COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 0 (2022) e13170

ISSN: 1551-6709 online
DOI: 10.1111/cogs.13170

Reasoning About *Want*

Hillary Harner,[a] Sangeet Khemlani[b]

[a]*Altamira Technologies Corporation*
[b]*US Naval Research Laboratory, Navy Center for Applied Research in Artificial Intelligence*

Received 2 June 2021; received in revised form 6 May 2022; accepted 11 May 2022

Abstract

No present theory explains the inferences people draw about the real world when reasoning about "bouletic" relations, that is, predicates that express desires, such as *want* in "Lee wants to be in love". Linguistic accounts of *want* define it in terms of a relation to a desirer's beliefs, and how its complement is deemed desirable. In contrast, we describe a new model-based theory that posits that by default, desire predicates such as *want* contrast desires against facts. In particular, *A wants P* implies by default that *P* is not the case, because you cannot want what is already true. On further deliberation, reasoners may infer that *A believes*, but does not know for certain, that *P* is not the case. The theory makes several empirical predictions about how people interpret, assess the consistency of, and draw conclusions from desire predicates like *want*. Seven experiments tested and validated the theory's central predictions. We assess the theory in light of recent proposals of desire predicates.

Keywords: Desire predicates; Want; Bouletic reasoning; Mental models; Dual processes

1. Introduction

Some desires cause people to act, such as the desire to eat or sleep or watch a movie. Others remain dormant or unrealized for the entirety of a person's lifetime, as with the plight of the would-be world traveler who never makes it abroad. While the state of desiring something

Authors' note: This work was supported by an NRC Research Associateship Award to HH and funding from the Office of Naval Research to SK. We are indebted to Krista Casler, Bokyung Mun, Tony Harrison, Laura Hiatt, Laura Kelly, Greg Trafton, conference attendees at CogSci 2020 and 2021, and this journal's anonymous reviewers for their advice and comments. We also thank Kalyan Gupta, Danielle Paterno, Kevin Zish, and the Knexus Research Corporation for their help in data collection.

Correspondence should be sent to Hillary Harner, Altamira Technologies Corporation, 8201 Greensboro Drive, Suite 800, McLean, VA 22102, USA. E-mail: hillaryjharner@gmail.com

does not guarantee any particular action or outcome, it can lead listeners to make inferences about the world. For instance, it seems reasonable to draw the conclusion below:

1) Jiro wants to be a pilot.
Therefore, Jiro is not a pilot.

The premise in (1) expresses a *bouletic* relation—that is, a relation that concerns an individual's desires—between Jiro and the complement of *want*, that is, "to be a pilot." Indeed, predicates such as *want*, *wish*, and *be glad* are desire predicates (see, e.g., Heim, 1992) since they all express bouletic relations. Desire predicates are part of a larger family of verbs known as propositional attitude verbs, namely those verbs that describe an individual's "attitudes" toward propositions (e.g., *know, believe, say, advise*). Attitudes roughly correspond to mental states, for example, knowledge or belief, but any verb that takes complements that express propositions—we use the term "complement" to refer to these propositions—is generally deemed an attitude verb, even if there is no distinct mental state that corresponds with that verb, as with, for example, *say* and *advise*.

Linguists and philosophers have examined the meaning and inferences of desire verbs, such as *want*. They have examined, for instance, an inference closely related to (1) above, namely that if *A wants P*—where *P* is a sentential-like complement—then *A* believes that both *P* and *not-P* are both possible. More concretely, if Jiro wants to be a pilot, then Jiro believes that it is possible for him to be a pilot, and likewise that it is possible for him not to be a pilot. The pattern follows from Karttunen's (1973, 1974) observation that *A wants P* can be true even if the presuppositions of *P* are false. As a general rule, a sentence cannot be true if its presuppositions are not themselves true. For example, the sentence *it stopped raining* presupposes that it was raining. If in fact it was never raining, then the sentence cannot be true. However, Karttunen notes that this general rule does not apply to attitude verbs: sentences with verbs, such as *want*, can be true even when their presuppositions are false, so long as *A* believes those presuppositions. So a sentence such as *Hannah wants it to stop raining* can be true if Hannah believes that it is raining; the presupposition need not be satisfied in the general context. As a consequence, linguists have argued that desires are grounded in people's beliefs: what we want is restricted by what we believe to be true or possible (von Fintel, 1999, 2018; Geurts, 1998; Giorgi & Pianesi, 1997; Grano & Phillips-Brown 2020; Harner, 2016; Heim, 1992; Homer, 2015; Portner, 1997; Portner & Rubinstein, 2012, 2020; Rubinstein, 2012; Schlenker, 2005; Staniszewski, 2019; Villalta, 2008). They propose that *A wants P* presupposes that *A believes that P is possible* and that *A believes that P is false* (e.g., Harner, 2016; Heim, 1992; Portner, 1997; Rubinstein, 2012; Schlenker, 2005; Villalta, 2008; von Fintel, 1999). In other words, under these accounts, the complements of desire verbs presuppose two things about a person's beliefs, not about reality itself. But as a consequence, linguistic theories have no account of (1) above, because the inference in (1) is about the world, not about beliefs about the world.

This paper's goal is to outline and present evidence that tests a novel theory of the meaning and mental representation of desire predicates, with a specific focus on *want*. It first provides an overview of desire predicates in the linguistics literature to identify gaps in the literature's account for how people reason about *want*. It then describes the novel account that aims to

address those gaps. The theory adopts a modal semantics such that reasoners comprehend the meaning of *want* by mentally simulating hypothetical possibilities (Khemlani, Byrne, & Johnson-Laird, 2018). The paper reports experiments that test several novel predictions of the theory: first, that desire verbs are treated similarly to verbs that cause people to draw negative conclusions (Experiment 1); second, that verbs such as *want* cause people to infer facts before they infer beliefs (Experiments 2 and 3); third, that reasoners are susceptible to illusions when reasoning about the desires expressed by *want* (Experiment 4); fourth, that people should distinguish *want*'s desires from intentions and plans (Experiment 5); fifth, that reasoners assess some conclusions as more consistent with *want*-premises than others (Experiment 6); and sixth, that reasoners use desire statements expressed by *want* to rule out disjunctive possibilities (Experiment 7). The paper concludes by assessing the novel theory in light of recent proposals of desire predicates.

2. Linguistic treatments of *want*

Linguists have examined a variety of patterns of how people comprehend and produce *want* and other desire predicates. For instance, a widely studied feature of *want* and other desire predicates concerns the fact that in Romance languages, such predicates variously require the verbs in their complements to be in the subjunctive mood (e.g., Anand & Hacquard, 2013; Bolinger, 1968; Farkas, 2003; Giorgi & Pianesi, 1997; Portner, 1997; Portner & Rubinstein, 2020; Schlenker, 2005; Villalta, 2008). The subjunctive mood does not exist in English, so it is often described as the verbal conjugation that is used with verbs in the complement of Romance attitude verbs like Spanish *querer* "want," *preferir* "prefer," *temer* "fear," *lamentarse* "regret," and *dudar* "doubt." All of these take complements in the subjunctive mood, as shown with *querer* "want" below.

2) Elena quiere que Fabián esté en la fiesta. [*esté* is in subjunctive mood]
Elena wants Fabián to be at the party.

The subjunctive is contrasted with the indicative mood, which Spanish verbs like *saber* "know," *pensar* "think," *creer* "believe," *decir* "say," and *comprender* "understand" select for in their complements, as with *saber* "know" below.

3) Elena sabe que Fabián está en la fiesta. [*está* is in indicative mood]
Elena knows that Fabián is at the party.

The subjunctive mood is described by example because there is no agreed upon definition of what the subjunctive mood is. However, regardless of the theory, the basic starting point for all of them is that some verbs take the subjunctive and other verbs do not. As this is true for *want*'s equivalent across all Romance languages, *want* and other desire verbs are commonly studied and theorized about in relation to the subjunctive mood (e.g., Anand & Hacquard, 2013; Bolinger, 1968; Farkas, 2003; Giorgi & Pianesi, 1997; Portner, 1997; Portner & Rubinstein, 2020; Schlenker, 2005; Villalta, 2008).

Another major topic involves reasoning and presupposition under *want*. For instance, *Nick took a free trip* implies that *Nick took a trip*. But if it is true that *Nick wants a free trip*, does that imply that *Nick wants a trip*? The matter is controversial and unsettled (cf. Anand & Hacquard, 2013; Asher, 1987; Blumberg & Hawthorne, 2021; Crnič, 2011; Heim, 1992; Levinson, 2003; Portner & Rubinstein, 2020; Staniszewski, 2019; Villalta, 2008; von Fintel, 1999, 2018). There are some points on which linguists have reached relative consensus. For instance, theorists observe that *want* can express visceral desires, that is, desires we may have against our better judgment, for example, *I want to eat that whole chocolate cake* (Bolinger, 1974; Harner, 2016; Portner & Rubinstein, 2012); that *want* can be used to give advice, for example, *no, no, you don't want to guess 12, you want to guess 20* (Jerzak, 2019); that individuals can hold inconsistent sets of desires (Lassiter, 2011b; Levinson, 2003; Portner & Rubinstein, 2012; see also the General discussion); and that *not wanting* something can be ambiguous (Gajewski, 2005, 2007; Homer, 2015; Staniszewski, 2019; von Fintel, 2021), for example, if *Paula doesn't want a beer*, it may mean that she lacks desire for a beer, or that her desire is to have no beer.

Semanticists have defined *want* in at least two ways: one way treats *want* as a quantifier over possible worlds, or similar entities like situations or events, and which compare those possible worlds to some alternative set (e.g., Anand & Hacquard, 2013; Blumberg & Hawthorne, 2021; Grano & Phillips-Brown, 2020; Harner, 2016; Heim, 1992; Portner, 1997; Portner & Rubinstein, 2012; Rubinstein, 2012; Villalta, 2008; von Fintel, 1999, 2018). On these accounts, when *A wants P*, it implies that *A* prioritizes those possible worlds in which *P* holds as more desirable to those in which *P* does not. Another approach uses a decision-theoretic semantics that bases the meaning of desire on both the desirability and probability of the considered alternatives: hence, *A wants P* means that *A's* desire for *P* exceeds some threshold (e.g., Jerzak, 2019; Lassiter, 2011, 2011a,b; Levinson, 2003; Phillips-Brown, 2021; van Rooij, 1999; Wrenn, 2010; we address these approaches in the General discussion). Both accounts, however, treat desires as fundamentally belief-oriented (see the Introduction). Hence, both proposals are incapable of explaining the inference introduced in (1) above.

So, what can explain the inference in (1), that *Jiro wants to be a pilot* implies that he is not currently a pilot? To explain this inference, we begin with its classification. There are three traditional ways to categorize inferences: as a presupposition, a conversational implicature, or a conventional implicature (Grice, 1975). The inference may be best described as a conventional implicature, since it does not behave like a presupposition or a conversational implicature. Presuppositions are marked by needing to be true in order for their containing premise to be true. If a presupposition turns out to be false, then the premise that triggered the presupposition is either false (Russell, 1905) or valueless, that is, neither true nor false (Strawson, 1950), as in (4):

4) The jester stole the candy.
(There is a Jester.) [presupposition]

If there is in fact no jester, (4) cannot be true—it must be either false or have no truth value. But the inference that *Jiro is not a pilot* does not likewise impact the truth value of *Jiro wants to be a pilot*. To illustrate, suppose that Jiro is a pilot but he has amnesia and cannot remember

this fact. In such a case, the conclusion in (1) is true even though the premise is false. Thus, the conclusion is not a presupposition.

