Desires can conflict with intentions; plans cannot

Hillary Harner[1,2] and Sangeet Khemlani[1]
hillary.harner.ctr@nrl.navy.mil, 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

While many formal frameworks distinguish between desires and intentions, and considerable empirical work shows that people interpret them differently, no studies examine how people reason about them. We extend Harner and Khemlani's (2020) model-based theory of relations describing desire. The theory holds that people represent desires, as in, e.g., *Pav wants to visit Angkor Wat*, by pairing a factual representation of the negation of the desire (e.g., that Pav is not currently visiting Angkor Wat) with a future possibility where the desire is realized. We propose that intentions, which people express using verbs like *plan*, are represented as future actions that agents seek to perform. A particular individual's plans must be consistent with one another, whereas desires can conflict with these plans. We show how the model theory distinguishes desires and intentions, namely that models can be coherent even when a desire and a plan are inconsistent with each other. The distinctions make predictions about how reasoners should assess the consistency of statements concerning desires and intentions, and we report on two experiments that validate them.

Keywords: bouletics; desires; plans; intentions; model theory

Introduction

People can desire an outcome without planning to achieve it, e.g.:

1) Lisa wants to quit her job to work on a ranch.

Lisa may have no serious intention to quit her job, and given the opportunity to work on a ranch, she might even refuse it. So long as the very thought of quitting her job and working on a ranch appeals to her, it is true that she has the desire. In contrast, goals imply an intention to act towards them (Bratman, 1987), so when people set goals for themselves, they intend to act to achieve them. Intentions about goals can be expressed with the verb *plan*:

2) Lisa plans to quit her job to work on a ranch.

If Lisa were to exhibit no signs of following through on her goal of quitting her job and working on a ranch, (2) would be considered false. Unlike desires, the mere contemplation of a goal is not sufficient to warrant a statement about having the goal; intent must accompany it.

Intentions further differ from desires in the objects they are directed to. A person may desire any number of events or outcomes they have no way of making happen, as in (3a):

3a) Lisa wants France to vote in a new president.
b) # Lisa plans for France to vote in a new president.
c) Lisa plans to vote for a new president.

The same is not true for the intention described in (3b), where '#' denotes an incoherent statement. Lisa can only intend those things she is able to accomplish, as in (3c).

Malle and Knobe (2001) performed a corpus analysis that support these distinctions: 98% of active uses of the goal-oriented verbs *plan*, *intend, decide* revealed that people's plans, intentions, and decisions were directed toward behavior they themselves would carry out. In contrast, 63% of the active uses of desire verbs *want, wish,* and *hope* were directed towards other people's behavior. Perugini and Bagozzi (2004) likewise show that people construe intentions as actions that can be performed more often than desires. And around the age of 5, children seem to be able to distinguish the differential satisfaction of desires and intentions (Schult, 2002).

Despite these differences, desires and intentions share many similarities: neither desires nor intentions, as Searle (1983) observed, are 'right' or 'wrong' in the same way that beliefs can be wrong – desires and intentions cannot be falsified by facts in the way that beliefs can. Desires and intentions are better characterized as being satisfied or not. Both desires and intentions can be predicates for action: the desire for, say, more money, may cause an individual to buy a gun because the would-be criminal intends to rob a bank. Hence, desire and intentionality are central to systems of moral and legal thinking (Kenny, 1973a, b; Malle & Nelson, 2003; Marshall, 1968; Williams, 1993). Reasoners often make the reverse inference to impute intentionality and desire on others, i.e., they might reason that an individual (with sufficient means) who hasn't carried out an action refrains from doing so because of a lack of desire or intention. For instance, a teetotaler may possess no desire for alcohol, while a pregnant woman may possess such a desire without the intention to pour the drink (see Baird & Astington, 2005, for an integrative review of intentionality).

