

The Neural Instantiation of Spontaneous Counterfactual Thought

Regan M. Bernhard¹, Fiery Cushman², Alara Cameron (Jessey Wright)³, and Jonathan Phillips⁴

¹ Department of Psychology and Neuroscience, Boston College

² Department of Psychology, Harvard University

³ Yellowknife, Northwest Territories, Canada

⁴ Program in Cognitive Science, Dartmouth College

Many of the most interesting cognitive feats that humans perform require us to consider not just the things that *actually occur* but also *alternative possibilities*. We often do this explicitly (e.g., when imagining precisely how a first date could have gone better), but other times we do it spontaneously and implicitly (e.g., when thinking, “I have to catch this bus,” implying bad alternatives if the bus is not caught). A growing body of research has identified a core set of neural processes involved in explicit, episodic counterfactual thinking. Little is known, however, about the processes supporting the spontaneous, possibly implicit representation of alternatives. To make progress on this question, we induced participants to spontaneously generate counterfactual alternatives by asking them to judge whether agents were forced to make a particular choice or chose freely—a judgment that implicitly depends on their alternative options. Using functional magnetic resonance imaging, we found 14 clusters that were preferentially engaged when participants were making force judgments (which elicit the spontaneous consideration of alternatives) compared to judgments of what actually occurred (which do not elicit alternatives). These clusters were widely distributed throughout the brain, including in the bilateral prefrontal cortex, bilateral inferior parietal lobule, bilateral middle and inferior temporal gyri, bilateral posterior cingulate, and bilateral caudate. In many of these regions, we additionally show that variability in the neural signal correlates with trial-by-trial variability in participants’ force judgments. Our findings provide a first characterization of the neural substrates of the spontaneous representation of counterfactual alternatives.

Public Significance Statement

Many of the most interesting cognitive feats that humans perform require us to consider not just the things that *actually occur* but also *alternative possibilities*. We often do this explicitly (e.g., when imagining precisely how a first date could have gone better), but other times we do it spontaneously and implicitly (e.g., when thinking, “I have to catch this bus,” implying bad alternatives if the bus is not caught). Little is known about how the brain engages in this type of spontaneous and often implicit counterfactual thought. In this study, we use functional neuroimaging to identify a set of brain regions that are preferentially engaged when spontaneously considering counterfactual alternatives.

Keywords: counterfactuals, force judgments, alternative possibilities

Supplemental materials: <https://doi.org/10.1037/xge0001676.supp>

This article was published Online First November 7, 2024.
Michele Diaz served as action editor.

Regan M. Bernhard  <https://orcid.org/0000-0002-8828-3046>

Preprints of this article are available on PsyArXiv at <https://osf.io/preprints/psyarxiv/zdf2w>. In addition, this work was shared in a virtual poster at the 2022 Annual Conference of the Social Affective Neuroscience Society.

The authors have no known conflicts of interest to disclose. This work was funded by grants from the John Templeton Foundation through the Summer Seminars in Neuroscience and Philosophy, Harvard University Mind Brain and Behavior Postdoctoral Fellow Research Award, and Harvard University Harvard Brain Science Initiative Young Scientist Transitions Award to Regan M. Bernhard.

Regan M. Bernhard played a lead role in data curation, formal analysis,

investigation, project administration, visualization, writing—original draft, and writing—review and editing and an equal role in conceptualization, funding acquisition, methodology, and resources. Fiery Cushman played a supporting role in formal analysis, funding acquisition, and writing—review and editing and an equal role in conceptualization, methodology, and supervision. Alara Cameron (Jessey Wright) played a supporting role in conceptualization, methodology, and resources. Jonathan Phillips played a supporting role in formal analysis, supervision, writing—original draft, and writing—review and editing and an equal role in conceptualization, funding acquisition, and methodology.

Correspondence concerning this article should be addressed to Regan M. Bernhard, Department of Psychology and Neuroscience, Boston College, McGuinn 430A, 140 Commonwealth Avenue, Chestnut Hill, MA 02467, United States. Email: regan.bernhard@gmail.com

Many of our most interesting cognitive feats are grounded not just in representations of what *actually* occurred, but in comparison to *alternative possibilities*. Consider, for example, the case of “duty-to-retreat” laws. In states with duty-to-retreat laws, you can kill a person who attacks you only if you have no alternative means of escape. In other words, consistent with many philosophical theories (Aquinas, 1273/1920; Aristotle, 340 BCE/2002; Locke, 1690/1975), legal (and moral) responsibility depends not just on what you did, but what you might have done alternatively (Phillips & Knobe, 2009; Woolfolk et al., 2006).

Although we may deliberate explicitly about counterfactuals in the courtroom—or when daydreaming, ruminating, or Monday-morning quarterbacking—we also seem to make judgments like this quite effortlessly and spontaneously. In fact, in many cases, people probably fail to notice the way that their thought and language depend on counterfactual representations. For instance, alternative possibilities play a role not only in determining responsibility but also in causal judgments: When deciding whether a match caused a fire, people spontaneously consider what *would have happened* if the match had never been lit (Gerstenberg et al., 2017; Kominsky et al., 2015). Yet presumably a person who remarks “the match made the candle ignite” could easily fail to notice the important work of a counterfactual in the background. Once cognitive scientists trained their eyes to discern the lurking presence of counterfactuals, however, they began to find them everywhere. Representations of nonactual events have been shown or argued to play a role in linguistic meaning and communication (Kratzer, 2012), conditional reasoning (Stalnaker, 1968), and many other aspects of high-level cognition and social judgments (Byrne, 2017; Phillips et al., 2015).

Despite the ubiquitous importance of alternative possibilities in our reasoning, judgment, and decision making, relatively little is known about how such thought is instantiated in the brain during judgment and decision making. Moreover, what little we do know is from work involving the explicit representation of counterfactuals (as in courtrooms), rather than the implicit counterfactual representations that shape our reasoning and language from moment to moment. Our focus is on these spontaneous, possibly implicit representations. We use functional neuroimaging to investigate the neural activation associated with the spontaneous representation of counterfactual alternatives.

Explicit Counterfactual Simulation in the Brain

An important and growing body of research has identified a core set of neural processes involved in explicit, episodic counterfactual thinking—that is, deliberative simulation of nonactual events, whether in the past or the future (Barbey et al., 2009; De Brigard et al., 2013, 2015, 2017; De Brigard & Parikh, 2019; Faul et al., 2020; Gomez Beldarrain et al., 2005; Kulakova et al., 2013; Nieuwland, 2012; Parikh et al., 2018; Schacter et al., 2015; St. Jacques et al., 2018; Urrutia et al., 2012; Van Hoeck et al., 2015). In much of this work, individuals are encouraged to reimagine past events in a way that produces an alternative outcome (De Brigard et al., 2013, 2015; St. Jacques et al., 2018; Urrutia et al., 2012; Van Hoeck et al., 2015). This work has found that deliberative, episodic counterfactual thought typically engages a core network often implicated in other types of mental simulation, including regions in the medial frontal and temporal lobes, the posterior cingulate cortex, precuneus, and the lateral parietal and temporal lobes (Benoit & Schacter, 2015;

De Brigard et al., 2013, 2015, 2017; Khoudary et al., 2022; St. Jacques et al., 2018; Van Hoeck et al., 2015). For example, this network of regions is activated both when individuals are asked to accurately remember past events and to reimagine them as having gone differently (De Brigard et al., 2013). Likewise, this network is preferentially engaged both when individuals imagine how events could have gone differently in the past or could go differently in the future (Benoit & Schacter, 2015; Van Hoeck et al., 2013). While activation in these regions is modulated somewhat by whether counterfactual simulation is self- versus other-directed (Barbey et al., 2009; De Brigard et al., 2015), involves internal or external events (Khoudary et al., 2022), or is episodic versus semantic (Parikh et al., 2018), there is considerable consistency even as these aspects of mental simulations are altered.

Default Representations of Possibilities

Explicit, episodic counterfactual reasoning shows clear behavioral dissociations with the spontaneous, implicit representations of alternative possibilities (Phillips et al., 2019; Phillips & Cushman, 2017). For example, unlike explicit episodic representations of possibilities, the spontaneous consideration of alternatives seems to be strongly biased toward alternatives that are “normal” (Phillips et al., 2019). In one study that demonstrated this effect, Phillips and Cushman (2017) asked participants to make judgments of whether an event was possible, either under time pressure or after reflecting, and found that when participants were not given enough time to explicitly deliberate, they tended to judge that immoral and irrational actions were actually impossible. They were significantly more likely to judge these sorts of “abnormal” events to be possible when given time to explicitly deliberate.

Moreover, this effect was not restricted to judgments of what is “possible.” Instead, it arose for a whole host of judgments in which people must generate possibilities of various kinds: judgments of what an individual could, may, might, should, or ought to do. When under time pressure, all of these judgments began to look more similar, specifically by picking out “normal” possibilities (those that are descriptively common and prescriptively good) and excluding “abnormal ones.” Thus, for instance, under time pressure people judge what an individual *might* do and what he *ought* to do in surprisingly similar ways. Both reflect a shared implicit or default representation of what is possible (Phillips & Cushman, 2017). However, when given time to reflect, these judgments come apart—participants clearly differentiate what should be done from what could be done.