Similarly, the conclusion in (1) is not a conversational implicature (Grice, 1975). Conversational implicatures are nonliteral meanings that come from the conversational context rather than from particular words. They can be cancelled without affecting the truth value of the containing clause, as in this example:

5) *Quentin*: How do you like the painting?
Amari: It has a nice frame.
Amari doesn't like the painting very much. [conversational implicature]

Amari's answer in (5) conveys something other than its literal meaning; literally, it means that the frame is nice. But because this is not a direct answer to the question, it carries the implicature that the painting is not nice. This meaning is not tied to any particular lexical item—Amari could convey it with different wording, such as "I like the frame." Likewise, this implicature is cancellable without affecting the truth of Amari's response—if Amari actually likes the painting but perhaps mentioned the frame first because it caught her attention, she could cancel the conversational implicature by adding, "oh and I do also like the painting itself." This cancellation would not affect the truth of Amari's claim: it would still be true that Amari thought the frame was nice. While the inference we discuss in (1) can likewise be cancellable, as in a case where the speaker clarifies that Jiro has amnesia, it cannot be treated as a conversational implicature because it seems to arise from the word *want* specifically, that is, it is tied to a particular lexical item.

Thus, we categorize the inference in (1) as a conventional implicature. The term "conventional implicature" originates with Grice (1975), but he introduced it without much of a definition, so it is variously refined by later researchers. Potts (2015) notes that conventional implicature may be a sort of catch-all category and that the properties of such implicatures are heterogeneous. Thus, one general definition is that it is pragmatic in nature, and that it conveys backgrounded information that is independent of the literal content of a sentence. So, it can be true or false without affecting the entailment of the sentence. Yet, unlike conversational implicature, conventional implicature arises from a particular lexical item, as in (6):

6a) Shaq is huge but agile.
6b) Being huge normally precludes being agile.

The literal meaning of (6a) is that Shaq is both huge and agile, but the conventional implicature, arising from *but*, is as given in (6b) (cf. Potts, 2015, p. 30). This implicature can be canceled—the speaker could clarify that they said *but* to contradict the assumption that is commonly made about Shaq, that it is his hugeness that precludes his agility. The inference we illustrate in (1) seems similar: it arises from the word *want* in the premise of (1) and is separate from the literal content of that sentence: it is possible for the inference to be false without affecting the truth of the *want*-sentence.

What gives rise to this conventional implicature? We propose that by default, reasoners draw inferences such as (1) because knowledge is simpler to mentally represent and more

cognitively primitive than reasoning about belief states (cf. Phillips et al., 2020 survey of a broad literature that suggests that knowledge is more basic than belief). Reasoners initially assume that a person's desires are relative to what they know, that is, to what is factual. So, reasoners infer from *A wants P* that the complement *P* is false. However, they may consider the alternative possibility that a desirer holds a false understanding of reality, and they can infer that the complement is false only in the desirer's belief state—in other words, they can make the assumption that is the core of many linguistic proposals. For example, if a reasoner knows that Jiro has severe amnesia and has forgotten he is a pilot, then they would not draw the default inference in (1), but would instead conclude that Jiro *believes* he is not a pilot. But, considering this possibility demands cognitive effort, that is, this inference is a result of deliberation: a reasoner overrides the default inference about what is factual to instead infer something about a person's beliefs. Because linguistic theories of desire predicates are focused on this deliberative inference and make no mention of the initial default inference, they have no account of it. In what follows, we describe a theory of the mental representations that underlie reasoning about desire.

3. The mental representation of desire

Linguists commonly define propositional attitude verbs such as *want* as "modal" (Harner, 2016; Heim, 1992; Portner, 1997; Portner & Rubinstein, 2020; Rubinstein, 2012; Schlenker, 2005; Villalta, 2008; von Fintel, 1999), meaning that a person's desire describes a possible way that things could be (for background on modality broadly, see Portner, 2009). Yet, despite linguists' wide reliance on possibilities as part of the meaning of desire verbs and all other modals, the use of possibilities as a psychological construct is a topic of debate within cognitive science; many cognitive scientists ignore possibilities altogether (see Johnson-Laird, Khemlani, & Goodwin, 2015, for a review). But recent research by cognitive scientists supports the claim that people base many higher-level thought processes, such as moral reasoning and counterfactual thinking, on the mental representation of possibilities (Carey, Leahy, Redshaw, & Suddendorf, 2020; Johnson-Laird & Ragni, 2019; Phillips, Morris, & Cushman, 2019). Possibilities are highly relevant to how people represent desire predicates, such as *want*, because when a person wants something, or reasons about what another person wants, they are capable of bringing to mind the state of affairs where their desires come true, that is, a bouletic possibility. For instance, if Tarek wants to visit Abu Dhabi, he must be able to envision a scenario, that is, possibility, where he is in Abu Dhabi.

One theory that is founded on the mental representation of possibilities is *mental model theory*—the "model" theory for short. It argues that all forms of reasoning depend on the mental simulation of sets of possibilities (Johnson-Laird, 2006; Khemlani et al., 2018). It rests on three fundamental principles:

- **The principle of iconicity: models represent iconic possibilities**. The structure of a mental model reflects the structure of the real-world scenario it represents (Peirce, 1931–1958, Vol. 4). Hence, an iconic model of the spatial relation, *the thief is to*

the left of the bank consists of two tokens, one for the *thief* and one for the *bank*, arranged in the same spatial configuration as described in the relation. Iconicity allows reasoners to mentally scan a model from one component to another to make inferences. Models can represent static possibilities or situations that unfold in time (see Khemlani, Mackiewicz, Bucciarelli, & Johnson-Laird, 2013). They can also include abstract symbols from concepts that cannot be represented iconically, such as the symbol for negation (Khemlani, Orenes, & Johnson-Laird, 2012).

- **The principle of parsimony: people prefer to reason based on one model**. When people reason about relations, they construct a single possibility—a situation that describes a finite alternative scenario—consistent with those relations (Johnson-Laird, 2006; Khemlani et al., 2018). The spatial relation, *the thief is next to the bank* is true in many different scenarios—the thief could be to the left or to the right of the bank—but reasoners tend to construct, maintain, and reason on the basis of a single possibility. If they deliberate, they can discover alternative possibilities, but doing so demands time and effort.

- **The principle of coherence: a single model cannot represent an impossible situation**. Models are coherent. For instance, there is no possibility in which a thief is simultaneously *to the left of the bank* and *not to the left of the bank*, and so there can be no single model of that scenario, either. A consequence is that when reasoners learn new information, they use it to update their model in a way that yields a coherent, consistent representation of the information available. When new information cannot be integrated into an existing model, people judge the information to be inconsistent with what came before it (Johnson-Laird, 2012; Johnson-Laird, Girotto, & Legrenzi, 2004) and often attempt to construct explanations of the inconsistency (Khemlani & Johnson-Laird, 2011, 2012, 2013).

The model theory explains reasoning about causal relations (Khemlani, Bello, Briggs, Harner, & Wasylyshyn, 2021), temporal relations (Kelly, Khemlani, & Johnson-Laird, 2020; Schaeken, Johnson-Laird, & d'Ydewalle, 1996), and other sorts of abstract relation (Cherubini & Johnson-Laird, 2004; Goodwin & Johnson-Laird, 2005). No theory of reasoning accounts for reasoning about bouletic relations, and so we extend the model theory to account for inferences such as (1) above.

A bouletic relation, for example, *Jiro wants to be a pilot*, concerns an agent, *Jiro*, and a desired possibility, *Jiro is a pilot*. People can express bouletic relations using desire verbs (e.g., *want* and *hope*) and they can be paired with infinitival complements, for example, they can express desires about events or states to be realized by other people or by the attitude holder, as in (7a–d):

7a) Lee wants Chris to buy a bike.
7b) Lee wants Chris to be a lawyer.
7c) Lee wants to fly a plane.
7d) Lee wants to be in love.
7e) Lee wants an espresso.

Infinitival complements lack any tense, and as Johnson-Laird and Ragni (2019) observe, additional information is needed to turn them into propositions—the complement, *to buy a bike* in (7a) is neither true nor false. It is thus a propositional function that transforms the infinitival *Chris to buy a bike* into a proposition, namely, *Chris buys a bike*.

Want can also take noun phrases as direct objects; no predicate is needed (cf. 7e). Yet, we generally understand such sentences to express a desire about a relevant action carried out on the object, for example, we interpret (7e) to mean that Lee wants *to drink* an espresso. Accordingly, we construe verbs of desire as a relation between an agent and a possibility, which can be either an event or a state.

As (1) illustrates, an important constraint on bouletic relations such as *Lee wants to be in love* is that they imply by default that the complement is false, for example, that Lee is not in love. In general, bouletic relations abide by the constraint that an agent cannot desire what the agent knows to be true. Hence, (8a) is unacceptable; (8b) is not:

8a) * Katy Perry wants to be an American this year.
8b) Katy Perry wants to be a billionaire this year.

Katy Perry is already American, and so the desire expressed in (8a) is redundant. There is a reading of *want* where it expresses pride; on such a reading, (8a) may be felicitous—Katy Perry may take pride in being American this year—but the present theory does not deal with such an interpretation of *want* and focuses instead on why (8b) seems more felicitous than (8a).

In sum, statements of the form *A wants P*, where *P* is a verb phrase, make the following assumption in default of information to the contrary:

i) *P* is not true.

By consequence, this assumption also yields the assumption:

ii) *P* is a counterfactual possibility.

Since when *P* is not true, but a desired situation, it is a counterfactual possibility. While this assumption is related to the topic discussed by linguists—that *A* believes that *P* is possible when *A* wants *P*—it is distinct in that it is not about what *A* believes but rather a possibility in the general context. Possibilities can include any coherent scenario, including those that are inconsistent with knowledge and not actually possible to achieve in the real world. For instance, a reasonable desire may be to travel faster than the speed of light: while the speed is practically impossible and inconsistent with theories of physics, it is nevertheless possible to imagine traveling at such speeds. In general, people are capable of envisioning scenarios that are impossible because they do not fully flesh out all the details or infer the consequences of such possibilities. So long as possibilities are internally consistent with the information an individual wishes to consider, the possibility is coherent, and hence, people can envision scenarios that may be incomplete, hyperbolic, or technically difficult to achieve, such as "I want to live on the moon," or "I want this weekend to last forever" (cf. Heim, 1992, p. 199).

The above constraints suffice to explain the models of the possibilities that bouletic relations refer to. Reasoners should interpret the statement, "Jiro wants to be a pilot," by keeping

track of two distinct states of affairs: the first state of affairs is the literal meaning expressed by the complement, that is, the possibility of Jiro being a pilot. The present tense asserts that the complement has not been realized, so we refer to such situations as *future possibilities*, that is, possibilities that could come about in a future state of the world. In general, the complement of *want* is a future possibility; it is strange to use *want* rather than *wish* for past possibilities, for example, "Bill wants it to be the case that Sue won" is less felicitous than "Bill wishes that Sue won" (cf. Harner, 2016, p. 138 et seq.).

