# A theory of bouletic reasoning

**Hillary Harner[1,2] and Sangeet Khemlani[1]**
{hillary.harner.ctr, sangeet.khemlani}@nrl.navy.mil
[1] Navy Center for Applied Research in Artificial Intelligence
U.S. Naval Research Laboratory, Washington, DC 20375 USA
[2] NRC Postdoctoral Fellow

**Abstract**

No present theory explains or models the inferences people draw about the real world when reasoning about "bouletic" relations, i.e., predicates that express desires, such as *want* in *Lee wants to be in love*. Linguistic accounts of such bouletic relations define them in terms of their relation to a desirer's beliefs, and how its complement is deemed desirable (cf. Heim, 1992; Villalta, 2008; Rubinstein 2012). In contrast, we describe a new model-based theory (cf. Johnson-Laird, 2006; Khemlani, Byrne, & Johnson-Laird, 2018) that posits that such predicates are fundamentally counterfactual in nature. In particular, *X wants P* should imply that *P* is not the case, because you cannot want what is already true. The theory makes empirical predictions about how people assess the consistency of bouletic relations as well as how they use such relations to eliminate disjunctive possibilities. Two experiments tested and validated the theory's central predictions. We assess the theory in light of alternative accounts of human reasoning.

**Keywords:** bouletic reasoning, desire, mental model

## 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 act of desiring something does not guarantee any particular action or outcome, the act of expressing a desire can lead listeners to make inferences about the world. For instance, it seems reasonable to draw the inference (1b) from (1a):

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

In (1a), the premise expresses a "bouletic" relation – i.e., a relation that concerns an individual's desires – between Jiro and the complement of *want*, i.e., "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 class of predicates known as propositional attitude verbs, namely those verbs (e.g. *know, say, believe, advise*) that express an attitude holder's "attitude" about sentence-like objects.

Linguists have examined the meaning and inferences of desire verbs such as *want*. Karttunen (1973b, 1974) proposed that in order for *X wants P* to be true, *X* must believe the presuppositions of *P*. Thus, for a sentence like *Hannah wants it to stop raining*, the presupposition that *it is raining* does not need to be true in the general context; it need only be believed by Hannah. Many theorists accordingly argue that desires are grounded in people's beliefs: what we want is restricted by what we believe to be true or possible (Harner, 2016; Heim, 1992; Rubinstein, 2012; Villalta, 2008). Linguists thus propose that *X* wants *P* presupposes that *X* believes that *P* is both possible and false (cf. von Fintel, 1999; Harner, 2016; Heim, 1992; Portner, 1997; Rubinstein, 2012; Schlenker, 2005; Villalta, 2008). The claim of *P*'s possibility is problematic, however, because of examples such as the following (from Heim, 1992, p. 199):

2) I want this weekend to last forever. (But I know, of course, that it will be over in a few hours.)

The speaker in (2) knows that *P* is impossible, but the utterance is nevertheless acceptable; in general, a theory of bouletic reasoning should not restrict people from desiring impossible things.

The inference in (1b) is different from that commonly discussed by linguists: it is not an inference about the beliefs of the desirer, but instead about what is true of the world. This is the default inference: without specification to the contrary, people assume that a desirer's beliefs are aligned with reality. New information can cancel it. For instance, suppose you learn that Jiro's amnesia prevents him from remembering that he is already a pilot. In such a case, reasoners may conclude instead, in line with linguistic proposals, that he merely believes he is not a pilot.

The inference in (1b) does not neatly classify. It can be false without affecting the truth of (1a). For this reason, it cannot be a presupposition, rather, it appears to be a conversational implicature. Yet by definition, conversational implicatures are not tied to specific words; they arise independent of the precise wording, whereas (1b) seems to derive from the meaning of *want*. Listeners appear to assume that when a speaker uses a desire verb without any stipulation, the desirer's beliefs align with reality.

Linguistic theories make no mention of inferences such as (1b), so they have no account of it. Likewise, while some theories identify the semantic properties of *want* and other desire predicates, they do not commit themselves to what people mentally represent when they reason about desire. One exception proposes a probabilistic account of bouletic relations (Lassiter, 2011a, 2011b); we examine it in the General Discussion.

