.1073/pnas.1619717114

Phillips, J., & Knobe, J. (2018). The psychological representation of modality. *Mind & Language, 33*(1), 65-94. https://doi.org/10.1111/mila.12165

Phillips, J., Luguri, J. B., & Knobe, J. (2015). Unifying morality's influence on non-moral judgments: The relevance of alternative possibilities. *Cognition, 145*, 30–42. https://doi.org/10.1016/j.cognition.2015.08.001

Phillips, J., Morris, A., & Cushman, F. (2019). How we know what not to think. *Trends in Cognitive Sciences, 23*(12), 1026–1040. https://doi.org/10.1016/j.tics.2019.09.007

Pizarro, D., Uhlmann, E., & Salovey, P. (2003). Asymmetry in Judgments of Moral Blame and Praise: The Role of Perceived Metadesires. *Psychological Science, 14*(3), 267-272. https://doi.org/10.1111/1467-9280.03433

Reeder, G. D., & Coovert, M. D. (1986). Revising an impression of morality. *Social Cognition, 4*(1), 1–17. https://doi.org/10.1521/soco.1986.4.1.1

Riskey, D. R., & Birnbaum, M. H. (1974). Compensatory effects in moral judgment: Two rights don't make up for a wrong. *Journal of Experimental Psychology, 103*(1), 171–173. https://doi.org/10.1037/h0036892

Saxe, R. (2005). Against simulation: the argument from error. *Trends in cognitive sciences*, *9*(4), 174–179. https://doi.org/10.1016/j.tics.2005.01.012

Skowronski, J. J., & Carlston, D. E. (1987). Social judgment and social memory: The role of cue diagnosticity in negativity, positivity, and extremity biases. *Journal of Personality and Social Psychology, 52*(4), 689–699. https://doi.org/10.1037/0022-3514.52.4.689

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

Wang, F., Aka, A., He, L., & Bhatia, S. (2024). Memory modeling of counterfactual generation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Advance online publication. https://dx.doi.org/10.1037/xlm0001335