Despite extensive work into the social and developmental aspects of intentionality – particularly work about how individuals attribute folk notions of intentionality onto others – no theory exists to explain how people mentally represent intentions or desires, and no existing psychological proposal of human reasoning explains the underlying processes by which individuals make inferences about desires and intentions. A recent account by Harner and Khemlani (2020) sought to explain how people mentally represent desires.

In what follows, we describe the account and extend it to distinguish intentions from desires. We describe two predictions that the new theory makes, and we report the results of experiments that corroborate the theory. Finally, we conclude by contrasting the account with common

frameworks that separate intentions and desires in the adjacent field of artificial intelligence.

The mental representation of desires and plans

A growing consensus among cognitive scientists is that many higher-level thought processes, such as moral reasoning, counterfactual thinking, and reasoning about physics, depend on the mental simulations of possibilities (Battaglia, Hamrick, & Tenebaum, 2013; Byrne, 2005; Carey, Leahy, Redshaw, & Suddendorf, 2020; Phillips, Morris, & Cushman, 2019; Ragni & Johnson-Laird, 2019). Possibility-based theories seem similarly relevant for a 'bouletic' verb – i.e., a verb that concerns desires – like *want* and an intention-based verb like *plan*, because when people want or plan something, or when they reason about what other people want or plan, they are capable of envisioning the possibilities in which their wants or plans come true.

Despite the utility of possibility-based accounts and a range of applied uses, many psychological accounts of human reasoning base it on, not the construction of possibilities, but the computation of proofs (e.g., Stenning & van Lambalgen, 2005, 2008; Rips, 1994) or probabilities (e.g., Elqayam et al., 2013; Evans, 2012; Oaksford & Chater, 2007; Over, 2009; Pfeifer, 2013; for a review, see Johnson-Laird, Khemlani, & Goodwin, 2015). A disadvantage of many proof- and probability-based accounts is that they provide no cognitively plausible way of characterizing how two propositions can conflict with one another (see, e.g., Johnson-Laird, 2012). For example, suppose that there is a 50% chance that a particular door is open (and a 50% chance that it's closed):

P(*the door is open*) = 50%
P(*the door is closed*) = 50%

There's nothing inconsistent about this scenario, but the conjunctive probability of the two situations is 0% because the two events can't be true at the same time:

P(*the door is open AND the door is closed*) = 0%.

In general, there is no sensible way to use probabilities alone to assess the consistency of a set of statements, each of which has a non-zero probability. As we will show below, desires and intentions can conflict, and people's assessments of their consistencies are systematic, so any viable theory about how people reason about desires and intentions must be able to account for how people assess those conflicts. Theories based on the construction of possibilities can do so.

Among different accounts of human reasoning, only one theory is based on the mental representation of possibilities: mental model theory. The model theory argues that all forms of reasoning depend on the mental simulation of sets of possibilities, i.e., mental models (Khemlani, Byrne, & Johnson-Laird, 2018). It rests on three fundamental principles:

- **People represent one model by default.** People typically reason by building a model of a single possibility consistent with the information they have at hand (Johnson-Laird, 2006; Khemlani, Byrne, & Johnson-Laird, 2018), even though that information may be compatible with a wide range of possibilities. In principle, people can deliberate to consider alternative models compatible with their information, but doing so demands additional mental resources and can lead to systematic errors.
- **Models are iconic.** Models are mental structures that mirror the real-world scenarios they represent, i.e., they are *iconic* (Peirce, 1931-1958, Vol. 4). For instance, reasoners represent a group of individuals by constructing a set of tokens that stand in place of those individuals (Khemlani & Johnson-Laird, under review). Other sorts of representations, such as Venn diagrams and Euler circles (see Khemlani & Johnson-Laird, 2012a) are not iconic because they use a single entity (a circle) to stand in place of several individuals. Some abstract concepts, such as the concept of negation, cannot be iconic because they have no inherent structure; models represent these abstract concepts as symbols (Khemlani, Orenes, & Johnson-Laird, 2012). They can also represent possibilities that unfold in time (see Khemlani, Mackiewicz, Bucciarelli, & Johnson-Laird, 2013) by simulating the step-by-step changes that occur.
- **Models are coherent.** Models cannot, e.g., represent a scenario in which a door is both open and closed at the same time – they represent only those possibilities that are coherent. Reasoners can update models with new information that they remember, learn, perceive, or imagine – when they do so, the resulting updated model is likewise coherent. When new information conflicts with a model, people judge the information to be inconsistent with what came before it (Johnson-Laird, 2012; Johnson-Laird, Girotto, & Legrenzi, 2004) and make inferences to eliminate the conflict (Khemlani & Johnson-Laird, 2012b).