In this paper, we present a novel account of the meaning and mental representation of desire predicates. The theory adopts a modal semantics such that reasoners model the meaning of *want* by mentally simulating hypothetical alternatives (Khemlani, Byrne, & Johnson-Laird, 2018). The paper describes two ramifications of the theory: first,

reasoners should assess some conclusions as more consistent with *want* than others; and second, *want* should prompt reasoners to make systematic inferences about what is true of the world. Two experiments bear out the theory. The paper concludes by assessing the theory in light of recent proposals of desire predicates.

## The mental representation of desire

Recent theorists have renewed 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; Phillips, Morris, & Cushman, 2019). Modal concepts seem highly relevant to how people represent desire predicates such as *want*, because when a person wants something, or reasons about what another person wants, at a minimum, they are capable of bringing to mind the situations in which their desires come true – such situations are known as bouletic possibilities. But many psychological accounts of human reasoning ignore possibilities altogether (for a review, see Johnson-Laird, Khemlani, & Goodwin, 2015). One theory that is founded on the mental representation of possibilities is mental model theory – the "model" theory for short. The theory argues that all forms of reasoning depend on the mental simulation of sets of possibilities (Khemlani, Byrne, & Johnson-Laird, 2018). It rests on three fundamental principles:

- **Models represent one possibility by default.** When people reason about relations, they construct a possibility – a situation that describes finite alternatives – consistent with those relations (Johnson-Laird, 2006; Khemlani, Byrne, & Johnson-Laird, 2018). Typically, reasoners tend to construct, maintain, and reason on the basis of a single possibility – the mental model – but in principle, they are capable of deliberating and discovering alternative possibilities.
- **Models are iconic.** 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. 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).
- **Models are coherent.** Models cannot represent impossible situations. 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 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).

The model theory explains reasoning about causal relations (Briggs & Khemlani, 2019), temporal relations (Kelly, Khemlani, & Johnson-Laird, under review; Schaeken et al., 1996), and other sorts of abstract relations (Goodwin & Johnson-Laird, 2005; Cherubini & Johnson-Laird, 2004). No theory of reasoning explains reasoning about bouletic relations, and so we extended the model theory to account for inferences such as (1b) above.

A bouletic relation, e.g., *Jiro wants to be a pilot,* concerns an agent, *Jiro,* and a desired possibility, *to be a pilot*. People express them using desire verbs (e.g., *want* and *hope*) and they can be paired with infinitival complements, e.g., they can express desires about events or states to be realized by other people or by the attitude holder (see the respective examples in 3a-d).

3a) Lee wants Chris to buy a bike.
b) Lee wants Chris to be a lawyer.
c) Lee wants to fly a plane.
d) Lee wants to be in love.
e) Lee wants an espresso.

*Want* is special as a desire predicate since it can also take nouns as complements; no predicate is needed (cf. 3e). Yet we generally understand such sentences to express a desire about an event or state, e.g., we take (3e) to mean that Lee wants *to drink* an espresso. Accordingly, we construe verbs of desire as a relation between an agent and a desired possibility, which can be either an event or a state, and can be expressed linguistically as an object.

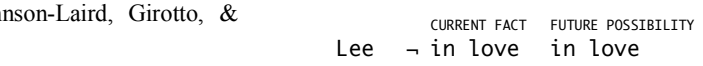
One constraint on bouletic relations such as *Lee wants to be in love* is that they imply that the complement is false, e.g., that Lee is not in love. In general, bouletic relations abide by the default constraint that an agent cannot have desires for what is already true. Hence, (4a) is unacceptable; (4b) is not:

4a) * Katy Perry wants to be an American this year.
b) Katy Perry wants to be a billionaire this year.

Katy Perry is already American, and so the desire expressed in (4a) is meaningless. (There is a reading of *want* that treats it as expressing pride; on such a reading, 4a 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 4b seems more plausible than 4a.) In sum, statements of the form *A wants B* make two assumptions in default of information to the contrary:

i) It is possible for A to have B.
ii) A does not have B.