A second state of affairs concerns what Jiro's desire implies about the complement, that is, it is a fact for Jiro that he is not presently a pilot. The information can be depicted in the following diagram:

```
       FACT         FUTURE POSSIBILITY
Jiro  ¬pilot        pilot
```

The diagram shows tokens that stand in place of a desiring agent, Jiro (who is also the agent of the desires in this example), a desired future state of affairs, as well as a current state of affairs, that is, one in which Jiro is not a pilot. The diagram uses "¬," that is, the symbol for negation, to denote the factual state of affairs (see, e.g., Khemlani et al., 2012). The model represents the temporal relation between the possibilities on a spatial axis (see, e.g., Kelly et al., 2020; Schaeken et al., 1996): it represents current information to the left of a future possibility since the former precedes the latter. Reasoners can construct the model piecemeal, for example, they can first represent the assertion of the *want*-clause, that is, the future possibility that represents Jiro as a pilot, and then add the inferred information, that is, Jiro is not currently a pilot. The treatment above provides an account of people's initial interpretations of statements of the form *A wants P*, that is, a mental model that represents the statement. It suggests that *A wants P* causes people to infer that *P* is not the case.

One way to show that people take the complement of *want* to be false is to contrast it with how they reason about other attitude verbs. For instance, verbs like *decline* often imply that their complements are false:

9a) Jiro declines to be a pilot. [counterfactive]

Decline is "counterfactive" here, since its complement is false: if Jiro declines to be a pilot, it is the case that he is not one. "Factive" verbs, that is, verbs whose complements concern a fact, such as *manage*, show the opposite pattern:

9b) Jiro manages to be a pilot. [factive]

If Jiro manages to be a pilot, it is a fact that Jiro is a pilot. Since we propose that the complement of *want* and other desire predicates are taken to be false, the model theory makes the following prediction:

> **Prediction 1**. Reasoners should interpret desire predicates (of the form *A wants P*) to imply that the proposition expressed by *P* is false. As a result, people should make inferences from desire predicates that mimic inferences from counterfactive verbs (e.g., *decline*) rather than from factive verbs (e.g., *manage*).

Experiment 1 tested the prediction.

Another prediction concerns differences in the inferences people make by default and those they make after deliberation. While people draw the default inference that *P* is false when they reason about *A wants P*, they may deliberate and modify their initial mental model to instead represent *A*'s belief states directly. In other words, reasoners may revise the default model so that it concerns, not a fact about the present state of affairs, but a belief about it. Hence, it may be possible for Jiro to *want* to be a pilot and to *be* a pilot, but only in the odd scenario where he does not know that he is already a pilot. Such a change would require the following alteration to the default model of the desire expressed in (1):

```
       BELIEF       FUTURE POSSIBILITY
Jiro  ¬pilot        pilot                [model of Jiro's belief]
```

It would also demand that reasoners keep track of an additional model of the actual state of affairs:

```
FACT
pilot                                    [model of the facts]
```

Hence, deliberation should force reasoners to keep track of three states of affair one model to represent Jiro's (possibly false) belief that he is not a pilot, and his subsequent desire to become a pilot in the future; another model to represent the state of the world. This interpretation should be much more difficult to process because of an increased load on working memory. Hence, their deliberation may demand additional effort and time, so reasoners, more often than not, should prefer the default model to the deliberated model. An immediate consequence of the theory is that if most reasoners default to interpreting *not-P* as a fact, reasoners should be more likely to describe *A* as "knowing" *P* instead of merely "believing" it. This distinction follows because *know* is factive and *believe* is not, and so the theory makes the following prediction:

> **Prediction 2**. Reasoners should be more likely to interpret *A wants P* as implying that *A knows that P is not true* than *A believes that P is not true*.

Experiments 2 and 3 tested this second prediction.

As we outlined above, a basic assumption of the model theory is that people build and scan representations of possibilities to reason about them. A corollary of the principle of parsimony is that by default, mental models represent only what is true of a particular situation, not what is both true and false. As a result, reasoners are susceptible to "illusory inferences" (see Khemlani & Johnson-Laird, 2017, for a review). The model theory predicts and explains such cognitive illusions: they occur whenever, for instance, a reasoner infers that an impossible conclusion is possible, or vice versa. Research has revealed illusory inferences across a wide variety of reasoning domains, such as with reasoning about probabilities and conditionals (Johnson-Laird & Savary, 1996), disjunctions (Khemlani & Johnson-Laird, 2009; Sablé-Meyer & Mascarenhas, 2021), and Boolean concepts (Goodwin & Johnson-Laird, 2010).

The theory accordingly predicts that illusions should occur in bouletic reasoning. Consider the following set of premises:

10) Christina wanted Argentina and Brazil to place.
Jeremiah wanted Brazil or else Chile to place.
Only one of them got what they wanted.
Is it possible that Argentina and Brazil placed but Chile did not?

If you answered the question affirmatively, then you fell prey to an illusion: the correct answer is "no," because if Argentina and Brazil placed but Chile did not, then both Christina and Jeremiah got what they wanted—but the third premise above stipulates that only one of them got what they wanted. The model theory predicts that reasoners should fall prey to the illusion because reasoners tend to build parsimonious models that represent only what is true. For example, consider the set of models reasoners may build to represent the dueling desires in (10) (we depict only future possibilities in these models, with the proviso that reasoners are keeping track of facts as well):

```
            FUTURE POSSIBILITY
Christina   Argentina     Brazil
            FUTURE POSSIBILITY
Jeremiah    Brazil
                  Chile
```

Jeremiah's desire is for the disjunction of either Brazil or Chile to place, and this disjunction is represented by the possibilities being on two separate lines. The following scenario:

```
Argentina       Brazil       ¬ Chile
```

would seem to match Christina's desires well, and the model theory predicts that reasoners should use this match to infer the erroneous conclusion: that the scenario is possible given the premises. To appreciate that the scenario is in fact impossible, reasoners must represent falsity. Hence, they should flesh out their models of desire as follows:

```
            FUTURE POSSIBILITY
Christina   Argentina ¬ Brazil
            FUTURE POSSIBILITY
Jeremiah                Brazil    ¬ Chile
                       ¬Brazil      Chile
```

By doing so, they can recognize that the scenario described in the question in (10) satisfies both Christina's and Jeremiah's desires. A list of all of the ways in which Argentina, Brazil, and Chile could place is as follows:

```
 Argentina     Brazil     Chile
 Argentina   ¬ Brazil     Chile
 Argentina     Brazil   ¬ Chile
 Argentina   ¬ Brazil   ¬ Chile
¬Argentina     Brazil     Chile
```

¬Argentina ¬ Brazil ¬ Chile
¬Argentina Brazil ¬ Chile
¬Argentina ¬ Brazil ¬ Chile

The bolded lines highlight those possibilities compatible with Christina and Jeremiah's joint desires, that is:

Argentina Brazil Chile
Argentina ¬ Brazil Chile
¬Argentina Brazil ¬ Chile
¬Argentina ¬ Brazil Chile

The first possibility satisfies Christina's desire and conflicts with Jeremiah's. The other three possibilities do not satisfy Christina's desire, but they all satisfy Jeremiah's desire of only Brazil or Chile placing. Thus, the model theory predicts that reasoners will commit illusions of possibility when reasoning about *want*:

> **Prediction 3**. Reasoners tend to represent what is true and not what is false, so they should make illusory inferences from models of desire.

Experiment 4 tested this third prediction.

Many formal frameworks and philosophical treatments separate between desires and intentions: desires concern what an agent wants, and intentions concern what an agent plans to do (see, e.g., Brand, 1984; Bratman, Israel, & Pollack, 1988; Galitsky, 2013; Kinny & Georgeff, 1991; Rao & Georgeff, 1995; Searle, 1983; Thalberg, 1984). Yet, there is scant work on the models of psychological representations of desires and intentions. Quillien and German's (2021) recent account of intention suggests that intentions are causally dependent on "positive attitudes," and that a "desire for X is simply a positive attitude toward X"—a point that bears resemblance to the Aristotelian account of intention that defines it in terms of desires and beliefs (Aristotle, 300BC/1926, 1110a et seq.; see also Hume, 1740/1978). But an earlier theory by Malle and Knobe (2001) explains how reasoners distinguish the two. Based on natural language corpus studies and experiments, they propose a theory where a reasoner interprets another person's attitude as an intention if the following conditions hold: the activity is actionable by that person, the activity is perceived to be an output of that person's reasoning, and they perceive that person to be committed to performing the activity. In contrast, if one or more of these criteria is not met and the activity is instead not actionable by a person, is perceived as an input to reasoning, or the person is perceived to have no commitment to that activity, then reasoners conclude that the person's attitude is a desire. The model theory is compatible with Malle and Knobe's account, though it also explains the inference in (1). People can assess a person's attitude as a desire or an intention, and then represent it accordingly in their mental model. The model theory treats desires as one or more future possibilities and intentions as future actions that the intending agent can perform (cf. Malle & Knobe, 2001; Portner, 2004). This example illustrates the difference between the two:

11a) Aliyah wants Steve to listen to Dorothy Ashby. [desire]

11b) Aliyah plans for Steve to listen to Dorothy Ashby.
11c) Steve plans to listen to Dorothy Ashby. [intention]

In these examples, the verbs *want* and *plan* do the work of establishing the person's attitude as a desire or intention, respectively. Aliyah's desire in (11a) entails a future possibility where Steve listens to Dorothy Ashby, and is true even in a case where Aliyah or Steve have no intention to make this happen. However, for a sentence with *plan* to be true, as in (11b), the subject of *plan* must have an intention to carry out an action personally. In other words, (11b) must mean that Aliyah has a future action in mind to get Steve to listen to the album. Notably, (11b) cannot mean that Steve has an intention to listen to the album, but Aliyah does not have this intention. A sentence like (11c) must be used to convey that Steve, but not necessarily Aliyah, has an intention for Steve to listen to Dorothy Ashby. This demonstrates that *plan* is an intentional verb, and it describes the intentions of the person making the plans. And it shows that in general, people can have intentions only if they concern actions that they can perform themselves. However, they can desire outcomes they may have no control over, as in (11a).

Thus, people can express desires that may have no related intentions using bouletic verbs, such as *want, hope*, and *wish*. They can express intentions using verbs like *plan, be going*, and *will*. Since the model theory argues that desires and intentions are distinct from one another, it also follows that it permits representation of desires for objects that are complete opposites of the objects of an agent's intention. For example, suppose that Jessica plans on going into the office to work over the weekend even though she does not want to. The model of her attitudes would be as follows.

	FUTURE POSSIBILITY	FUTURE ACTION
Jessica	¬ work on weekend	work on weekend

This model is coherent because it keeps desires separate from intentions and thereby allows them to conflict. Malle and Knobe (2001) define desires as inputs to reasoning and stipulate that people resolve conflicting desires to establish the actions they intend to perform, so it might seem to follow that a person would not have desires for objects that conflict with the objects of their intentions. However, we propose that although a person may have reasoned over a variety of conflicting desires to establish an intent, that person may maintain representations of those desires. Just because Jessica has decided to work on the weekend does not entail that she no longer has the desire to not work on the weekend (cf. Davis, 1984, 1986, 2005; Harner, 2016). Because the model theory distinguishes desires from intentions, it asserts that people build different possibilities to represent desires and intentions, and makes the following prediction as a consequence:

> **Prediction 4**. Reasoners should consider sentences of the following form: *A wants to P* and *A plans to Q*, where *P* and *Q* conflict, as compatible with one another, because the objects of desires and intentions can conflict without being incoherent. In contrast, they should consider sentences of the form *A plans to P* and *A plans to Q*, where *P* and *Q* conflict, as inconsistent.