The model theory explains reasoning about causal relations (Khemlani, Bello, Briggs, Harner, & Wasylyshyn, 2021), temporal relations (Kelly, Khemlani, & Johnson-Laird, 2020; Schaeken et al., 1996), and other sorts of abstract relations (Goodwin & Johnson-Laird, 2005; Cherubini & Johnson-Laird, 2004). Harner and Khemlani (2020) recently extended it to bouletic reasoning; they proposed that reasoners represent the statement *A wants P* by assuming that *P* is false and that it is a possibility that can occur in the future. For example, a model of a sentence such as "Marcus wants to run the marathon" is:

	CURRENT FACT	FUTURE POSSIBILITY
Marcus	¬ marathon	marathon

The above diagram represents a token for an individual, a token that represents a current fact about that individual, and a token that represents a future possibility for that individual. The diagram uses words, e.g., "marathon", as a shorthand to represent a mental simulation, e.g., the simulation in which Marcus runs the marathon, though the simulation can contain explicit symbols for negation (i.e., '¬') to represent the

scenario in which Marcus does not run the marathon. Hence, the diagram represents that Marcus is not currently running the marathon and that it is a possibility in the future.

We extend the theory to reasoning about intentions by proposing that they are future actions that a person intends to perform (cf. Altmann & Trafton, 2002). So reasoners should represent the statement, "Marcus plans to run the marathon", as:

	CURRENT FACT	FUTURE ACTION
Marcus	¬ marathon	marathon

Models of future actions are limited to those actions an individual can perform personally. And, like any other kind of model, models of future actions must be coherent: one cannot simultaneously maintain a future action of running and not running a marathon.

Because future action models are distinct from models of future possibilities, the two can conflict, e.g., the following statement seems sensible:

4) Marcus wants to run a marathon, but he doesn't plan on doing so.

Reasoners can build the following model to capture the meaning of (4):

	CURRENT FACT	FUTURE POSSIBILITY	FUTURE ACTION
Marcus	¬ marathon	marathon	¬ marathon

The integrated model shows that the models of desires and goals are independent from one another and thus can conflict with each other, even though the models of desires and goals must be internally consistent. As a consequence, the model theory makes the following prediction:

Prediction 1. Reasoners should consider sentence pairs of the form *A plans P* and *A plans not-P* as incompatible with one another. In contrast, reasoners should consider sentences of the form *A plans P* and *A plans Q*, where *P* and *Q* do not conflict, as consistent.

The prediction concerns the verbs *plan* and *want* specifically since we take them as expressions of intentions and desires, respectively (cf. Malle & Knobe, 2001). A corollary is that because intentions can conflict with desires without resulting in incoherence, reasoners should judge the following two sentences to be consistent with one another:

5) Marcus wants to not run the marathon.
Marcus plans to run the marathon anyway.

which yield the following model:

	CURRENT FACT	FUTURE POSSIBILITY	FUTURE ACTION
Marcus	¬ marathon	¬ marathon	marathon

Thus the model theory makes this prediction concerning the relation between *plan* and *want*:

Prediction 2. Reasoners should consider sentences of the following form: *A wants P* and *A plans not-P*, as compatible with one another, because desires and intentions can conflict without being incoherent.