The above constraint suffices to explain the models of the possibilities that bouletic relations refer to. By default, reasoners should interpret a relation of the form, *Lee wants to be in love*, as a set of two distinct possibilities about Lee, as depicted in this diagram:



The diagram shows tokens that stand in place of an agent and a state of affairs, as well as '¬', i.e., the symbol for negation. It depicts both a desired future state of affairs as well as a current state of affairs, that is, one in which Lee is not in love. The model represents the temporal relation between the possibilities on a spatial axis (see, e.g., Schaeken et al., 1996; Kelly et al., under review): it represents current information to the left of a future possibility since the former precedes the latter. Reasoners may build the model by first constructing the future possibility, i.e., the assertion of the *want*-clause, that represents Lee as being in love, and then adding the inferred fact, i.e., that Lee is not currently in love.

Sentences can conflict with other sentences; consider the following:

5a) Aria visited Addis Ababa last year.
b) Aria did not visit Ethiopia last year.

Provided that the first premise refers to the capital of Ethiopia, the two premises are inconsistent, i.e., they cannot be true at the same time. 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're inconsistent. Hence, reasoners should fail to build a model of the premises in (5), 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 (6a) to be consistent more often than (6b):

6a) Amy has a black belt in karate.
Amy wants to be good at telling jokes.
b) Amy has a black belt in karate.
Amy wants to be good at a martial art.

In (6a), the model of the first premise is:

| | CURRENT FACT |
|---|---|
| Amy | black-belt |

and the model of the second premise is:

| | CURRENT FACT | FUTURE POSSIBILITY |
|---|---|---|
| Amy | ¬ jokes | jokes |

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

| | CURRENT FACT | FUTURE POSSIBILITY |
|---|---|---|
| Amy | karate | jokes |
| | ¬ jokes | |

that depicts the current state of Amy's abilities as well as a future possibility. In contrast, an integrated model of the premises in (6b) should yield the following:

| | CURRENT FACT | FUTURE POSSIBILITY |
|---|---|---|
| Amy | karate | martial-art |
| | ¬ martial-art | |

and many reasoners should find such a model incoherent because of its current facts: to have a black-belt in karate is to be good at a martial art. Hence, the theory makes the following prediction:

**Prediction 1.** Reasoners should be more likely to treat the following pair of statements as inconsistent: *X is P* and *X wants to be P'* (where *P* implies *P'*). In contrast, they should judge the following pair of statements to be consistent: *X is P* and *X wants to be Q* (where *P* does not imply *Q*).

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

7) 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, i.e., it yields the following model:

| | CURRENT FACT | FUTURE POSSIBILITY |
|---|---|---|
| Matt | ¬ radiologist | radiologist |

Reasoners should infer 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 additional prediction:

**Prediction 2.** When reasoning about a disjunction of the form *X is P or X is Q*, desire predicates of the form *X wants to be P* should rule out one of the clauses in the disjunction. Hence, such statements should cause reasoners to infer that *X is Q* follows.

One caveat with the treatment above is that it provides an account of people's default interpretations of statements of the form *A wants B*. The context people understand *A wants B* in may call on them to deliberate and modify their initial model. People can do so in at least two ways. First, they can revise their default inference to concern, not current facts, but current beliefs. Hence, it may be possible for Matt to *want* to be a radiologist and to *be* a radiologist, but only in the odd scenario in which, unbeknownst to him, he was already a radiologist. Such a change would require the following alteration to the default model of the desire expressed in (7):

| | CURRENT BELIEF | FUTURE POSSIBILITY |
|---|---|---|
| Matt | ¬ radiologist | radiologist |

Deliberation may also call on reasoners to elaborate on the contents of the possibilities. Consider (3e) above, "Lee wants an espresso." The model theory posits that the default model should be the following:

| | CURRENT FACT | FUTURE POSSIBILITY |
|---|---|---|
| Lee | ¬ espresso | espresso |

Of course, it is plausible to desire an espresso after already enjoying a cup, and so pragmatic constraints may require reasoners to modify the desire to concern, not a desire for espresso in the abstract, but the desire for a fresh cup:

| | CURRENT FACT | FUTURE POSSIBILITY |
|---|---|---|
| Lee | ¬ fresh-espresso | fresh-espresso |

In general, the model theory can account for a variety of ways in which people interpret bouletic expressions. The remainder of the paper presents two experiments that test and corroborate the predictions of the model theory.