This prediction is unique to the model theory but incompatible with recent theories of intention (Quillien & German, 2021). Experiment 5 tested it.

The model theory makes additional predictions about when sentences should conflict with one another; consider the following:

12a) Aria visited Addis Ababa last year.
12b) Aria did not visit Ethiopia last year.

Provided that the first premise refers to the capital of Ethiopia, the two premises are inconsistent, that is, they cannot be true at the same time (Johnson-Laird, Legrenzi, Girotto, & Legrenzi, 2000). The model theory posits that reasoners without any background in logic can detect inconsistencies: they do so by building a model of a possibility in which every premise is true. If they can build such a model, the premises are consistent; otherwise, they are inconsistent (e.g., Kelly et al., 2020). Hence, reasoners should fail to build a model of the premises in (12), and then judge the premises to be inconsistent. Often, the detection of an inconsistency prompts reasoners to spontaneously construct explanations to figure out why the inconsistency arose in the first place (Khemlani & Johnson-Laird, 2011, 2012).

The model theory of bouletic reasoning accordingly predicts that reasoners should judge (13a) to be consistent more often than (13b):

13a) Amy has a black belt in karate.
Amy wants to be good at telling jokes.
13b) Amy has a black belt in karate.
Amy wants to be good at a martial art.

In (13a), the model of the first premise is:

```
      FACT
Amy   black-belt
```

and the model of the second premise is:

```
      FACT              FUTURE POSSIBILITY
Amy ¬ jokes             jokes
```

The two models can be combined to yield a single model:

```
      FACT              FUTURE POSSIBILITY
Amy ¬ jokes             jokes
      black-belt
```

that depicts the current state of Amy's abilities as well as a future possibility. In contrast, an integrated model of the premises in (13b) should yield the following:

```
      FACT              FUTURE POSSIBILITY
Amy ¬ martial-art       martial-art
      black-belt
```

The model shows that Amy is not good at a martial art while having a black belt—and reasoners who know that a black belt denotes competence in a martial art should consider the model incoherent, since its two facts cannot be integrated into a single mental simulation in which Amy has a black belt without being good at a martial art. Hence, the theory makes the following prediction:

> **Prediction 5**. Reasoners should be more likely to treat statements of the following form as inconsistent: *A is X* and *A wants to be X'* (where *X* implies *X'*). In contrast, they should judge the following pair of statements as consistent: *A is X* and *A wants to be Y* (where *X* does not imply *Y*).

Experiment 6 tested this prediction.

A corollary of the treatment above is that reasoners should be able to use representations of future, desired possibilities to make inferences about the present. Consider the possibilities established by the following statement:

14) Matt is a doctor.
Matt wants to be a radiologist.
Which is more likely to be true?
[ ] Matt is a radiologist.
[ ] Matt is an oncologist.
[ ] Both statements are equally likely to be true.

The second premise establishes a desire that implies that Matt is not a radiologist, that is, it yields the following model:

	FACT	FUTURE POSSIBILITY
Matt	¬ radiologist	radiologist

Reasoners should be more likely to conclude that Matt is an oncologist. In doing so, they eliminate a possibility out of a disjunctive set of alternatives. So, the model theory makes the following prediction:

> **Prediction 6**. When reasoning about a disjunction of the form *A is X or A is Y*, desire predicates of the form *A wants to be X* should rule out the first clause in the disjunction. Hence, such statements should cause reasoners to infer that *A is Y*.

which we tested in Experiment 7.

In what follows, we describe the seven experiments that tested predictions 1–6 above.

4. Experiment 1

Experiment 1 tested the model theory's prediction that reasoners should interpret desire predicates (e.g., *A wants P*) to imply that *P* is not the case. Hence, the theory posits that reasoners should treat such verbs similar to how they treat counterfactive verbs, such as

decline—which imply the falsity of the complements—instead of factive verbs, such as *manage*—which imply the truth of their complements. Thus, we tested how people rated the truth value of desire verbs' complements in comparison to those of counterfactual verbs and factive verbs. To more robustly evaluate desire verbs, we also tested how they compared to appearance verbs, such as *claim*, that imply neither truth nor falsity of their complements. The experiment accordingly provided participants with statements such as:

15) Alice wants to own a robot.
To what extent does this sentence strike you as true or false:
Alice currently owns a robot.

The first sentences were all of the same format: *A [verb] to own X*, where the matrix verb (in brackets) was a randomly assigned verb from one of four categories. Participants rated the truth of the statement using a slider scale.

4.1. Method

4.1.1. Participants

Fifty participants (mean age = 36.26 years; 27 males and 23 females) volunteered through the Amazon Mechanical Turk online platform (AMT; see Paolacci, Chandler, & Ipeirotis, 2010 for a review) for monetary compensation. All participants reported English as their native language.

4.1.2. Open science and preregistration

Effects were preregistered prior to data collection. Data, materials, and experimental code for this study and all subsequent studies are available through the Open Science Framework. See Appendix A for links.

4.1.3. Design, procedure, and materials

Participants carried out 12 problems. Each trial presented participants with a statement describing an individual (e.g., Alice) and an object that can be owned, as linked by an attitude verb. The study manipulated the verb used in each statement: verbs could imply facts (factives: *manage*, *happen*, *turn out*), negations of facts (counterfactives: *refuse*, *fail*, *decline*), perceived facts (appearance verbs: *seem, appear, claim*), and desires (desire verbs: *want*, *wish*, *hope*). The problems in the study were constructed by randomly pairing a pool of 12 names to the set of 12 verbs and 12 objects (see online Appendix A). Hence, no two participants saw the same set of problems. Participants then assessed whether the agent's ownership of the object was true or false by rating the truth of a statement such as "Alice currently owns a robot" on a 7-point Likert scale that ranged from –3 (the sentence is definitely false) to 0 (I cannot be certain) to +3 (the sentence is definitely true).

4.2. Results and discussion

Fig. 1 shows participants' mean ratings for the four types of verbs in Experiment 1. A Friedman nonparametric analysis of variance showed that participants' tendency to infer

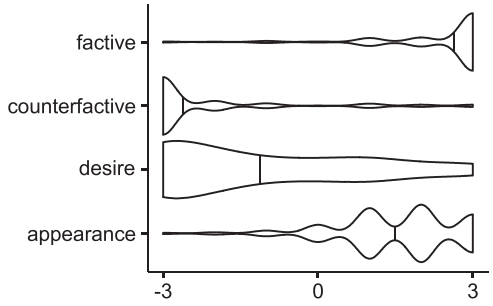

Fig. 1. Participants' truth-ratings (–3 = definitely false, 0 = I cannot be certain, +3 = definitely true) for the four different types of verbs in Experiment 1. Circles denote participants' individual ratings, and violin plots provide a smoothed out distribution of participant responses; bars at the center of each plot indicate the mean response.

embeddings as true differed as a function of the verbs in the sentences ($\chi^2 = 96.82$; $p < .001$). An analogous generalized mixed-model (GLMM) regression treated the materials as random effects and the four types of verb as a fixed effect; the regression further validated the differences between the four types of verb in the study ($|\ \beta\ |$'s > 0.89, $|\ t\ |$'s > 4.66, p's $< .002$).

The figure shows that participants responded sensibly: for factive verbs in statements, such as "Alice manages to own a robot," they rated statements describing the current state of the complement, for example, "Alice currently owns a robot," to be true ($M = 2.42$); these ratings were reliably greater than chance performance (i.e., a mean of 0; Wilcoxon test, $z = 6.23$, $p < .001$, Cliff's $\delta = 0.92$). For counterfactive verbs, they rated such statements as false ($M = –2.23$), and their ratings were reliably lower than chance (Wilcoxon test, $z = 5.98$, $p < .001$, Cliff's $\delta = 0.76$).

Participants yielded intermediate truth-ratings for appearance verbs ($M = 1.54$) and desire verbs ($M = –1.08$). The model theory posits that desire verbs should pattern similarly to counterfactive verbs, since desire verbs imply that their embeddings are false. The results of Experiment 1 were mixed: on the one hand, desire verbs' complements were rated as false overall, at a rate significantly lower than chance (Wilcoxon test, $z = 3.86$, $p < .001$, Cliff's $\delta = 0.44$); reasoners were more likely than not to conclude that "Alice currently owns a robot" is false given that "Alice wants to own a robot." On the other hand, as Fig. 1 shows, desire verbs produced a weaker inference than counterfactive verbs (M_{desire} vs. $M_{\text{counterfactive}}$, Wilcoxon test, $z = 4.79$, $p < .001$, Cliff's $\delta = 0.46$). Hence, the results of Experiment 1 only partially corroborated prediction 1; participants rated both desire verbs' and counterfactive verbs' complements as false, but their rejection was much stronger for counterfactive verbs.

Somewhat similarly, but on the other side of the scale, participants inferred that the complement of an appearance verb, for example, *appear*, was true, as in "Alice appears to own a robot." This finding was not fully in line with our expectations, since *appear* does not guarantee anything about whether Alice owns the robot or not, that is, it is neither factive nor counterfactive. Perhaps, this rating stems from these verbs having an evidential component to their meaning (cf. Gisborne & Holmes, 2007). Evidentials include those words people use to

indicate how they acquired the information they are reporting on, for example, *evidently*. *Evidently* communicates that the speaker infers the presented information from some perceived evidence, for example, your wet coat leads me to conclude that "evidently it's raining." Thus, evidentials allow us to hedge about information being true. If appearance verbs are evidential, it is possible that participants believed that they were likewise hedging that Alice's ownership of the robot was true as far as the evidence permitted. Thus, Experiment 1 corroborated the first prediction of the model theory, with some caveats. While participants inferred the falsity of facts from statements about people's desires, they did so more strongly for counterfactive verbs. Perhaps, this is due to the study design that asked reasoners to consider the same complement under 12 different verbs. Comparing desire verbs to straightforward classes like factives and counterfactives and a more nuanced class like the appearance verbs may have encouraged participants to slow down their reasoning process on the whole. Thus, they might have switched from concluding the default inference to the deliberated inference under desire verbs. Or, it may be the case that the complements of desire verbs are simply not perceived as strongly false as the complements of counterfactives.

Experiments 2 and 3 used a different task and a stronger methodology to explore participants' reasoning behavior. And they tested the theory's second prediction, that reasoners should be more likely to infer knowledge than beliefs when reasoning about desires.

5. Experiment 2

Experiment 2 tested prediction 2: reasoners should interpret the sentence, "Jiro wants to be a pilot" to imply by default that Jiro is not a pilot. Some reasoners may deliberate over the possibility that Jiro merely believes that he is not a pilot, but they should do so less often. As a consequence, reasoners should be more likely to conclude that Jiro *knows* that he is not a pilot rather than that Jiro merely *believes* that he is not a pilot, since *know* is a factive and *believe* is not. To test the idea, Experiment 2 provided participants with premises such as the following:

16) Jackie wants Naomi to get her driver's license, which means that…
[ ] she knows that Naomi does not have her driver's license.
[ ] she believes that Naomi does not have her driver's license.

They selected the option that best completed the sentence. On some trials, participants chose between the verb *know* and the verb *believe;* on other trials, they chose between the verb *know* and the verb *think*.

Half of the problems in Experiment 2 concerned comparisons as in the example above. The other half of the problems were used to mask the intent of the study by providing fillers that tested their engagement in the study. Those problems provided participants with general knowledge questions that always have correct answers, such as the following:

17) Miriam picked a peach, which means that…
[ ] she picked a fruit.
[ ] she picked a vegetable.