We designed two experiments to test each of these predictions.

Experiment 1

Experiment 1 tested prediction 1: reasoners will accept pairs of consistent plans, but will reject pairs of inconsistent plans. It provided participants with two sentences that described either plans that conflicted or those that didn't. For example, the pair of sentences below describes a consistent pair of plans:

6) Keegan plans to water the indoor plants.
Keegan plans to water the outdoor plants.

Keegan can water both indoor and outdoor plants without any conflict. The other half of the sentence pairs concerned inconsistent plans, e.g.:

7) Colleen plans to spend the next hour alone.
Colleen plans to spend the next with friends.

The experiment also provided participants with sentences that paired desires and intentions, e.g., problems in which one sentence used the verb *want* and the other used the verb *plan*, e.g.:

8) Jacob wants to watch a movie tonight.
Jacob plans to watch a movie tomorrow.

The theory predicts that people will consider any *want/plan* pair, like (8), to be consistent, regardless of whether their complements conflict or not. In contrast, it predicts that reasoners should rate *plan-plan* pairs as consistent only when their complements are consistent, as with (6).

Method

Participants. 50 participants (mean age = 40.5 years; 22 females and 28 males) performed the study using the Amazon Mechanical Turk online platform (see Paolacci, Chandler, & Ipeirotis, 2010, for a review). All participants reported being native English speakers.

Design, procedure, and materials. Participants read 12 sentence pairs, one at a time. 8 of the pairs were of the form *A plans [complement], A plans [complement]*. For 4 of these 8 pairs, the complements of the two sentences were consistent; for the other 4, they were inconsistent. The remaining four sentence pairs were of the form *A wants P, A plans Q*; the complements *P* and *Q* were consistent. Thus there were 8 consistent complement pairs, and the experiment randomly assigned them to a *want*/*plan* sentence pair or a *plan/plan* sentence pair. It also randomly assigned the name of an agent to each sentence pair, half of which were female and the other half male. The experiment randomized the order of presentation for all sentence pairs. In sum, the study yielded a nested within-participants design.

After reading each sentence pair, participants typed out their response to the question, "Can both of these sentences be true at the same time?", which is a task used to elicit consistency judgments from participants without extensive training in logic (see, e.g., Johnson-Laird et al., 2004). Participants were required to type 'yes' or 'no' and they could elaborate on their response if they wanted. The first author

coded participants' responses on whether they responded affirmatively or negatively, i.e., whether they thought the two sentences were consistent or not, and did so blind to the condition of the study.

Open science. Data, materials, experimental code, and analysis scripts are available online (https://osf.io/kx74s/).

Results and discussion

Participants based their responses in Experiment 1 on the consistency of the complements. They judged *plan/plan* sentence pairs with consistent complements as consistent more often than *plan/plan* pairs with inconsistent complements (80% vs. 22%, Wilcoxon test, $z = 5.73$, $p < .001$, Cliff's $\delta = 0.79$), which corroborates prediction 1. Their judgments of consistency were not reliably different between *want/plan* pairs and the consistent *plan/plan* pairs: 82% of *want/plan* pairs were judged consistent and 80% of consistent *plan/plan* pairs were judged as such (Wilcoxon test, $z = 0.41$, $p = 0.68$, Cliff's $\delta = 0.02$). Accordingly, the differences in ratings between *want/plan* pairs (whose complements were all consistent) and between the *plan/plan* pairs with inconsistent complements were similar to the differences between both sets of *plan/plan* pairs (82% v. 22%, Wilcoxon test, $z = 5.65$, $p < .001$, Cliff's $\delta = 0.82$).

