AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: General

ISSN: 0096-3445

2024, Vol. 153, No. 3, 837–863
https://doi.org/10.1037/xge0001540

Psychological Freedom, Rationality, and the Naive Theory of Reasoning

Corey Cusimano[1], Natalia Zorrilla[2], David Danks[3], and Tania Lombrozo[2]
[1] School of Management, Yale University
[2] Department of Psychology, Princeton University
[3] Halıcıoğlu Data Science Institute and Department of Philosophy, University of California, San Diego

To make sense of the social world, people reason about others' mental states, including whether and in what ways others can form new mental states. We propose that people's judgments concerning the dynamics of mental state change invoke a "naive theory of reasoning." On this theory, people conceptualize reasoning as a rational, semi-autonomous process that individuals can leverage, but not override, to form new rational mental states. Across six experiments, we show that this account of people's naive theory of reasoning predicts judgments about others' ability to form rational and irrational beliefs, desires, and intentions, as well as others' ability to act rationally and irrationally. This account predicts when, and explains why, people judge others as psychologically constrained by coercion and other forms of situational pressure.

Public Significance Statement
Judgments of others' psychological freedom and constraint are pivotal to a wide range of important social and political judgments. For instance, people believe that others are morally and legally responsible only when those individuals act freely. And, people evaluate political and economic institutions, and practices such as nudges, price gouging, and sweatshop labor, partly based on how these institutions and practices affect people's freedom. We demonstrate that people's judgments of freedom and constraint draw principally on an intuitive theory of reasoning. This model shows how judgments about others' freedom may be a part of people's mature, allocentric, and rational theory of mind. Our model supplants prior theories according to which people make these judgments in a moralized or motivated way.

Keywords: theory of mind, control, rationality, reasoning, freedom

Consider someone who is living in dire poverty and just sold their kidney for a large sum of money. Were they free to walk away? How people answer this question reflects how they think about psychological freedom and constraint, and specifically, how they think about others' capacity to exert control over their own minds in the face of situational pressure. Attributions of this kind are pivotal for an important set of social judgments and behaviors. In the domain of law and moral judgment, for instance, people excuse others for their bad attitudes or harmful behaviors when they think that those individuals lack control over them (Alicke, 2000; Cusimano & Goodwin, 2022; Weiner, 1995). Attributions of freedom and constraint are also important in political theory because people evaluate political institutions and economic systems based in large part on whether citizens are free under them (Nozick, 1974; Rawls, 1971; Wertheimer, 1987). Indeed, people's opinions on nudges, price gouging, compensation in medical studies, sweatshop labor, and consent, to name a few, depend on their beliefs about whether individuals in these situations can think and act freely (Baron, 1998; Fischel, 2019; Powell & Zwolinski, 2012; Sommers, 2019). For these reasons, understanding how people attribute mental state control to others is important for understanding why they hold others responsible and how they navigate complex social and political issues.

Corey Cusimano https://orcid.org/0000-0002-4980-7839

A subset of this work was presented at the 2021 Meeting of the Cognitive Science Society and published in the 2021 Proceedings of the Cognitive Society (Cusimano et al., 2021). This project was made possible through the support of a grant from the Templeton World Charity Foundation. The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the Templeton World Charity Foundation. The authors would like to thank members of the Princeton University Concepts and Cognition Lab, members of the Princeton University Center for Human Values, and Clayton Critcher and Megan Saylor for their valuable feedback on this project. David Danks was at Carnegie Mellon University, and Corey Cusimano was at Princeton University, when some experiments were conducted. All preregistrations, materials, data, analysis scripts, and supplemental materials are available at https://researchbox.org/398.

Corey Cusimano served as lead for conceptualization, formal analysis, methodology, visualization, writing–original draft, and writing–review and editing and served in a supporting role for funding acquisition. Natalia Zorrilla served in a supporting role for methodology and writing–review and editing. David Danks served as a lead for funding acquisition. Tania Lombrozo served as a lead for funding acquisition and in a supporting role for methodology. David Danks and Tania Lombrozo contributed equally to supervision and writing–review and editing.

Correspondence concerning this article should be addressed to Corey Cusimano, School of Management, Yale University, 165 Whitney Avenue, New Haven, CT 06511, United States. Email: corey.cusimano@yale.edu

In the current article, we propose and test a theory of how people reason about others' psychological freedom and constraint. Specifically, we propose that everyday judgments about psychological freedom primarily draw on one component of people's naive theory of mind, namely, their naive theory of reasoning. Across six experiments, we show how our theory predicts judgments about others' control over their beliefs, desires, intentions, and intentional behavior, and how it explains extant puzzles in everyday attributions of freedom and constraint.

Prior Work on Attributions of Freedom and Constraint

There is general agreement about some of the conditions under which people think others lack control over their mental states or behavior. People believe that someone who lacks the capacity to think rationally is, by virtue of this incapacity, unable to think and act freely (Alicke, 2000; Gray et al., 2007; Malle, 2019; Nelkin, 2011; Weisman et al., 2017; Wolf, 1990). Likewise, people seem to lack freedom when physical forces override what they want to do, or when their bodies move in ways that do not reflect their choices (Murray & Lombrozo, 2017; Woolfolk et al., 2006). But there is widespread disagreement about whether, and on what grounds, people think of others as psychologically constrained in the sense that their situation prevents them from behaving in certain ways (even if they possess rational capacities and are physically unconstrained). Resolving how people think about constraint in this sense is necessary to understand how people reason about coercion, manipulation, situational pressure, and moral responsibility.

According to one view, the "no constraint" view, people do not treat situational pressure as a genuine constraint on others. For example, when thinking about the person who sells their kidney, this view predicts that people will believe that this person did so freely and could have chosen not to. The fact that they would starve to death without the money does not prevent them from doing otherwise. Indeed, according to no constraint theories, people think that others have the capacity to change their mind however they please, including in highly arbitrary, irrational, and self-destructive ways (Aristotle, 1985; Kalish, 1998; Kushnir et al., 2015; Reeder, 2009). For instance, Kalish (1998) and Reeder (2009) argue that people distinguish between laws about what is "impossible" and laws about what is "impermissible" and suggest that these two laws "involve two types of conformity: 'automatic' when conformity is physically necessary, and 'voluntary' when conformity is enjoined but optional" (p. 706, Kalish, 1998). Kushnir et al. (2015) likewise assume that people believe that "our free choice may be impossible because of [physical constraints] but … we can exercise our free choice even when our options are not equally desirable" (p. 81). Consistent with this view, people often report that others can think and act in destructive and irrational ways if they want to (e.g., Cusimano & Goodwin, 2019).

A challenge for these "no constraint" theories is that they do not explain intuitions about coercion and other cases of extreme situational pressure. Many people cite coercion as a threat to free will (Monroe & Malle, 2010; Woolfolk et al., 2006). And in law, defendants who acted under duress, or who faced life-threatening situations, are excused for breaking the law on the grounds that they were unfree (P. H. Robinson, 1997). Extreme cases of situational pressure have a common feature, namely, that they seem to present people with only one good choice. Consider for instance someone in a high-pressure situation, such as a ship captain caught in a storm who can only prevent the ship from sinking by throwing something overboard (Phillips & Knobe, 2009; Young & Phillips, 2011). When that person takes the only good option available—throwing some cargo overboard—observers tend to say that they could not have done otherwise (even though bad options, like refusing to save the ship, or saving the ship by throwing people overboard, were physically available; Phillips & Knobe, 2009; Young & Phillips, 2011). Judgments of this kind appear to extend broadly to how people think about mental state formation. For instance, people think that others have less control over what they believe when their beliefs are strongly recommended by the evidence (e.g., Cusimano & Goodwin, 2019, 2020). These data present a challenge to theories according to which people view others as always free: When someone only has one good choice, it seems like they have no choice at all.

Some recent work has tried to explain these intuitions about constraint by appeal to quirks of cognition such that attributions of control are contaminated by normative judgments. According to one proposal, people heuristically replace the question of what someone "can" do with the question of what that person "ought" to do (Phillips & Cushman, 2017; Phillips & Knobe, 2009). According to another, people overlook or disregard irrational options (but not rational options) when they think about the different ways someone can act (Phillips et al., 2015). And according to others, people conflate what they think others "can" and "should" do as a result of motivated reasoning: they want others to be able to behave how they (as observers) prefer because they (as observers) want to hold people responsible for behaving poorly (Clark et al., 2014, 2019, 2021; Everett et al., 2021; but see Monroe & Ysidron, 2021). The unifying feature of all of these theories is that attributions of control are moralized such that observers draw a direct connection between the quality of someone's decision and that person's agency in that decision.[1] Accordingly, people treat rational options as "fundamentally open" (such that people are free to choose them) and irrational options as "fundamentally closed" (such that people are constrained from choosing them; Phillips & Knobe, 2009, p. 35). Evidence for these theories comes from the observation that people do not think that situations uniformly constrain people's capacity to think and act. Even though people judge the ship captain as unable to do otherwise when he throws cargo overboard, they judge him as able to do otherwise when he (immorally and irrationally) throws a person overboard instead (Phillips & Knobe, 2009).

These two families of theories expose conflicting intuitions in everyday reasoning about psychological freedom and constraint. "No constraint" theories explain the observation that people often judge others as capable of behaving irrationally. But these theories underpredict attributions of psychological constraint: people also sometimes judge that someone who lacks good options thereby lacks the ability to behave freely. By contrast, theories

[1] According to these theories, intuitions about freedom reflect egocentric reasoning: the actions that the observer judges as bad are ones that people have less capacity and freedom to do, while the actions that the observer judges as good are ones that others have more capacity to do. Our proposed model posits that agency judgments reflect allocentric reasoning: attributions of freedom and constraint reflect judgments regarding what reasons, from the perspective of the target, rationalize the mental state or behavior.

that moralize agency account for commonplace judgments about coercion (and related phenomena). But these theories overpredict attributions of psychological constraint: judgments of rationality and control are sometimes not conflated, such as when people say that others can form irrational mental states or act contrary to what they think would be good for them.

The complementary successes and failures of these two views illuminate the main puzzle that we aim to resolve. Any theory of how people reason about psychological freedom and constraint must explain why people sometimes judge others to be constrained in light of what they have strong reasons to do, while also predicting when, and explaining why, people sometimes think others are not constrained by rationality in this way. Finally, while most prior work has focused on people's ability to intentionally control their behavior, the same tension reveals itself in everyday judgments about mental states like beliefs and desires. For example, the person in our introductory example is not physically prevented from desiring starvation over kidney donation, but also (intuitively) seems unable to do so. Any theory of how people think about freedom and constraint must explain both how people evaluate intentional behavior and how they evaluate mental states like beliefs and desires.

The Naive Theory of Reasoning

We propose that intuitions about psychological freedom and constraint largely invoke one component of everyday theory of mind which we call the naive theory of reasoning. Here we outline its main features, demonstrate how it resolves the puzzle identified above, and derive the main predictions that we test.

In lay theory of mind, "reasoning" is the mental process that takes reasons as inputs and outputs new mental states such as beliefs, desires, and intentions. We propose that people conceptualize reasoning as having two important properties: First, people believe that (unless defective) reasoning produces beliefs, desires, and intentions that are rational[2] in light of the reasons that enter into reasoning. Put another way, people expect the output of others' reasoning to reflect whatever is subjectively rational for that person, given the reasons considered by that person at the time. And second, people believe that reasoning is a semi-autonomous process such that people can choose whether to initiate reasoning (and thus initiate mental state change), but they cannot directly override or alter the mental states that reasoning produces. These two properties jointly entail that, when people think about others forming mental states through reasoning, they expect this process to uncontrollably eliminate mental states that would be irrational (from the perspective of the reasoner) and replace them with mental states that would be rational (from the perspective of the reasoner; Figure 1A).

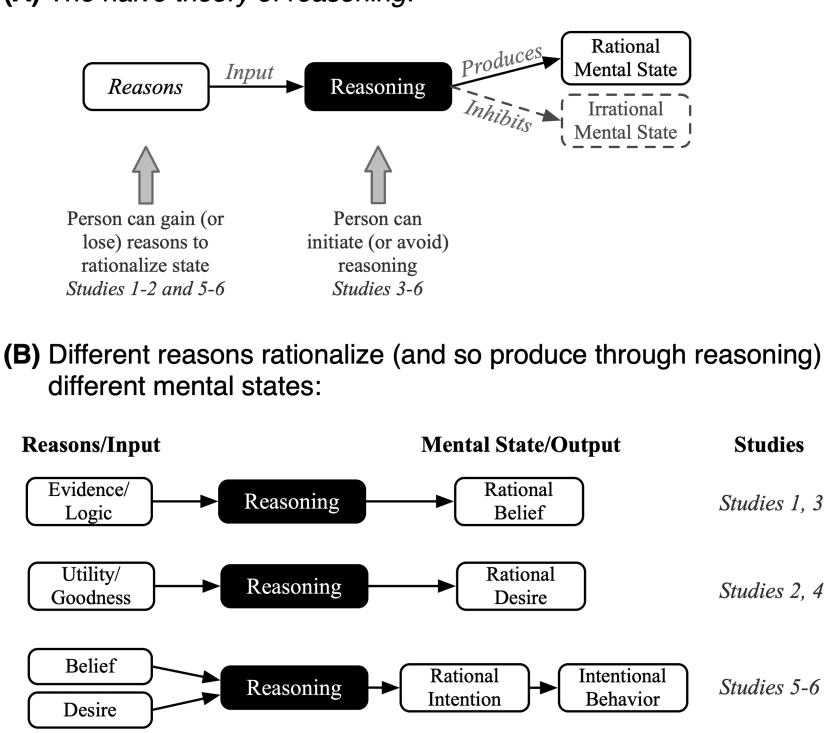
The proposed naive theory of reasoning explains why people think that others are more capable of thinking and acting rationally compared to irrationally. On this account, people are free to think and act rationally because all they have to do is reason. People should expect the process of reasoning, once it is initiated, to automatically output beliefs, desires, and intentions that are rational based on the reasons consciously available to the reasoner. However, people should judge it to be difficult or impossible for someone to form irrational mental states, or make irrational choices, by thinking about what to believe, desire, or do. The differential ease of adopting rational and irrational mental states through reasoning explains why people are often seen as free to do the former but constrained from doing the latter. Situations involving coercion (or other forms of pressure) appear to constrain others because they limit which mental states are rational, and thereby limit which states can be produced by reasoning. We thus predict in the studies below that judgments about whether someone can voluntarily adopt a new mental state should be commensurate with how rationalizable that state is (based on the reasons to which that person has access).

The naive theory of reasoning also predicts when, and explains why, people will think others are unconstrained. On our proposal, people believe that the constraints of rationality apply specifically to the process of reasoning rather than more broadly to a person or to all their mental processes. Other reactions that someone might have to a situation invoke mental processes related to attention and memory and include avoiding thinking, searching memory for specific information, and forgetting or suppressing information. Here we focus on two specific mental maneuvers (Figure 1A): first, people can react to situations by selectively thinking, and second, people can try to rationalize particular states by seeking or suppressing conscious access to certain reasons. Because people identify reasoning as the source of rational constraint, people should judge others as capable of irrational behavior when they think others can react and manipulate reasoning in either or both of these ways. In the studies below, we test this prediction by measuring perceived mental state control in situations in which the target person can selectively forget or suppress information (Studies 1–2) or choose whether they scrutinize their extant irrational attitudes (Studies 3–6).

Finally, the naive theory of reasoning applies to all products of reasoning, including (but not limited to) beliefs, desires, intentions, and intentional behavior (Figure 1B). However, the source and character of psychological constraint change across different mental states because different mental states are sensitive to different kinds of reasons (D'Andrade, 1987). For instance, people treat belief formation as primarily sensitive to logic and evidence, preference and desire formation as primarily sensitive to goodness, and intention formation as primarily a product of individuals' beliefs and desires (e.g., Gergely & Csibra, 2003; Jara-Ettinger et al., 2016; Malle, 1999, 2004). We hypothesize that people treat evidence and goodness not only as grounds for predicting and inferring mental state formation (as established by prior work), but also as rational constraints on the reasoning process that generates those states. Accordingly, we predict that people consider the evidence available to someone when determining what beliefs they can form (Studies 1 and 3), consider what people have reason to think about which options are good or bad when determining which desires others can form (Studies 2 and 4), and consider

[2] "Rational" means different things for different types of mental states (e.g., beliefs vs. desires). We conjecture that people view belief formation as constrained by "epistemic" or "intellectual" rationality such that beliefs seem easier to form when they are supported by evidence and logic. By contrast, we conjectured that people view desire formation as constrained by "practical" or "instrumental" rationality. Such that desires seem easier to form when someone has reasons to think that something is good. In general, we use the terms rationality and reason without qualification since the context will always disambiguate: epistemic rationality (or epistemic reasons) when applied to belief formation, and instrumental rationality (or instrumental reasons) when applied to desire formation.