Any participant who paid attention to the problem would choose the former over the latter. Those who got more than one filler question wrong were dropped from the analysis.

5.1. Method

5.1.1. Participants

One hundred and six participants (mean age = 35.51 years; 62 males, 43 females, 1 preferred not to say) volunteered through AMT. All but two participants reported being native English speakers; these two were dropped from further analysis. Likewise, any participant who responded incorrectly on more than one filler problem was dropped from analysis (26 in total). We analyzed the remaining 78 participants.

5.1.2. Open science and preregistration

Predicted effects were preregistered prior to data collection.

5.1.3. Design, procedure, and materials

Participants carried out 16 trials in total; half of them were test problems, and half were filler questions. On each trial, participants read a statement, such as "Mason wants the Olympics to be held this summer, which means that…." The test problems concerned desire predicates, and the filler questions concerned general trivia. On each trial, participants chose between two options to select which one best completed the sentence. For the eight test problems, one of the responses concerned what Mason knew, and the other concerned what Mason thought or believed, for example,

[ ] he knows that the Olympics are not being held this summer.
[ ] he believes that the Olympics are not being held this summer.

For the eight filler items, for example, "Sophia has a degree in physics, which means that…," one of the two options was factually correct, and other was factually incorrect, for example,

[ ] she studied physics.
[ ] she studied accounting.

The experiment randomized the names of the individuals as well as the order of the two options for all the trials. Online Appendix A provides the link to all experiment material, including the fillers as well as the test problems.

5.2. Results and discussion

Answers to filler questions were 88% correct across the study as a whole, but some participants provided many incorrect answers and other participants responded accurately to all the filler questions. Hence, we report data on test problems from only those participants who responded correctly to at least seven of the eight filler questions.

One reviewer pointed out a confound with some of the test items, namely that for some of them, a person normally has the relevant knowledge about whether their desire is already realized or not. For example, if "Jackie wants it to be Tuesday," Jackie normally knows whether

it is already Tuesday or not, so it is a bad candidate to test whether an inference about *know* or *believe/think* can be drawn from the desire statement. Thus, we report on the results for those items that do not exhibit the confound (they mirror the results for all items), as well as all eight of the filler items. Participants selected the *know* option on 69% of test problems, which was significantly greater than chance performance (Wilcoxon test, $z = 4.82$, $p < .001$, Cliff's $\delta = 0.55$), and 58 out of 78 participants exhibited the pattern (binomial test, $p < .001$ with a prior probability of .5). The result corroborates the theory's second prediction. Participants' tendency to prefer the *know* option did not differ as a function of whether the verb in the alternative option was *think* or *believe* (70% vs. 68%, respectively; Wilcoxon test, $z = 0.82$, $p = .41$, Cliff's $\delta = 0.06$).

One possible consequence of the model theory concerns the relative amount of time it takes to process default and deliberative mental models of bouletic relations. If reasoners need to deliberate in order to override their default inference about bouletic relations, that is, if they need to process the model to recognize that a person may simply *believe* that a desired outcome is not the case, then perhaps people should be faster at selecting *know* responses than *think* or *believe* responses. The data do not bear out this prediction: the experimental methodology was not sensitive enough to detect a reliable difference between when people selected a *know* response (12.06 s on average) versus when they did not (11.97 s on average; Mann–Whitney test, $z = 0.83$, $p = .41$, Cliff's $\delta = 0.04$).

One concern with the study is that it might have imposed an artificial distinction on participants: some of them might have treated knowledge as implying belief: if a person knows *P*, it could be taken to imply that the person also believes or thinks *P*. The study design forced participants to pick between the *know* and *think/believe* when those options could have been compatible with one another. Experiment 3 ruled out the issue as a concern.

6. Experiment 3

Experiment 3 differed from Experiment 2 in that it replaced the two problematic test items with new material. It also provided every question with four possible answers. For the test items, the two added possible answers were of the form (1) x both knows and thinks not y, and (2) x does not know and does not think not y, such as with the following:

18) Daymond wants the book to be on the top shelf, which means that…
[ ] he knows that the book is not on the top shelf.
[ ] he thinks that the book is not on the top shelf.
[ ] he both knows and thinks that the book is on the top shelf.
[ ] he doesn't know and doesn't think that the book is not on the top shelf.

The third option allowed participants to not have to choose between *know* and *think* options. The fourth option allowed them to reject any inference of knowledge or thought from the *want* statement. As with Experiment 2, half the test items randomly used *think* as the alternate verb to *know*; the other half used *believe*. As well, half of the problems were test items and half were controls. Control items tested general knowledge and had four

Table 1
Percentage of trials on which participants selected the four available options in Experiment 3

Option selected	Percentage
knows	54%
thinks/believes	31%
both knows and thinks/believes	11%
doesn't know and doesn't think/believe	4%

possible answers, where only one was correct. Participants had to select one option of the four provided.

Since the theory predicts that people infer knowledge from desire statements before they infer beliefs, it predicts that participants should choose either the first or the third options—with *know* or with *know* and *think/believe*—more often than the second and fourth options.

6.1. Method

6.1.1. Participants

Fifty-three participants (mean age = 35.1 years; 25 males, 28 females) volunteered through AMT. All but one participant reported being native English speakers; he was dropped from further analysis. Likewise, any participant who responded incorrectly on more than two filler problems was dropped from analysis (17 in total). We analyzed the remaining 35 participants.

6.1.2. Design, procedure, and materials

The experiment was identical to Experiment 2 in design and procedure except that all problems had to be answered from a set of four possible answers, as described above. The materials varied from Experiment 2 in that two test items were replacements of the two confounded test items in Experiment 2. All control items were identical, except that all had an additional two incorrect choices. All problems and their answers were randomized for all trials.

6.2. Results and discussion

Participants responded correctly to 75% of the test items in Experiment 3. Since this average was lower than Experiment 2, we eliminated participants less stringently, evaluating responses of participants who incorrectly answered up to two of the eight filler questions. Table 1 provides the percentages of times that participants picked from the four test choices. Participants selected the *know* option the most often—54% of the time—which was significantly greater than chance performance (Wilcoxon test, $z = 4.49$, $p < .001$, Cliff's $\delta = 0.71$), and 25 out of 35 participants exhibited the pattern (binomial test, $p < .001$ with a prior probability of .25). And, since it is possible that reasoners may think that to know something entails belief or thought, the finding that participants chose the *both* option 11% of the time further reinforces that participants inferred knowledge from desire statements more often than they inferred belief or thoughts. These findings thus replicate and extend the results from Experiment 2, and they support the theory's second prediction.

Analysis of participants' responses times reflected no reliable difference between when people selected the *know* option (15.19 s on average) over any of the other three options (18.35 s on average; Mann–Whitney test, $z = 0.83$, $p = .41$, Cliff's $\delta = 0.04$), despite trending in the predicted direction. However, participants took reliably longer to select the *both* or *neither* options (22.83 s) compared to the two others (15.65 s; Mann–Whitney test, $z = 2.22$, $p = .03$, Cliff's $\delta = 0.22$). Given that all four options were presented for each problem, and that the order in which they were presented was randomized, the results suggest an increase in processing time separate from the time it takes to initially process the longer sentences needed for the *both* and *neither* options.

Experiment 1 provided evidence for prediction 1 and Experiments 2 and 3 provided evidence for prediction 2. The results of the studies concern how people interpret *A wants P*, and the inferences associated with its interpretation. But people can reason about desires as well. Experiment 4 accordingly tested the prediction concerning people's tendency to represent only what is true at the expense of neglecting what is false. Experiment 5 tested people's reasoning about desire and intention. Experiments 6 and 7 sought to test two predictions based on how people assess the consistency of bouletic relations and use those relations to eliminate alternative possibilities.

7. Experiment 4

Experiment 4 tested the third prediction of the model theory, that reasoners should succumb to illusions of possibility, where a description of a set of desires makes reasoners conclude that a situation is possible when in fact it is impossible. The experiment provided participants with problems, such as:

19) Rebekah wanted the Hawks and the Cubs to win.
Derek wanted the Cubs or the Bills, but not both, to win.
Only one of them got what they wanted.

The study varied whether the second premise described an inclusive or an exclusive disjunction. Half of the problems concerned illusory conclusions, that is, conclusions that reasoners should draw if they represent only what is true in the premises, for example,

Is it possible that the Hawks and the Cubs won? [illusory conclusion]

The other half concerned control conclusions which reasoners should get right whether or not they represent what is false in the premises, for example,

Is it possible that the Cubs, but not the Hawks or the Bills, won? [control conclusion]

The theory predicts that the participants should be more accurate in their responses to control problems and illusions.

7.1. Method

7.1.1. Participants

Forty-nine participants (mean age = 36.1 years; 21 males, 28 females) volunteered on AMT. Every participant self-identified as a native English speaker.

7.1.2. Open science

Data, materials, experimental code, and analysis scripts are available online.

7.1.3. Design, procedure, and materials

Participants completed eight problems. Each problem consisted of three sentences that described two individuals' desires and a question, as in (19). For half the problems, the second person's desires concerned an inclusive disjunction (B or C or both) and for the other half, they concerned exclusive disjunctions (B or C but not both). The materials were drawn from a pool of objects and corresponding verbs, for example, biological events (e.g., "…wanted the roses and daffodils to bloom") and social events (e.g., "…want St. Joe's and Assumption Church to have a festival this weekend"). The experiment randomly assigned materials, such as the names of individuals, and objects/verb sets to experimental and control trials, such that no participant ever received the same set of materials in the same arrangement of experimental and control problems. The order of the eight problems was randomized for each participant. Participants had to choose between "yes" and "no" to answer each problem.

7.2. Results and discussion

Participants provided the correct answer for control items more often than they did for experimental items (71% vs. 35%, Wilcoxon test, $z = 4.78$, $p < .001$, Cliff's $\delta = 0.58$). Accuracy was not affected by whether desires for B or C were expressed as an inclusive or an exclusive disjunction for control problems (73% vs. 69%, respectively, Wilcoxon test, $z = 0.66$, $p = .51$), or experimental problems (37% vs. 33%, Wilcoxon test, $z = 0.54$, $p = .58$). A posthoc by-item analysis found that participants tended to make fewer errors when the premises in the problem sets were longer. For example, participants performed worse on problem sets containing sentences like, "Danielle wanted the roses and the daffodils to bloom" than they did for problem sets with sentences like, "Peter wanted St. Joe's and Assumption Church to have a festival this weekend" (respectively, they gave the correct answer 16% vs. 43% of the time for these examples). It may be that the length of the item slowed participants down and caused them to consider the options more carefully, or it may be that the meaning of the materials had an analogous effect. Nevertheless, the study confirmed the third prediction of the model theory, that reasoners systematically reason that sets of premises are possible, that is, compatible with desires, when they are in fact impossible.

8. Experiment 5

Experiment 5 tested prediction 4: people should judge sentence pairs of the form *A wants to P* and *A plans to Q*, where *P* and *Q* conflict, as consistent, but in contrast, they should judge pairs of *A plans to P* and *A plans to Q*, where *P* and *Q* conflict, as inconsistent. Control sentence pairs used *plan* as the matrix verb in both sentences, with the complements *P* and *Q* conflicting with each other:

20) Keegan plans to spend the next hour alone.
Keegan plans to spend the next hour with friends. [control]

Experimental sentence pairs were of the form *A wants to P* and *A plans to Q*, where *P* and *Q* also conflicted:

21) Lucy wants to wake up at 10am tomorrow.
Lucy plans to wake up at 8am tomorrow. [experimental]

Even though the complements of the sentences in (21) conflict with each other, the theory predicted that participants would rate pairs like (21) as consistent since Lucy's desires may imply no intentions, unlike her plans. In contrast, (20) is problematic since the complements do conflict and *plan* implies intention, so participants should reject control sentence pairs.