## Experiment 1

Experiment 1 tested whether people make the inference that when *X* wants *P*, *P* is not already true. 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 1. For control pairs, the *want*-sentence reported on a desire that had no relation to the first sentence:

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

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

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

In the example above, by definition a person who has written 3 best-selling books is an author. Participants then evaluated whether both sentences could be true at the same time – an intuitive way of evaluating the consistency of a statement (Johnson-Laird et al., 2004). If prediction 1 is true, reasoners should judge that both sentences are true at the same time more often for control pairs than experimental pairs.

### Method

*Participants*. 49 participants (mean age = 36.9 years; 27 males and 22 females) volunteered through the Amazon Mechanical Turk online platform (see Paolacci, Chandler, & Ipeirotis, 2010, for a review). All participants reported English as their native language.

*Design, procedure, and materials.* Participants were presented with 8 pairs of sentences, one pair at a time. 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 8 premises were used as the first sentence on each trial. Half of the second sentences were controls, i.e., 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 a control or an experimental one from a pool of 16 materials, 8 control and 8 experimental. Each sentence pair was randomly assigned one of 8 unique male or female names to serve as its subject. The order of presentation for the 8 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.

*Open science.* Data, materials, experimental code, and analysis scripts are available online (https://osf.io/njve4/).

### Results and discussion

Participants' responses were coded for whether they responded affirmatively or negatively, i.e., 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, $z = 3.55$, Cliff's $\delta = .43$), i.e., their behavior corroborated prediction 1. A follow-up generalized mixed-model (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 ($\beta = 1.22$, $z = 5.04$, $p < .0001$).

A post-hoc analysis of participants' natural responses examined the spontaneous use of the word "already" to explain their consistency judgments. It found that they used the word "already" 28% of the time for experimental items but only 0.5% of the time for control items (Wilcoxon test, $z = 5.09$, Cliff's $\delta = 0.54$). For example, one participant responded: "No, both sentences cannot be true because Elizabeth is *already* an author." Hence, the experiment not only confirmed prediction 1, i.e., that participants would reject experimental pairs at a higher rate than control pairs, but it substantiated the notion that their rejection was because they interpreted *want* to mean that its complement, i.e., the object of desire, is not already realized.

One curiosity of the present experiment is that participants, on balance, judged that experimental problems were consistent more often than not. A strong version of prediction 1 would have suggested that they should judge those problems as inconsistent, but we suspect that many participants interpreted the premises in a cooperative way. Indeed, reasoners may have initially judged the premises to be inconsistent, and then they may have "explained away" the inconsistency (see Khemlani & Johnson-Laird, 2012 for evidence of such behavior). Hence, in the example problem about May being an author, participants may have believed that she was once was a writer, gave up the job for some other profession, and then longed to return to the career. Such cooperative interpretations may obscure participants' interpretation of *want*. Experiment 2 therefore provided only neutral information that could not be reinterpreted. It sought to test prediction 2 above.

## Experiment 2

Experiment 2 tested whether people interpret *want* to mean that its complement is false. Such an interpretation should affect the way they reason about disjunctive choices. In particular, a statement of the form *X wants to be P* should make reasoners believe that *X is Q* instead of *X is P*. The type of inference is analogous to a valid pattern of reasoning known as a disjunction elimination, as in:

P or Q.
Not P.
Therefore, Q.

Hence, Experiment 2 served as a test of principle 2 above. It presented participants with a sentence describing a fact as well as a second sentence describing a *want*-statement, as follows:

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

Participants were asked to press a button on the screen to select the most likely of two options, e.g.,

Option 1. David is wearing a yellow hat.
Option 2. David is wearing a blue hat.
Both sentences are equally likely.

If participants select either of the first two options above, it would reflect a disjunctive syllogism. In contrast, if they judge that the two options are equally likely, it would reflect no disjunctive syllogism. Prediction 2 above predicts that for control problems, participants should avoid inferring a disjunction elimination. Experimental problems, in contrast, were of the following format:

David is wearing a hat.
David wants to wear a yellow hat. [experimental]

*Which sentence is most likely?*
Option 1. David is wearing a yellow hat.
Option 2. 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 premises should rule out the possibility.

### Method

*Participants*. 49 native English speakers (mean age = 36.3 years, 31 males, 17 females, 1 preferred not to say) volunteered through Mechanical Turk.

*Open science.* The predicted effects and analyses were preregistered via the Open Science Framework (https://osf.io/3ftr6/).