This “default” representation of possibility seems to serve as a common engine for the generation of *alternative possibilities* across the many diverse cognitive operations that likely elicit quick, spontaneous representations of alternatives. Specifically, across cognitive operations thought to rely on spontaneous comparisons to alternative possibilities, one tends to see the hallmarks of normality constraints in the focus on what is descriptively common and what is prescriptively good (Bear & Knobe, 2017; Kominsky et al., 2015). To illustrate with one example, whether an agent is considered to have caused an outcome is influenced by both the morality of the action and how typical it is (Halpern & Hitchcock, 2015; Icard et al., 2017; Kominsky et al., 2015; Kominsky & Phillips, 2019; Phillips et al., 2015). In a similar fashion, violations of social/conventional norms (Uttich & Lombrozo, 2010) and moral norms (Knobe, 2003;

Pettit & Knobe, 2009) have the same impact on whether or not an agent was viewed as having acted intentionally. Violations of normality also affect people's moral judgments (Acierno et al., 2022). Finally, the perceived prudential value, moral value, or social normality of counterfactual alternatives predicts the degree to which agents are perceived as having acted freely (Bernhard et al., 2022; Phillips et al., 2015; Phillips & Knobe, 2009; Young & Phillips, 2011).

The vulnerability of these types of judgments to the normality of counterfactual alternatives suggests that they do not engage an explicit or deliberative simulation of counterfactual alternatives. Rather, they likely rely on the spontaneous representation of alternative possibilities—more similar to those that are revealed under time pressure. One might wonder then, in what ways these two types of representations of alternatives overlap. As the research described above suggests, there are clear behavioral dissociations between spontaneous, default, and explicit, deliberative counterfactual thought. On the other hand, both processes require representing hypothetical past states of the world. Comparing the neural activation associated with each type of counterfactual representation may allow us to gain traction on identifying the shared and distinct processes underlying each type of thought. While there is now a considerable body of research on the neural substrates of episodic counterfactual simulation, there is very little known about how spontaneous counterfactual thought is instantiated in the brain. We aim to fill this gap.

The Present Research

In order to study the spontaneous neural representations of alternative possibilities, it is necessary to have a task that reliably evokes them. Building on extensive prior research, we used a task in which people must judge an agent to have been *forced* or *free* to perform various actions. Intuitively, a person is forced to perform an action if they have no suitable alternatives, whereas they are free to perform it (or not) if suitable alternatives are available. Many prior studies show that ordinary people make judgments of force and freedom in precisely this way: by spontaneously considering alternative possibilities (Bernhard et al., 2022; Phillips et al., 2015; Phillips & Knobe, 2009; Young & Phillips, 2011). Moreover, the specific alternative possibilities that they consider are influenced by normality—that is, by what is descriptively common and by what is prescriptively good (Bernhard et al., 2022; Phillips & Cushman, 2017). In other words, people are judged “free” to choose an action as long as relatively common and good alternatives to that action exist. This provides reason to think that the alternative possibilities evoked by making judgments of force and freedom are precisely the default, spontaneous representations that we intend to study.

The next challenge is to ensure that the neural representations we identify during functional neuroimaging are just the ones we intend: representations of alternative possibilities and not unintended or spurious correlates. We address this challenge in two ways. First, at a coarse level, we contrast neural activation associated with judgments of what an agent was *forced* to do with activation associated with judgments of what an agent *actually* did. For instance, imagine that a traveler to a desert island has no more room in her pack and must remove one of three objects: a water filter, a teddy bear, or nail polish (see Table 1).

Table 1
Possible Options When Considering What to Remove From Pack

Decision	Option
The traveler must remove one item from her pack	<ul style="list-style-type: none">• Water filter• Teddy bear• Nail polish

Suppose that the traveler removes the teddy bear. Was the traveler *forced* to remove the teddy bear? Prior work suggests that people answer this question by spontaneously considering that she had the *alternative possibility* of removing the nail polish (and thus was not forced). In contrast, if we asked if the traveler *actually* removed the teddy bear, this question can be answered without representing any counterfactual state of affairs.

Of course, judgments of force versus judgments of actual behavior differ in other ways as well, and so we complemented this approach with a more fine-grained analysis. Consistent with prior research, we show that participants' force judgments are, on average, sensitive to the availability of good alternatives: People tend to judge that a person is more “forced” to act a certain way if they have no good alternative and less forced if a good alternative is available. However, there is considerable variability, on a trial-by-trial level, in the force judgments that people make. Therefore, we model the trial-by-trial residual deviations of the blood oxygen level dependent response across all force trials using participants' behavioral ratings of whether the agent was forced as a predictor. Evidence for such a relationship would indicate a tight coupling between a psychological model of how alternative options influence force judgments, behavior, and neural activation.

Our overall strategy, then, is to use the univariate contrast between judgments of force versus judgments of actual behavior to identify candidate neural substrates for the representation of alternative possibilities, and then within these regions to interrogate the relationship between this blood oxygen level dependent response and force judgments, which rely on the spontaneous representation of alternatives, at a trial-by-trial level.

Method

Participants

Forty participants participated in this study. Participants were recruited through the Harvard University Participant Study Pool or through participation in previous neuroimaging studies. All were native English speakers, right-handed, and had no history of psychological or neurological conditions. Of these, two were excluded because they were unable to complete the task, and two were excluded for excessive motion (defined as 2 SDs from the group mean on two of the three following motion parameters: average absolute motion per run, average number of movements greater than 0.5 mm per run, and average per run slice-wise signal to noise ratio). The remaining 36 participants (20 identified as female, 15 as male, and one as nonbinary; ages 20–38 years old; average age 25.5 years; 20 identified as White, seven as Asian, two

as Black, three as multiracial, and three as White/Hispanic/Latino) were included in all subsequent analyses.

Stimuli

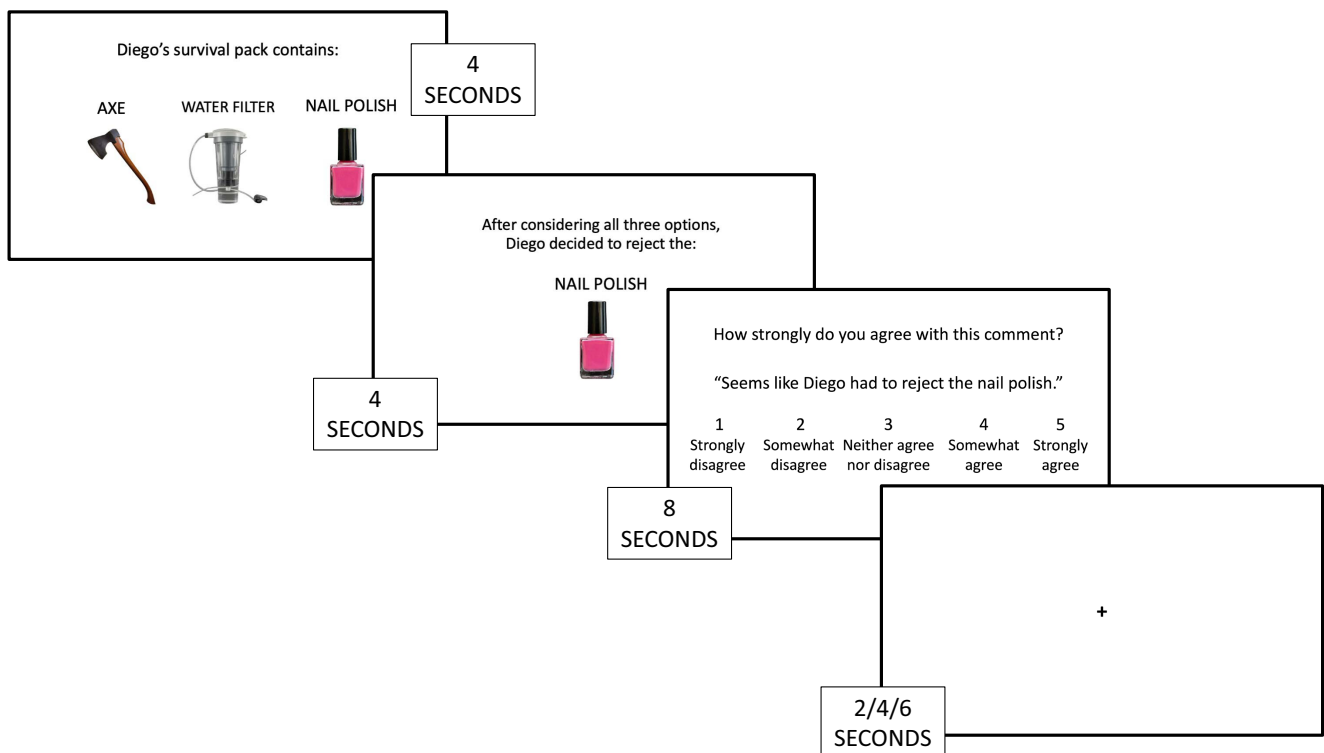
We developed a novel task designed to elicit the spontaneous representation of counterfactuals. To do this, we leveraged the previously demonstrated tendency for individuals to consider alternative possibilities when judging whether an agent was forced or acted freely (Bernhard et al., 2022; Phillips et al., 2015; Phillips & Knobe, 2009; Young & Phillips, 2011). In this task, participants were told that they were to imagine there was a reality television show where the goal was for the contestants to survive living for 1 month alone on a desert island. The only thing the contestants could bring with them on the island were the contents of a “survival pack.” The survival pack always contained three items, but contestants were only able to bring a subset of those items with them to the island. The contestant must choose which items from the survival pack to bring to the island and which items to leave behind.

Participants engaged with two different types of games. In the *reject game*, participants were told that each contestant must reject one item from the survival pack but can take the other two to the desert island. In the *keep game*, participants were told that each

contestant may only keep one item from the survival pack and must leave the other two behind. Participants played one game for an entire scanning run (14 trials), alternating games between runs. Which game they played first was randomly determined.