Figure 1
The Naive Model of Reasoning and Its Application to Common Mental States and Intentional Behavior



Note. (A) Reasoning is constrained in that it only enables rational mental states and inhibits irrational mental states. People can sometimes flexibly manipulate reasoning, but features of the situation may also limit their ability to do so. (B) Consciously available epistemic reasons enable and constrain belief formation. Goodness/utility enable and constrain desire formation. Beliefs and desires enable and constrain intention formation (and as a result, intentional behavior).

which beliefs and desires others have (or can lose or adopt) when determining which intentions others are capable of forming (Studies 5–6).

The Current Studies

Six studies, reported below, support predictions derived from the proposed naive theory of reasoning. In accordance with this theory, people think that others can easily adopt beliefs (through reasoning) when they possess epistemic reasons (like good evidence) that make it rational to do so. However, people think that others cannotadopt beliefs (through reasoning) when it would be epistemically irrational to do so unless they can manipulate what would seem (to them) to be epistemically rational to believe (Study 1). Likewise, people think that others can easily adopt desires when those others possess reasons that make the desires practically or instrumentally rational (i.e., when the desire would be for something that they can understand as being good or good for them). However, people also think that others are incapable of adopting a desire for something that they know would be bad for them unless they have some means of manipulating what (to them) appears to be good or bad for them (Study 2). Studies 3 and 4 demonstrate that people treat reasoning as both largely necessary and sufficient for rationality to enable and constrain mental state change. In Studies 5 and 6, we extend the naive theory of reasoning to judgments of someone's ability to form intentions and make intentional choices. In these studies, we demonstrate how the naive theory of reasoning explains commonplace attributions of freedom and constraint in situations involving duress and coercion, manipulation, and situational necessity. Taken as a whole, these studies document the precise ways in which people think situations psychologically constrain others: People think that features of a situation, as understood through evidence about its properties and the expected utility of different options, constrain belief and desire formation (Studies 1–4), and these states in turn constrain intention formation and intentional action (Studies 5–6).

Our primary interest in this project is to understand how people think about others' freedom and constraint in general terms. To this end, across studies, we operationalize perceived freedom and constraint in different ways. In Studies 1 and 2, participants rated how "easy or difficult" it would be for an individual to adopt a given mental state; in Studies 3 and 4, participants reported whether a character "can" adopt certain beliefs or desires; and in Studies 5 and 6, participants reported whether a character "has the ability to" form a certain intention and perform a certain action. Attributions of freedom and constraint are implicated in all of these judgments, and we verify our predictions across all of these studies.

Transparency and Openness

All studies reported in this article were preregistered. All study materials, preregistrations, data, and code are available on ResearchBox (see link in Appendix A). All studies reported in this article were approved by the Office of Research Ethics at Princeton University and Yale University Institutional Review Board (IRB). All studies were conducted on Prolific. We limited recruitment to people currently living in the United States with at least 100 prior completed tasks and an approval rating ≥95%. We further removed participants who failed simple attention checks (see details below).

Studies 1–2: Reasons Enable and Constrain Mental State Change Through Reasoning

Studies 1–2 examined how observers think about the role of reasons and reasoning in enabling someone to adopt a new belief (Study 1) or desire (Study 2). According to the naive theory of reasoning, participants should believe that a person can easily adopt a new belief or desire if that person has reasons that rationalize that mental state and they can reason. However, when people lack rationalizing reasons, either because no such reasons are available in the person's environment or because they lost or suppressed conscious access to those reasons, then participants should think that it will be extremely difficult for that person to adopt the desired mental state. To test these predictions, we manipulated whether a character obtains new information that rationally favors some desired belief or desire and then asked participants to judge how easy or difficult it would be for that character to form the desired belief or desire through a specified process.

We predicted that participants would judge reasoning as an effective process for changing one's beliefs or desires when the character gained new information that rationally favored the new state. We also predicted that if the character did something that inhibited their conscious access to information (i.e., forgetting or not thinking about it), then it would seem harder for them to adopt the desired state even if the facts available in their environment still favored that state. Finally, Studies 1–2 test the naive theory of reasoning across beliefs and desires, which we hypothesized would be seen as responsive to, and constrained by, different kinds of reasons. In Study 1, we test whether people believe that belief change is enabled by strong evidence. In Study 2, we test whether people believe that desire change is enabled by information that an action or outcome is good for the character.

Method

Participants

Data collection for Studies 1 and 2 occurred simultaneously and participants were randomly assigned to either Study 1 or 2. For Study 1, we recruited 599 participants. After excluding participants who failed at least one attention check, our final sample for Study 1 comprised 570 participants (44% reported female, 54% reported male, 1% unreported or other, $M_{age} = 35$ years). For Study 2, we recruited 601 participants. After attention check exclusions, our final sample for Study 2 comprised 545 participants (48% reported female, 51% reported male, 1% unreported or other, $M_{age} = 34$ years).

Design and Vignette Construction

Studies 1 and 2 each used a 2 (reason strength: weak vs. strong) × 2 (reaction type: reasoning vs. suppression) × 4 (vignette) mixed between-within design.

In both studies, participants read a vignette about a character who forms both a mental state—either a belief (Study 1) or a desire (Study 2)—and a desire to not possess that mental state. For instance, in Study 1, some participants read about a ship captain who forms the belief that he will soon be caught in a terrible storm. This belief makes him anxious and, on that basis, he wishes that he did not believe that he will soon be caught in a terrible storm. In Study 2, the ship captain forms a desire to turn his ship away from the imminent storm. At the same time, he wishes that he were braver (like the captains that he admires) such that he wanted to pilot through the storm.

In both studies, participants learned that the character gains new information that provides either a strong reason or a weak reason to change their mind (reason manipulation: weak vs. strong). In Study 1, participants read that the target received either strong evidence or weak evidence in favor of their desired belief. For instance, in the "Storm" scenario, participants read that the ship captain hears a weather station report that there will be clear weather. In the strong reason condition, the captain has prior evidence that the weather station is reliable. In the weak reason condition, the captain knows the weather station to be unreliable. In Study 2, the character learns new information that either signals that the outcome will confer high utility (strong reason condition) or does not signal anything new about the utility of the desired outcome (weak reason condition). For instance, Study 2 participants who were assigned to the strong reason condition read that the ship captain, who does not want to pilot his ship through the storm but wishes that he had that desire, hears over his radio that piloting his ship through the storm will net him a large bonus payment on his current job. Participants assigned to the weak reason condition read that the captain learns that there will be no bonus payment for piloting through the storm (and so the utility is unchanged).

All participants then attributed mental state control. There were two types of mental state control attributions based on two different kinds of reactions the character could have to the new information (reaction type manipulation). One reaction was to consider the new reasons ("reasoning" reaction). For instance, in Study 1, participants reported how easy or difficult it would be for the ship captain to change his belief if he "thinks about whether it is true or false that there will be a storm." In Study 2, participants reported how easy or difficult it would be for the captain to change his desire after he "thinks about the costs and benefits of piloting his ship into the storm." Both studies also included another "reasoning" item, which asked how difficult it would be if the character "considers all the information" that he or she has. This item was identical across Studies 1 and 2. The two reason items correlated highly with each other in both Study 1 ($r = .50$) and Study 2 ($r = .69$).

The second set of items probed attributions of mental state control via reactions that suppress access to information in the environment ("suppression" reaction). Accordingly, one item asked participants to rate how difficult or easy it would be for the captain to change his belief or desire if he "forgets" what he just heard. The second item asked participants to make a similar judgment if the character "does not think about" what he or she just heard. Both of these reactions stop the

characters' reasons for mental state change from entering into their reasoning about what to believe or desire. And moreover, they do so without modifying any other features of the vignette (including whether it is generally rational or not to adopt the new mental state). The two suppression items correlated highly with each other in both Study 1 ($r = .69$) and Study 2 ($r = .68$).

Participants were randomly assigned to read one of four vignettes that instantiated this design in different contexts (vignette manipulation). In Study 1, vignettes described situations involving (a) a captain's beliefs about a storm, (b) a new employee's beliefs about the health benefits of working out, (c) a student's beliefs about their upcoming grade on a paper, and (d) a cop's beliefs about whether he will be picked to go undercover. In Study 2, participants read about (a) a captain's desire to pilot his ship into bad weather, (b) a new employee's desire to go to the gym, (c) a student's desire to do homework, and (d) a cop's desire to go undercover. See the Supplemental Materials, linked in Appendix A, for the full text of all eight vignettes. Figure 2 provides a schematic overview of this design and procedure.

Procedure

At the beginning of the study, participants were randomly assigned to either the weak or strong reason condition, and then to one of the four vignettes. After reading the setup of the vignette, participants answered two comprehension questions. The first question asked what mental state the main character currently holds (e.g., a belief that there will be a storm), and the second question asked what mental state the main character wished that they held (e.g., a belief that there will not be a storm). As preregistered, participants who answered either of these questions incorrectly were excluded from data analysis.

Participants then read that the character learned new information that either constitutes a strong or a weak reason to change their mind. On the next page, participants answered two manipulation checks that tested whether the reason strength manipulation was successful. In Study 1, participants responded to two questions about the amount of evidence that the character now had for the desired belief, including "How much evidence does Jeremiah have that he will not be caught in a severe storm?" and "How much do you agree or disagree with the following statement? Based on how reliable the weather station has been, Jeremiah will probably not be caught in a severe storm." In Study 2, participants responded to two questions about the utility of the desired outcome, including, "How much better off will Jeremiah be if he pilots his ship through the storm?" and, "How much do you agree or disagree with the following statement? If Jeremiah pilots his ship through the storm, he has a chance to make things better off for himself." Participants answered all questions using 7-point rating scales. Questions were shown in a random order for each participant.

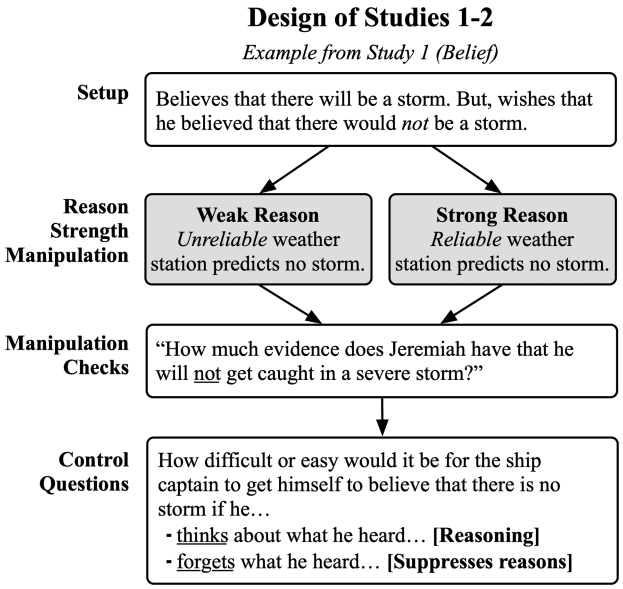
Participants then attributed control to the character over their mental states. Across both studies, participants read that the main character still held the undesired attitude (belief or desire) and that the character still wished that they held the opposing attitude. For instance, in Study 1, participants read:

> Jeremiah still believes that he will be caught in a severe storm. However, Jeremiah still also prefers not to have this belief, but instead to have the belief that he will not get caught in a severe storm. In the following questions, we want to know how easy you think it would be for Jeremiah to adopt that belief.

Participants reported the ease or difficulty of changing one's mind based on four reactions the character could have. These reactions included two "reasoning" reactions and two "suppression" reactions, described in detail above. Participants answered these questions using 7-point rating scales anchored at 1 (*extremely difficult*) and 7 (*extremely easy*). The order of the four control questions was randomized for each participant. Study 2 used nearly identical language in the prompt and control measures but adapted for desires (see the Supplemental Materials, linked in Appendix A).

Lastly, participants reported sex and age and were debriefed about the study.

Figure 2
Schematic Overview of Design of Studies 1–2



Note. Example text is drawn from the "Storm" vignette from Study 1 (Belief). Dark gray boxes represent between-subjects manipulation.

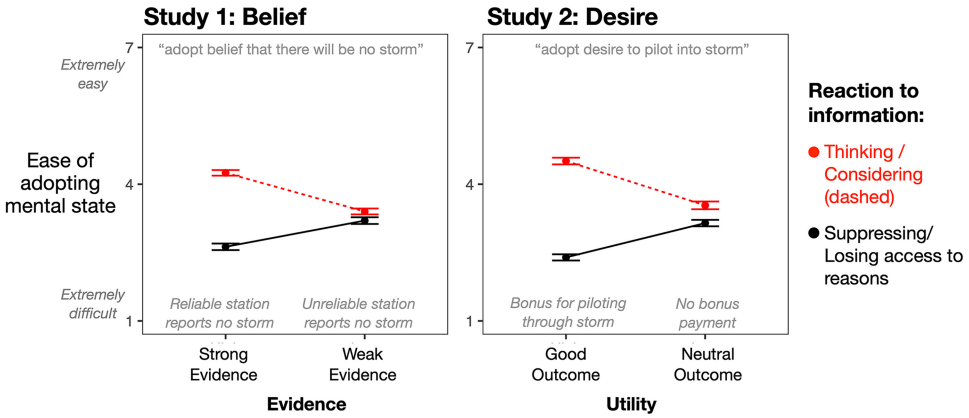
Results

Our reason strength manipulation was effective. We created composite evidence ($r = .60$) and utility ($r = .83$) ratings and submitted these ratings to 2 (reason strength) × 4 (vignette) fully crossed analyses of variance (ANOVAs). The reason strength manipulations were effective in both Studies 1 and 2. In Study 1, participants judged that the characters had more evidence favoring the desired belief in the strong evidence condition ($M = 4.85$, $SD = 1.25$) compared to the weak evidence condition ($M = 2.74$, $SD = 1.14$), $F(1, 727) = 612.75$, $p < .001$. And in Study 2, participants judged that the target behavior was associated with higher utility in the strong reason condition ($M = 5.72$, $SD = 1.19$) compared to the weak reason condition ($M = 3.48$, $SD = 1.91$), $F(1, 684) = 587.49$, $p < .001$.

We then analyzed participants' control judgments. To do this, we averaged together the two "reason" reaction items, and separately, the two "suppression" items. We then submitted participants' belief and desire control judgments to 2 (reason strength: weak vs. strong) × 2 (reaction: reasoning vs. suppression) × 4 (vignette) mixed within-between fully crossed ANOVAs (see Figure 3). Below we

Figure 3

Results From Studies 1 and 2



Note. Mean (and standard error of the mean) control judgments for belief (1) and desire (2) across weak and strong reason conditions, and reasoning versus intentional forgetting conditions. Sample stimuli are displayed in smaller font. See the online article for the color version of this figure.

report results for two key predictions: First, when reporting control via reasoning, participants should attribute greater control to the characters in the strong reason condition compared to the weak reason condition, and second, in the suppression condition, the difference between the strong and weak reasons conditions should attenuate. See the Supplemental Materials, linked in Appendix A, for full model output.

Study 1: Belief

We observed the predicted Reason Strength × Reaction interaction, $F(1, 727) = 147.84$, $p < .001$. When judging control via reasoning, participants attributed greater control in the strong evidence condition ($M = 4.25$, $SD = 1.18$) compared to the weak evidence condition ($M = 3.40$, $SD = 1.24$), $F(1, 727) = 87.51$, $p < .001$. However, this relationship reversed in the suppression condition: Participants now attributed lower ability to adopt the desired belief in the strong evidence condition ($M = 2.63$, $SD = 1.38$) compared to the weak evidence condition ($M = 3.20$, $SD = 1.40$), $F(1, 727) = 33.34$, $p < .001$. Put another way, participants thought it would be extremely difficult for the character to adopt their desired belief by reasoning when they lacked evidence favoring the belief, or by suppressing evidence that rationally favored the belief.

Study 2: Desire

We observed a similar pattern of results for desire. First, we observed the predicted Reason Strength × Reaction interaction, $F(1, 684) = 164.47$, $p < .001$. When judging control through reasoning, participants attributed greater control in the good outcome condition ($M = 4.51$, $SD = 1.37$) compared to the neutral outcome condition ($M = 3.54$, $SD = 1.75$), $F(1, 684) = 82.50$, $p < .001$. However, this pattern reversed in the suppression condition: participants now said it would be much more difficult to adopt the wanted desire in the strong reason condition ($M = 2.40$, $SD = 1.26$) compared to the weak reason condition ($M = 3.15$, $SD = 1.31$), $F(1, 684) = 57.29$, $p < .001$. In other words, and as in Study 1, participants thought it would be extremely difficult for the character to adopt their wanted desire by reasoning when their information suggested the outcome would be bad for them, or when suppressing any information that suggested otherwise.