*Design, procedure, and materials.* All participants were presented with the same 8 problems, each consisting of two premises and 3 options to choose from as most likely. Half of the problems were controls in that the premises did not eliminate either of the first two options. The other four problems were experimental because one of the two options was incompatible with the *want*-sentence, leaving the other as more likely. Each problem was randomly assigned one of 8 male or female names and the problem order was randomized for each participant, and the order in which the options were displayed was randomized on each trial. Participants were required to choose one of the 3 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.

| | **Control:** *A wants to wear X* | **Experimental:** *A wants to wear B* |
|---|---|---|
| Option 1: *A is wearing B.* | 6% | 22% |
| Option 2: *A is wearing C.* | 7% | **51%** |
| Neither | **87%** | 27% |

**Table 1.** Participants' percentages of responses for which option is most likely for control and experimental problems in Experiment 2. 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, i.e., they reflect the order provided before randomization.

### Results and discussion

Table 1 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, $z = 6.10$, Cliff's $\delta = .86$). Experiment 2 therefore confirmed prediction 2. A follow-up generalized mixed-model (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' tendency to eliminate a disjunctive alternative ($\beta = 3.13$, $z = 10.79$, $p < .0001$). The frequency data in Table 1 were subjected to a Fisher's exact test, which showed a reliable difference in responses as a function of the type of problem and the three different response options (Fisher's exact test, $p < .0001$).

The data suggest that people infer that the complement of *want* is not realized, i.e. false, which causes them to select choices that are consistent with *want*'s complement when the other choice is inconsistent with *want*'s complement, in line with prediction 2. In cases where either choice is consistent with *want*'s complement, participants have no preference for one choice over the other.

## 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; Harner, 2016; Heim, 1992; Portner, 1997; Rubinstein, 2012; Schlenker, 2005; Villalta, 2008). Because such theories are about the desirer's beliefs, they cannot explain how reasoners can use a statement of the form *X wants P* to infer that *P* isn't a truth about the actual state of affairs – such an inference is not "doxastic" in nature. Contra semantic theories, the present account argues that desire reports convey information beyond an attitude holder's desires or beliefs. Thus we developed a psychological account of bouletic reasoning in which reasoners interpret desire predicates as a set of two possibilities: a default possibility that represents a current state of affairs and a desired future possibility. The theory yields predictions that two experiments validated. Experiment 1 found that reasoners are more likely to judge the following description to be inconsistent:

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

more often when it is completed by "a stringed instrument" vs. "soccer". Experiment 2 gave participants premises of the following form:

9) Elizabeth wants to be reading fiction.

and found that they were more likely to infer that Elizabeth was reading non-fiction than reading fiction. Both of these inferences concern, not just the mental states of the desirers, but also facts abouts the activities they do. And they corroborate the central predictions of the model theory.

In general, mental models present a coherent set of possibilities. In the present case, coherence implies that reasoners cannot build a model where *X* wants *P* and *X* wants not-*P* at the same time. Yet *want* is well-known to permit conjunctions of contradicting desires (see e.g. Levinson, 2003; Lassiter, 2011b; and Portner & Rubinstein, 2013), e.g.,

10) 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, e.g., *Opal wants to run the marathon to visit Boston, but also, she doesn't want to run the marathon because she wants to be lazy and not train*. In cases where *want*s 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 (10) as expressing two desires, e.g., *I want to visit Boston* and *I want to be lazy*, using a single model of the form:

| | CURRENT FACT | FUTURE POSSIBILITY |
|---|---|---|
| Opal | ¬ Boston | Boston |
| | ¬ lazy | lazy |

An alternative approach treats a person's desires as incompatible by representing them with separate models:

| | CURRENT FACT | FUTURE POSSIBILITY |
|---|---|---|
| Opal | ¬ marathon | marathon |

| | CURRENT 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 *want* as inherently probabilistic (cf. Lassiter 2011a, 2011b) – bouletic relations operate by enumerating a set of possible worlds, attributing to each world an estimated probability of its occurrence and a utility measure, producing expected utilities to compare the complement to alternatives. The goal of the account is to explain the graded difference between, e.g., *want* and *desperately want*, because *desperately want* seems to imply a higher utility than *want*. Probabilistic approaches treat *want* and other modals as inherently gradable and comparative such that *X wants P* is equivalent to saying:

*X* attributes a higher utility to those situations in which *P* is true than those in which *P* is false.