Participants tended to reject inconsistent plans and to accept consistent plans, in line with the theory's first prediction. A post hoc analysis of the comments participants generated supported this interpretation: they explicitly noted the conflicts between two inconsistent plans, e.g., in response to sentences like (7), one participant explained "No, Helen cannot be alone and with friends at the same time". Less often, participants rated inconsistent plans as consistent; when they did so, their explanations tended to explain away the conflict, e.g., in response to a person's plans to go to bed at 9pm and to go to bed at 11pm, one participant justified it with, "Yes. One can sleep from 9:00 to 10:30, possibly completing an entire REM cycle, and subsequently returning to bed at 11:00." Hence, participants on occasion provided cooperative interpretations that mitigated the inconsistency. Another kind of explanation showed a lack of association between *plan* and intention, e.g., a participant who received the sentence pair in (7) said:

> *Colleen could be making plans to go either way and be waiting for a phone call from her friends to see if they can get together.*

Responses such as these suggest that the participant treated the verb *plan* as equivalent to *prepare for,* which could indicate that *plan* has a reading that does not entail intention.

As for the *want/plan* sentence pairs, participants generally found them to be consistent. In cases where they didn't, their explanations often elaborated that a person could not perform an action and its negation at the same time, perhaps indicating some overlap between the meaning of *want* and the meaning of *plan*. This possibility reveals a limitation in the design of Experiment 1: it did not directly compare situations in which *want/plan* pairs were inconsistent. If people accept such pairs, as prediction 2 holds, then it would indicate that people tend to distinguish desires from intentions. Experiment 2 remedied this limitation.

Experiment 2

Experiment 2 tested prediction 2: people should accept *want/plan* sentence pairs of the form *A wants P* and *A plans not P* as consistent, e.g.,

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

In contrast, participants should judge *plan/plan* pairs such as *A plans P* and *A plans not-P*, as inconsistent, e.g.:

10) Jimmy plans to get a full refund on the movie ticket.
Jimmy plans to exchange the movie ticket for a different showing. [control]

Participants in Experiment 1 had done so, and we expected participants in Experiment 2 to provide similar judgments, so sentence pairs such as (10) served as controls.

Method

Participants. 48 participants (mean age = 36.1 years; 21 females and 27 males) volunteered through Amazon Mechanical Turk. All but one participant reported English as their native language; we dropped their data from our analysis.

Design, procedure, and materials. Participants responded to 12 problems – 6 experimental and 6 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 (9). 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 (10). 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. The experiment also randomly assigned each sentence pair a unique male or female name to serve as its subject. The order of presentation for the 12 problems was shuffled for each participant.

As in Experiment 1, participants read the sentence pairs and then typed out their response to the question, "Can both sentences be true at the same time?" They had to respond with 'yes' or 'no' and could elaborate further if they chose. Data were coded in a manner similar to Experiment 1.

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

Results and discussion

Participants judged experimental *want-plan* sentence pairs to be compatible 65% of the time and control pairs to be

compatible 22% of the time (Wilcoxon test, $z = 4.68$, $p < .001$, Cliff's $\delta = 0.65$). Experiment 2 accordingly confirmed prediction 2: people consider a scenario consistent even when it describes desires and intentions that conflict. This supports the model theory's claim that the desires expressed by *want* have no necessary connection to intentionality, as desire is distinct from intention. In contrast, since the verb *plan* tends to suggest intentionality (as Experiment 1 shows), people tended to reject incompatible plans – a result that replicates the previous study.

Not every participant response conformed to these overarching trends; we reviewed written responses in a posthoc analysis for some insight. When participants rejected *want/plan* pairs as incompatible, several gave explanations indicating incompatibility of the complements, e.g., for the following problem:

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

one participant responded: "no, because if Henry gets a refund then he can't also get an exchange". Thus it is possible that their rejection was based on the complements themselves and not inferences about the sentences in their entirety as *want/plan* pairs. An alternative hypothesis is that the participant had read the problem too quickly, and had failed to see that the verbs in the two statements were different. Analysis of participants' acceptance of *plan/plan* pairs likewise reveals an occasional tendency to accommodate incompatibilities, e.g., one participant responded to a pair like (11) by saying: "Yes. First plan Teagan get[s a] refund movie ticket but next Teagan exchange[s] a ticket so it is possible."