In both games, each trial proceeded over three stages (Figure 1). In the first stage, participants were presented with the three items in the survival pack. Items were presented both as an image and with the object’s name (e.g., for the fishing pole, both an image of a fishing pole and the words “fishing pole” were displayed on the screen). In the second stage, participants were told which item the contestant decided to reject or keep. For example, participants read, “After considering all three options, Sarah decided to keep the fishing pole.” Again, both the item’s name and an image of the item were presented. Finally, in the third stage, participants made one of two judgments. In the “force” condition, participants made a force judgment by responding on a 1–5 scale how strongly they agreed with a statement like, “Seems like Sarah *had to* keep the fishing pole.” In the “actual action” condition, participants made a judgment about what the contestant actually did by responding on a 1–5 scale how strongly they agreed with a statement such as, “Seems like Sarah kept the fishing pole.” To control for rote responding, participants were given either an “affirmative” version of these statements, like the one described above, or a “negative” version of these statements (e.g., “Seems like Sarah *didn’t have to* keep the

Figure 1
Sample Stimuli



Note. Each trial consisted of three prompts displayed sequentially. In the first prompt, the contents of the survival pack were presented. In the second prompt, participants learned the contestant’s decision. Finally, in the third prompt, participants made their judgment. All neuroimaging analyses were done on the hemodynamic response corresponding to the 8-s period that participants were viewing the third prompt. Each trial was followed by a 2, 4, or 6 s intertrial interval. Images used in the figure were generated by Gemini AI. AI = artificial intelligence. See the online article for the color version of this figure.

fishing pole.”). Including both the affirmative and negative versions of the statements allowed us to disentangle the level of agreement from the condition (see details on conditions below). To that end, all imaging analyses were collapsed across the affirmative and negative statement versions.

There were six possible items that could be in a contestant’s survival pack. These items were selected based on pretesting with a separate set of subjects. Specifically, three items were rated in pretesting as those that would be very useful (high-value) for survival on a desert island (an ax, a fishing pole, and a water filter) and three items were rated as those that would be less useful (low-value) for survival on a desert island (a stuffed animal, a ring, and nail polish). As described above, prior research has shown that individuals are more likely to judge an agent as having acted freely when there is an alternative action of equal or better value the agent could have done instead (Bernhard et al., 2022). Therefore, by systematically varying the items the contestant rejected or kept and the items the contestant could have rejected or kept but did not, we were able to manipulate the likelihood that a participant would agree that the agent was forced or acted freely. For example, consider the case where the survival pack contains two low-value items—the stuffed animal and the ring—and one high-value item—the ax—and the contestant decides to reject the stuffed animal. If the participant is considering the value of the counterfactual alternatives (actions the contestant could have taken instead) when making their force judgment, they are unlikely to say the contestant was forced to reject the stuffed animal because there is an alternative action of equal value the contestant could have taken—the contestant could have rejected the ring instead. Alternatively, consider the case where the survival pack contains only one low-value item—the stuffed animal—and two high-value items—the ax and the water filter—and the contestant decides to reject the stuffed animal. In this case, participants considering the value of the counterfactual alternatives are more likely to say that the contestant was forced to reject the stuffed animal because there is no alternative action the contestant could have taken of comparable value. Rejecting the ax or the water filter instead of the stuffed animal would have been irrational, and thus, it might feel as though the contestant had no other choice.

Importantly, whether a low-value item is associated with a rational or irrational action varies depending on whether the participant is engaging with the reject game or the keep game. In the reject game, the rational action is to reject a low-value item, and a rational alternative would be to reject a different low-value item as well. In the keep game, the opposite is true: The rational action is to keep a high-value item, and the rational alternative would be to keep a different high-value item as well. Thus, by including both the reject and keep games, we were able to eliminate any systematic relationship between the rationality of an action and the value of the items (see Table 2). As a consequence, any association we find between variability in neural activation and force judgment cannot be explained by systematic differences in the value of the items under consideration.

In what we call the “critical trials” of the task, the contestant always makes a rational decision (either rejecting a low-value item in the reject game or keeping a high-value item in the keep game). Moreover, in these critical trials, there is always a rational alternative action (another low-value item the contestant could have rejected instead in the reject game, or another high-value item the contestant could have kept instead in the keep game). Table 3 provides a

concrete illustration of the expected responses for critical trials on which the contestant rejects a ring.

On every run, participants completed eight critical trials, four in the actual action condition (where subjects made judgments about what the contestant actually did) and four in the force condition (where subjects made judgments about what the contestant had to do). For each participant, for every trial in which they were making judgments about the reject game, two of the three possible low-value items were designated as the rational alternatives and two of the three possible high-value items were designated as the irrational alternatives. Likewise, for each participant, for every trial in which they were making judgments about the keep game, two of the three possible high-value items were designated as rational alternatives, and two of the three possible low-value items were designated as the irrational alternatives.















Establishing these critical trials provides several advantages. First, we anticipated that if subjects were considering counterfactual alternatives when making their force judgments, they would most likely do so in these critical trials (when there is something else the contestant could have rationally done instead). Therefore, the bulk of the trials on each run consisted of critical trials. Additionally, by constraining the number of details that vary within critical trials (e.g., the specific items that appear in the survival pack), we were able to perfectly match the features of critical trials across the force and actual action conditions. In each run, for every critical trial force judgment participants were asked to make, they made an actual action judgment for an identical scenario. Given that these trials (a) are designed to elicit maximal consideration of relevant counterfactual alternatives in the force condition and (b) are uniquely matched to trials in the actual action condition, we constrained our initial univariate contrast comparing the force and actual action conditions to just these trials.

Finally, on four of the eight critical trials on each run (two actual action trials and two force trials), the statement presented on the third screen used the affirmative wording (e.g., “Seems like Sarah *had to* keep the fishing pole.”) and the other four critical trials used the negative wording (e.g., “Seems like Sarah *didn’t have to* keep the fishing pole.”). Including both types of wording addresses any confounds between motor responses, or level of agreement, and neural activation. It is important that any association between trial-by-trial residual deviations in neural activation and judgment is not driven by selecting “strongly agree” versus “strongly disagree” or pressing the button box with the pinky versus the thumb. Including the negative wording trials (effectively reversing the conceptual orientation of participants’ responses) disentangles motor action or response label (i.e., “strongly agree” vs. “strongly disagree”) from force judgment, ensuring that these extraneous variables are not driving our effects.

As described above, the four actual action and four force trials in each run were perfectly matched for the actual action and possible alternative actions, as well as affirmative versus negative wording. An identical set of critical trials was used in every reject game run, and a separate, identical set of critical trials was used in every keep game run.¹

¹ This study structure, and specifically the carefully constructed set of critical trials, was designed to support the use of multivoxel pattern analysis to identify brain regions in which we might be reliably able to decode the identity of alternative actions. However, our multivariate analyses proved to be either nonsignificant or inconclusive, possibly due to a lack of power. For transparency and completeness, the details of these analyses are included in our Supplemental Material.

Table 2*Example Survival Pack Contents for the Reject and Keep Games in the Rational and No Rational Alternative Conditions*

Condition	Reject game				Keep game			
	Package contents	Rejected item	Rational alternative item	Irrational alternative item	Package contents	Kept item	Rational alternative item	Irrational alternative item
Rational alternative								
No rational alternative			None				None	

Note. In the *critical trials*, participants always take a rational action (rejecting a low-value item in the reject game and keeping a high-value item in the keep game). The value of the rational and irrational alternative actions then depends on which game participants are engaging with. Images used in this table were generated by Gemini AI. AI = artificial intelligence. See the online article for the color version of this table.

In addition to the critical trials, participants engaged in six noncritical trials on each run. Noncritical trials differed from critical trials in several important ways. On two of these trials, the participants were told that the contestant made an irrational choice (e.g., keeping the stuffed animal) when there were two possible rational options (e.g., the survival pack also contained a fishing pole and an ax). On two other noncritical trials, participants were told that the contestant made a rational choice (e.g., keeping the fishing pole) but there were no possible rational alternatives (e.g., the only other items in the survival pack were the stuffed animal or the ring). Finally, the remaining two noncritical trials had a similar structure to the critical trials (the contestant takes a rational action and there is one rational alternative), but the objects that served as the actual action in the critical trials were now alternative actions, and the objects that served as alternative actions were now actual actions. Noncritical trials could either be in the actual action or force condition and could use the affirmative or negative wording. Finally, for some randomly selected noncritical trials, participants were asked how strongly they agreed with a statement that referred to the

nonactual action (e.g., “Seems like [the contestant] had to reject the ring” when the stuffed animal was actually the item that was rejected).

These noncritical trials served two purposes. First, they allowed us to avoid rote responding and minimized the likelihood that we would see effects associated with participants anticipating the contents of the next trial. Second, it gave us some trials where, if the participant was considering the available alternatives when making force judgments, they should perceive the contestant as forced. Recall that in the critical trials, the contestant always takes the rational action and always has a rational alternative action available. Therefore, if participants’ force judgments depend on the counterfactual alternatives, their response should always be that the contestants were *not* forced to take the action they did. Conversely, some noncritical trials were designed to elicit the opposite response—that the contestant was forced. Including what we call these *high-force noncritical trials* should generate response variability and allow us to measure the degree to which participants’ force judgments are influenced by

Table 3*Anticipated Behavioral Results by Condition, Judgment, and Wording Type, if Participants’ Force Judgments Are Driven by the Availability of a Rational Alternative Action*

	Force judgments: “Seems like the contestant <i>HAD</i> to reject the ring”		Actual action judgments: “Seems like the contestant rejected the ring”	
	Affirmative version: “had to”	Negative version: “didn’t have to”	Affirmative version: “rejected”	Affirmative version: “didn’t reject”
The contestant rejects the ring				
Rational alternative: The survival pack also contained a <i>stuffed animal</i> and an <i>ax</i> .	Low agreement	High agreement	High agreement	Low agreement
No rational alternative: The survival pack also contained a <i>water filter</i> and an <i>ax</i> .	High agreement	Low agreement	High agreement	Low agreement

the available alternatives. In addition, these high-force noncritical trials allow us to measure the degree to which variability in the neural response to force trials is predicted by judgment. Therefore, these high-force noncritical trials, as well as our critical trials, were included in our parametric modulation (PMOD) analysis described below.