The Relationship Between Reason Strength and Control Across Studies 1 and 2

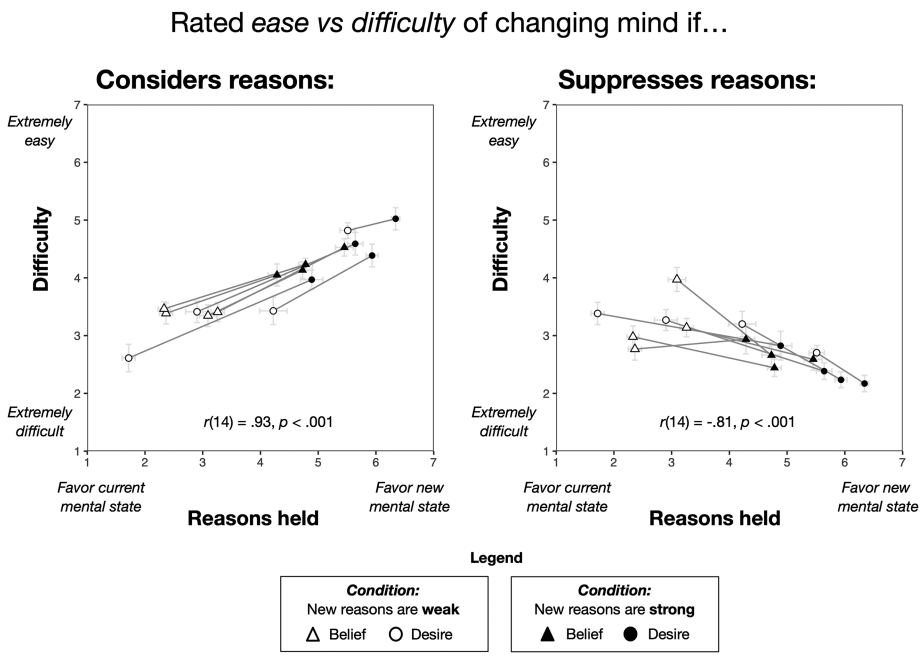
The analyses reported above demonstrate two things. First, manipulating perceived reason strength affects perceived mental state control. Second, manipulating perceived reason strength affects perceived control via reasoning differently compared to control that might be enacted through other reactions (like suppressing thoughts). These analyses do not address another claim of our theory. We conjectured that people think that the rationality of mental state change is the primary consideration when evaluating whether someone can change their mind through reasoning. Accordingly, people's judgments about how rational it would be to change one's mind should be a very strong predictor of their attributions of mental state control. One way to test this claim is to take advantage of the fact that reason strength (as measured by our manipulation checks) varied substantially across vignettes and conditions. Attributions of mental state control should also vary substantially across vignettes and conditions by closely tracking perceived reason strength. To test this prediction, we averaged participants' judgments of reason strength (i.e., information about evidence or utility as measured by our manipulation checks) and their attributions of mental state control for each condition and for each vignette for both belief (Study 1) and desire (Study 2). This resulted in 16 different average reason strength and control judgments—displayed in Figure 4.[3]

Examining our findings this way provides additional evidence for our model. Across cells, attributions of control via thinking nearly perfectly correlated with how well those reasons rationalized the desired mental state, $r(14) = .93$, $p < .001$. Indeed, when the reasons very weakly favored the desired proposition, as they did in the "Storm" scenario when the ship captain wished that he could want to pilot into the

[3] Because we collected data for Studies 1 and 2 at the same time, participants were effectively randomly assigned to one of these 16 cells (4 vignettes × 2 reason strength × 2 belief vs. desire).

Figure 4

Mean Judgments of Reason Strength (x-Axis) and Mental State Control (y-Axis) for Each Condition of Each Vignette in Studies 1 (Triangles) and 2 (Circles)



Note. Gray lines connect the "weak" and "strong" reason conditions of the same vignette: Slope of line represents how the change in perceived reason strength related to change in perceive control within that vignette. Data are divided by the character's reaction to the new information: reasoning (left panel) and suppressing/losing reasons (right panel). Error bars correspond to standard errors.

storm but lacked any reason to ($M = 1.90$), participants on average judged that it would be extremely difficult for him to change his desire ($M = 2.40$). But when the character's reasons strongly favored their desired mental state, as they did in the "Gym" scenario when the character wanted to want to go the gym and knew that it would improve his health ($M = 6.35$), participants thought it would be very easy for the character to change his desire ($M = 5.02$).[4] This relationship between reason strength and control reversed when participants contemplated the character suppressing their reasons, $r(14) = -.74$, $p < .001$. For instance, participants thought it would be extremely difficult for the character to adopt his desired gym desire if he suppressed the reasons that rationalized it ($M = 2.29$).

Replications and Extensions

Studies 1 and 2 test our proposed model by manipulating the strength of reasons (evidence or utility) available to a person and then measuring attributions of mental state control via reasoning or reactions that suppress conscious access to those reasons. We have conducted several other studies that use this paradigm and that replicate and extend the findings reported above. We report these replications in the Supplemental Materials, linked in Appendix A, and briefly describe them and their theoretical significance here.

Studies S1 and S2 replicated Studies 1 and 2 with another way that the character could suppress their conscious access to reasons, namely, by taking a pill that would immediately and effectively cause them to forget the (weak or strong) reasons they just learned. Taking a pill to forget information had the same impact on the character's perceived capacity to change their mind as naturally forgetting the information and actively suppressing their good reasons.

Studies S3 and S4 extended Studies 1 and 2 by using four new vignettes and by changing the structure of the vignettes such that the characters start off with strong reasons to change their mind, and then learn either weak or strong inhibiting reasons against changing their mind. These studies replicate all the results above.

Using data combined from Studies S1–S4, Appendix B reports a replication of the relationship between average reason strength judgments and control judgments across vignette. We observe the same strong relationship between attributions of reason strength and mental state control via reasoning, $r(30) = .83$, $p < .002$, and the same attenuated, negative relationship between mental state control and reason suppression/loss, $r(30) = -.65$, $p < .001$.

Finally, Study S5 addresses one limitation of these studies, namely, their relatively simple construction. In Studies 1–2 and Studies S1–S4, evidence manipulations were paired with attributions of belief control (Studies 1, S1, and S3), and utility

[4] The same conclusions follow when analyzing the data in disaggregated form. In another set of analyses, we regressed attributions of control via reasoning on perceived reason strength (our manipulation check), mental state, vignette, and the interaction of these variables. These analyses revealed that, for every 1-point increase in reason strength, participants attributed 0.40 points ($SE = 0.02$, $t = 18.2$) more control. This association did not vary across beliefs and desires ($b = 0.20$, $SE = 0.20$, $t = 0.96$) or vignette ($ts < 1.53$). A similar analysis shows a decrease in perceived control via suppression ($b = -0.12$, $SE = 0.02$, $t = -4.83$), which similarly did not vary by mental state ($b = 0.12$, $SE = 0.24$, $t = 0.51$) or vignette ($ts < 1.10$).

manipulations were paired with attributions of desire control (Studies 2, S2, and S4). Thus, the relationship between the strength of evidence and belief, and between utility and desire, is especially salient to participants. In Study S5, we again manipulated evidence and utility across participants, but now participants rated controllability over both beliefs and desires. This study replicated the constraining impact of evidence on belief and utility on desire. Moreover, this study demonstrated a lack of impact of evidence on desire, and a lack of impact of utility on belief. The types of reasons that constrain mental state change are specific to the mental state in question.

Discussion

Studies 1 and 2, and their replications and extensions, confirmed several predictions of the naive theory of reasoning posited earlier. As expected, participants did not always think that the characters could easily adopt the beliefs or desires that they wanted to. Indeed, participants thought the characters would have considerable difficulty when they lacked reasons that rationalized those states. However, participants thought it would be relatively easy for the characters to adopt the belief or desire that they wished to adopt when that characters had reasons that rationally favored those states. In Study 1, participants judged that the character could more easily adopt their desired belief when they had evidence that supported doing so. And in Study 2, participants judged that the character could more easily adopt the desire that they wished to when the outcome was associated with high utility.

Studies 1 and 2 also confirmed our predictions that the perceived ease and difficulty of adopting the mental state additionally depended on what the characters did in reaction to those new reasons. Participants only thought it would be relatively easy for the character to adopt the desired state if they had conscious access to the reason they had that rationalize those states. The mere presence of good reasons in their environment did not suffice for participants to attribute high mental state control. Indeed, participants thought that it would be extremely difficult for the character to adopt their desired attitude if the character forgot or suppressed these reasons. In other words, in the naive theory of reasoning, good reasons to change one's beliefs or desires only enable (and inhibit) mental state change when those reasons are consciously available to the reasoner and so available as inputs to reasoning.

These findings support the proposed naive theory of reasoning over the two extant theories described above. First, against "no constraint" theories, participants did not report that others could always easily form the belief or desire they wanted to. Instead, participants' responses implied that they thought that people's desire for certain mental states has only an indirect impact on what states they hold. But second, participants did not simply substitute what they thought would be "easy" or "difficult" to do with what they thought would be rational to do. If participants had attributed control in this way, then they would have always said that the character could adopt the state when it would be rational for them to do so, and always difficult when it was irrational to do so. Instead, rationality only predicted control when participants considered the person trying to change their mind by reasoning about what to believe or desire.

In the studies that follow, we extend our results in two ways. In Studies 3 and 4, we provide additional evidence that the proposed naive theory of reasoning predicts when people will judge that others can hold rational versus irrational mental states. And then in Studies 5 and 6, we demonstrate how access—or a lack of access—to rationalizing reasons for what to intend and what to do affects attributions of control over intentions and action.

Studies 3–4: Reasoning Is Necessary and Sufficient to Rationally Constrain Mental States

Our starting point for Studies 3 and 4 was the observation that people encounter others who hold beliefs and desires that vary in their apparent rationality. For instance, in response to the same evidence, one person might hold an irrational belief that does not match the evidence, while another person holds a rational belief that matches it. Likewise, in response to the same options, someone might irrationally desire something bad or immoral, while another person might desire something good. The naive theory of reasoning makes clear predictions about how people judge another's capacity to change or maintain mental states given the rational and irrational attitudes that they already hold.

According to the naive theory of reasoning that we propose, reasoning is the primary psychological process that produces rational mental states and inhibits irrational mental states. Accordingly, people should judge others as able to adopt rational mental states as long as the target can think about the reasons that are available to them. And if someone stops and thinks about what to believe or desire, observers should judge that person as incapable of holding irrational beliefs and desires because the mere act of reasoning should inhibit those irrational states. These predictions apply whether the target's current attitudes are rational or irrational. However, we propose that observers should make completely different judgments about what mental states others can hold if those individuals refuse to think. When someone does not think, they should seem capable of keeping whatever mental state they currently hold, whether that state is rational or irrational. For instance, people with irrational mental states should seem capable of keeping those states because only reasoning replaces irrational mental states with rational ones. Finally, we predicted that refusing to reason would also affect which mental states observers thought that others could not hold. Accordingly, people who currently possess irrational states and refuse to reason should seem incapable of adopting rational mental states because reasoning is the only process that enables rational mental state change. And because reasoning is the main process of mental state change in general, people who currently possess rational states, who refuse to reason, should seem incapable of adopting irrational states.

We tested these predictions using four new vignettes. One vignette is based on the "Storm" scenario used in Studies 1 and 2. The other three describe situations that impose pressure due to financial hardship (the "Medicine" scenario) or a poor job market (the "Job" scenario), or involve one person coercing another (the "Mugging" scenario). Thus, Studies 3–4 not only provide additional support for the naive theory of reasoning in several new scenarios, but also demonstrate how this theory explains people's judgments about others in canonically coercive and constraining situations.

Study 3: Reasoning to Change (or Keep) One's Rational or Irrational Belief

Method

Participants

We recruited 789 participants. After excluding participants who failed at least one attention check, our final sample for Study 3

comprised 721 participants (54% reported male, 44% reported male, 1% unreported or other, $M_{age} = 35$ years).

Design and Vignette Construction

Study 3 comprised a 2 (current belief rationality; between participants) × 2 (target belief rationality; within participants) × 2 (process: reasoning vs. avoiding reasoning; within participants) × 4 (vignette; between participants) design. At the beginning of the study, participants were randomly assigned to either the irrational current belief or rational current belief condition and then to one of four vignettes. We describe each of our manipulations below. A schematic of this design is provided in Figure 5 (left panel).

Setup and Current Belief Manipulation. All participants read scenarios featuring a character in a constraining situation. All scenarios were designed such that there was clearly one rational thing to believe and clearly one rational preference to have. For instance, in the "Storm" vignette (adapted from the same vignette used in Studies 1–2), the captain is piloting an old rickety boat and knows that there is a storm approaching. The captain has strong evidence that it would be dangerous for him to pilot his boat into the storm, as his boat is old, has not been repaired in a long time, is clearly damaged and worn, and takes on water even in ideal boating conditions. In all scenarios, the character "immediately and spontaneously" forms a belief. Half of participants read that the character formed a belief that made sense in light of the information available (current belief: rational). The other half of the participants read that the main character adopts a belief that violates the strong evidence that they have (current belief: irrational). For instance, the captain either formed the belief that it would be dangerous to pilot his boat into the storm (current belief: rational), or that he would be safe piloting his boat into the storm (current belief: irrational). Manipulation checks verified that, in all conditions and vignettes, participants considered the character's rational belief to be more rational than the character's irrational belief.

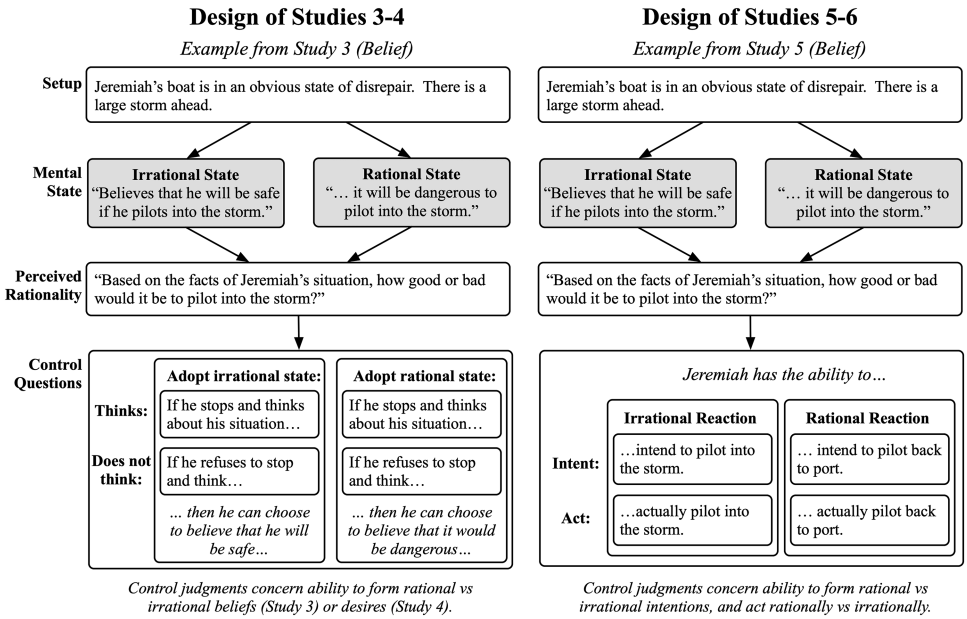
Target Belief and Process Manipulations. Participants reported whether they thought the character could hold either of two beliefs, reflecting a within-participants "target belief rationality" manipulation. The two target beliefs were identical to the two beliefs that comprised the rationality manipulation. So for instance, in the Storm vignette, one target belief was that it "would be dangerous to pilot into the storm" (rational target belief) while the other was that it "would be safe to pilot into the storm" (irrational target belief).

Participants made two judgments for each target rational and irrational belief. The two judgments corresponded to two different reactions the character could have in the scenario and reflected our "control process" manipulation. The first reaction question asked participants to indicate whether they agreed or disagreed that the target can adopt the target belief "if [he] stops and thinks about [his] situation" (thinking reaction). And the second reaction question asked participants to indicate whether they agreed or disagreed that the target can adopt the target belief "if [he] refuses to stop and think about [his] situation" (no thinking reaction).

Vignettes. We replicated this design across four vignettes. The "Storm" vignette, in which someone's life is threatened by natural circumstances, describes an instance of situational constraint that would likely give rise to a necessity defense in court. The other vignettes reflected situations that are commonly discussed in legal debates about freedom, coercion, and exploitation. The "Mugging" vignette reflects a standard case of duress in which the character is threatened by an armed mugger who demands that she hand over her backpack. This person spontaneously forms either the belief that she is weaker than the mugger (rational) or that she is stronger than the mugger (irrational). The remaining scenarios described cases of exploitation and economic constraint. In "Job," the character

Figure 5

Design of Studies 3–4 (Left) and Studies 5–6 (Right)



Note. Example text is drawn from the "Storm" vignette from Study 3 and Study 5. Dark gray boxes indicate the between-subject "current mental state" manipulation.

works a terrible job but in a town with high unemployment. She forms the belief either that there are no other jobs available (rational) or that there are plenty of other jobs available (irrational). And in "Medicine," the character is making barely enough money to pay for food and rent when her daughter suddenly falls ill. She believes her doctor's prognosis that her daughter will only get better with medicine (rational) or believes that the doctor is wrong and that her daughter will get better without medicine (irrational).

Procedure

At the start of the study, participants read a description of the scenario establishing the details that make it unambiguous what a rational person would desire and believe. For instance, in the "Storm" vignette, the captain has strong evidence that his boat is in poor condition, knows that there is a storm descending, and knows that there is a safe harbor nearby. After reading the scenario, participants answered two questions that served as comprehension checks. As preregistered, participants who answered at least one of these incorrectly were excluded from data analysis. Participants then learned that the character spontaneously forms either the belief warranted by their evidence (rational belief condition) or the belief that violated their evidence (irrational belief condition). Participants were not supplied any explanation for why the character forms the rational or irrational belief. However, participants were told that the main character was about to make a decision based on the belief that they just formed. For instance, in the storm scenario, participants in the irrational belief condition read that, "Based on this belief, Jeremy is about to pilot his boat into the storm." We used this language to signal that the belief was sufficiently strong in both conditions to impact the character's choices.