Despite these responses, on the whole, the data from Experiment 2 supports the claim that people do not associate *want* with intention, but they do so for *plan*.

General discussion

Do people associate desire with intention when they reason? Philosophical accounts (Bratman, 1987, Bratman, 1988, Searle, 1983) and computational models (e.g., Rao and Georgeff, 1995) often distinguish desires from intentions theoretically and in application. And studies reveal that children distinguish between the two by the age of 5 (see, e.g., Schult, 2002). But few studies have examined how people reason about a person's desires and intentions to resolve conflicts between them, and no contemporary theory of reasoning takes their differences into account.

We ran two studies to show that the goals expressed by *plan* are associated with intention, as opposed to the desires expressed by *want*, which are not associated with intention. Experiment 1 showed that *plan* is associated with intention, as participants rejected pairs of *plan/plan* sentences whose complements were inconsistent, e.g.:

12) Ella plans to settle permanently in Ohio.
Ella plans to settle permanently in Utah.

This rejection was not based on their reticence to accept pairs of *plan/plan* sentences. They accepted such pairs as long as the complements were consistent; likewise, they accepted *want/plan* sentence pairs which also had consistent complements:

13) Marcus [plans / wants] to read the comics.
Marcus plans to read the news.

Experiment 2 showed that people accept *want/plan* pairs even when their complements are inconsistent.

14) Henry wants to skip the play tonight.
Henry plans to attend the play tonight.

These findings support the model theory, as they reveal that people distinguish desires from intentions. A basic premise of the theory is that models must be coherent, which requires the situations they represent, e.g., the model of future possibilities, to be consistent. But the theory posits that people keep separate simulations of future possibilities and simulations of future actions, so those models can conflict without yielding an inconsistency. Hence, reasoners should treat *A wants P* and *A plans not-P* as consistent, because they can build an integrated model that represents both assertions:

	FUTURE POSSIBILITY	FUTURE ACTION
Marcus	read-comics	read-comics
		read-news

In contrast, the theory predicts reasoners' rejection of inconsistent plans since plans are modeled together, i.e., there is no coherent model in which Henry both plans to attend the play and intends to skip it. Other theoretical frameworks in cognitive science and AI call for different representations for desires and intentions, and it is not clear how they yield the predictions of the model theory. For example, a common software architecture in AI separates between the beliefs, desires, and intentions (BDI) of a particular simulated agent. Some of the systems that implement this architecture explicitly stipulate that desires and plans must overlap (Rao & Georgoff, 1995). Others treat desires as necessary precursors to intentions (e.g., Woodridge, 1999). These systems serve to model how humans incorporate desires and intentions into reasoning and decision-making processes, but they do not (and are not intended to) predict how people reason about desires. Indeed, the systems cannot explain how desires and plans differentially conflict with one another. In contrast, the model theory predicts that desires can be inconsistent with plans; and the experiments and data we report on validate these 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 Bill Adams, Gordon Briggs, Tony Harrison, Andrew Lovett, Laura Kelly, and Greg Trafton for their advice and comments. We also thank Kalyan Gupta, Danielle Paterno, and the Knexus Research Corporation for their help in data collection.

References

Baird, J. A., & Astington, J. W. (2005). The development of the intention concept: From the observable world to the unobservable mind. In R. Hassin, J. Uleman, & J. Bargh (Eds.), *The new unconscious* (pp. 256-276). Oxford: Oxford University Press.

Battaglia, P., Hamrick, J., & Tenenbaum, J. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences,* 110, 18327-18332.

Bratman, M. E. (1987). *Intentions, plans, and practical reason.* Cambridge, MA: Harvard University Press.

Bratman, M. E., Israel, D., & Pollack, M. E. (1988). Plans and resource-bounded practical reasoning. *Computational Intelligence, 4,* 349-355.