Experimental Procedure

Prior to beginning the experiment, participants completed a functional magnetic resonance imaging screening form and provided demographic information. Participant gender and race/ethnicity information were collected by asking participants to write in their gender, race, and ethnicity on a screening form. Once participants entered the scanner, they were thoroughly trained on the task. The training included detailed, step-by-step instructions as well as several practice trials. After the training, each participant was verbally debriefed to ensure complete task comprehension. During the experiment, participants completed 16 runs of 14 trials each, for a total of 224 trials. One hundred twenty-eight of these were critical trials (64 actual action trials and 64 force trials). As illustrated in Figure 1, on each trial, Prompt 1 (the contents of the survival pack) was presented for 4 s, Prompt 2 (the contestant's choice) was presented for 4 s, and Prompt 3 (participants' judgments) was presented for 8 s. The prompts advanced automatically after their display time was up. Each trial was concluded with a pseudorandomly selected 2, 4, or 6 s intertrial interval. Each run lasted a total of approximately 4 min and 45 s, and the entire experiment took 76 min to complete. Between runs, participants rested until they indicated that they were ready to begin the next run, so participants' total time in the scanner was variable, but rarely exceeded 90 min.

Functional Magnetic Resonance Imaging Acquisition and Preprocessing

Neuroimaging was performed using a Siemens Prisma 3.0T scanner with a 32-channel head coil at the Harvard Brain Sciences Center in Cambridge, Massachusetts. A high-resolution structural scan was performed prior to functional data acquisition using a 3-D multi-echo magnetization prepared rapid gradient echo sequence (repetition time [TR] = 2,530 ms, echo time = 1.69 ms, flip angle = 7°, field of view = 256 mm, slice thickness = 1.0 mm, 176 slices). The echo-planar imaging pulse sequence for functional scans used a 2,000-ms TR with 146 TRs per functional run (echo time = 28 ms, flip angle = 80°, field of view = 208 mm, slice thickness = 2.0 mm, 93 slices). Stimuli were presented using Psychtoolbox (Brainard, 1997; Kleiner et al., 2007; Pelli, 1997) software for Matlab.

Data preprocessing was performed using an adaptation of Analysis of Functional Neuroimages Software's (AFNI's) `afni_proc.py` program (https://afni.nimh.nih.gov/pub/dist/doc/program_help/afni_proc.py.html). The first five TRs were removed from each run. After performing despiking and slice time correction, each subject's echo-planar imaging images were spatially registered to the first volume of the second run using cubic polynomial interpolation. Data were scaled and then smoothed for the univariate analyses with a Gaussian kernel at 2 mm full width at half maximum (the equivalent of 1 voxel). Finally, a mask was created to remove any voxels with more than 12

TRs of missing data and was used for all subsequent single-subject analyses.

Region of Interest (ROI) Identification

To identify the neural activation associated with participants' judgments about the contestants' actual actions versus participants' force judgments on critical trials, the hemodynamic response function deconvolution for the 8-s period when participants were asked to make their judgments was performed using AFNI's `3dDeconvolve` with a block model set for an 8-s event duration. This deconvolution was performed on the 2-mm smoothed data. An ordinary least squares regression was performed, with condition (actual action vs. force) as the primary regressor and motion parameters entered as regressors of no interest. A general linear test (GLT) on the contrast of interest was also included in this analysis. Finally, the resultant individual subject whole-brain β -maps were warped to Talairach space using Advanced Normalization Tools (Avants et al., 2009).

To identify regions that, across subjects, were preferentially engaged by the force condition versus the actual action condition, we ran a one-sample t test (with AFNI's `3dttest++` function). This allowed us to identify clusters of voxels in which the average of the betas from the individual-subject force > actual action GLTs was significantly different from zero. Using the `-Clustsim` flag in `3dttest++`, we ran Monte Carlo simulation to perform cluster-wise correction for multiple comparisons on the resultant t -maps.

PMOD

In this study, we are particularly interested in those regions that are preferentially activated by force trials precisely because they are engaged by the spontaneous consideration of counterfactual alternatives. To investigate which of our ROIs are playing this role in particular, we conducted a PMOD. PMOD allows us to test whether the trial-by-trial activation in each of these regions is linearly associated with the degree to which participants perceive the contestant as forced. Because our behavioral results indicate that force judgments are highly dependent on the consideration of counterfactual alternatives (see below), we can infer that those regions whose activation covaries with force judgments are likely engaged by the consideration of these alternatives.

To this end, we identified participants' mean-centered judgments for each trial in the force condition. Importantly, we used judgments from force trials in which there was a relevant alternative action available (and we expect low force judgments) and those in which there was not (and we expect high force judgments). This strategy has two advantages. First, it provides within-subject variability in force judgment. Second, it means that the data used in these analyses are only partially overlapping with the data used to identify our ROIs (which included only data from the critical trials).

At the individual subject level, the trial-by-trial mean-centered judgments were input as parametric modulators in a `3dDeconvolve` analysis in AFNI. As with the original ROI identification analysis, we used a BLOCK model set for an 8-s event duration beginning at the start of the point in each trial when participants were making their force judgments. We used the `-stim-times_AM2` argument to include participants' judgments on each trial as a modulator. This analysis yielded whole-brain voxel-wise coefficients that provide a

measure of the degree to which activation in each region is linearly related to force judgment.

To identify which of our ROIs' force-trial associated activation was significantly modulated by force judgment, we averaged each participant's coefficient maps across all of the voxels in each ROI. This produced one average coefficient per participant, per ROI. We then conducted a series of one-sample *t* tests to evaluate the degree to which the average coefficient across participants was significantly different from zero in each ROI.

Transparency and Openness

For this study, sample size was determined based on sample sizes used in similar neuroimaging studies by the authors and on available funding. No formal power analyses were conducted. Sample size was determined a priori, without intermittent data analyses. All materials, stimuli, group-level neuroimaging data, and analysis scripts are available on the Harvard Dataverse (<https://doi.org/10.7910/DVN/VY6SNO>). Preprocessed and subject-level processed neuroimaging data are available by contacting the first author.

Results

Behavioral Results

To measure the effect of our manipulation on participants' force judgments, we computed separate average rates of agreement for force trials when there was a rational alternative available and when there was no rational alternative available to the contestant. In order

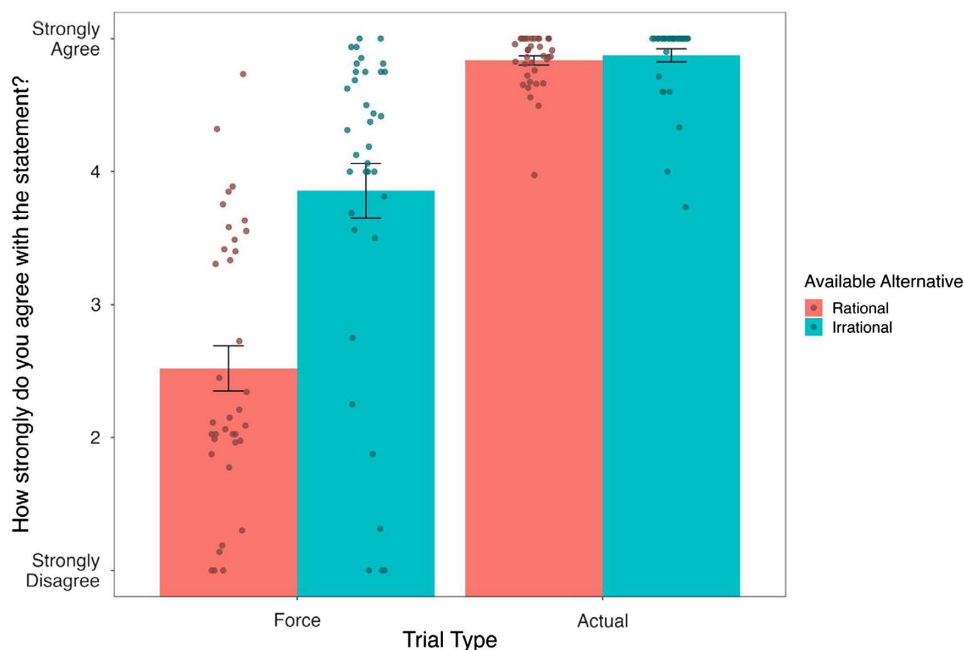
to get a single score per participant per condition, we reverse-scored the ratings in response to statements that used the "negative" wording (e.g., "Seems like the contestant *didn't have to* reject the ring"). Further, we included both the critical force trials and the high-force noncritical trials in these analyses. Therefore, this behavioral data only partially overlaps with the data used for the ROI selection described above, which exclusively used the critical trials.

To test the effect of the availability of rational alternatives on force judgment, we used the *lmer* function from the package *lme4* (Bates et al., 2014) in R to run a mixed effects regression predicting subjects' force judgment from condition (rational alternative vs. no rational alternative), with random slopes and intercepts for subject. We then compared this model to a reduced model that included only the random effects. We found that our original model that included condition predicted force judgments significantly better than the reduced model, $\chi^2(1, N = 36) = 32.08, p < .00001, b = 1.34$. Subjects more strongly agreed that contestants had to take the action they did when there was no rational alternative available ($M = 3.86, SD = 1.23$) than when there was a rational alternative available ($M = 1.97, SD = 1.02$; see Figure 2). We found no meaningful relationship between the availability of a rational alternative and judgment in the actual action condition (rational alternative: $M = 3.97, SD = 0.21$; irrational alternative: $M = 3.73, SD = 0.3$; see Figure 2).