On the following page, participants evaluated the quality of two different choices the character could make. This was our measure of the perceived rationality of the belief. Participants responded to two questions: "Based on the facts of X's situation, how good or bad would it be for X to…" (a) perform the behavior based on the irrational belief (e.g., "pilot his boat into the storm") and (b) perform the behavior based on the rational belief (e.g., "pilot his boat back to port"). Participants responded using 7-point rating scales anchored at 1 (*extremely bad*) and 7 (*extremely good*). These questions were presented in a random order. These manipulation checks played an important role. If participants spontaneously explained the character's irrational belief by positing some reason why it might be rational, then these rationalizations would affect what is rational to intend (and do), and so would confound judgments of what the character can intend (and do). However, these ad hoc rationalizations would reveal themselves in these judgments about what would be good or bad for the character to do. Thus, these questions allowed us to check whether participants rationalized the irrational belief that the main character spontaneously adopted. We use these manipulation checks a similar way in Studies 4–6, below.

Participants then reported what they thought about the character's ability to hold different beliefs. They made four judgments based on the 2 (current belief: rational vs. irrational) × 2 (target belief: rational vs. irrational) design described above. Specifically, participants reported their agreement or disagreement with statements that the character can form the (a) rational belief by thinking; (b) rational belief by refusing to think; (c) irrational belief by thinking; and (d) irrational belief by refusing to think. Participants responded using 7-point rating scales anchored at 1 (*completely disagree*) and 7 (*completely agree*). These items were always presented in pairs such that participants evaluated the ability to form one target belief first by thinking and then by refusing to think, and then made the same pair of judgments for the other target belief (in the same order). We counterbalanced order of target belief rating pairs (rational vs. irrational) across participants.

At the end of the study, participants reported their age and sex and were debriefed.

Results

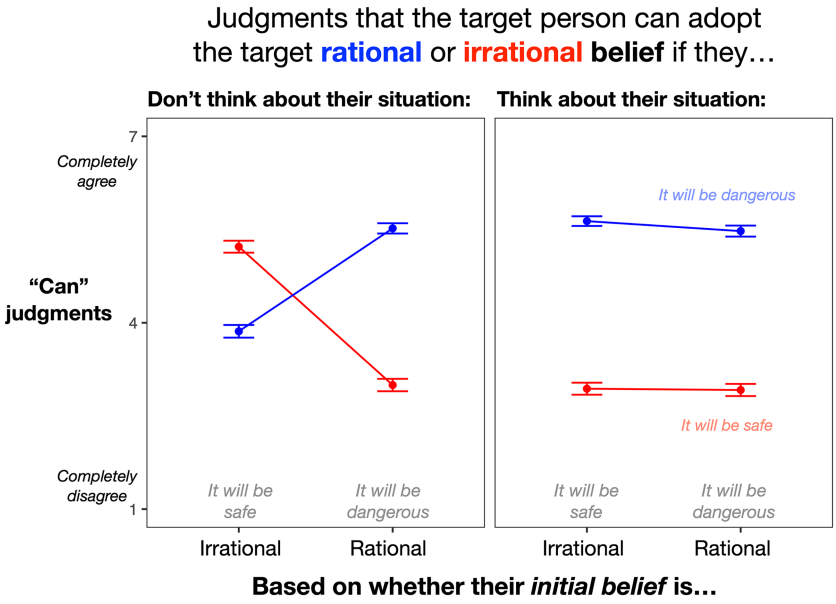
Our rationality manipulation operated as we intended. Participants thought that, based on the target's situation, it would be extremely bad for the character to act on their "irrational" belief ($M = 1.89$, $SD = 1.34$), but extremely good for them to act on their "rational" belief ($M = 5.72$, $SD = 1.53$), $F(1, 713) = 2{,}078.8$, $p < .001$, $\eta^2_G = .66$. The "current belief" manipluation did not significantly affect judgments of the irrational reaction, $F(1, 713) = 3.06$, $p = .081$, $\eta^2_G < .01$, or the rational reaction, $F(1, 713) = 3.09$, $p = .079$, $\eta^2_G < .01$. In other words, participants did not rationalize the character's irrational belief. This finding means that our description of the target's current belief as "irrational" or "rational" is apt, and none of the results below can be explained by participants rationalizing the target's current belief.

Participants' average control judgments across conditions are displayed in Figure 6. We submitted participants' judgments to a 2 (current belief) × 2 (target belief) × 2 (control process) × 4 (vignette) mixed-design ANOVA. As expected, participants' control judgments varied based on whether (a) participants were attributing freedom to hold a rational versus irrational belief (current belief manipulation), (b) the character started with rational or irrational beliefs (target belief manipulation), and (c) participants were considering the person thinking or avoiding thinking (control process manipulation). The three-way interaction between these factors was statistically significant: $F(3, 713) = 17.54$, $p < .001$, $\eta^2_G = .02$.

Participants' control judgments jointly depended on whether the character thought or avoided thinking, and whether the character's current belief was rational or irrational, $F(1, 713) = 257.53$, $p < .001$, $\eta^2_G = .08$. Participants thought that the character was more able to adopt rational beliefs when they thought ($M = 5.55$, $SD = 1.59$) compared to when they did not think ($M = 4.70$, $SD = 1.94$), $F(1, 713) = 117.64$, $p < .001$, $\eta^2_G = .07$. However, participants thought that others were more able to hold irrational beliefs by avoiding thinking ($M = 4.10$, $SD = 2.17$) compared to thinking ($M = 2.93$, $SD = 1.84$), $F(1, 713) = 216.87$, $p < .001$, $\eta^2_G = .10$. This finding supports our proposal that thinking seems necessary for rational mental states and is sufficient to inhibit irrational mental states.

We next examined control judgments separately for when the character thought about their situation and when the character avoided thinking about their situation. Turning first to the "thinking" condition (Figure 6, right panel), participants attributed more freedom to the character to adopt a rational state ($M = 5.55$, $SD = 1.59$) compared to an irrational state ($M = 2.93$, $SD = 1.84$), $F(1, 713) = 820.78$, $p < .001$, $\eta^2_G = .40$. Control attributions did not vary depending on which belief the target started with, $F(1, 713) = 1.19$, $p = .276$, nor was there an interaction between starting belief and target belief, $F(1, 713) < 0.01$, $p = .986$. When the target held a rational belief, and thought about their situation, participants thought they could keep their belief ($M = 5.48$, $SD = 1.69$) but not change it ($M = 2.92$, $SD = 1.86$), $F(1, 360) = 383.82$, $p < .001$, $\eta^2_G = .38$. We observed a similar difference in perceived ability to hold a

Figure 6
Results From Study 3



Note. Points represent average control judgments and error bars represent standard error around the mean. Blue points depict judgments that the character can adopt (or hold) a rational belief (e.g., "that piloting into the storm would be dangerous"). Red points depict judgments that the character can adopt (or hold) an irrational belief (e.g., "that piloting into the storm would be safe"). The left panel depicts control judgments when participants consider the character refusing to think about their situation. The right panel depicts control judgments when participants consider the character thinking about their situation. See the online article for the color version of this figure.

rational versus irrational belief when the target started off with an irrational belief and then thought about their situation. In this condition, participants thought that the target could change their belief to be rational ($M = 5.61$, $SD = 1.47$), but if they started thinking, they could not keep their irrational belief ($M = 2.94$, $SD = 1.83$), $F(1, 353) = 439.04$, $p < .001$, $\eta^2_G = .41$.

Turning next to when the character refuses to think, participants' judgments now jointly depended on whether the target belief was rational or irrational, and whether the current belief was rational versus irrational, $F(1, 713) = 352.65$, $p < .001$, $\eta^2_G = .23$. The predicted cross-over interaction, shown in the left panel of Figure 6, depicts this finding. When the character held a rational belief, participants thought that they could hold on to that belief ($M = 5.52$, $SD = 1.58$) more than they thought they could change it ($M = 3.00$, $SD = 1.90$), $F(1, 360) = 311.47$, $p < .001$, $\eta^2_G = .35$. But their ability to possess a rational versus irrational belief reversed when they held an irrational belief and did not think. Now, participants judged the character as more able to hold an irrational belief ($M = 5.22$, $SD = 1.82$) than a rational one ($M = 3.86$, $SD = 1.93$), $F(1, 353) = 83.04$, $p < .001$, $\eta^2_G = .12$.

Study 4: Reasoning to Change One's Rational or Irrational Desire

Method

Participants

We recruited 787 participants. After excluding participants who failed at least one attention check, our final sample for Study 3 comprised 726 participants (49% reported male, 48% reported male, 2% unreported or other, $M_{age} = 39$ years).

Design and Procedure

Study 4 used the same design, procedure, and set of vignettes as Study 3 but manipulated the rationality of the character's desire instead of the character's belief. That is, Study 4 comprised a 2 (current desire: rational vs. irrational) × 2 (target desire: rational target vs. irrational target) × 2 (process: thinking vs. refusing to think) × 4 (vignette) design. At the beginning of the study, participants were randomly assigned to one of the two "current desire" conditions and then to one of the four vignettes.

To manipulate current desire rationality, the character either "immediately and spontaneously" adopted a desire that, if fulfilled, would result in the best outcome for them (rational current desire condition) or "immediately and spontaneously" adopted a desire that, if fulfilled, would result in a terrible outcome (irrational current desire condition). For instance, in the rational desire condition in the "Storm" vignette, the captain cares more about his personal safety than about delivering his packages swiftly, and so forms the rational desire to pilot his boat back to port. In the irrational desire condition, the captain cares more about making a swift delivery than about his personal safety and forms the irrational desire to pilot his boat into the storm despite the significant costs. The characters in the vignettes always held rational beliefs (e.g., belief that the storm is dangerous) such that the only irrational mental state that the character possessed was the desire in the irrational desire condition. As in Study 3, the target desires were either the same (rational

or irrational) desire that the character had spontaneously formed, or the opposing (rational or irrational) desire. We also used the same ordering, randomization, and counterbalancing procedure from Study 3.

The other vignettes also matched those from Study 3. In "Mugging," the character spontaneously forms the desire to protect either their life (rational) or their bag (irrational). In "Job," the character forms the desire either to show up to work (rational) or to quit her job (irrational). And in "Medicine," the character forms the desire to care for either her daughter instead of her cat (rational desire) or her cat instead of her daughter (irrational desire).

Results

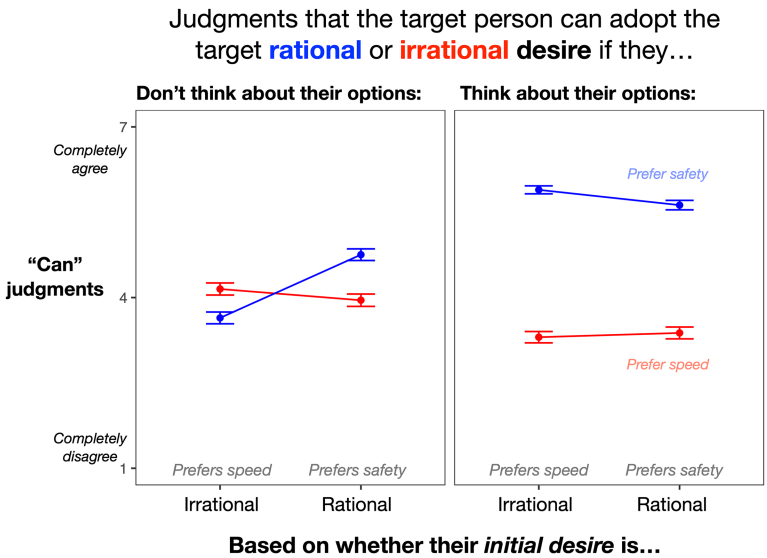
We successfully manipulated the perceived rationality of the "rational" and "irrational" current desire. Participants thought that, based on the facts of the character's situation, it would be bad for them to act on their "irrational" desire ($M = 1.85$, $SD = 1.23$), but good for them to act on their "rational" desire ($M = 5.65$, $SD = 1.60$), $F(1, 718) = 2{,}325.43$, $p < .001$, $\eta^2_G = .68$. Participants did not judge the irrational behavior differently based on whether the character adopted a rational ($M = 1.80$, $SD = 1.27$) or irrational ($M = 1.89$, $SD = 1.19$) desire, $F(1, 718) = 1.88$, $p = .171$, $\eta^2_G < .01$. However, unlike in Study 3, participants judged the rational target desire a little less positively when the target held an irrational ($M = 5.30$, $SD = 1.78$) compared to rational ($M = 6.00$, $SD = 1.29$) desire, $F(1, 718) = 45.93$, $p < .001$, $\eta^2_G = .06$. Even when the character spontaneously adopted an irrational desire, participants thought that the rational target desire was much better ($M = 5.30$) than the irrational target desire ($M = 1.89$), $F(1, 362) = 849.40$, $p < .001$, $\eta^2_G = .61$. Our results below replicate even if we restrict our analyses to only those participants who viewed the irrational desire as highly irrational (ratings < 3) and the rational desire as highly rational (ratings > 5). In this subsample, not only are the irrational and rational desires (necessarily) extremely different from each other, but there is also, as in Study 3, no difference in perceived rationality across the rationality manipulation. Despite some unintended variation in the perceived rationality of the character's desire, the results below cannot be explained by participants rationalizing the character's current desire, and our description of the character's current desire as "irrational" or "rational" is still apt.

Participants' average control judgments across condition are displayed in Figure 7. We submitted judgments to a 2 (current desire) × 2 (target desire) × 2 (control process) × 4 (vignette) mixed-design ANOVA. Again, control judgments reflected a complex pattern that depended on whether (a) participants were attributing the ability to hold a rational versus irrational desire (target desire manipulation), (b) the character started with rational or irrational desires (current desire manipulation), and (c) participants were considering the person thinking or avoiding thinking (control process manipulation). The three-way interaction between these factors was statistically significant: $F(1, 718) = 45.17$, $p < .001$, $\eta^2_G = .01$.

Participants' control judgments jointly depended on whether they considered the target adopting a mental state by thinking or avoiding thinking, and whether their current desire was rational or irrational, $F(1, 718) = 352.46$, $p < .001$, $\eta^2_G = .09$. Participants thought that others were more able to adopt rational desires when they reasoned ($M = 5.76$, $SD = 1.47$) compared to when they did not reason ($M = 4.19$,

Figure 7

Results From Study 4



Note. Points represent average control judgments; error bars represent standard error around the mean. Blue points depict judgments that the character can adopt (or hold) a rational desire (e.g., "caring more about safety than speed"). Red points depict judgments that the character can adopt (or hold) an irrational belief (e.g., "caring more about speed than safety"). The left panel depicts control judgments when participants consider the character refusing to think about their situation. The right panel depicts control judgments when participants consider the character thinking about their situation. See the online article for the color version of this figure.

$SD = 2.03$), $F(1, 718) = 390.85$, $p < .001$, $\eta^2_G = .18$. However, participants thought that others were more able to hold irrational desires by avoiding reasoning ($M = 4.05$, $SD = 2.05$) compared to reasoning ($M = 3.34$, $SD = 1.94$), $F(1, 718) = 52.3$, $p < .001$, $\eta^2_G = .03$. This result replicates Study 3 and supports our proposal that reasoning is largely necessary for rational mental states, and sufficient to inhibit irrational mental states.

We next examined control judgments separately for when the character is thinking about their options and when the character is avoiding thinking about their options. Turning first to the "thinking" condition (Figure 7, right panel), participants attributed greater ability to adopt a rational state ($M = 5.76$, $SD = 1.47$) compared to an irrational state ($M = 3.34$, $SD = 1.94$), $F(1, 718) = 678.12$, $p < .001$, $\eta^2_G = .34$. Control attributions did not vary depending on which desire the character started with, $F(1, 718) = 1.36$, $p = .243$, nor was there an interaction between starting desire and target desire, $F(1, 718) = 3.42$, $p = .065$. When the character held a rational desire, and thought about their options, participants thought they could keep their desire ($M = 5.62$, $SD = 1.58$) but not change it ($M = 3.38$, $SD = 1.97$), $F(1, 356) = 277.15$, $p < .001$, $\eta^2_G = .29$. We observed a similar difference in perceived ability to hold a rational versus irrational desire. In these conditions, participants thought that the character could change their desire to be rational ($M = 5.89$, $SD = 1.34$), but if they started thinking, they could not keep their irrational desire ($M = 3.30$, $SD = 1.91$), $F(1, 362) = 410.2$, $p < .001$, $\eta^2_G = .40$.