8.1. Method

8.1.1. Participants

Forty-eight participants (mean age = 36.1 years; 21 females and 27 males) volunteered through AMT. All but one participant reported English as their native language; we dropped their data from our analysis.

8.1.2. Open science

Data, materials, experimental code, and analysis scripts are available online.

8.1.3. Design, procedure, and materials

Participants responded to 12 problems—six experimental and six control. Experimental problems consisted of sentence pairs, where the first sentence described a person's desire and the second a plan that was incompatible with this desire, as in (21). The control problems were similar in form except that the matrix verb of the first sentence was *plan* instead of *want;* the complements of the verbs were likewise incompatible with each other, as in (20). The experiment randomly assigned a pair of complements to have *want/plan* or *plan/plan* as their matrix verbs; no complement pair was designed for a particular matrix verb pairing. Each sentence pair was also randomly assigned a unique male or female name to serve as its subject. The order of presentation for the 12 problems was shuffled for each participant.

After reading a sentence pair, participants typed out their response to the question, "Can both sentences be true at the same time?" They responded with "yes" or "no" and could elaborate further if they wanted. We used this wording in the question as a way to assess

participant's consistency rating, since individuals without training in logic can be confused by the word *consistent* and interpret it in different ways (Johnson-Laird et al., 2004).

8.2. Results and discussion

Participants judged experimental *want/plan* sentence pairs to be consistent 65% of the time and control pairs to be consistent 22% of the time (Wilcoxon test, $z = 4.68$, $p < .001$, Cliff's $\delta = 0.65$). Thirty-two out of 48 participants yielded the pattern (binomial test, $p = .03$ with a prior probability of .5). Thus, Experiment 5 confirmed prediction 4: *wants* can be in conflict with *plans* without being inconsistent. People judge scenarios as consistent when they describe desires and intentions with conflicting complements, but not so when they describe intentions with conflicting complements. We performed a posthoc analysis of the answers that did not conform to this general trend. In cases where participants rated inconsistent plans as consistent, they tended to explain away the conflict, for example, one participant justified sentences, such as (21) by writing, "Yes. Technically, he could wake up at 8 am for a few moments, go back to sleep, and wake up again at 10 am." In cases where participants rated consistent *want/plan* pairs as inconsistent, their responses relied on an explanation of the complements as inconsistent. For instance, one participant read the following problem:

22) Henry wants to get a full refund on the movie ticket.
Henry plans to exchange the movie ticket for a different showing.

and denied its consistency by explaining, "no, because if Henry gets a refund then he can't also get an exchange." This could suggest that they assigned intention to *want*, or that they did not read the sentences closely. As an anonymous reviewer noted, a possible explanation is that the participant considered the fact that both of Henry's options are desirable and meet the general goal of him not having to use a ticket he did not want. So, for Henry to want a refund but not act on it is strange. This contrasts with an example like (21), where Lucy may only have the desire to get up at 10am, and not the desire to get up at 8am, so for her to want to get up at 10am but plan to get up at 8am may not seem inconsistent.

Overall, the study supports the theory's claim that the desires expressed by *want* have no necessary connection to intentionality, as desire is distinct from intention. The model theory represents this distinction by modeling desires—represented as future possibilities—of *want* separately from intentions, as expressed by verbs like *plan*, which it represents as future actions that the agent can perform. Thus, a model of the following pair of sentences:

23) Lia wants to reside solely in Norway.
Lia plans to reside solely in the US.

is as follows:

	FACT	FUTURE POSSIBILITY	FUTURE ACTION
Lia	¬ Norway	Norway	US

This model is coherent, since the possibility of residing solely in Norway is a future possibility, and the possibility of residing solely in the US is a future action. In other words, they do

not conflict with each other since they are kept separate in the model of Lia's mental states. If both sentences in (23) were about Lia's plans to reside solely in either Norway or the US, the model would be incoherent since the future actions conflict with each other.

9. Experiment 6

Experiment 6 tested prediction 5, that is, that people should judge statements of the form *A wants to be X* to be inconsistent with statements of the form *A is X*. It provided participants with pairs of sentences where the first sentence reported on a person's status or an activity they had completed, and the second sentence reported that person's desire using *want*. Half of the sentence pairs were controls, and the other half were designed to test prediction 5. For control pairs, the *want*-sentence reported on a desire that had no relation to the first sentence:

24) May has written 3 best-selling books.
May wants to be a doctor. [control problem]

For experimental pairs, the *want*-sentence reported on a desire whose complement is implied as already true by the first sentence.

25) May has written 3 best-selling books.
May wants to be an author. [experimental problem]

A person who has written three best-selling books is an author, so if the sentence *May wants to be an author* implies that May is not an author, then it conflicts with the first sentence. In general, if *A wants P* implies that *P* is not the case, then reasoners should consider the sentences as inconsistent. As with Experiment 4, we asked participants whether both sentences could be true at the same time. If prediction 5 is accurate, reasoners should judge that both sentences could be true at the same time more often for control pairs than experimental pairs.

9.1. Method

9.1.1. Participants

Forty-nine participants (mean age = 35.7 years; 27 males and 22 females) volunteered through AMT. All participants reported English as their native language.

9.1.2. Design, procedure, and materials

Participants carried out eight problems, four experimental problems and four controls, where each problem consisted of a pair of sentences. The first sentence described a fact about an individual's status or an activity they had engaged in, and the second sentence described some desire held by the individual. The same eight premises were used as the first sentence on each trial. Half of the second sentences were controls, that is, they concerned a desire that was irrelevant to the first sentence, and the other half were experimental sentences that described a desire to do or be something that the first sentence implied was already the case. The experiment randomly assigned whether the second sentence was control or experimental

from a pool of 16 materials, eight control and eight experimental. Each sentence pair was randomly assigned one of eight unique male or female names to serve as its subject. The order of presentation for the eight problems was shuffled for each participant.

After reading a sentence pair, participants typed out their response to the question, "Can both sentences be true at the same time?" They were asked to respond with "yes" or "no" and to elaborate on their response if they wanted.

9.1.3. Open science

Data, materials, experimental code, and analysis scripts are available online.

9.2. Results and discussion

Participants' responses were coded for whether they responded affirmatively or negatively, that is, whether they thought the two sentences were consistent or not. They judged control pairs to be consistent more often than experimental pairs (84% vs. 60%, Wilcoxon test, $z = 3.55$, $p < .001$, Cliff's $\delta = 0.43$). A follow-up GLMM regression treated the materials as random effects and the type of problem (control vs. experimental) as a fixed effect; it corroborated the difference between control and experimental pairs ($\beta = 1.22$, $z = 5.04$, $p < .001$). Nevertheless, they judged experimental patterns to be consistent reliably more than chance (i.e., a mean of 0.5; Wilcoxon test, $z = 2.08.$, $p = .038$, Cliff's $\delta = 0.27$), whereas the theory's fifth prediction hypothesized that reasoners should treat such statements as inconsistent. Hence, the study lends only partial support to the model theory: the results yielded the directional pattern described in prediction 5—participants judged control problems to be more consistent than experimental problems—but the results ran against the qualitative prediction that participants should judge experimental problems to be inconsistent.

It is not entirely clear why reasoners judged experimental items to be consistent as often as they did, but a post-hoc analysis of participants' natural responses showed that they distinguished reasoning about inconsistent experimental items from control items. This post-hoc analysis examined the spontaneous use of the word *already* in participants' written responses. It found that they used *already* 28% of the time for experimental items but only 0.5% of the time for control items (Wilcoxon test, $z = 5.09$, $p < .001$, Cliff's $\delta = 0.54$). For example, one participant responded: "No, both sentences cannot be true because Elizabeth is *already* an author." Usage of *already* suggests that participants may have interpreted *want* to mean that the proposition implied by its complement is not already realized.

As a reviewer noted, one explanation for why participants rated experimental items as consistent may be because they interpreted the premises in a cooperative way, that is, they "explained away" the inconsistency (see Khemlani & Johnson-Laird, 2012 for evidence of such behavior). For example, in the problem about May's authorship in (23), participants may have reasoned about her status in terms of dual character concepts (cf. Knobe, Prasada, & Newman, 2013). Dual character concepts associate members with (1) a set of concrete features and (2) abstract values for the concept. For example, the dual character concept for

an author could be (1) writing books or articles and (2) being so devoted to writing that the person has no other professional pursuits. In other words, participants may have reasoned that May was technically an author since she had the concrete feature of having written books, but she longed to be an author in the "true" sense, as someone completely devoted to writing. Or, reasoning without reliance on dual character concepts, participants may have interpreted the scenario to mean that she was once was a writer, gave up the job for some other profession, and then hoped to return to the career. Such cooperative interpretations may obscure participants' interpretation of *want*. To eliminate this possibility, Experiment 7, therefore, provided only neutral information that could not be reinterpreted.

10. Experiment 7

Experiment 7 tested whether people interpret *want* to mean that the proposition implied by its complement is false. Such an interpretation should affect the way they reason about disjunctive alternatives. In particular, if it is the case that Max is an astronaut or that he is an astrologer, a statement such as "Max wants to be an astronaut" should make reasoners believe that Max is an astrologer instead. The type of inference bears similarity to a valid pattern of reasoning known as an "*or*-elimination," as in:

26) P or Q.
Not P.
Therefore, Q.

Hence, Experiment 7 served as a test of prediction 6 above. It presented participants with a sentence describing a fact as well as a second sentence describing a desire, as follows:

27) David is wearing a hat.
David wants to wear a green scarf. [control]

Participants pressed a button on the screen to select the most likely of two options held, for example,

[ ] David is wearing a yellow hat.
[ ] David is wearing a blue hat.

or else to select a button that indicated that both sentences are equally likely. The two options implicitly serve to articulate a set of disjunctive alternatives, that is, David is wearing either a yellow or a blue hat. If participants select either of the first two options above, it would reflect an *or*-elimination. In contrast, if they judge the two options as equally likely, it would reflect no *or*-elimination. Prediction 6 above predicts that for control problems, participants should avoid making the inference—no information about David's desire to wear a green scarf gives evidence to his choice of hat color. Experimental problems, in contrast, were of the following format:

28) David is wearing a hat.
David wants to wear a yellow hat. [experimental]
Which sentence is most likely?
[ ] David is wearing a yellow hat.
[ ] David is wearing a blue hat.
[ ] Both sentences are equally likely.

Such problems should promote disjunctive elimination so that participants should avoid inferring that David is wearing a yellow hat, since the *want*-premise should rule out the possibility.

10.1. Method

10.1.1. Participants

Forty-nine native English speakers (mean age = 36.3 years, 31 males, 17 females, 1 preferred not to say) volunteered through Mechanical Turk.

10.1.2. Open science

The predicted effects and analyses were preregistered on OSF, and the data, analyses, and experimental code are also available online.

10.1.3. Design, procedure, and materials

All participants were presented with the same eight problems, each consisting of two premises and three options to choose from as most likely. Half of the problems were controls in that the *want*-premises did not eliminate either of the two presented options. The other four problems were experimental because one of the two options was incompatible with the *want*-premise, leaving the other option as more likely. Each problem was randomly assigned one of eight male or female names and the problem order was randomized for each participant. The order the options were displayed in was randomized on each problem as well. Participants were required to choose one of the three responses before they could proceed to the next problem.