But such an account has difficulty explaining why participants decided that some *want* descriptions are inconsistent (Experiment 1) or why they yielded disjunction elimination inferences (Experiment 2). Indeed, probabilistic accounts of cognition (see, e.g., Baratgin et al., 2015; Elqayam & Over, 2013) have difficulty explaining people's inconsistency judgments more generally, because a set of statements can be inconsistent even though each individual statement has a probability > 0 (Johnson-Laird et al., 2004).

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 he has an espresso to those in which he doesn't. 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 it is compatible with 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, e.g., children – to learn (Lagattuta, 2005). Comparative readings are more complex and subtle, and therefore harder to compute. Once 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. Both Harner's (2016) account and

the one presented above serve as viable theoretical foundations.

In sum, this paper proposed a model-based theory of how people mentally represent desire predicates such as *want, wish,* and *hope*. It sought to show how such predicates 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.

## Acknowledgments

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, and Greg Trafton 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.

## References

Baratgin, J., Douven, I., Evans, J. S. B. T., Oaksford, M., Over, D., & Politzer, G. (2015). The new paradigm and mental models. *Trends in Cognitive Sciences*, 19, 547-548.

Briggs, G. & Khemlani, S. (2019). A cognitively plausible algorithm for casual inference. In T. Stewart (Ed.), *Proceedings of the 17th International Conference on Cognitive Modeling*.

Carey, S., Leahy, B., Redshaw, J., & Suddendorf, T. (2020). Could it be so? The cognitive science of possibility. *Trends in Cognitive Science, 24*, 3-4.

Cherubini, P., & Johnson-Laird, P. N. (2004). Does everyone love everyone? The psychology of iterative reasoning. *Thinking & reasoning, 10*, 31-53.

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). Precedent Pub, Chicago.

Davis, W. (2005). Reasons and psychological causes. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition, 122*, 51-101.

Elqayam, S., & Over, D. E. (2013). New paradigm psychology of reasoning: An introduction to the special issue edited by Elqayam, Bonnefon, and Over. *Thinking & Reasoning, 19,* 249–265.

von Fintel, K. (1999). NPI licensing, Strawson entailment, and context dependency. *Journal of Semantics, 16*, 97-148.

Goodwin, G.P., & Johnson-Laird, P.N. (2005). Reasoning about relations. *Psychological Review, 112,* 468-493.

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.

Johnson-Laird, P.N. (2006). *How we reason*. Oxford University Press, NY.

Johnson-Laird, P. N. (2012). Inference with mental models. *The Oxford Handbook of Thinking and Reasoning*, 134-145.

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., Girotto, V., & Legrenzi, P. (2004). Reasoning from inconsistency to consistency. *Psychological Review*, 111.

Karttunen, L. (1973b). The last word. Mimeograph, University of Texas, Austin.

Karttunen, L. (1974). Presupposition and linguistic context. *Theoretical Linguistics, 1,* 181-94.

Kelly, L., Khemlani, S., & Johnson-Laird, P.N. (under review). Reasoning about durations. Manuscript under review.

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. (2011). The need to explain. *Quarterly Journal of Experimental Psychology, 64,* 2276-88.

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

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.

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 Ashton, Neil, Chereches, Anca, and Lutz David (Eds.). *Semantics and Linguistic Theory (SALT) XXI*, 694-711.

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

Levinson, D. (2003). Probabilistic model-theoretic semantics for want. In R. Young and Y. Zhou (Eds.), *Semantics and Linguistic Theory (SALT) XIII*, 222-239.

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. In Hartshorne, C., Weiss, P., and Burks, A. (Eds.). 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.

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

Portner, P., & Rubinstein, A. (2013). Mood and contextual commitment. In A. Chereches (Ed.), *Semantics and Linguistic Theory (SALT) XXII*, 461-487.

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

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 Geerts, T., van Gynneken, I., and Jakobs, H. (Eds.), *Romance Languages and Linguistic Theory 2003* (pp. 269-309). Amsterdam/Philadelphia: John Benjamins.

Villalta, E. (2008). Mood and gradability: An investigation of the subjunctive mood in Spanish. *Linguistics and Philosophy, 31,* 467-52.