Byrne, R. M. J. (2005). *The rational imagination: How people create alternatives to reality*. Cambridge, MA: The MIT Press.

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.

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.

Evans, J.St.B.T. (2012). Questions and challenges for the new psychology of reasoning. *Thinking & Reasoning, 18*, 5-31.

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

Harner, H., & Khemlani, S. (2020). A theory of bouletic reasoning. In B. Armstrong, S. Denison, M. Mack, and Y. Xu (Eds.), *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Johnson-Laird, P.N. (2006). *How we reason.* NY: OUP.

Johnson-Laird, P. N. (2012). Inference with mental models. *The Oxford handbook of thinking and reasoning*, 134-145.

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

Johnson-Laird, P. N., Khemlani, S., & Goodwin, G. (2015). Logic, probability, and human reasoning. *Trends in Cognitive Sciences, 19,* 201-214.

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

Kenny, A. (1973a). Intention and *mens rea* in murder. In P. M. S. Hacker, & J. Raz (Eds.), *Law, morality, and society: Essays in honour of H. L. A. Hart* (pp. 161–174). Oxford: Clarendon.

Kenny, A. (1973b). The history of intention in ethics. In A. Kenny (Ed.), *The anatomy of the soul: Historical essays in the philosophy of mind* (pp. 129–146). Oxford: Blackwell.

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

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. (2012a). Hidden conflicts: Explanations make inconsistencies harder to detect. *Acta Psychologica, 139,* 486-491.

Khemlani, S., & Johnson-Laird, P. N. (2012b). Theories of the syllogism: A meta-analysis. *Psychological Bulletin, 138*.

Khemlani, S., & Johnson-Laird, P. N. (2021). Reasoning about properties: A computational theory. Manuscript in press at *Psychological Review*.

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 Twelfth International Joint Conference on Artificial Intelligence (IJCAI-91),* Sydney, Australia: Morgan Kaufmann.

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

Malle, B. F., & Nelson, S.E. (2003). Judging *mens rea*: The tension between folk concepts and legal concepts of intentionality. *Behavioral Sciences and the Law,* 21, 563-580.

Marshall, J. (1968). *Intention in law and society.* New York: Funk and Wagnall.

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning.* OUP.

Over, D. E. (2009). New paradigm psychology of reasoning. *Thinking & Reasoning, 15(4),* 431.

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

Perugini, M., & Bagozzi, R. (2004). The distinction between desires and intentions. *European Journal of Social Psych,* 34, 69-84.

Peirce, C. S. (1931-1958). *Collected papers of Charles Sanders Peirce.* 8 vols. In C. Hartshorne, P. Weiss, and A. Burks (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.

Pfeifer, N. (2013). The new psychology of reasoning: A mental probability logical perspective, *Thinking & Reasoning, 19*, 329-345.

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

Ragni, M., & Johnson-Laird, P.N. (2019). Possibilities as the foundation of reasoning. *Cognition,* 193, 1039-50.

Rips, L. J. (1994). *The psychology of proof: Deductive reasoning in human thinking.* Cambridge, MA: The MIT Press.

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

Schult, C. (2002). Children's understanding of the distinction between intentions and desires. *Child Development,* 73, 1727-1747.

Searle, J. R. (1983). *Intentionality: An essay in the philosophy of mind.* New York: Cambridge University Press.

Stenning, K., & van Lambalgen, M. (2005). Semantic interpretation as computation in nonmonotonic logic: The real meaning of the suppression task. *Cognitive Science, 29(6),* 919–960.

Stenning, K. & van Lambalgen, M. (2008). *Human reasoning and cognitive science.* Cambridge, MA: The MIT Press.

Williams, G. (1965). *The mental element in crime.* Jerusalem: Magnes.

Woodridge, M. (1999). Intelligent Agents. In G. Weiss (Ed.), *Multi-agent systems*: *a modern approach to distributed artificial intelligence* (pp. 27-77). Cambridge, MA: MIT Press.