ROI Identification Results

As described above, to identify regions that were preferentially engaged for force judgments versus judgments of the actual action, we ran a one-sample *t* test on the betas from the force > actual action

Figure 2
Participants' Degree of Agreement in Force Versus Actual Action Trials When There Was (Red Bars) and Was Not (Blue Bars) a Rational Alternative Action Available



Note. Points represent participant means. Error bars represent one standard error. See the online article for the color version of this figure.

GLTs from each subject. For the force > than actual action contrast, we found 14 clusters of voxels that exceeded a voxel-wise significance threshold of $p < .001$ and a cluster-wise corrected significant threshold of $p < .05$ (see Table 4/Figure 3). These clusters were widely distributed throughout the brain, including in the bilateral prefrontal cortex, bilateral inferior parietal lobule, bilateral middle and inferior temporal gyri, bilateral posterior cingulate, and bilateral caudate. The actual action > force contrast yielded no significant clusters that met our threshold of a voxel-wise $p < .001$ and a cluster-wise corrected $p < .05$.

PMOD Results

While the univariate contrast served to identify regions that were more responsive when participants were making force judgments than thinking about actual actions, significant effects from this analysis could be due to any number of incidental features associated with differences between the two conditions. One might imagine, for example, that making force judgments is more effortful than judging what the participant actually did. If this were the case, we would find regions of increased activation for force relative to actual action trials, not because participants were considering counterfactuals, but because they were engaging in a more effortful task. Alternatively, if the increased activation in our ROIs for force trials is specifically driven by the consideration of counterfactual alternatives during force judgments, we would expect activation in these regions to covary with participants' trial-by-trial force judgments.

To test the latter hypothesis, we ran a PMOD within each ROI and then assessed the degree to which the resulting average coefficient in each ROI was significantly different from zero. In every one of the 14 regions, the coefficient was greater than zero, and in nine of the 14 regions, using one-sample t tests, it was significantly above 0 ($p < .05$ uncorrected; see Table 5/Figure 4). These results suggest that, considered as a group, this collection of ROIs tends to exhibit positive linear relationships between the strength of the force-trial

related activation and the degree to which they perceived each participant as forced. Moreover, given that our behavioral results indicate that force judgments are strongly dependent on the availability of a rational alternative action, these findings suggest that activation in these regions is closely tied to the proposed spontaneous consideration of counterfactual alternatives and not merely an extraneous difference between force and actual action trials.

A separate question concerns the evidence that we have found for this relationship in any specific individual ROI, which would be important if one wanted to draw strong inferences, for example, based on the location of a given ROI. While our primary focus is not on this form of inference, the four regions that survive Bonferroni correction for multiple comparisons are the left middle temporal gyrus, right inferior/middle temporal gyrus, right middle temporal gyrus, and right caudate.

Finally, it is important to consider any potential bias of ROI definition on the results of the PMOD analysis reported here. Since the ROIs were identified by selecting regions where activation to Force/Rational Alternative trials exceeded activation to Actual Action/Rational Alternative trials (the critical trials), we should expect an inflated estimate of activation on Force/Rational Alternative trials. Meanwhile, our PMOD analysis asks whether behavioral judgments of greater forcedness predict increased activation within these regions on Force trials, including both Rational Alternative and No Rational Alternative trials (both the critical and noncritical trials). As described above, average behavioral forcedness ratings are higher for No Rational Alternative trials. Thus, inflated estimates of neural activation on Rational Alternative trials would tend to produce the opposite of the result we observe here: greater neural activation corresponding with higher behavioral ratings on No Rational Alternative trials. In short, nonindependence between the ROI definition and the PMOD tests we conducted make the tests we performed particularly conservative and cannot explain the observed pattern of results.

Table 4

Univariate Activation Associated With Force Judgments (Significantly More Activation for Force Trials Relative to Actual Action Trials)

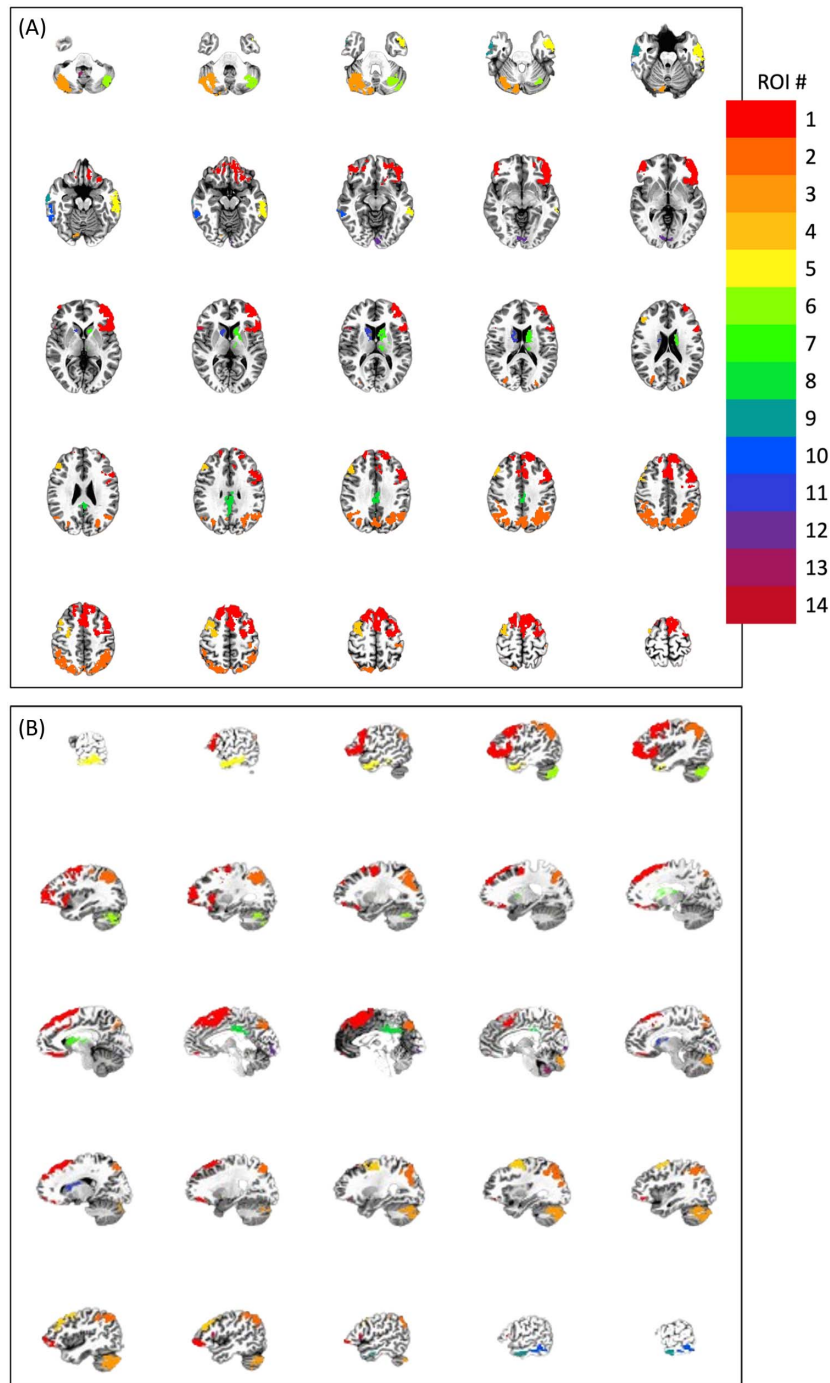
ROI No.	L/R ^a	Anatomical region ^b	TLRC coordinate ^c			Peak z-score	Cluster size (voxel)
			x	y	z		
1	L	Middle frontal gyrus	-11.5	18.5	56.5	6.62	7,550***
2	R/L	Precuneus; inferior frontal lobule	-41.5	-59.5	34.5	6.06	3,868***
3	R	Pyramus	40.5	-59.5	-37.5	6.45	1,681***
4	R	Middle frontal gyrus	26.5	-13.5	42.5	5.05	779***
5	L	Middle temporal gyrus	-45.5	4.5	-25.5	5.47	770***
6	L	Cerebellar tonsil	-43.5	-61.5	-35.5	5.40	688***
7	L	Caudate	-13.5	4.5	18.5	6.32	395***
8	R/L	Cingulate	-1.5	-29.5	36.5	5.67	367***
9	R	Inferior/middle temporal gyrus	52.5	-11.5	-19.5	4.83	222***
10	R	Middle temporal gyrus	58.5	-39.5	-9.5	5.08	199***
11	R	Caudate	14.5	12.5	12.5	5.47	198***
12	R/L	Lingual gyrus	0.5	-85.5	-7.5	4.74	143**
13	R	Cerebellar tonsil	6.5	-53.5	-41.5	5.25	70*
14	R	Inferior frontal gyrus	54.5	16.5	8.5	4.49	55*

Note. All clusters surpass a voxel-wise significance threshold of $p < .001$. ROI = region of interest; TLRC = Talairach.

^aLocalized in the left versus right hemisphere. ^bIndicates anatomical region containing the largest proportion of voxels, although in most cases clusters extend through additional regions. ^cTLRC coordinates for voxels of peak activation.

* Cluster-wise corrected $p < .05$. ** Cluster-wise corrected $p < .01$. *** Cluster-wise corrected $p < .005$.

Figure 3
Regions That Were Significantly More Active for Force Trials Than Actual Action Trials



Note. (A) Axial view. (B) Sagittal view. All clusters surpass a voxel-wise significance threshold of $p < .001$ and a voxel-wise corrected threshold of $p < .05$. These 14 clusters were used as ROIs for parametric modulation. ROI = region of interest. See the online article for the color version of this figure.