Participants' responses were coded to assess whether they made a disjunctive elimination or not. Hence, any trial on which a participant selected one of the two initial options was marked as producing a disjunctive elimination.

10.2. Results and discussion

Table 2 provides the proportions of participants' three responses. The results showed that they eliminated one of the two disjuncts more often for experimental problems than control problems (73% vs. 13%, Wilcoxon test, $z = 6.10$, $p < .001$, Cliff's $\delta = 0.86$). Experiment 6, therefore,

confirmed prediction 6. A follow-up GLMM regression treated the materials as random effects and the type of problem (control vs. experimental) as a fixed effect; the regression further validated the difference between experimental and control problems in participants'

Table 2
Participants' percentages of responses for which option is most likely for control and experimental problems in Experiment 6

	Control	Experimental
Option 1	6%	22%
Option 2	7%	**51%**
Neither	**87%**	27%

Note. Option 1 denotes the option provided to participants that was incompatible with the premises in the experimental condition. For the control condition, there was no conceptual difference between option 1 and option 2, that is, they reflect the order provided before randomization.

tendency to eliminate a disjunctive alternative ($\beta = 3.13$, $z = 10.79$, $p < .001$). The frequency data in Table 2 were subjected to a Fisher's exact test, which showed a reliable difference in response as a function of the type of problem and the three different response options (Fisher's exact test, $p < .001$).

The results suggest that people infer that the complement of *want* is not realized, that is, false, which causes them to select choices that are consistent with *want*'s complement when the other choice is inconsistent with it, in line with prediction 6. In cases where either choice is consistent with *want*'s complement, participants have no preference for one over the other.

In sum, Experiments 1–7 provide converging evidence for the model theory, which provides an account of what people mentally represent and how they process those representations when thinking about desire relations.

11. General discussion

What does it mean for an individual to want something? Previous linguistic accounts argue that a person's wants are restricted by what they believe to be true or possible (von Fintel, 1999, 2018; Geurts, 1998; Giorgi & Pianesi, 1997; Grano & Phillips-Brown, 2020; Harner, 2016; Heim, 1992; Homer, 2015; Portner, 1997; Portner & Rubinstein, 2012, 2020; Rubinstein, 2012; Schlenker, 2005; Staniszewski, 2019; Villalta, 2008). Because such theories are about the desirer's beliefs, they cannot explain why reasoners tend to treat statements of the form *A wants P* as inferring that *P* is presently false. We report studies that show, in contrast to semantic theories, that desire reports convey information beyond an attitude holder's desires or beliefs. A psychological account of bouletic reasoning posited that reasoners interpret *want* as a set of two possibilities: a desired future outcome, and a default possibility that represents a current, factual state of affairs. The theory yields a set of predictions validated by seven separate studies.

Experiment 1 showed that reasoners treat desire verbs similarly to the way they treat counterfactive verbs that imply the falsity of their complement. That is, reasoners treat *A refuses P* to imply that *P* is not the case, and they likewise—but to a lesser degree—treat *A wants P* to imply that *P* is not the case. Experiments 2 and 3 showed that reasoners use the desire

verb *want* to make inferences about what an individual knows—they take the statement *A wants P* to imply that *A knows P is false* instead of *A thinks P is false*. Experiment 4 showed that reasoners perceive statements as possible, that is, compatible with desires, when they are in fact impossible. This supports the model theory's prediction that reasoners neglect falsity and represent only what is true when reasoning about desires, which implicitly supports the broader claim of model theory, that reasoners represent desires as possibilities. Experiment 5 found evidence that desires can conflict with intentions, whereas intentions must be consistent. Participants rated sentences pairs like the following:

29) Katie [plans/wants] to finish reading the book now.
Katie plans to finish reading the book tomorrow.

as consistent more often when the first sentence's matrix verb is *want* than when it is *plan*. Experiment 6 found that reasoners are more likely to judge the following description to be inconsistent:

30) Katie plays the guitar.
Katie wants to play [a stringed instrument/soccer].

more often when it is completed by "a stringed instrument" versus "soccer." Experiment 6 gave participants premises of the following form:

31) Elizabeth wants to be reading fiction.

and found that they were more likely to infer that Elizabeth was reading nonfiction than reading fiction. Both of these inferences concern, not just the mental states of the desirers, but also facts about the activities they do. Experiment 7 showed how *want* can guide the way individuals make *or*-elimination inferences. All seven experiments corroborate the central predictions of the model theory.

Mental models are representations of coherent possibilities. For descriptions of desires, coherence implies that reasoners cannot build a model where *A wants P* and *A wants not-P* at the same time. Yet, many scholars observe that *want* permits conjunctions of contradicting desires (see Lassiter, 2011b, p. 133; Levinson, 2003, p. 227 et seq.; Portner & Rubinstein, 2012, p. 472), for example,

32) Opal wants to run the Boston marathon and she doesn't want to run the Boston marathon.

In contrast, the factive verb *know* permits no such conjunctions. This presents a challenge to the present theory of bouletic reasoning: mental models cannot represent conflicting possibilities in a single model. One way to overcome the challenge is to treat *want* as an expression of a desire relative to a certain set of interests, goals, or inclinations, for example, Opal wants to run the marathon to visit Boston, but also, she does not want to run the marathon because she wants to be lazy and not train. In cases where wants contradict, reasoners maintain separate models, not of the person's stated desires, but of their underlying goals, reasons, or motivations. Such an account can treat (32) as expressing two desires, for example, *Opal wants to visit Boston* and *Opal want to be lazy*, using a single model of the form:

	FACT	FUTURE POSSIBILITY
Opal	¬ Boston	Boston
	lazy	¬ lazy

An alternative approach treats Opal's desires about running the Boston marathon as incompatible by representing them with separate models:

	FACT	FUTURE POSSIBILITY
Opal	¬ marathon	marathon

	FACT	FUTURE POSSIBILITY
Opal	¬ marathon	¬ marathon

Such extensions to the present theory can help explain how people construe contradictory desires.

Can other approaches explain how people interpret *want*? One approach in formal semantics treats modals, including *want*, as a probabilistic comparison operator (cf. Lassiter, 2011a, 2011b). It, like other accounts of *want*, assumes that the verb is comparative in nature, that is, it serves to highlight a comparison between its complement and a set of alternatives. And a proposition, such as that denoted by the complement of *want* and of the alternatives it is compared to, is a set of possible worlds. The contribution of this proposal is that all worlds in these propositions are assigned not just a measure of desirability, but also an estimated probability of occurrence, thus yielding expected utilities, which are summed to generate expected utilities of the relevant propositions. Thus, *A wants P* is equivalent to saying:

> *A* attributes a higher expected utility to those situations where *P* is true than those where the alternatives to *P* are true.

But such an account has difficulty explaining, in general, how statements about desire affect people's interpretation about a current state of affairs (Experiment 1), why people make epistemic inferences from *want* (Experiments 2 and 3), why they fall prey to illusions when reasoning about desire (Experiment 4), why they decide that some *want* descriptions are inconsistent (Experiment 5), or why they yield *or*-elimination inferences (Experiments 6 and 7).

While most semantic accounts treat *want* as comparative, Harner (2016) argues that *want* has a reading that is not comparative (see also Davis, 1984, 1986, 2005). In this reading, to say that *Lee wants an espresso* does not imply that Lee compares situations in which she has an espresso to some contextually defined alternatives. It means instead that Lee's interest in having an espresso exceeds some threshold of desirability. No reference to alternatives is invoked on this meaning, and so the account undergirds the model theory of bouletic reasoning outlined above. Indeed, a threshold interpretation of *want* may align with the default representation of desire proposed above. Such an interpretation is simpler to compute and easier—for example, for children—to learn (Lagattuta, 2005). Comparative readings are more complex and subtle, and, therefore, harder to compute. One central constraint for a plausible cognitive theory of bouletic reasoning is to be algorithmically economical: the theory should not demand that reasoners engage in intractable mental operations in order to understand and

reason about seemingly simple and commonplace concepts, and it should rely on computations that minimize working memory in a way that makes learning *want* easy for young children, particularly since *want* is among the earliest mental state verbs for children to acquire (Bartsch & Wellman, 1995; Ferres, 2003; Moore et al., 1995). Both Harner's (2016) account and the one presented above serve as viable theoretical foundations.

The theory presented here is meant to account for how people reason about the mental state of desire in particular. But its central predictions can help build theories of other kinds of mental states, such as those expressed using verbs, such as *think, know, plan, discover*, and *forget*, namely that descriptions of mental states are true only in certain situations, and these situations allow reasoners to make inferences about the real world. For instance, if it is true that an individual *discovers* that Jiro is a pilot, then it implies an extended period during which the fact of Jiro's occupation was unknown to the individual. Factive verbs, such as *know* and *discover*, imply facts about the world, but the mechanisms and representations by which reasoners infer those facts remain unknown. The present theory offers a tractable account that relates mental states to reasoning about the world, and it can serve as the basis for future explorations of mental state reasoning.

There are several avenues of interest to continue exploring concerning *want*. One involves the claim that by default, reasoners infer knowledge over belief from desire statements. Experiments 2 and 3 were not sensitive enough to reveal any difference in responses times, primarily because the experimental paradigm was not suitable for granular analyses of these inferences. Future methodologies can address when individuals make such default inferences. Another promising route is to extend the results to other languages to study *want* cross-linguistically: all of our experiments were in English, though the theoretical predictions above apply to any language that includes some equivalent of *want* under the assumption that desire is a cognitive primitive. Cross-linguistic analyses may confirm or reject this assumption.

In sum, this paper proposed a comprehensive theory of how people mentally represent desires, as expressed by verbs like *want*, *wish*, and *hope*. It showed how reasoning about these desires can yield systematic inferences, not just about the states of desire of an individual who wants something, but about information in the world as well. We want, wish, and hope for additional studies to bear out its central predictions.

Open Research Badges

 This article has earned Open Data and Open Materials badges. Data and materials are available at https://osf.io/7ewym/.

References

Anand, P., & Hacquard, V. (2013). Epistemics and attitudes. *Semantics and Pragmatics*, *6*, 1–59.

Aristotle. (300BC/1926). *Nicomachean ethics*. Cambridge, MA: Harvard University Press.

Asher, N. (1987). A typology for attitude verbs and their anaphoric properties. *Linguistics and Philosophy*, *10*, 125–197.

Bartsch, K., & Wellman, H. (1995). *Children talk about the mind*. New York: Oxford University Press.

Blumberg, K., & Hawthorne, J. (2022). Wanting what's not best. *Philosophical Studies*, *179*, 1275–1296.