By contrast, when the character refuses to stop and think about their options, their ability to adopt a desire jointly depended on whether the target desire was rational or irrational, and how it related to their current desire, $F(1, 718) = 55.17$, $p < .001$, $\eta^2_G = .03$. The predicted cross-over interaction, shown in the left panel of Figure 7, depicts this result. When the character held a rational desire, participants thought that they could hold on to that desire ($M = 4.75$, $SD = 1.94$) more than they thought they could change it ($M = 3.34$, $SD = 1.94$), $F(1, 356) = 42.52$, $p < .001$, $\eta^2_G = .04$. Thus, when someone holds a rational desire, and they refuse to think, it seems like they are more able to stay rational than they are to become irrational. The character's relative ability to be rational (versus irrational) reverses when they refuse to think and they hold an irrational attitude. Now, they seemed more able to hold an irrational desire ($M = 4.15$, $SD = 2.03$) than a rational one ($M = 3.64$, $SD = 1.98$), $F(1, 362) = 17.18$, $p < .001$, $\eta^2_G = .02$.

Studies 3 and 4 Discussion

Studies 3 and 4 demonstrate two things. First, participants believed that, when someone reasons, they are more free to adopt a rational mental state than they are an irrational one. This finding supports our proposal that people view reasoning as constrained by rationality. If reasoning were not constrained, then participants would say that others could adopt irrational desires when they reason, and participants did not say that (Figures 6 and 7, right panels). Second, our participants believed that, when others do not think, they are more free to keep their current attitude than they are to change it (Figures 6 and 7, left panels). This second finding supports our claim that people think of mental state change as primarily the output of reasoning and that people cannot directly override mental states (or otherwise reliably change them via nonreasoning mechanisms). Together, these two findings demonstrate that, in commonplace theory of mind, the psychological act of reasoning is both largely necessary and sufficient to rationally change and constrain belief and desire formation. Accordingly, people believe others are free to adopt rational desires and beliefs as long as they can stop to think. Likewise, people believe others are free to hold irrational beliefs and desires when those states spontaneously occur and people shield themselves from the disinfecting force of reason by refusing to think about them.

Results from Studies 1 to 4 reveal a pattern in people's attributions of psychological freedom and constraint that both rules out existing theories and provides direct support for our proposed model. People do not attribute to others the general capacity (or lack of capacity) to adopt any potential belief or desire. For instance, in Studies 1 and 2, participants did not indiscriminately say that others could adopt the belief or desire that the target person most wanted to adopt. And in Studies 3 and 4, participants did not indiscriminately say that others could keep the belief or desire they currently held. Instead, participants attributed control by considering what psychological mechanisms the character could trigger that would cause those states. So, when the character had access to reasons that would make a certain belief or desire rationalizable, participants thought that the character could initiate reasoning, and in so doing, adopt that state. However, participants believed that others were unable to leverage reasoning to adopt irrational mental states. If someone starts thinking about what to believe or desire, then participants by and large thought that this person could not avoid the rational states that would result (Figure 4). One way to overcome this constraint is to do something—such as forgetting or suppressing information—that modifies the reasons that are used as input into reasoning such that the desired mental state is (momentarily, subjectively) rational. Another way to overcome this constraint is to avoid it—that is, refuse to reconsider one's extant irrational attitudes.

Results from Studies 1–4 are also incompatible with theories according to which people moralize freedom and control. Consider the proposal that people think good options are "fundamentally open" and bad options are "fundamentally closed" (Phillips & Knobe, 2009), or the related proposal that people attribute control commensurate with what options they (as observers) find desirable (Clark et al., 2014). If people thought about psychological constraint in this way, then the impact of rationality on perceived mental state control would not depend on the specific process of exercising control that observers consider. In other words, if people implicitly replace the question of what mental state someone "can" form with the question of what mental state would be "good" to form, then people should attribute the same capacity to adopt a rational (or irrational) mental state no matter whether they consider that person thinking, avoiding thinking, trying to suppress reasons, and so on. But as just reviewed, participants' judgments heavily depended on the specific mental process that they considered. Indeed, under the right circumstances, people attribute to others a greater capacity to be irrational than rational (Figures 6 and 7, left panels) (Figures 4 and 5, left panels). We will return to this point in the General Discussion after examining results from Studies 5 and 6, which provide further evidence against moralized theories of psychological freedom and constraint.

This pattern of control attributions replicates across a wide variety of vignettes and across different ways of measuring perceived freedom and constraint. However, the scenarios we used in Studies 3 and 4 are particularly important to examine because, in these scenarios, the characters were in situations involving coercion, economic constraint, exploitation, and situational necessity. These studies therefore illuminate how observers think about what beliefs and desires someone can

hold in these situations. For instance, when someone is being mugged, and they form rational beliefs (e.g., "I will die unless I hand the bag over") and desires (e.g., "I prefer to lose my bag than lose my life") about the situation, observers are likely to think that those individuals are constrained such that they can only keep those particular attitudes. Likewise, Studies 3 and 4 demonstrate the unique conditions under which others can hold irrational attitudes in the face of constraining circumstances, namely, when they spontaneously start off with such states and do not interrogate them. Studies 5 and 6 directly build on these findings by demonstrating that the combination of (a) people's judgments about which beliefs and desires people can form in these situations and (b), their naive theory of reasoning, predicts what people think others can physically do in these situations.

Studies 5–6: Rationality, Intention Formation, and Intentional Action

Studies 5–6 applied the proposed naive theory of reasoning to predict judgments about others' abilities to control their intentions and actions. In lay theory of mind, intentions mediate the influence of beliefs and desires on intentional behavior (Cushman, 2015; Malle & Knobe, 1997, 2001). Specifically, reasoning takes as inputs one's current beliefs and desires and outputs an intention that rationally reflects those beliefs and desires (Figure 1B). We predicted that people will think that others are constrained to forming only intentions that are rationalizable with respect to the beliefs and desires that those individuals can hold.

If beliefs and desires are necessary inputs to intentions, then constraints on beliefs and desires should entail constraints on intentions. Consider a scenario (such as one from Studies 3 and 4) in which someone holds a set of beliefs and desires, these states are rational, and no other states are rationalizable. In such a scenario, people judge others as unable to change their beliefs and desires—that is, to turn them into irrational states (Studies 3–4). If intentions are constrained by one's beliefs and desires, then someone in this situation should be limited to form intentions that rationally reflect the rational beliefs and desires they are constrained to hold (Figure 8A). For instance, the ship captain who rationally believes that a storm is dangerous, and rationally desires to save his own life, can intend to return to port. But, he cannot intend to pilot through the storm as that would require that he can insert into his thinking an irrational belief (e.g., that he would be safe) or desire (e.g., not caring about his safety), and he cannot form such states. And because intentions are necessary to cause intentional behavior, constraints on intentions should entail constraints on behavior: If the ship captain cannot intend to pilot into the storm, then he cannot intentionally pilot into the storm.

In light of findings from Studies 1—4, we further predicted a discrepancy in attributions of control for people who hold rational beliefs and desires compared to people who hold irrational beliefs and desires (see Figure 8). As we just saw, someone who holds rational beliefs and desires is constrained: Although they can intend and act one way, they cannot easily adopt irrational beliefs and desires and thereby intend and act in alternative, irrational ways. However, someone who currently holds an irrational belief or desire should be viewed as less constrained in the sense that they have the ability to think, intend, and act in a wider variety of ways. First, as shown in Studies 3 and 4, these individuals can maintain their irrational states by avoiding scrutinizing them. As a result, they should seem capable of forming intentions that take those irrational states as inputs. However, because these individuals possess strong reasons for alternative rational beliefs and desires, they can also adopt those alternative states by thinking. These new states in turn enable them to form a different intention, and so, perform a different intentional behavior. Put succinctly, someone with irrational beliefs or desires can intend (and act) based either on those irrational states or instead based on new, rational states that they can form by scrutinizing their irrational ones (Figure 8B). We tested these predictions in Study 5 by manipulating whether the agent holds a rational (vs. irrational) belief, and in Study 6 by manipulating whether the agent holds a rational (vs. irrational) desire.

These predictions, shown in Figure 8, are unique to the proposed naive theory of reasoning. Some alternative theories predict that people will judge characters in constraining situations as equally and completely capable of acting both rationally (e.g., handing over the bag when being robbed) and irrationally (e.g., fighting back) regardless of the beliefs and desires they hold (e.g., Kalish, 1998; Kushnir et al., 2015; Reeder, 2009).[5] Other theories predict that people will judge others as more capable of acting rationally compared to irrationally (also regardless of the beliefs and desires the agent happens to hold; e.g., Phillips & Cushman, 2017; Phillips & Knobe, 2009). In contrast to both theories, we predict that participants will attribute to others the capacity to intend (and act) based on whether their initial beliefs and desires are rational, and so whatever mental states are enabled and inhibited according to the naive theory of reasoning. We tested these predictions using the vignettes from Studies 3 and 4.

Study 5: Intentions and Intentional Behavior in Light of Rational and Irrational Beliefs

Method

Participants

We recruited 600 participants. After exclusions, our final sample comprised 580 participants (50% reported male, 47% reported female, 3% unreported or other, $M_{age} = 34$ years).

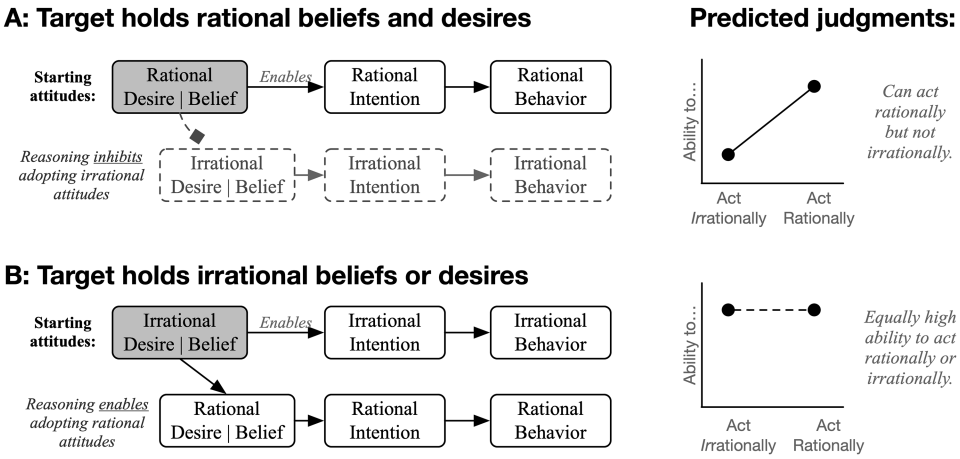
[5] In Studies 5–6 we refer to beliefs, desires, intentions, and actions as being either "rational" or "irrational." This may be confusing given that one of our claims is that people conceptualize mental states as a product of a reasoning process that is constrained by rationality. One might think that we are claiming that all mental states are necessarily rational. However, we contend that reasoning is constrained to produce rational mental states based on the reasons that it takes as input at the specific moment of reasoning. In other words, the reasoning is rational in a "local" sense. It is constrained by the reasons that are salient at the time and place of reasoning.

We can distinguish this local sense of rationality from a "global" sense in which a state is the product of all the reasons available to that person in a situation. These can be best distinguished by thinking about someone who seeks out evidence in a one-sided way in order to rationalize some desired belief. This individual is constrained by local rationality but not global rationality. In particular, this person's resulting belief is locally rational (i.e., rationalized by the narrow set of reasons provided to reasoning) but globally irrational (i.e., if they reflected on all of their evidence, then they would rationally form a different belief).

This distinction is important for understanding "rationality" in the context of intention and action. Someone might have beliefs and desires that are globally irrational (i.e., they do not make sense in light of all of someone's reasons). But, according to the naïve theory of reasoning, those globally irrational states can still be combined in reasoning to produce a locally rational intention (i.e., it rationally derives from the belief-desire set used as input). In Studies 5–6, when we say that an intention (or action) is "rational" or "irrational," this label applies in the "global" sense.

Figure 8

According to the Proposed Naive Theory of Reasoning, Belief and Desire Change Entail Constraints on Intention and Action



Note. (A) Someone who holds rational attitudes is inhibited from adopting irrational ones. Because beliefs and desires enable intentions, these individuals are capable of intending and acting rationally but not irrationally. (B) Someone who holds an irrational belief or desire is capable of forming the corresponding irrational intentions. However, they can also rationally change their belief or desire through reasoning and so adopt the alternative intention.

Experimental Design and Vignette Construction

Study 5 comprised a 2 (belief rationality; between participants) × 2 (behavior rationality; within participants) × 4 (vignette; between participants) design. Participants were randomly assigned to either the irrational belief or rational belief condition and then to one of four vignettes. We describe each of our manipulations below. A schematic of this design is shown in Figure 5 (right panel).

All participants read scenarios featuring a character in a constraining situation. For instance, in the "Storm" vignette (adapted from the same vignette used in Studies 1–4), the captain is piloting an old rickety boat and knows that there is a terrible storm descending. Half of the participants read that the character adopts a belief that reflects the evidence that he has (rational belief condition). The other half of the participants read that the character adopts a belief that violates the evidence (irrational belief condition). For instance, the captain either formed the belief that his ship was not storm-ready and so getting caught in the storm would be dangerous (rational belief), or that his ship would handle the rough storm just fine and so he would be safe (irrational belief). Manipulation checks verified that participants considered the rational belief to be more rational than the irrational belief in all conditions and vignettes. The characters in the vignettes always held a rational desire (e.g., a desire to preserve one's life) such that the only irrational mental state that the character possessed was the belief in the irrational belief condition.

Participants reported what they thought the character "had the ability to do." They reported whether the character could react rationally, and also seperately, irrationally, reflecting a within-participants manipulation (behavior manipulation: rational vs. irrational). Rational reactions included both rational intentions and actions (e.g., piloting the boat back to port). Irrational reactions comprised irrational intentions and actions (e.g., piloting the boat into the storm). In total, participants made four judgments. They reported whether the character had the ability to adopt a (1) rational and (2) irrational intention, as well as whether the character had the ability to perform the corresponding (3) rational and (4) irrational actions.

We replicated this design across four vignettes identical to those from Studies 3 and 4. In "Storm," participants judged whether the captain has the ability to pilot back to port (rational) or to pilot into the storm (irrational). In "Mugging," the character forms either the belief that she is weaker than the mugger (rational) or that she is stronger than the mugger (irrational). Participants then judged whether she had the ability to hand her bag over (rational) or fight (irrational). In "Job," the character believes either that there are no other jobs available (rational) or that there are plenty of other jobs available (irrational). Participants then judged whether she has the ability to keep her job (rational) or quit her job (irrational). And in "Medicine," the character believes her doctor's prognosis that her daughter will only get better with medicine (rational) or believes that the doctor is wrong and that her daughter will get better without medicine (irrational). Participants then judged whether she has the ability to buy medicine for her daughter (rational) or spend the money on something else (irrational).

Procedure

At the start of the scenario, participants read establishing details that made it unambiguous what a rational person would desire and believe. For instance, in the "Storm" vignette, the captain has strong evidence that his boat is in poor condition, that there is a storm descending, and that there is a safe harbor nearby. After reading the scenario, participants answered two questions that tested comprehension of this information. As preregistered, participants who answered at least one of these incorrectly were excluded from data analysis. Participants then learned that the main character spontaneously formed either the belief warranted by their evidence (rational belief condition) or the belief that violated their evidence (irrational belief condition). Participants were not supplied any explanation for why the character adopts their rational or irrational belief.

On the following page, participants reported which intentions and behaviors would be good or bad for the character. Specifically, participants responded to the question, "Based on the facts of X's situation, how good or bad would it be for X to…" (a) form the irrational intention (e.g., "intend to pilot his boat into the storm"), (b) form the rational intention (e.g., "intend to pilot his boat back to port"), (c) perform the irrational behavior (e.g., "pilot his boat into the storm"), and (d) perform the rational behavior (e.g., "pilot his boat back to port"). The text "X" and the corresponding intention and behavior descriptions changed across vignettes to reflect the character's designation and the options presented in the scenario. Participants responded using 7-point rating scales anchored at 1 (*extremely bad*) and 7 (*extremely good*). These questions were presented in a random order. These manipulation checks played the same important role from Studies 3 and 4. If participants spontaneously explained the character's irrational belief by positing some reason why it might be rational, then these rationalizations would affect what is rational to intend (and do), and so would confound judgments of what the character can intend (and do). However, these ad hoc rationalizations would reveal themselves in these judgments about what would be good or bad for the character to do. Thus, these questions allowed us to check whether participants rationalized the irrational belief that the main character spontaneously adopted.

Participants next rated which intentions the main character could form and which actions the main character could perform. Specifically, participants reported their agreement or disagreement with statements that the character "has the ability to" form the (a) irrational intention and (b) rational intention, and perform the (c) irrational behavior and (d) rational behavior. These four statements were presented in a random order. Participants responded using 7-point rating scales anchored at 1 (*completely disagree*) and 7 (*completely agree*).

At the end of the study, participants reported their age and sex and were debriefed.