Table 5
Average Coefficient for Each ROI From the Parametric Modulation

ROI No.	L/R	Anatomical region ^a	Mean Coefficient	Coefficient <i>SD</i>	<i>t</i> -stat (<i>df</i> = 32)	<i>p</i>
1	L	Middle frontal gyrus	0.007	0.044	0.9	.375
2	R/L	Precuneus; inferior frontal lobule	0.003	0.04	0.369	.714
3	R	Pyramus	0.008	0.046	0.991	.329
4	R	Middle frontal gyrus	0.01	0.038	1.463	.153
5	L	Middle temporal gyrus	0.02	0.038	2.964	.006*
6	L	Cerebellar tonsil	0.008	0.039	1.128	.268
7	L	Caudate	0.017	0.046	2.089	.045
8	R/L	Cingulate	0.026	0.064	2.313	.027
9	R	Inferior/middle temporal gyrus	0.027	0.048	3.201	.003*
10	R	Middle temporal gyrus	0.02	0.038	3.011	.005*
11	R	Caudate	0.024	0.048	2.921	.006*
12	R/L	Lingual gyrus	0.026	0.058	2.582	.015
13	R	Cerebellar tonsil	0.0281	0.063	2.559	.015
14	R	Inferior frontal gyrus	0.0222	0.054	2.355	.025

Note. ROI = region of interest; L/R = Localized in the left versus right hemisphere.

^a Indicates anatomical region containing the largest proportion of voxels, although in most cases clusters extend through additional regions.

* Survives Bonferroni correction for multiple comparisons at $p < .05$.

Discussion

Many important kinds of decisions, from judgments of moral responsibility to judgments of causal reasoning, evoke spontaneous counterfactual thought. Yet there has been little research on how such thought is instantiated in the brain. Here, we used functional magnetic resonance imaging to investigate how the brain is engaged by spontaneous counterfactual thinking when making force judgments. We found a network of 14 brain regions that were preferentially engaged when participants were making force judgments relative to considering what was actually done. Critically, we found that activation in the majority of these regions was specifically related to variation in participants' force judgments on a trial-by-trial basis. Given that our behavioral results indicate that force judgments in our task are directly dependent on the nature of the available alternatives, our neuroimaging findings suggest that these regions are engaged in the spontaneous consideration of counterfactual alternatives.

Our study design leverages several key features that allow us to rule out alternative explanations for these findings. First, including both the reject and keep games allowed us to orthogonalize the value of the items associated with alternative actions and the rationality of the alternatives that drives force judgments. In the reject game, rational alternative actions involved low-value items and irrational alternatives involved high-value items, while the opposite was true in the keep game. This feature of our design makes it unlikely that trial-by-trial variability in activation in our ROIs is associated with item value rather than force judgment in particular. Second, it is possible that rather than tracking the consideration of alternatives during force judgments, activation in our ROIs is being modulated by some incidental feature associated with responding during force trials. For example, it could be that variability in activation in these regions is driven by specific motor actions required to indicate a judgment or the degree to which participants agree with the statement presented during the trial. However, in our study, participants made judgments about both whether the contestant was forced *and* whether the contestant was *not* forced. Providing both the affirmative and negative versions of the force (and actual action)

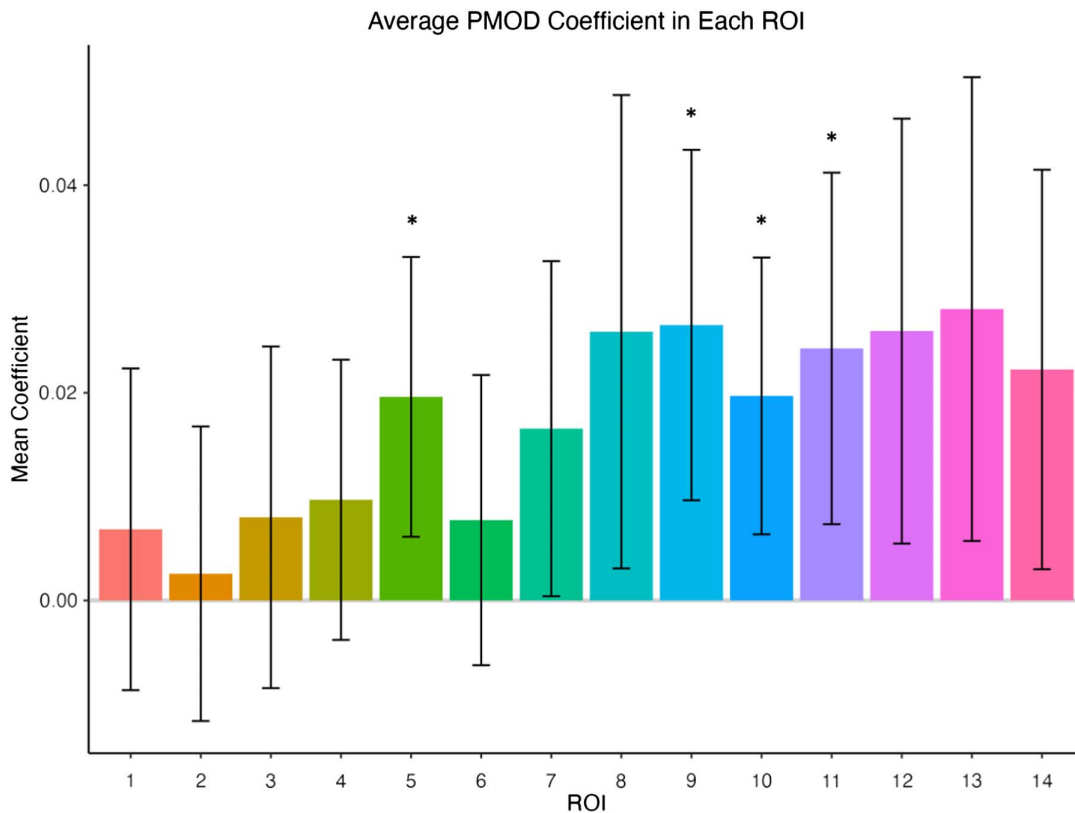
statements ensured that a given judgment (i.e., that the contestant was forced) required indicating both agreement and disagreement and required using distinct motor responses to do so. Finally, asking participants to make judgments in response to both of these types of statements decreases the likelihood that our findings were merely a byproduct of thinking about agents being forced. Rather, because participants were also considering whether agents were *not* forced, any effects we see across wording types are likely due to the shared features of the affirmative and negative trials: whether or not there was a rational alternative action. Put simply, the careful control afforded by our study design suggests that the observed association between trial-by-trial neural activation in many of our ROIs and force judgment is driven by the consideration of counterfactual alternatives rather than low-level or incidental features of our task.

The Neural Instantiation of Spontaneous Counterfactual Thought

Our results implicate a set of 14 brain regions as candidate substrates for the spontaneous representation of alternative possibilities. Many overlap significantly with those identified in prior studies of explicit, episodic simulation of nonactual events (Barbey et al., 2009; De Brigard et al., 2013, 2015, 2017; De Brigard & Parikh, 2019; Faul et al., 2020; Gomez Beldarrain et al., 2005; Kulakova et al., 2013; Nieuwland, 2012; Parikh et al., 2018; Schacter et al., 2015; St. Jacques et al., 2018; Urrutia et al., 2012; Van Hoeck et al., 2013, 2015). These include the lateral temporal lobe, medial prefrontal cortex, cingulate cortex, and lateral/posterior parietal cortex (De Brigard et al., 2013). This "core" network overlaps substantially with the default mode network (Buckner et al., 2008) and is also preferentially engaged during episodic simulation of future events and episodic memory (Addis et al., 2007; Hassabis et al., 2007; Okuda et al., 2003; Schacter et al., 2007; Szpunar et al., 2007).

Our results depart from this prior work in several important respects, however. First, we did not find increased activation in the medial temporal lobe (MTL; Addis et al., 2007, 2011; Hassabis et al., 2007; Okuda et al., 2003; Szpunar et al., 2007; Weiler et al., 2010). Given that a null result does not necessarily indicate an absence of an

Figure 4
Average Coefficient for Each ROI From Our Parametric Modulation



Note. ROIs are identified by number as indicated in Table 5 and are ordered by the size of the ROI. Error bars represent 95% confidence intervals. ROI = region of interest; PMOD = parametric modulation. See the online article for the color version of this figure.

* Survives Bonferroni correction for multiple comparisons at $p < .05$.

effect, interpreting this finding should be done cautiously. However, engagement of the hippocampus and parahippocampal gyrus occurs during many types of explicit episodic simulation, making the absence of such activation in the present study noteworthy. MTL engagement during episodic simulation has been interpreted as allowing individuals to recombine previously encoded memories into novel counterfactual thoughts or imagined futures (Chua et al., 2004; Eichenbaum, 2001; Schacter & Wagner, 1999; Sullivan Giovanello et al., 2004) or as encoding these simulations for future recall (Martin et al., 2011). The absence of MTL activation associated with force judgments in our study suggests that memory may play a less significant role in spontaneous counterfactual thought relative to episodic counterfactual simulation. Therefore, despite the overlap of many regions between spontaneous counterfactual thought in our task and in explicit episodic simulation, it suggests that they may be supported by at least partially divergent processes. However, as previously mentioned, this inference depends on the absence of above-threshold activation in the MTL in our study and therefore should be treated with some caution.