Bolinger, D. (1968). *Aspects of language*. New York: Harcourt, Brace & World.
Bolinger, D. (1974). Meaning and form. *Transactions of the New York Academy of Sciences*, *36*, 218–233.
Brand, M. (1984). *In tending and acting: Toward a naturalized action theory*. Cambridge, MA: MIT Press.
Bratman, M. E., Israel, D., & Pollack, M. E. (1988). Plans and resource-bounded practical reasoning. *Computational Intelligence*, *4*, 349–355.
Carey, S., Leahy, B., Redshaw, J., & Suddendorf, T. (2020). Could it be so? The cognitive science of possibility. *Trends in Cognitive Sciences*, *24*, 3–4.
Cherubini, P., & Johnson-Laird, P. N. (2004). Does everyone love everyone? The psychology of iterative reasoning. *Thinking & Reasoning*, *10*, 31–53.
Crnič, L. (2011). Getting even. Doctoral Dissertation. MIT Press.
Davis, W. (1984). The two senses of desire. *Philosophical Studies*, *45*, 181–195.
Davis, W. (1986). The two senses of desire. In J. Marks (Ed.), *The ways of desire: New essays in philosophical psychology on the concept of wanting* (pp. 63–82). Chicago, IL: Precedent Publications.
Davis, W. (2005). Reasons and psychological causes. *Philosophical Studies*, *122*, 51–101.
Farkas, D. (2003). Assertion, belief and mood choice. Paper presented at The Workshop on Conditional and Unconditional Modality 2002. Vienna: ESSLLI.
Ferres, L. (2003). Children's early theory of mind: Exploring the development of the concept of desire in monolingual Spanish children. *Developmental Science*, *6*, 159–165.
von Fintel, K. (1999). NPI licensing, Strawson entailment, and context dependency. *Journal of Semantics*, *16*, 97–148.
von Fintel, K. (2018). *On the monotonicity of desire ascriptions*. MIT Press.
von Fintel, K. (2021). *How weak is your want?* MIT Press.
Gajewski, J. (2005). *Neg-raising: Polarity and presupposition*. Doctoral Dissertation. MIT Press.
Gajewski, J. (2007). Neg-raising and polarity. *Linguistics and Philosophy*, *30*, 289–328.
Galitsky, B. (2013). Exhaustive simulation of consecutive mental states of human agents. *Knowledge-Based Systems*, *43*, 1–20.
Geurts, B. (1998). Presuppositions and anaphors in attitude contexts. *Linguistics and Philosophy*, *21*, 545–601.
Giorgi, A., & Pianesi, F. (1997). *Tense and aspect. From semantics to morphosyntax*. New York/Oxford: Oxford University Press.
Gisborne, N., & Holmes, J. (2007). A history of English evidential verbs of appearance. *English Language and Linguistics*, *11*, 1–20.
Goodwin, G. P., & Johnson-Laird, P. N. (2005). Reasoning about relations. *Psychological Review*, *112*, 468–493.
Goodwin, G., & Johnson-Laird, P. N. (2010). Conceptual illusions. *Cognition*, *114*, 253–265.
Grano, T., & Phillips-Brown, M. (2020). *Counterfactual want ascriptions and conditional belief*. MIT Press.
Grice, H. P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and semantics, vol. 3: Speech acts* (pp. 43–58). New York: Academic Press.
Harner, H. (2016). *Focus and the semantics of desire predicates and directive verbs*. Doctoral Dissertation. Georgetown University.
Heim, I. (1992). Presupposition projection and the semantics of attitude verbs. *Journal of Semantics*, *9*, 183–221.
Homer, V. (2015). Neg-raising and positive polarity: The view from modals. *Semantics and Pragmatics*, *8*, 1–88.
Hume, D. (1740/1978). *A treatise of human nature*. London: Clarendon.
Jerzak, E. (2019). Two ways to want? *Journal of Philosophy*, *116*, 65–98.
Johnson-Laird, P. N. (2006). *How we reason*. New York: Oxford University Press.
Johnson-Laird, P. N. (2012). Inference with mental models. In K. J. Holyoak and R. G. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 134–145).
Johnson-Laird, P. N., Girotto, V., & Legrenzi, P. (2004). Reasoning from inconsistency to consistency. *Psychological Review*,*111*, 640–661.
Johnson-Laird, P. N., Khemlani, S., & Goodwin, G. (2015). Logic, probability, and human reasoning. *Trends in Cognitive Sciences*, *19*, 201–214.

Johnson-Laird, P. N., Legrenzi, P., Girotto, P., & Legrenzi, M. S. (2000). Illusions in reasoning about consistency. *Science*, *288*, 531–532.

Johnson-Laird, P. N., & Ragni, M. (2019). Possibilities as the foundation of reasoning. *Cognition*, *193*, 103950.

Johnson-Laird, P. N., & Savary, F. (1996). Illusory inferences about probabilities. *Acta Psychologica*, *93*, 69–90.

Karttunen, L. (1973). *The last word. Mimeograph.* Austin, TX: University of Texas.

Karttunen, L. (1974). Presupposition and linguistic context. *Theoretical Linguistics*, *1*, 181–194.

Kelly, L., Khemlani, S., & Johnson-Laird, P. N. (2020). Reasoning about durations. *Journal of Cognitive Neuroscience*, *32*, 2103–2116.

Khemlani, S., Bello, P., Briggs, G., Harner, H., & Wasylyshyn, C. (2021). Much ado about nothing: The mental representation of omissive relations. *Frontiers in Psychology*, *11*, 609–658.

Khemlani, S., Byrne, R. M. J., & Johnson-Laird, P. N. (2018). Facts and possibilities: A model-based theory of sentential reasoning. *Cognitive Science*, *42*, 1887–1924.

Khemlani, S., & Johnson-Laird, P. N. (2009). Disjunctive illusory inferences and how to eliminate them. *Memory & Cognition*, *37*, 615–623.

Khemlani, S., & Johnson-Laird, P. N. (2011). The need to explain. *Quarterly Journal of Experimental Psychology*, *64*, 2276–2288.

Khemlani, S., & Johnson-Laird, P. N. (2012). Hidden conflicts: Explanations make inconsistencies harder to detect. *Acta Psychologica*, *139*, 486–491.

Khemlani, S., & Johnson-Laird, P. N. (2013). Cognitive changes from explanations. *Journal of Cognitive Psychology*, *24*, 139–146.

Khemlani, S., & Johnson-Laird, P. N. (2017). Mental models and causation. In M. Waldmann (Ed.), *Oxford handbook of causal reasoning* (pp. 1–42). Academic Press.

Khemlani, S., Mackiewicz, R., Bucciarelli, M., & Johnson-Laird, P. N. (2013). Kinematic mental simulations in abduction and deduction. *Proceedings of the National Academy of Sciences*, *110*, 16766–16771.

Khemlani, S., Orenes, I., & Johnson-Laird, P. N. (2012). Negation: A theory of its meaning, representation, and use. *Journal of Cognitive Psychology*, *24*, 541–559.

Kinny, D., & Georgeff, M. P. (1991). Commitment and effectiveness of situated agents. In J. P. Mylopoulos & R. Reiter (Eds.), Proceedings of the 12th International Joint Conference on Artificial Intelligence (IJCAI-91). Sydney: Morgan Kaufmann, 82–88.

Knobe, J., Prasada, S., & Newman, G. (2013). Dual character concepts and the normative dimension of conceptual representation. *Cognition*, *127*, 242–257.

Lagattuta, K. H. (2005). When you shouldn't do what you want to do: Young children's understanding of desires, rules, and emotions. *Child Development*, *76*, 713–733.

Lassiter, D. (2011a). Nouwen's puzzle and a scalar semantics for obligations, needs, and desires. In N. Ashton, A. Chereches, & D. Lutz (Eds.), *Semantics and linguistic theory* (pp. 694–711).

Lassiter, D. (2011b). Measurement and modality: The scalar basis of modal semantics. Doctoral Dissertation. New York University.

Levinson, D. (2003). Probabilistic model-theoretic semantics for want. In R. Young & Y. Zhou (Eds.), *Semantics and linguistic theory* (pp. 222–239).

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).

Moore, C., Jarrold, C., Russell, J., Lumb, A., Sapp, F., & MacCallum, F. (1995). Conflicting desire and the child's theory of mind. *Cognitive Development*, *10*, 467–482.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, *5*, 411–419.

Peirce, C. S. (1931–1958). *Collected papers of Charles Sanders Peirce. 8 vols*. Cambridge, MA: Harvard University Press.

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

Phillips, J., Buckwalter, W., Cushman, F., Friedman, O., Martin, A., Turri, J., … Knobe, J. (2021). Knowledge before belief. *Behavioral and Brain Sciences*. https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/abs/knowledge-before-belief/B434EF04A3EA77018384EABEB4973994

Phillips-Brown, M. (2021). What does decision theory have to do with wanting? *Mind*, *130*, 413–437.

Portner, P. (1997). The semantics of mood, complementation, and conversational force. *Natural Language Semantics*, *5*, 167–212.

Portner, P. (2004). The semantics of imperatives within a theory of clause types. In K. Watanabe & R. Young (Eds.), *Semantics and linguistic theory (SALT) XIV* (pp. 235–252).

Portner, P. (2009). *Modality*. Oxford: Oxford University Press.

Portner, P., & Rubinstein, A. (2012). Mood and contextual commitment. In A. Chereches (Ed.), *Semantics and linguistic theory* (pp. 461–487).

Portner, P., & Rubinstein, A. (2020). Desire, belief, and semantic composition: Variation in mood selection with desire predicates. *Natural Language Semantics*, *28*, 343–393.

Potts, C. (2015). Presupposition and implicature. In S. Lappin & C. Fox (Eds.), *The handbook of contemporary semantic theory* (2nd edition, pp. 168–202). Oxford: Wiley-Blackwell.

Quillien, T., & German, T. (2021). A simple definition of 'intentionally'. *Cognition*, *214*, 104806.

Rao, A. S., & Georgeff, M. P. (1995). BDI-agents: From theory to practice. In L. Gasser & V. Lesser (Eds.), Proceedings of the 1st International Conference on Multiagent Systems. San Francisco, CA: AAAI Press, 312–319.

van Rooij, R. (1999). Some analyses of pro-attitudes. In H. de Swart (Ed.), *Logic, game theory and social* choice (pp. 534–548). Tilburg: Tilburg University Press.

Rubinstein, A. (2012). Roots of modality. Doctoral dissertation. University of Massachusetts Amherst.

Russell, B. (1905). On denoting. *Mind*, *14*, 479–493.

Sablé-Meyer, M., & Mascarenhas, S. (2021). Indirect illusory inferences from disjunction: A new bridge between deductive inference and representativeness. *Review of Philosophy and Psychology.* https://link.springer.com/article/10.1007/s13164-021-00543-8#citeas

Schaeken, W., Johnson-Laird, P. N., & d'Ydewalle, G. (1996). Mental models and temporal reasoning. *Cognition*, *60*, 205–234.

Schlenker, P. (2005). The lazy Frenchman's approach to the subjunctive: Speculations on reference to worlds and semantic defaults in the analysis of mood. In T. Geerts, I. van Gynneken, & H. Jakobs (Eds.), *Romance languages and linguistic theory* (pp. 269–309). Amsterdam/Philadelphia: John Benjamins.

Searle, J. R. (1983). *In tentionality*. New York: Cambridge University Press.

Staniszewski, F. (2019). Wanting, acquiescing, and neg-raising. In M. Baird, D. Göksu, & J. Pesetsky (Eds.), Proceedings of the North East Linguistics Society (NELS).

Strawson, P. F. (1950). On referring. *Mind*, *59*, 320–344.

Thalberg, I. (1984). Do our intentions cause our intentional actions? *American Philosophical Quarterly*, *213*, 249–260.

Villalta, E. (2008). Mood and gradability: An investigation of the subjunctive mood in Spanish. *Linguistics and Philosophy*, *31*, 467–452.

Wrenn, C. (2010). A puzzle about desire. *Erkenntnis*, *73*, 185–209.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Supporting Information