Results

As expected, participants judged the rational intentions to be better for the character ($M = 5.65$, $SD = 1.46$) compared to the irrational intentions ($M = 2.31$, $SD = 1.48$), $F(1, 572) = 1{,}210.39$, $p < .001$, $\eta^2_G = .61$. Likewise, participants judged the rational actions as better for the character ($M = 5.68$, $SD = 1.48$) compared to the irrational actions ($M = 1.98$, $SD = 1.30$), $F(1, 572) = 1{,}626.05$, $p < .001$, $\eta^2_G = .68$. We observed minor attenuation in the difference between the rational and irrational intentions and behaviors. Specifically, participants in the irrational belief condition judged the difference in the quality of rational and irrational intentions, $F(1, 572) = 24.69$, $p < .001$, $\eta^2_G = .03$, and actions, $F(1, 572) = 36.49$, $p < .001$, $\eta^2_G = .05$, smaller compared to participants in the rational belief condition. However, the results below cannot be accounted for by these minor differences (see the Supplemental Materials, linked in Appendix A). The rational intentions and behaviors remained far superior options for the character compared to the irrational intentions and behaviors whether the character was said to hold the rational or irrational belief (Fs > 385, ps $< .001$, η^2_Gs $> .49$).

Our key results are displayed in Figure 9. We submitted participants' intention judgments, and separately their action judgments, to 2 (behavior rationality: rational vs. irrational) × 2 (belief rationality: rational vs. irrational) × 4 (vignette) mixed within-between ANOVAs. As predicted, judgments of what the character was capable of intending depended on both the belief they currently held (specifically, whether it was rational or irrational), and also whether the target intention or action was rational or irrational, $F(1, 572) = 74.9$, $p < .001$, $\eta^2_G = .04$. We found the same predicted interaction for judgments regarding the actions the character could perform, $F(1, 572) = 52.9$, $p < .001$, $\eta^2_G = .03$. Below we present results for the rational belief condition and then compare those to results in the irrational belief condition.

Rational Belief

We submitted participants' judgments in the rational belief condition to a 2 (behavior rationality) × 4 (vignette) mixed ANOVA. Participants judged the characters as more capable of forming the rational intention ($M = 6.22$, $SD = 1.25$) than the irrational intention ($M = 4.64$, $SD = 2.24$), $F(1, 282) = 149.47$, $p < .001$, $\eta^2_G = .16$. We observed the same pattern for intentional behavior. Participants judged the character as more able to perform the rational behavior ($M = 6.25$, $SD = 1.12$) than the irrational behavior ($M = 4.36$, $SD = 2.26$), $F(1, 282) = 188.93$, $p < .001$, $\eta^2_G = .22$. These findings extend results from Studies 1–4 showing that people believe others are inhibited from adopting irrational attitudes.

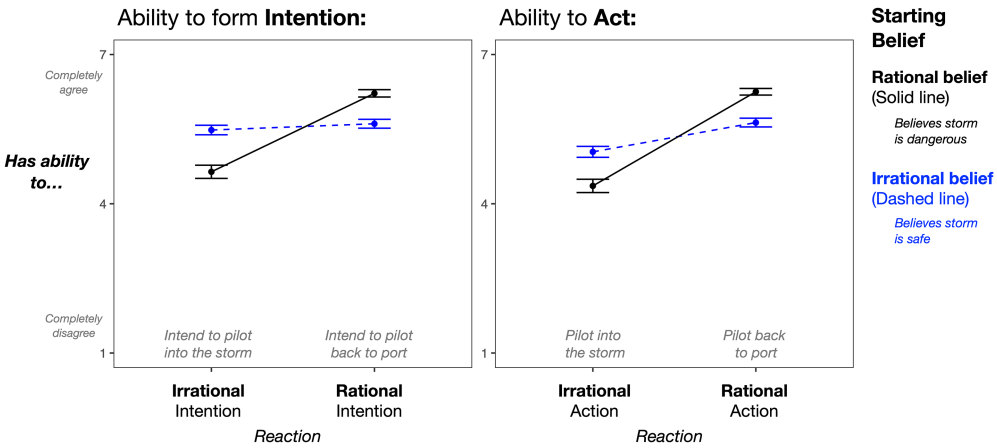
Irrational Belief

The key test of our theory concerned participants' judgments in the irrational belief condition. If their judgments reflected simply what they judged would be best for the character to do, then their judgments should be identical to those in the rational belief condition. After all, as confirmed by the manipulation checks, even when the character holds an irrational belief, they are judged to be much better off forming the opposing, rational intention and acting in the opposing, rational manner. However, we predicted that, because the character now possessed an irrational belief, they would thereby be enabled to input this belief into their reasoning to form the corresponding (irrational) intention. We also predicted that, because the current belief is irrational, participants would judge the target as able to change their belief, and therefore as able to intend and act otherwise. Taken together, we predicted that participants would judge the characters as capable of forming either intention, and so relatedly, performing either behavior (see Figure 8B).

As noted above, we observed this predicted Belief Rationality × Behavior Rationality interaction. When the character held the irrational belief, participants judged them as roughly equally capable of forming the rational ($M = 5.61$, $SD = 1.53$) and irrational intentions ($M = 5.48$, $SD = 1.64$), $F(1, 290) = 1.39$, $p = .239$, $\eta^2_G < .01$. Turning to judgments about the character's ability to perform certain actions, participants reported that the character was more capable of performing the rational behavior ($M = 6.63$, $SD = 1.51$) compared to the irrational behavior ($M = 5.04$, $SD = 1.88$), $F(1, 290) = 27.16$, $p < .001$, $\eta^2_G = .03$. However, this difference was much smaller compared to judgments in the rational belief condition. Thus, as predicted, participants thought that the character with the irrational belief was capable both of acting on that irrational belief (reflected by high ratings for the irrational behaviors) and of changing their mind (reflected by nearly equal high ratings for the rational behaviors).

Figure 9

Mean (and Standard Error) "Ability To" Ratings for Intention Formation (Left Panel) and Intentional Action (Right Panel) From Study 5



Note. Sample stimuli are displayed in smaller font. Compare to predictions in Figure 8. See the online article for the color version of this figure.

Irrational Beliefs Enable Irrational Intentions

According to the naive theory of reasoning, people should judge others as capable of adopting irrational intentions, and acting irrationally, if they can input irrational beliefs and desires into their reasoning about what to do. To test this prediction, we submitted participants' judgments in the irrational behavior condition to 2 (belief rationality) × 4 (vignette) fully crossed ANOVAs. As expected, participants thought that the agent with the irrational belief was more capable of adopting the irrational intention compared to the agent with the rational belief, $F(1, 572) = 25.88$, $p < .001$, $\eta^2_G = .04$. Likewise, participants judged that the agent with the irrational belief was more capable of acting irrationally compared to the agent with the rational belief, $F(1, 572) = 15.28$, $p < .001$, $\eta^2_G = .03$. Note that these results cannot be explained by a belief that, in general, others are more capable of intending and acting in line with their current beliefs compared to against them. In the irrational belief condition, this alternative theory would predict lower judgments for the rational intention (and action) compared to the irrational intention (and action). However, as noted above, in the irrational belief condition, we observe the opposite pattern of results when participants rate actions, and we observe no difference when participants rate intentions.

Individual Differences in Control Attributions

One unexpected finding was that participants on average attributed middling-to-high control to act irrationally in the condition where the target holds rational mental states. One possible explanation for this result is that these average capacity judgments reflected most participants thinking that rationality comprises only a very weak constraint on what the target could intend and do. This explanation would pose a problem for our theory. After all, if people primarily think about psychological constraint by drawing on an intuitive theory of reasoning (as we propose), then we should expect a stronger relationship between rationality and control. Examining these responses further revealed an apparent individual difference in how participants think about control. Specifically, judgments in this condition formed a bimodal distribution with peaks at opposite ends of the response scale: Although the most common response (31%) was to *completely agree* ("7") that the character could intent (or act) irrationally, the second most common response (15%) was to *strongly disagree* ("1"). By contrast, when the character held an irrational belief, the most common response (40%) was still to completely agree that they could intend and act irrationally, but only 4% of participants now completely disagreed—the least common response. See Appendix C. Our results appear to reflect the combination of two kinds of participants: Some participants appear to deny that someone ever lacks the ability to act irrationally—these individuals in effect deny that canonically coercive situations are genuinely constraining. Our remaining participants, however, attribute freedom and constraint by drawing on a naive theory of reasoning. These individuals think that, in canonically coercive situations, someone with an irrational belief is entirely capable of intending and acting irrationally, while also thinking that someone with a rational belief is nearly entirely incapable of doing so.

Study 6: Intentions and Intentional Behavior in Light of Rational and Irrational Desires

Method

Participants

We recruited 600 participants. After exclusions, our final sample comprised 569 participants (44% reported male, 55% reported female, 1% unreported or other, $M_{age} = 34$ years).

Design and Procedure

Study 6 used the same design, procedure, and set of vignettes as Study 5, but manipulated the rationality of the character's desire instead of the character's belief. Thus, Study 6 comprised a 2 (desire rationality) × 2 (behavior rationality) × 4 (vignette) design. At the beginning of the study, participants were randomly assigned to

one of the two desire rationality conditions and then to one of four vignettes.

To manipulate desire rationality, the character either adopted a desire that, if fulfilled, would result in the best outcome for them (rational desire condition) or adopted a desire that, if fulfilled, would result in a terrible outcome (irrational desire condition). For instance, in the rational desire condition in the "Storm" vignette, the captain cares more about his personal safety than about delivering his packages swiftly, and so forms the rational desire to pilot his boat back to port. In the irrational desire condition, the captain cares more about making a swift delivery than his personal safety and forms the irrational desire to pilot his boat into the storm despite the significant costs. The characters in the vignettes always held rational beliefs (e.g., belief that the storm is dangerous) such that the only irrational mental state that the character ever possessed was the desire in the irrational desire condition.

The other vignettes described the same coercion and exploitation vignettes from Study 5. In "Mugging," the character spontaneously forms the desire to protect either their life (rational) or their bag (irrational). In "Job," the character forms the desire either to show up to work (rational) or to quit her job (irrational). And in "Medicine," the character forms the desire to care for either her daughter instead of her cat (rational desire) or her cat instead of her daughter (irrational desire). The target intentions and behaviors were unchanged from Study 5.

Results

We adopted the same, preregistered analytic procedure from Study 5. As expected, participants judged the rational intentions to be better for the character ($M = 5.48$, $SD = 1.53$) compared to the irrational intentions ($M = 2.35$, $SD = 1.51$), $F(1, 561) = 1{,}109.88$, $p < .001$, $\eta^2_G = .56$. Likewise, participants judged the rational actions as better for the character ($M = 5.46$, $SD = 1.61$) compared to the irrational actions ($M = 1.96$, $SD = 1.31$), $F(1, 561) = 1{,}368.39$, $p < .001$, $\eta^2_G = .62$. The rational intentions and behaviors remained far superior options for the character compared to the irrational intentions and behaviors, regardless of whether the character was said to hold the rational or irrational desire ($fs > 407$, $ps < .001$, $\eta^2_G s > .49$). We observed some variation in perceived rationality in one vignette, but as in Study 5, the results below cannot be explained by appeal to this variation (see the Supplemental Materials, linked in Appendix A).

We submitted participants' intention judgments, and separately their action judgments, to 2 (behavior rationality: rational vs. irrational) × 2 (desire rationality: rational vs. irrational) × 4 (vignette) mixed within-between ANOVAs. Main results are displayed in Figure 10. As predicted, we observed a significant interaction between desire rationality (whether the current desire was rational or irrational) and behavior rationality (whether the target intention or action was rational or irrational), for both intention formation, $F(1, 561) = 90.24$, $p < .001$, $\eta^2_G = .05$, and action, $F(1, 561) = 54.98$, $p < .001$, $\eta^2_G = .04$. Below we present results for the rational desire condition and then compare those to results in the irrational desire condition.

Rational Desire

We submitted intention and action control judgments in the rational desire condition to 2 (behavior rationality) × 4 (vignette) mixed within-between ANOVAs. Results in this condition mirrored results from Studies 1–5: Participants judged the characters as more capable of forming the rational intention ($M = 6.29$, $SD = 1.04$) than the irrational intention ($M = 4.63$, $SD = 2.16$), $F(1, 277) = 166.92$, $p < .001$, $\eta^2_G = .20$. We observed the same pattern for intentional behavior. Participants judged the character as more able to perform the rational behavior ($M = 6.26$, $SD = 1.07$) than the irrational behavior ($M = 4.26$, $SD = 2.25$), $F(1, 277) = 224.72$, $p < .001$, $\eta^2_G = .25$. Thus, the character who starts off with a rational desire is partially constrained in that they can execute one action but are less capable of executing an alternative.

Irrational Desire

We next submitted control judgments in the irrational desire condition to 2 (behavior rationality) × 4 (vignette) mixed within-between ANOVAs. When the character held the irrational desire, participants judged them as roughly equally capable of forming the rational ($M = 5.96$, $SD = 1.39$) and irrational intentions ($M = 5.83$, $SD = 1.56$), $F(1, 284) = 1.80$, $p = .180$, $\eta^2_G < .01$. Turning to judgments about action, participants reported that the character was more capable of performing the rational action ($M = 6.00$, $SD = 1.37$) compared to the irrational action ($M = 5.13$, $SD = 1.88$), $F(1, 284) = 36.32$, $p < .001$, $\eta^2_G = .04$. However, this difference was much smaller compared to judgments in the rational desire condition. Thus, as predicted, participants thought that the character with the irrational desire was capable both of acting on that irrational desire (reflected by high ratings for the irrational behaviors) and of changing their mind (reflected by nearly equal high ratings for the rational behaviors).

Irrational Desires Enable Irrational Intentions

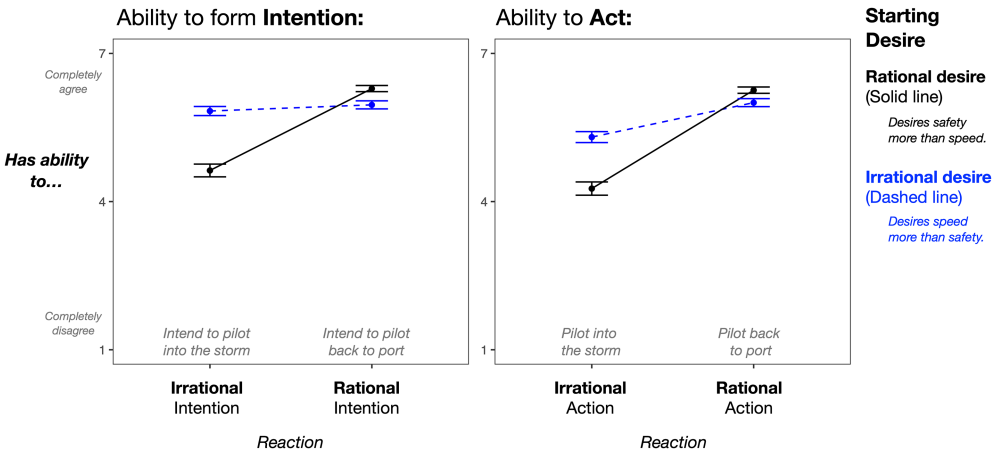
As in Study 5, we tested whether participants judged others as more capable of adopting irrational intentions and actions if they could input irrational beliefs and desires into their reasoning about what to do. To do this, we submitted participants' capacity judgments in the irrational behavior condition to 2 (desire rationality) × 4 (vignette) fully crossed ANOVAs. As expected, participants thought that the agent with the irrational desire was more capable of adopting the irrational intention compared to the agent with the rational desire, $F(1, 561) = 55.7$, $p < .001$, $\eta^2_G = .09$. Likewise, participants judged that the agent with the irrational desire was more capable of acting irrationally compared to the agent with the rational desire, $F(1, 561) = 36.47$, $p < .001$, $\eta^2_G = .06$. As in Study 5, these results cannot be explained by a belief that others are, in general, more capable of intending or acting in line with their current mental states than against them.

Individual Differences in Control Attributions

As in Study 5, the condition in which the target holds a rational desire resulted in a bimodal distribution of participants either attributing maximum or minimum control to the target. Again, in this condition, the most common response (30%) was to *completely agree* ("7") that the character could intend (or act) irrationally while the second most common response (14%) was *to completely disagree* ("1"). By contrast, when the character held an irrational desire, the most common response (45%) was still to completely agree that they could intend and act irrationally, but only 4% of participants

Figure 10

Mean (and Standard Error) Possibility Ratings for Intention Formation (Left Panel) and Intentional Action (Right Panel) From Study 6



Note. Sample stimuli are displayed in smaller font. Compare to predictions in Figure 8. See the online article for the color version of this figure.

now completely disagreed that they could intend or act irrationally ("1"). See Appendix C for additional details.

Discussion of Studies 5 and 6

Studies 5 and 6 verified predictions from the proposed naive theory of reasoning. When participants read about a character in a constraining situation with rational beliefs and desires, participants judged that character as relatively unfree. Specifically, they reported that the character could adopt rational intentions and perform the corresponding rational intentional action, but that this character was relatively incapable of changing their mind and acting irrationally. This result replicates prior findings that people think others lack freedom in canonical cases of coercion, manipulation, and situational pressure. However, it goes beyond this prior work by demonstrating perceived constraints on intention formation, not just on intentional behavior.