Additionally, we found that residual deviations in blood oxygen level dependent response were significantly predicted by force judgment in the right inferior frontal gyrus (right IFG). The IFG is

typically engaged by inhibitory or attentional control processes (Aron et al., 2004, 2014; Hampshire et al., 2010). While any conclusions we draw about the role of the IFG in the present study are speculative, it has been implicated in prior neuroimaging work as critical to entertaining hypothetical states of the world (Bernhard et al., 2023; Goel & Dolan, 2003). A central feature of counterfactual thought is exactly this: considering nonactual states of the world. Our findings provide further support for the idea that the IFG may play a special role in this process.

Finally, we found that activation in the bilateral caudate varied with force judgment. Given this region's established role in value representation and value-based affective responding (e.g., Grahn et al., 2008), one possible explanation for its recruitment in our task is that it is merely an artifact produced by our task's explicit manipulation of value. While participants may more explicitly represent value in the force condition than the actual action condition, it is worth noting that value is manipulated in both conditions suggesting this might not entirely explain this effect. Moreover, because we had both the reject and keep game versions of our task, object value was orthogonal to the availability in rational alternatives, making it unlikely that variability in caudate activation associated with judgment was merely tracking object value.

Alternatively, it could be that value representation is a critical component of spontaneous counterfactual thought, as theoretically proposed in other work (e.g., Phillips et al., 2019). Consistent with the latter possibility, past work has frequently implicated the striatum in processes related to counterfactual thought (see De Brigard & Parikh, 2019; Van Hoeck et al., 2015, for reviews). For example, changes in striatal activation have been associated with the influence of regret on future choices (Camille et al., 2010; Nicolle et al., 2011), and in comparing outcomes with possible alternatives (Henderson & Norris, 2013; Koehlin & Hyafil, 2007; Tobia et al., 2014). Possibly, then, the role of the caudate we identified here reflects its general contribution to representing counterfactuals, and not an incidental feature of our design.

Constraints on Generalizability

Although we believe that we provide strong evidence to support the role of many of our ROIs in spontaneous counterfactual thought during force judgment, there are some constraints on the generalizability of our findings. First, we used a convenience sample, consisting of primarily Harvard University undergraduates. Therefore, although spontaneous counterfactual thought is likely a ubiquitous process, the specific brain regions involved may vary across populations. Second, because evaluating actions and their alternatives in this task requires thinking about objects, some of the neural activation we see in our study may be specific to object representation. Consequently, differences in neural activation may arise when individuals are considering other types of counterfactual events. Finally, because spontaneous counterfactual thought was elicited in our study by asking participants to make force judgments, it is possible that in other contexts, such thought engages distinct (or only partially overlapping) sets of brain regions. For example, spontaneous counterfactual thought is a key feature of causal judgment, yet making judgments about causation is different in many ways from making judgments about force. Therefore, we may expect to find some differences in the neural activation associated with spontaneous counterfactual thought when making causal judgments and when making force judgments. This likely applies to the many other domains in which we spontaneously consider counterfactual alternatives as well.

Conclusions

While prior work has provided behavioral evidence for the role of spontaneous and implicit representations of possibilities that are constrained by normality (Acierno et al., 2022; Bernhard et al., 2022; Phillips & Cushman, 2017), the present research offers the first neural evidence that corroborates this emerging picture. This project offers a first step in illuminating our understanding of the neural basis for spontaneous counterfactual thought and points toward important avenues for continuing research on this topic.

References

- Acierno, J., Mischel, S., & Phillips, J. (2022). Moral judgements reflect default representations of possibility. *Philosophical Transactions of the Royal Society B*, 377(1866), Article 20210341. <https://doi.org/10.1098/rstb.2021.0341>
- Addis, D. R., Cheng, T., Roberts, R. P., & Schacter, D. L. (2011). Hippocampal contributions to the episodic simulation of specific and general future events. *Hippocampus*, 21(10), 1045–1052. <https://doi.org/10.1002/hipo.20870>
- Addis, D. R., Wong, A. T., & Schacter, D. L. (2007). Remembering the past and imagining the future: Common and distinct neural substrates during event construction and elaboration. *Neuropsychologia*, 45(7), 1363–1377. <https://doi.org/10.1016/j.neuropsychologia.2006.10.016>
- Aquinas, T. (1920). *The summa theologiae of St. Thomas Aquinas* (second and revised edition). Translated by Fathers of the English Dominican Province. Encyclopedia Britannica. (Original work published 1273).
- Aristotle. (340 BCE/2002). *Nicomachean ethics*. Oxford University Press.
- Aron, A. R., Robbins, T. W., & Poldrack, R. A. (2004). Inhibition and the right inferior frontal cortex. *Trends in Cognitive Sciences*, 8(4), 170–177. <https://doi.org/10.1016/j.tics.2004.02.010>
- Aron, A. R., Robbins, T. W., & Poldrack, R. A. (2014). Inhibition and the right inferior frontal cortex: One decade on. *Trends in Cognitive Sciences*, 18(4), 177–185. <https://doi.org/10.1016/j.tics.2013.12.003>
- Avants, B. B., Tustison, N., & Song, G. (2009). Advanced normalization tools (ANTS). *The Insight Journal*, 2(365), 1–35. <https://doi.org/10.54294/uvnhin>
- Barbey, A. K., Krueger, F., & Grafman, J. (2009). Structured event complexes in the medial prefrontal cortex support counterfactual representations for future planning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1521), 1291–1300. <https://doi.org/10.1098/rstb.2008.0315>
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). *Fitting linear mixed-effects models using lme4*. arXiv. <https://doi.org/10.48550/arXiv.1406.5823>
- Bear, A., & Knobe, J. (2017). Normality: Part descriptive, part prescriptive. *Cognition*, 167, 25–37. <https://doi.org/10.1016/j.cognition.2016.10.024>
- Benoit, R. G., & Schacter, D. L. (2015). Specifying the core network supporting episodic simulation and episodic memory by activation likelihood estimation. *Neuropsychologia*, 75, 450–457. <https://doi.org/10.1016/j.neuropsychologia.2015.06.034>
- Bernhard, R. M., LeBaron, H., & Phillips, J. (2022). It's not what you did, it's what you could have done. *Cognition*, 228, Article 105222. <https://doi.org/10.1016/j.cognition.2022.105222>
- Bernhard, R. M., Phillips, J. S., Cushman, F. A., & Cameron, A. (2023). *The neural instantiation of spontaneous counterfactual thought*. <https://doi.org/10.31234/osf.io/zdf2w>
- Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, 10(4), 433–436. <https://doi.org/10.1163/156856897X00357>
- Buckner, R. L., Andrews-Hanna, J. R., & Schacter, D. L. (2008). The brain's default network: Anatomy, function, and relevance to disease. *Annals of the New York Academy of Sciences*, 1124(1), 1–38. <https://doi.org/10.1196/annals.1440.011>
- Byrne, R. M. (2017). Counterfactual thinking: From logic to morality. *Current Directions in Psychological Science*, 26(4), 314–322. <https://doi.org/10.1177/0963721417695617>
- Camille, N., Pironi, V. A., Dodds, C. M., Aitken, M. R. F., Robbins, T. W., & Clark, L. (2010). Striatal sensitivity to personal responsibility in a regret-based decision-making task. *Cognitive, Affective & Behavioral Neuroscience*, 10(4), 460–469. <https://doi.org/10.3758/CA.BN.10.4.460>
- Chua, E. F., Rand-Giovannetti, E., Schacter, D. L., Albert, M. S., & Sperling, R. A. (2004). Dissociating confidence and accuracy: Functional magnetic resonance imaging shows origins of the subjective memory experience. *Journal of Cognitive Neuroscience*, 16(7), 1131–1142. <https://doi.org/10.1162/0898929041920568>
- De Brigard, F., Addis, D. R., Ford, J. H., Schacter, D. L., & Giovanello, K. S. (2013). Remembering what could have happened: Neural correlates of episodic counterfactual thinking. *Neuropsychologia*, 51(12), 2401–2414. <https://doi.org/10.1016/j.neuropsychologia.2013.01.015>