These findings were only half the story. When participants read about a character with an irrational belief (Study 5) or desire (Study 6), participants now judged that character as relatively free. This relative freedom reflects the combination of two results. First, participants judged the character as capable of intending and acting on their irrational beliefs and desires: Participants attributed a greater capacity to intend and act irrationally to the character with irrational beliefs and desires compared to the character with rational beliefs and desires. And second, participants judged that this character could easily change their mind and adopt (and act in line with) intentions that corresponded to rational beliefs and desires. This finding reflects the lack of difference in judgments that this character has the capacity to form rational and irrational intentions (and perform the rational and irrational acts). This pattern of findings is not accounted for by prior theories of freedom and constraint, but this pattern of control attributions is consistent with results in Studies 1–4 and is predicted by our proposed account of the naive theory of reasoning. The naive theory of reasoning does the best job of describing how average attributions of freedom and control change across situations.

Finally, these studies also revealed individual differences in how people think about freedom in canonically constraining situations. About one third of our sample denied that canonically coercive situations constrain people. Our best guess is that these individuals responded to questions about freedom over behavior similarly to how participants in Cusimano and Goodwin (2020) responded to questions about freedom over belief. Cusimano and Goodwin (2020) found that many people, when they consider another person's ability to voluntarily change their belief, ignore information about that person's situation or psychology, and simply draw on a conception of belief as something that is in principle free and voluntary. Some of our participants may have approached these studies in a similar way by calling to mind a rule-of-thumb that people can make any choice, even in constraining situations, simply because they have "free will." The rest of our participants, however, thought about freedom and constraint in the manner that we hypothesized. And indeed, the most common response among these participants was to completely disagree that others had the ability to react irrationally (when the target individual started off with rational mental states). These results suggest that, insofar as people think that others are not free because of situational pressure, they do so because they rely on the naive theory of reasoning that we propose.

General Discussion

When do people judge that someone can exert control over their own mind, and so believe, desire, intend, or do otherwise? We propose that people draw on a naive theory of reasoning wherein reasoning is a rational, semi-autonomous process that people can leverage to produce new mental states, but which they cannot directly override. In support of this model, participants did not judge others as capable of adopting whatever belief or desire they wanted to. Instead, participants thought that one way people could think and do otherwise was by reasoning their way to rational mental states. Across six studies, participants judged others as capable of adopting new beliefs, desires, and intentions only when those states could be

rationalized by information or other mental states that they possessed or could easily acquire.

Our proposal readily explains why people judge others as lacking the ability to think and do otherwise in cases of coercion or situational constraint. Prior work has demonstrated that people believe others are less capable of acting contrary to strong incentives (e.g., Baron, 1998), coercive threats (e.g., Woolfolk et al., 2006), social norms, moral norms (e.g., Phillips & Cushman, 2017; Phillips & Knobe, 2009) and dire circumstances (Young & Phillips, 2011). When viewed through the lens of our model, we can see that a unifying feature of these situations is that they afford few beliefs and desires that someone can rationally adopt. Accordingly, people are constrained in the sense that reasoning, because it is uncontrollably rational, can only produce one set of beliefs, desires, and intentions. Even if people want to think or desire otherwise, these desires cannot override the rational reasons they have to think and desire a particular way. And when someone can only form one set of beliefs, desires, and intentions, people then believe that there is only one way they can act. Thus, the naive theory of reasoning explains commonplace judgments that situations constrain others by appeal to the observation that situations prevent people from forming, through reasoning, the beliefs, desires, and intentions that are necessary to produce those behaviors.

At the same time, the naive theory of reasoning accommodates the observation that, in more mundane circumstances, people think that others are capable of thinking and acting irrationally (e.g., Cusimano & Goodwin, 2019). Our model accommodates these observations by positing that the constraint of rationality applies only to the process of reasoning itself, not to the broader suite of reactions a person might have to their circumstances. Included in this broader suite of reactions are decisions to avoid reasoning or to manipulate or circumvent reasoning. By avoiding or manipulating reasoning, people can sidestep the constraint of rationality. Accordingly, people judge others as free to think and act irrationally when they think that others have, for instance, an opportunity to selectively ignore or forget reasons, or to selectively reconsider irrational mental states. Indeed, in mundane circumstances, people may think that the environment is ambiguous with respect to the beliefs, desires, and plans that people can rationalize. For instance, observers might judge others as able to choose whether to believe in God because they think that those others have the power to generate and attend either to evidence that God exists or to arguments about why God does not exist. Judgments that others can flexibly manipulate their reasoning should only increase as people attribute to others additional opportunity and drive to acquire favorable evidence, forget unwanted information, or think of new ways to rationalize certain attitudes.

This observation affords a novel insight into the situations in which people think others are constrained. Specifically, canonical cases of situational pressure combine tight constraints on what is rational for someone to do with tight constraints on their opportunity to change the reasons they have or manipulate their reasoning in other ways. Consider, for instance, someone held at gunpoint who must make an immediate decision about whether to comply. The fact that, in that moment, the only rational thing for them to do is to hand the bag over is one part of why they seem constrained. But the fact that they have to make their decision immediately is another important part of why they seem constrained: They cannot take the time to successfully manipulate their reasoning and so are bound by the initial and immediate rational beliefs, desires, and intentions that they form.

Relationship to Apparent Biases and Asymmetries in Attributions of Control

The naive theory of reasoning predicts that people will judge others as free to change their mind when they think that others can reason, and in so doing, change an irrational mental state to a rational one. This process of attributing freedom and constraint to others may interact with well-known biases to produce biased attributions of freedom and constraint. For instance, people tend to believe that their own beliefs rationally reflect the available evidence while the beliefs of those who disagree with them do not (Pronin et al., 2004; Reeder et al., 2005; Ross & Ward, 1996). It follows from this observation that people should think that others who agree with them are inhibited from changing their minds while others who disagree with them are enabled to change their minds. Many studies document this pattern of judgments. However, as we now review, prior studies explain these findings by appeal to psychological mechanisms that are superseded by the naive theory of reasoning.

Cusimano and Goodwin (2020, Study 1) found that people judge others who hold different beliefs to be more capable of voluntarily changing their mind than they themselves are. This self-other difference is readily explained by our model: Participants likely thought that others could voluntarily change their minds because they could attend to reasons why they should believe otherwise, which would then enable belief change. For instance, they may have thought that someone who does not believe in anthropogenic climate change can attend to the salient evidence in favor of anthropogenic climate change, and in so doing, rationally come to believe in anthropogenic climate change.

Everett et al. (2021) found that people judge others to be more free when they perform behaviors that they judge as immoral compared to moral (see also Clark et al., 2014, 2021). For instance, conservatives (relative to liberals) think that drug addicts are more free to stop using drugs, and liberals (relative to conservatives) judge that someone who discriminated against transgender people was more free not to. Those authors argued that increased attributions of free will are motivated by a desire to punish others for bad behavior. However, motivated reasoning fails to explain results from our studies. It cannot explain why people think that someone who can choose to suppress or forget information can irrationally change their mind (Studies 1–2), or why someone who starts off with an irrational belief or desire can keep it (Studies 3–4) and act on it (Studies 5–6). Most importantly, in our studies participants reported what someone could do before they made any choices, and so before they made any bad choices that engender a motive to punish them. By contrast, the naive theory of reasoning readily explains differences in free will attributions for perceived moral and immoral behavior. People who moralize a behavior likely also believe that there are especially strong rational reasons for that behavior. For instance, someone who believes that it is immoral to discriminate against transgender people likely believes they have strong reasons against such discrimination. Thus, these individuals should think that someone who was aware of arguments against discrimination, but had not fully considered those reasons, could realize their mistake if they thought about it more. This opportunity to change one's mind is not available to someone who already holds the view that they take to be rational (i.e., that discriminating against transgender people is wrong), thereby explaining the difference in control attributions.

The same line of reasoning applies to related findings that people tend to view immoral behaviors as more free than moral behaviors (e.g., Phillips & Knobe, 2009; Young & Phillips, 2011). In these studies, participants reported that someone who committed an immoral act (e.g., throwing a person overboard to save their crew) was free to do otherwise while someone who committed a morally good act (e.g., throwing cargo overboard to save their crew) was not. These authors explain these judgments by arguing that people conflate judgments of morality and agency. The studies reported here rule out this theory while accommodating the pattern of judgments reported in those articles. Consider: When the ship captain throws the cargo overboard, no amount of thinking about what to do will rationally lead him to believe that there is a better course of action. However, when the ship captain throws the person overboard, it seems that he could have stopped to think about what he was doing, and in so doing, rationally realized that there were better options (such as throwing cargo overboard instead). Our view also explains why this difference disappears when the behavior is accidental (Young & Phillips, 2011): When the outcome is an accident, the reasons for or against an action are absent from reasoning, and so are unable to play any enabling or inhibiting role in mental state formation.

One common feature of these prior theories is that attributions of control stem from people's egocentric judgments of what is good or bad to do. By contrast, our theory explains how these differences can stem from entirely allocentric reasoning about others' mental states. Allocentric reasoning about others' mental states is a feature of mature and rational theory of mind. Accordingly, we propose that the naive theory of reasoning provides a "rational" alternative explanation for the apparently "biased," "motivated," or "moralized" attributions of freedom documented in these prior studies. That said, biases can still play a role in our account: Biases affect what people think others have rational reason to think, want, and do. Future work investigating apparent biases in how people think about freedom should first check whether discrepancies can be explained by our proposed naive theory of reasoning.

Application to Prior Work on Situational Attributions of Behavior

Our findings contrast with prior work documenting that people neglect situational constraints on others' behavior (Gilbert & Malone, 1995; Jones & Harris, 1967). This discrepancy reflects the fact that, in the current studies, participants read short stories that were relatively simple and in which the constraining evidence and utility were made salient. Absent these features, it is likely that participants would give less consideration to the reasons the characters had to believe and desire as they did. As a result of being unaware of the reasons that others have, participants would likely default to assuming some flexibility about which mental states the characters in the situation could rationalize. Consistent with this line of reasoning, prior work shows that people tend not to spontaneously think about the constraining evidence that others have for their beliefs, and because of this tendency, assume others are free to change their beliefs (Cusimano & Goodwin, 2020). This line of reasoning leads to the following general observation about how people think about psychological freedom and constraint: In theory, observers judge others as constrained by rationality, and so constrained by the strong reasons that underly their mental states; in practice, observers do not probe deeply enough into others' reasoning to reliably uncover the reasons that constrain them.

Application to Attributions of Rationality and Agency

Our results also illuminate why the capacity to be rational is seen as a precondition for autonomy and self-control. In our studies, participants believed that rational reasoning about the information available to the characters enabled those individuals to adopt new beliefs and desires. It stands to reason that, if someone lacked the capacity to comprehend that information, then that information would be unavailable as an input to that person's reasoning. For instance, people may believe that a particular robot, in virtue of its inability to think rationally about what makes something moral or immoral, lacks the ability to adopt morally good attitudes. In this case, the robot is not able to adopt a new, specific mental state the same way that a human would—namely by stopping and reflecting on the reasons available to them.

But while a lack of rationality is often seen as a precondition for autonomy and self-control, our results also suggest that a lack of rationality may sometimes be seen as enabling someone to adopt a wider range of mental states, too. For instance, perhaps people think that children are more capable of adopting irrational beliefs about the world because their capacity to think rationally is underdeveloped, and so is a less effective constraint on their mental state formation. Likewise, people may expect drunk or mentally ill individuals to be capable of voluntarily adopting a wider range of mental states because those individuals do not reason in a way that is subject to ordinary constraints of rationality. Uncovering the relationship between observers' attributions of rationality and the enabling or constraining impact of reasons on their mental state formation is an important direction for future research.

Constraints on Generality and Directions for Future Work

One virtue of the current investigation is that it readily makes sense of prior work on perceived rationality and control. However, this investigation also focuses on the usual narrow population, and it is not clear whether individuals from non-Western cultures possess the naive theory of reasoning that we propose. On the one hand, some work suggests commonality across cultures in conceptualizations of free will (e.g., Chernyak et al., 2013) and mental states (e.g., Thornton et al., 2020). On the other hand, prior work has observed considerable variation across cultures regarding how much mental agency people attribute to others (Cohen & Rozin, 2001). Future work should investigate variation in people's conception of reasoning to determine similarities and differences across cultures and individuals.

Another important direction for future research is to document the origins of this naive theory of reasoning. Some work has found that children intuitively think that others are constrained by strong social and moral reasons (e.g., Kalish, 1998; Kushnir et al., 2015; see discussion in Kushnir, 2018). However, we speculate that the lay theory of reasoning may also reflect people's experience with both their own and others' decision making. After all, people often experience their reasoning as constrained, and report that they feel like they cannot believe other than how they do, or cannot choose otherwise, when they face strong evidence and lopsided choices (Cusimano & Goodwin, 2020; Kouchaki et al., 2018; see also discussion in Wolf, 1980). Likewise, scholars have argued that belief, desire, and emotion change are constrained by their environment (e.g.,

Kunda, 1990; Lazarus, 1991). For these reasons, we speculate that the naive theory of reasoning is partially supported by people's experience with their own and others' reasoning and decision making.

The suggestion that people's naive theory of reasoning reflects their own experience speaks to another important issue, namely, whether people's judgments about others' capacity to flexibly change their beliefs, desires, and intentions are accurate. For instance, in our studies, participants judged that others could not adopt beliefs that directly violated strong evidence. Based on accounts of both motivated reasoning (Kunda, 1990) and people's own experience of constraint (Cusimano & Goodwin, 2020), people's attribution of constraint may be accurate. However, this line of reasoning is speculative and future work should investigate the relationship between people's experience of their own and others' reasoning, and the naive model of reasoning they acquire and deploy to understand and predict others.

Conclusion

We have argued that people reason about others' capacity to control their mind and behavior by drawing on a naive theory of reasoning. Accordingly, people conceptualize reasoning as a process that can be leveraged to bring about rational mental states or manipulated to yield irrational alternatives. This theory explains common intuitions about others' ability to control their beliefs, desires, intentions, and intentional actions. As a result, the naive theory of reasoning explains everyday judgments about coercion and situational constraint. People's naive theory of reasoning may form the foundation of their intuitive theories of autonomy, self-control, responsibility, and persuasion.

References

Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological Bulletin*, *126*(4), 556–574. https://doi.org/10.1037/0033-2909.126.4.556

Aristotle. (1985). *Nicomachean ethics* (T. Irwin, Trans.). Hackett.

Baron, J. (1998). *Judgment misguided: Intuition and error in public decision making*. Oxford University Press.

Chernyak, N., Kushnir, T., Sullivan, K. M., & Wang, Q. (2013). A comparison of American and Nepalese children's concepts of freedom of choice and social constraint. *Cognitive Science*, *37*(7), 1343–1355. https://doi.org/10.1111/cogs.12046

Clark, C. J., Luguri, J. B., Ditto, P. H., Knobe, J., Shariff, A. F., & Baumeister, R. F. (2014). Free to punish: A motivated account of free will belief. *Journal of Personality and Social Psychology*, *106*(4), 501–513. https://doi.org/10.1037/a0035880

Clark, C. J., Winegard, B. M., & Baumeister, R. F. (2019). Forget the folk: Moral responsibility preservation motives and other conditions for compatibilism. *Frontiers in Psychology*, *10*, Article 215. https://doi.org/10.3389/fpsyg.2019.00215

Clark, C. J., Winegard, B. M., & Shariff, A. F. (2021). Motivated free will belief: The theory, new (preregistered) studies, and three meta-analyses. *Journal of Experimental Psychology: General*, *150*(7), e22–e47. https://doi.org/10.1037/xge0000993

Cohen, A. B., & Rozin, P. (2001). Religion and the morality of mentality. *Journal of Personality and Social Psychology*, *81*(4), 697–710. https://doi.org/10.1037/0022-3514.81.4.697

Cushman, F. (2015). Deconstructing intent to reconstruct morality. *Current Opinion in Psychology*, *6*, 97–103. https://doi.org/10.1016/j.copsyc.2015.06.003

Cusimano, C., & Goodwin, G. P. (2019). Lay beliefs about the controllability of everyday mental states. *Journal of Experimental Psychology: General*, *148*(10), 1701–1732. https://doi.org/10.1037/xge0000547

Cusimano, C., & Goodwin, G. P. (2020). People judge others to have more voluntary control over beliefs than they themselves do. *Journal of Personality and Social Psychology*, *119*(5), 999–1029. https://doi.org/10.1037/pspa0000198

Cusimano, C., & Goodwin, G. P. (2022). Mental states and control-based theories of responsibility. In T. Nadelhoffer & A. Monroe (Eds.), *Advances in experimental philosophy of free will and responsibility* (pp. 45–65). Bloomsbury Publishing.

Cusimano, C., Zorilla, N., Danks, D., & Lombrozo, T. (2021, July 26–29). *Reason-based constraint in theory of mind* [Paper presentation]. Proceedings of the Annual Meeting of the Cognitive Science Society.

D'Andrade, R. (1987). A folk model of the mind. In D. Holland & N. Quinn (Eds.), *Cultural models in language and thought* (pp. 112–148). Cambridge University Press.