- De Brigard, F., Nathan Spreng, R., Mitchell, J. P., & Schacter, D. L. (2015). Neural activity associated with self, other, and object-based counterfactual thinking. *NeuroImage*, 109, 12–26. <https://doi.org/10.1016/j.neuroimage.2014.12.075>
- De Brigard, F., & Parikh, N. (2019). Episodic counterfactual thinking. *Current Directions in Psychological Science*, 28(1), 59–66. <https://doi.org/10.1177/0963721418806512>
- De Brigard, F., Parikh, N., Stewart, G. W., Szpunar, K. K., & Schacter, D. L. (2017). Neural activity associated with repetitive simulation of episodic counterfactual thoughts. *Neuropsychologia*, 106, 123–132. <https://doi.org/10.1016/j.neuropsychologia.2017.09.022>
- Eichenbaum, H. (2001). The hippocampus and declarative memory: Cognitive mechanisms and neural codes. *Behavioural Brain Research*, 127(1–2), 199–207. [https://doi.org/10.1016/S0166-4328\(01\)00365-5](https://doi.org/10.1016/S0166-4328(01)00365-5)
- Faul, L., St. Jacques, P. L., DeRosa, J. T., Parikh, N., & De Brigard, F. (2020). Differential contribution of anterior and posterior midline regions during mental simulation of counterfactual and perspective shifts in autobiographical memories. *NeuroImage*, 215, Article 116843. <https://doi.org/10.1016/j.neuroimage.2020.116843>
- Gerstenberg, T., Peterson, M. F., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2017). Eye-tracking causality. *Psychological Science*, 28(12), 1731–1744. <https://doi.org/10.1177/0956797617713053>
- Goel, V., & Dolan, R. J. (2003). Explaining modulation of reasoning by belief. *Cognition*, 87(1), B11–B22. [https://doi.org/10.1016/S0010-0277\(02\)00185-3](https://doi.org/10.1016/S0010-0277(02)00185-3)
- Gomez Beldarrain, M., Garcia-Monco, J. C., Astigarraga, E., Gonzalez, A., & Grafman, J. (2005). Only spontaneous counterfactual thinking is impaired in patients with prefrontal cortex lesions. *Cognitive Brain Research*, 24(3), 723–726. <https://doi.org/10.1016/j.cogbrainres.2005.03.013>
- Grahn, J. A., Parkinson, J. A., & Owen, A. M. (2008). The cognitive functions of the caudate nucleus. *Progress in Neurobiology*, 86(3), 141–155. <https://doi.org/10.1016/j.pneurobio.2008.09.004>
- Halpern, J. Y., & Hitchcock, C. (2015). Graded causation and defaults. *The British Journal for the Philosophy of Science*, 66(2), 413–457. <https://doi.org/10.1093/bjps/axt050>
- Hampshire, A., Chamberlain, S. R., Monti, M. M., Duncan, J., & Owen, A. M. (2010). The role of the right inferior frontal gyrus: Inhibition and attentional control. *NeuroImage*, 50(3), 1313–1319. <https://doi.org/10.1016/j.neuroimage.2009.12.109>
- Hassabis, D., Kumaran, D., & Maguire, E. A. (2007). Using imagination to understand the neural basis of episodic memory. *The Journal of Neuroscience*, 27(52), 14365–14374. <https://doi.org/10.1523/JNEUROSCI.4549-07.2007>
- Henderson, S. E., & Norris, C. J. (2013). Counterfactual thinking and reward processing: An fMRI study of responses to gamble outcomes. *NeuroImage*, 64, 582–589. <https://doi.org/10.1016/j.neuroimage.2012.08.078>
- Icard, T. F., Kominsky, J. F., & Knobe, J. (2017). Normality and actual causal strength. *Cognition*, 161, 80–93. <https://doi.org/10.1016/j.cognition.2017.01.010>
- Khoudary, A., O'Neill, K., Faul, L., Murray, S., Smallman, R., & De Brigard, F. (2022). Neural differences between internal and external episodic counterfactual thoughts. *Philosophical Transactions of the Royal Society B*, 377(1866), Article 20210337. <https://doi.org/10.1098/rstb.2021.0337>
- Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). What's new in Psychtoolbox-3? *Perception*, 36(14), 1–16.
- Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis*, 63(3), 190–194. <https://doi.org/10.1093/analys/63.3.190>
- Koechlin, E., & Hyafil, A. (2007). Anterior prefrontal function and the limits of human decision-making. *Science*, 318(5850), 594–598. <https://doi.org/10.1126/science.1142995>
- 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), Article e12792. <https://doi.org/10.1111/cogs.12792>
- Kominsky, J. F., Phillips, J., Gerstenberg, T., Lagnado, D., & Knobe, J. (2015). Causal superseding. *Cognition*, 137, 196–209. <https://doi.org/10.1016/j.cognition.2015.01.013>
- Kratzer, A. (2012). *Modals and conditionals: New and revised perspectives* (Vol. 36). Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199234684.001.0001>
- Kulakova, E., Aichhorn, M., Schurz, M., Kronbichler, M., & Perner, J. (2013). Processing counterfactual and hypothetical conditionals: An fMRI investigation. *NeuroImage*, 72, 265–271. <https://doi.org/10.1016/j.neuroimage.2013.01.060>
- Locke, J. (1975). *An essay concerning human understanding* (Vol. 3). Oxford University Press. (Original work published 1690)
- Martin, V. C., Schacter, D. L., Corballis, M. C., & Addis, D. R. (2011). A role for the hippocampus in encoding simulations of future events. *Proceedings of the National Academy of Sciences*, 108(33), 13858–13863. <https://doi.org/10.1073/pnas.1105816108>
- Nicolle, A., Bach, D. R., Driver, J., & Dolan, R. J. (2011). A role for the striatum in regret-related choice repetition. *Journal of Cognitive Neuroscience*, 23(4), 845–856. <https://doi.org/10.1162/jocn.2010.21510>
- Nieuwland, M. S. (2012). Establishing propositional truth-value in counterfactual and real-world contexts during sentence comprehension: Differential sensitivity of the left and right inferior frontal gyri. *NeuroImage*, 59(4), 3433–3440. <https://doi.org/10.1016/j.neuroimage.2011.11.018>
- Okuda, J., Fujii, T., Ohtake, H., Tsukiura, T., Tanji, K., Suzuki, K., Kawashima, R., Fukuda, H., Itoh, M., & Yamadori, A. (2003). Thinking of the future and past: The roles of the frontal pole and the medial temporal lobes. *NeuroImage*, 19(4), 1369–1380. [https://doi.org/10.1016/S1053-8119\(03\)00179-4](https://doi.org/10.1016/S1053-8119(03)00179-4)
- Parikh, N., Ruzic, L., Stewart, G. W., Spreng, R. N., & De Brigard, F. (2018). What if? Neural activity underlying semantic and episodic counterfactual thinking. *NeuroImage*, 178, 332–345. <https://doi.org/10.1016/j.neuroimage.2018.05.053>
- Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, 10(4), 437–442. <https://doi.org/10.1163/156856897X00366>
- Pettit, D., & Knobe, J. (2009). The pervasive impact of moral judgment. *Mind & Language*, 24(5), 586–604. <https://doi.org/10.1111/j.1468-0017.2009.01375.x>
- 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. (2009). Moral judgments and intuitions about freedom. *Psychological Inquiry*, 20(1), 30–36. <https://doi.org/10.1080/10478400902744279>
- 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>
- Schacter, D. L., Addis, D. R., & Buckner, R. L. (2007). Remembering the past to imagine the future: The prospective brain. *Nature Reviews Neuroscience*, 8(9), 657–661. <https://doi.org/10.1038/nrn2213>
- Schacter, D. L., Benoit, R. G., De Brigard, F., & Szpunar, K. K. (2015). Episodic future thinking and episodic counterfactual thinking: Intersections between memory and decisions. *Neurobiology of Learning and Memory*, 117, 14–21. <https://doi.org/10.1016/j.nlm.2013.12.008>
- Schacter, D. L., & Wagner, A. D. (1999). Medial temporal lobe activations in fMRI and PET studies of episodic encoding and retrieval. *Hippocampus*, 9(1), 7–24. [https://doi.org/10.1002/\(SICI\)1098-1063\(1999\)9:1<7::AID-HIPO2>3.0.CO;2-K](https://doi.org/10.1002/(SICI)1098-1063(1999)9:1<7::AID-HIPO2>3.0.CO;2-K)

- St. Jacques, P. L., Carpenter, A. C., Szpunar, K. K., & Schacter, D. L. (2018). Remembering and imagining alternative versions of the personal past. *Neuropsychologia*, 110, 170–179. <https://doi.org/10.1016/j.neuropsychologia.2017.06.015>
- Stalnaker, R. C. (1968). A theory of conditionals. In N. Rescher (Ed.), *Studies in logical theory* (pp. 98–112). Blackwell.
- Sullivan Giovanello, K., Schnyer, D. M., & Verfaellie, M. (2004). A critical role for the anterior hippocampus in relational memory: Evidence from an fMRI study comparing associative and item recognition. *Hippocampus*, 14(1), 5–8. <https://doi.org/10.1002/hipo.10182>
- Szpunar, K. K., Watson, J. M., & McDermott, K. B. (2007). Neural substrates of envisioning the future. *Proceedings of the National Academy of Sciences*, 104(2), 642–647. <https://doi.org/10.1073/pnas.0610082104>
- Tobia, M. J., Guo, R., Schwarze, U., Boehmer, W., Gläscher, J., Finckh, B., Marschner, A., Büchel, C., Obermayer, K., & Sommer, T. (2014). Neural systems for choice and valuation with counterfactual learning signals. *NeuroImage*, 89, 57–69. <https://doi.org/10.1016/j.neuroimage.2013.11.051>
- Urrutia, M., Gennari, S. P., & de Vega, M. (2012). Counterfactuals in action: An fMRI study of counterfactual sentences describing physical effort. *Neuropsychologia*, 50(14), 3663–3672. <https://doi.org/10.1016/j.neuropsychologia.2012.09.004>
- Uttich, K., & Lombrozo, T. (2010). Norms inform mental state ascriptions: A rational explanation for the side-effect effect. *Cognition*, 116(1), 87–100. <https://doi.org/10.1016/j.cognition.2010.04.003>
- Van Hoeck, N., Ma, N., Ampe, L., Baetens, K., Vandekerckhove, M., & Van Overwalle, F. (2013). Counterfactual thinking: An fMRI study on changing the past for a better future. *Social Cognitive and Affective Neuroscience*, 8(5), 556–564. <https://doi.org/10.1093/scan/nss031>
- Van Hoeck, N., Watson, P. D., & Barbey, A. K. (2015). Cognitive neuroscience of human counterfactual reasoning. *Frontiers in Human Neuroscience*, 9(9), Article 420. <https://doi.org/10.3389/fnhum.2015.00420>
- Weiler, J. A., Suchan, B., & Daum, I. (2010). Foreseeing the future: Occurrence probability of imagined future events modulates hippocampal activation. *Hippocampus*, 20(6), 685–690. <https://doi.org/10.1002/hipo.20695>
- Woolfolk, R. L., Doris, J. M., & Darley, J. M. (2006). Identification, situational constraint, and social cognition: Studies in the attribution of moral responsibility. *Cognition*, 100(2), 283–301. <https://doi.org/10.1016/j.cognition.2005.05.002>
- Young, L., & Phillips, J. (2011). The paradox of moral focus. *Cognition*, 119(2), 166–178. <https://doi.org/10.1016/j.cognition.2011.01.004>

Received March 10, 2023

Revision received August 7, 2024

Accepted August 15, 2024 ■