Everett, J. A. C., Clark, C. J., Meindl, P., Luguri, J. B., Earp, B. D., Graham, J., Ditto, P. H., & Shariff, A. F. (2021). Political differences in free will belief are associated with differences in moralization. *Journal of Personality and Social Psychology*, *120*(2), 461–483. https://doi.org/10.1037/pspp0000286

Fischel, J. (2019). *Screw consent: A better politics of sexual justice*. University of California Press. https://doi.org/10.1525/9780520968172

Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: The naïve theory of rational action. *Trends in Cognitive Sciences*, *7*(7), 287–292. https://doi.org/10.1016/S1364-6613(03)00128-1

Gilbert, D. T., & Malone, P. S. (1995). The correspondence bias. *Psychological Bulletin*, *117*(1), 21–38. https://doi.org/10.1037/0033-2909.117.1.21

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, *315*(5812), 619–619. https://doi.org/10.1126/science.1134475

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences*, *20*(8), 589–604. https://doi.org/10.1016/j.tics.2016.05.011

Jones, E. E., & Harris, V. A. (1967). The attribution of attitudes. *Journal of Experimental Social Psychology*, *3*(1), 1–24. https://doi.org/10.1016/0022-1031(67)90034-0

Kalish, C. (1998). Reasons and causes: Children's understanding of conformity to social rules and physical laws. *Child Development*, *69*(3), 706–720. https://doi.org/10.1111/j.1467-8624.1998.tb06238.x

Kouchaki, M., Smith, I. H., & Savani, K. (2018). Does deciding among morally relevant options feel like making a choice? How morality constrains people's sense of choice. *Journal of Personality and Social Psychology*, *115*(5), 788–804. https://doi.org/10.1037/pspa0000128

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480–498. https://doi.org/10.1037/0033-2909.108.3.480

Kushnir, T. (2018). The developmental and cultural psychology of free will. *Philosophy Compass*, *13*(11), Article e12529. https://doi.org/10.1111/phc3.12529

Kushnir, T., Gopnik, A., Chernyak, N., Seiver, E., & Wellman, H. M. (2015). Developing intuitions about free will between ages four and six. *Cognition*, *138*, 79–101. https://doi.org/10.1016/j.cognition.2015.01.003

Lazarus, R. S. (1991). *Emotion and adaptation*. Oxford University Press.

Malle, B. F. (1999). How people explain behavior: A new theoretical framework. *Personality and Social Psychology Review*, *3*(1), 23–48. https://doi.org/10.1207/s15327957pspr0301_2

Malle, B. F. (2004). *How the mind explains behavior: Folk explanations, meaning, and social interaction*. MIT Press.

Malle, B. F. (2019). How many dimensions of mind perception really are there? In A. K. Goel, C. M. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st annual meeting of the cognitive science society* (pp. 2268–2274). Cognitive Science Society.

Malle, B. F., & Knobe, J. (1997). The folk concept of intentionality. *Journal of Experimental Social Psychology*, *33*(2), 101–121. https://doi.org/10.1006/jesp.1996.1314

Malle, B. F., & Knobe, J. (2001). The distinction between desire and intention: A folk-conceptual analysis. In B. F. Malle, L. J. Moses, & D. A. Baldwin (Eds.), *Intentions and intentionality: Foundations of social cognition* (pp. 45–67). MIT Press.

Monroe, A. E., & Malle, B. F. (2010). From uncaused will to conscious choice: The need to study, not speculate about people's folk concept of free will. *Review of Philosophy and Psychology*, *1*(2), 211–224. https://doi.org/10.1007/s13164-009-0010-7

Monroe, A. E., & Ysidron, D. W. (2021). Not so motivated after all? Three replication attempts and a theoretical challenge to a morally motivated belief in free will. *Journal of Experimental Psychology: General*, *150*(1), e1–e12. https://doi.org/10.1037/xge0000788

Murray, D., & Lombrozo, T. (2017). Effects of manipulation on attributions of causation, free will, and moral responsibility. *Cognitive Science*, *41*(2), 447–481. https://doi.org/10.1111/cogs.12338

Nelkin, D. K. (2011). *Making sense of freedom and responsibility*. Oxford University Press.

Nozick, R. (1974). *Anarchy, state, and utopia*. Basic Books.

Phillips, J., & Cushman, F. A. (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

Powell, B., & Zwolinski, M. (2012). The ethical and economic case against sweatshop labor: A critical assessment. *Journal of Business Ethics*, *107*(4), 449–472. https://doi.org/10.1007/s10551-011-1058-8

Pronin, E., Gilovich, T., & Ross, L. (2004). Objectivity in the eye of the beholder: Divergent perceptions of bias in self versus others. *Psychological Review*, *111*(3), 781–799. https://doi.org/10.1037/0033-295X.111.3.781

Rawls, J. (1971). *A theory of justice*. Harvard University Press.

Reeder, G. D. (2009). Mindreading: Judgments about intentionality and motives in dispositional inference. *Psychological Inquiry*, *20*(1), 1–18. https://doi.org/10.1080/10478400802615744

Reeder, G. D., Pryor, J. B., Wohl, M. J. A., & Griswell, M. L. (2005). On attributing negative motives to others who disagree with our opinions. *Personality and Social Psychology Bulletin*, *31*(11), 1498–1510. https://doi.org/10.1177/0146167205277093

Robinson, P. H. (1997). *Structure and function in criminal law*. Oxford University Press.

Ross, L., & Ward, A. (1996). Naive realism in everyday life: Implications for social conflict and misunderstanding. In E. S. Reed, E. Turiel, & T. Brown (Eds.), *The Jean Piaget symposium series. Values and knowledge* (pp. 103–135). Lawrence Erlbaum Associates.

Sommers, R. (2019). Commonsense consent. *Yale Law Journal*, *129*(8), 2232–2325.

Thornton, M. A., Wolf, S., Reilly, B. J., Slingerland, E., & Tamir, D. (2020). *The 3D Mind Model characterizes how people understand mental states across modern and historical cultures*. PsyArXiv. https://doi.org/10.31234/osf.io/m5p74

Weiner, B. (1995). *Judgments of responsibility: A foundation for a theory of social conduct*. Guilford Press.

Weisman, K., Dweck, C. S., & Markman, E. M. (2017). Rethinking people's conceptions of mental life. *Proceedings of the National Academy of Sciences*, *114*(43), 11374–11379. https://doi.org/10.1073/pnas.1704347114

Wertheimer, A. (1987). *Coercion*. Princeton University Press.

Wolf, S. (1980). Asymmetrical freedom. *The Journal of Philosophy*, *77*(3), 151–166. https://doi.org/10.2307/2025667

Wolf, S. (1990). *Freedom within reason*. Oxford University Press.

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

(*Appendices follow*)

Appendix A

Additional Materials, Data, and Code

Link to study materials, data, and code: https://researchbox.org/398.

Open Practices Statement

All studies reported in this article were preregistered. All study materials, preregistrations, data, and code are available at the links provided in Appendix A. All studies reported in this article were approved by the IRBs at Princeton University and Yale University (Table A1).

Table A1

Index of Supplemental Materials (Available on ResearchBox)

Section	Page
Supplement 1: Studies 1–2 Additional Analyses	2
Supplement 2: Within-Vignette Analyses	3–8
Supplement 3: Studies S1 and S2	9–15
Supplement 4: Studies S3–S4	16–23
Supplement 5: Study S5	24–33
Supplement 6: Study 5 Additional Analyses	34–37
Supplement 7: Study 6 Additional Analyses	38–40

Appendix B

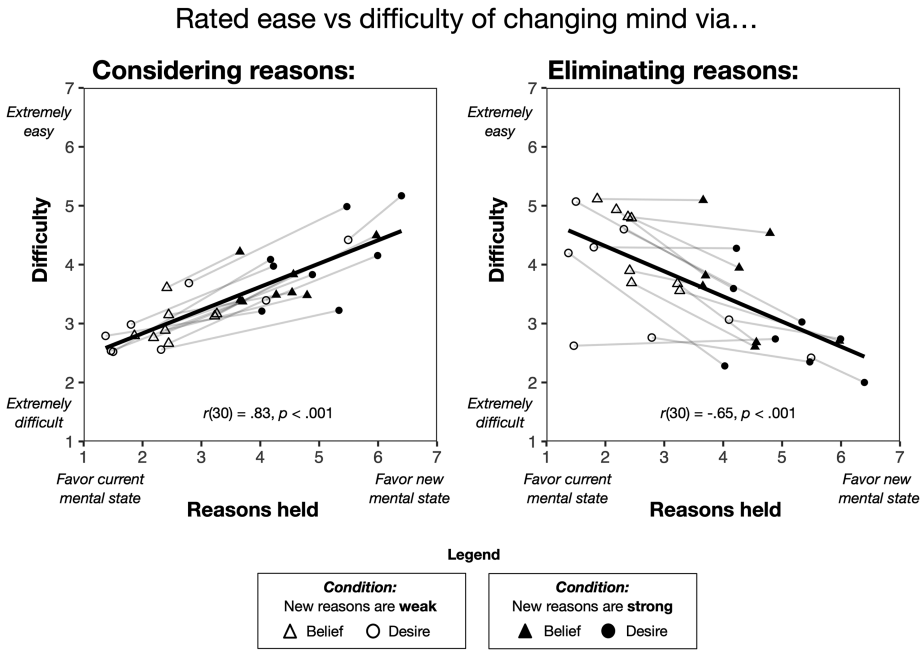
Association Between Perceived Reason Strength and Control in Studies S1–S4

In Studies 1 and 2, we examined the relationship between perceptions of how weakly or strongly the character's reasons favored their current or desired mental state, and how easy or difficult it seemed that the character could adopt their desired mental state. Here we report the same analysis—for each vignette, condition, and reaction type (reasoning vs. eliminating reasons)—for supplemental Studies S1–S4. The results of this procedure are displayed in Figure B1.

Studies S1–S4 replicated Studies 1 and 2. Across vignettes and conditions, we observed a strong—near perfect—relationship between how rationalizable the desired mental state is (based on the characters' reasons) and how difficult it seems for someone to change their mind by reasoning, $r(30) = .83$, $p < .001$. But this relationship reverses when participants consider how easy or difficult it would be if the agent actively suppresses the reasons they have, $r(30) = -.65$, $p < .001$.

Figure B1

Mean (and Standard Error) Judgments of Reason Strength (x-Axis) and Mental State Control (y-Axis) for Each Condition of Each Vignette in Studies S1 and S3 (Triangles) and S2 and S4 (Circles)



Note. Gray lines connect the "weak" and "strong" reason conditions of the same vignette: Slope of line represents how the change in perceived reason strength related to change in perceive control within that vignette. Data are divided by the characters reaction to the new information: reasoning (left panel) and intentionally forgetting reasons (right panel).

(*Appendices continue*)

In Studies S1–S4, this latter reaction involved the character taking a pill that (they know) would cause them to forget the information that they had just learned. Although a memory modification pill of this kind is highly unrealistic, it nevertheless serves a useful purpose. One potential limitation of the two suppression reactions in Studies 1 and 2—"forgetting" and "trying not think about the information"—is that in realistic situations these are either highly unreliable or they occur over a long time. Thus, participants might either attribute control to the character assuming that the suppression technique was unsuccessful, or attribute control by assuming that something about their situation has changed (that then alters the reasons available). The memory modification pill both works immediately and perfectly, thereby avoiding this ambiguity. And indeed, when the character's reasons strongly favored their current mental state, instead of their desired one, participants appeared to attribute greater control via memory modification (Studies S1–S4) compared to more mundane forms of suppression (Studies 1–2).

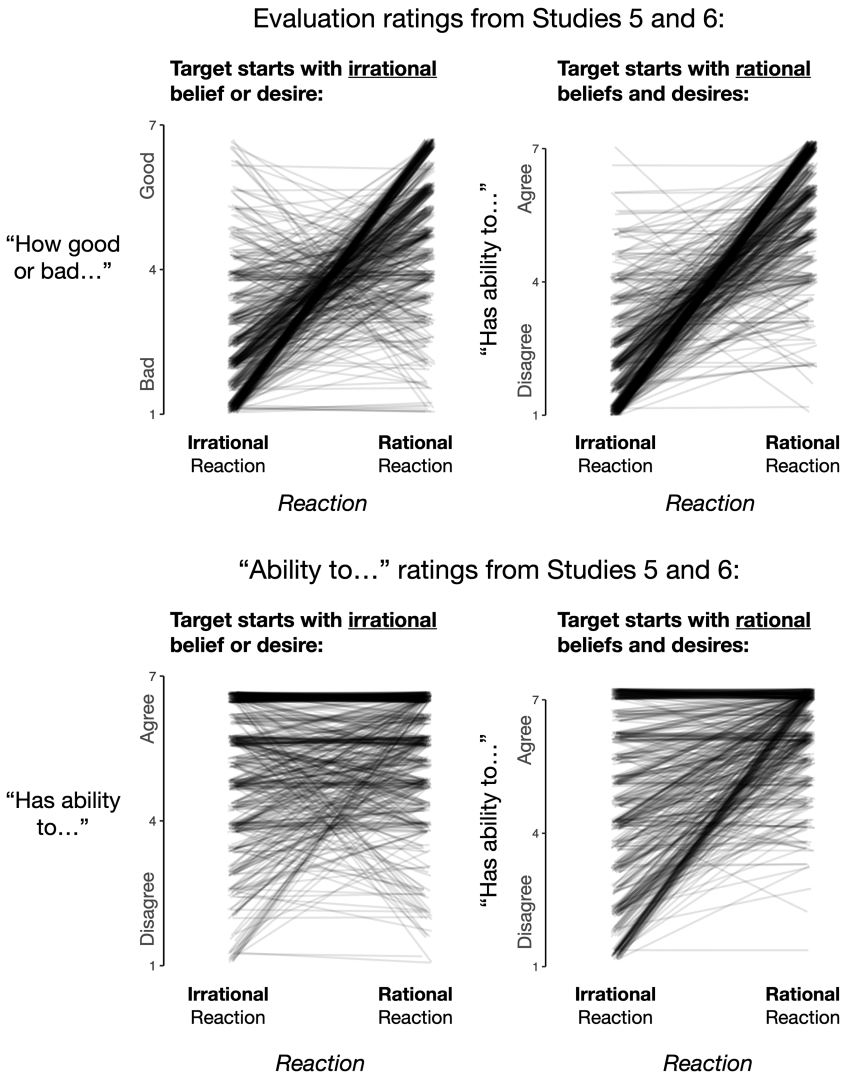
Appendix C

Detailed Analyses for Studies 5 and 6

In Studies 5 and 6, we noted that in one condition the data reflected a bimodal distribution. Specifically, in the "rational starting mental states" condition, participants split in their judgments concerning the character's ability to react irrationally. Some strongly agreed that the character could, but others strongly disagreed (the opposite end of the response scale) that the character could. Figure C1 displays

Figure C1

Evaluations (Top Row) and "Ability" Judgments (Bottom Row) From Studies 5 and 6



Note. Each line represents one participant's judgments regarding the character's potential intention formation and potential action. Lines are slightly jittered.

(*Appendices continue*)

Table C1
Most Common Response Pairs for the Rational and Irrational Reactions in Studies 5 and 6

		Holds rational mental states		Holds irrational mental state	
Study	Ranking	Rating pair (Rat–Irr)	*N* (%)	Rating pair (Rat–Irr)	*N* (%)
Study 5 (Belief)	1	7–7	164 (29)	7–7	161 (27)
	2	7–1	67 (12)	6–6	71 (12)
	3	6–6	36 (6)	5–5	45 (8)
	4	7–5	29 (5)	4–4	27 (5)
Study 6 (Desire)	1	7–7	158 (28)	7–7	229 (41)
	2	7–1	56 (10)	6–6	62 (11)
	3	6–6	43 (8)	5–5	30 (5)
	4	7–2	32 (6)	5–6	24 (4)

Note. Rat = rational; Irr = irrational.

participants' responses, across Studies 5 and 6. This figure also displays how each participant's judgments of the target irrational reaction (e.g., "piloting into the storm") and rational reaction ("piloting back to port") compared to each other.

Examining participants' judgments this way provides additional evidence that participants can be grouped into roughly two camps: (a) those who see no difference in someone's ability to act rationally or irrationally, and think others are capable of both, and (b) those whose judgments regarding what someone has the ability to do are isomorphic with their judgments of what is rational or good for them to do. When the character holds an irrational belief or desire (Figure C1, left column), most of the sample looks like Group 1 with people varying primarily with how confident they are that the character has the ability to do either. As shown in the Table C1, in both Studies 5 and 6, the majority of participants both reported the same capacity to react rationally and irrationally and gave a judgment of 4 or higher.

The distribution of responses changes when we examine responses to characters who hold only rational mental states. One distribution of responses looks again like Group 1: The most common, and third most common, reactions are to attribute the same high capacity to the character to react both rationally and irrationally. These reactions are the dark horizontal bands at the top of the scale in the lower left panel of Figure C1. But now the second most common reaction, and in Study 6 the fourth most common reaction, was to attribute maximum capacity to react rationally while simultaneously attributing a minimum (or near-minimal) capacity to react irrationally. This reaction—which corresponds to the thick, slanted band in the lower right panel of Figure C1—displays judgments of what the character has a capacity to do that are isomorphic with their judgments of what is good or bad to do (Figure C1, upper right panel).

Received November 3, 2021
Revision received November 29, 2023
Accepted December 9, 2023 ■