Reasoning about Durations

Laura Jane Kelly[1], Sangeet Khemlani[1], and P. N. Johnson-Laird[2,3]

Abstract

■ A set of assertions is consistent provided they can all be true at the same time. Naive individuals could prove consistency using the formal rules of a logical calculus, but it calls for them to fail to prove the negation of one assertion from the remainder in the set. An alternative procedure is for them to use an intuitive system (System 1) to construct a mental model of all the assertions. The task should be easy in this case. However, some sets of consistent assertions have no intuitive models and call for a deliberative system (System 2) to construct an alternative model. Formal rules and mental models therefore make different predictions. We report three experiments that tested their respective merits. The participants assessed the consistency of temporal descriptions based on statements using "during" and "before." They were more accurate for consistent problems with intuitive models than for those that called for deliberative models. There was no robust difference in accuracy between consistent and inconsistent problems. The results therefore corroborated the model theory. ■

INTRODUCTION

In Florida, a police officer stopped a driver on suspicion that he was driving while drunk. As the officer spoke to the driver, he noticed an open bottle of Jim Beam on the passenger's seat. The driver explained that he had not, in fact, been drinking while driving—because he drank only when the car was stopped at traffic lights. He was arrested after failing a sobriety test (Simmons, 2018). The temporal relations people can reason about are complex. English encodes some temporal relations in the tense and aspect of every statement, which relate its temporal reference to a separate reference time and to the time of the statement's utterance (Reichenbach, 1947). These relations enable individuals to make inferences about the order of events (Schaeken, 1996). They can also make such inferences from explicit temporal relations, as in "John shaved before he cooked breakfast" (e.g., Ye et al., 2012; Münte, Schiltz, & Kutas, 1998; Schaeken, Johnson-Laird, & d'Ydewalle, 1996; Clark, 1971). Likewise, verbs themselves differ in their temporal implications (see, e.g., Steedman, 2019; Moens & Steedman, 1988; Miller & Johnson-Laird, 1976, Sec. 6.2.4; Vendler, 1967, Chap. 4).

Logicians have formulated temporal logics, and researchers in artificial intelligence (AI) have developed calculi for reasoning about temporal relations (e.g., Øhrstrøm & Hasle, 1995; Freksa, 1992; Allen, 1983, 1991; for reviews, see Fisher, Gabbay, & Vila, 2005; Goranko, Montanari, & Sciavicco, 2004). Temporal logics sometimes neglect Reichenbach's (1947) distinction among three sorts of temporal referents (e.g., Prior's, 1967; tense logic). Some of the AI systems posit relations that do not map into simple everyday English. For instance, Allen's (1983) system has a primitive relation that can be expressed in English only in multiple clauses: "Event A and event B began simultaneously, but event A ended before event B did." None of these systems captures the full flexibility of natural language. For instance, tense and aspect can describe the same event as lasting over an extended period or as a point in time. It is exemplified in the difference between "Marie was doing the dishes" and "Marie did the dishes" (see, e.g., Kamp, 2017). Perhaps the AI systems most relevant to psychology are event calculi (see, e.g., Kowalski & Sergot, 1986), because they have axioms governing temporal relations, and so they are compatible with theories of reasoning based on orthodox logic (e.g., Rips, 1994). A typical commonsense principle in event calculi is that a condition is true at a particular time if and only if something happened earlier that initiated it and nothing has happened since to terminate it. Likewise, the following sort of axiom can establish a consequence of temporal relations:

> If X happened during Y and Y happened before Z, then X happened before Z.

Various logical systems use such axioms to avoid producing invalid temporal deductions. However, because these systems yield only valid deductions, they have difficulty in explaining why reasoners make systematic mistakes in temporal inferences.

This study focuses on temporal relations expressed with "during," for example:

1a. The car broke down during the road trip.
b. Breckinridge graduated during the Progressive Era.

[1]U.S. Naval Research Laboratory, [2]Princeton University, NJ, [3]New York University

These statements each describe an event—(1a) the breakdown and (1b) the graduation—that can be construed as "punctate," that is, as happening at a single point in time within the span of a longer "durative" event—(1a) the road trip and (1b) the Progressive Era. The sentential connective "while" can yield similar interpretations, as in:

2a. The man slept *while* the neighbors fought.
b. The neighbors fought *while* the man slept.

Durative relations are temporal and concern at least one event that has a duration. The examples in (2) illustrate a subtle difference: (2a) could mean that the neighbors fought for longer than the man slept, whereas (2b) could mean the converse. However, both assertions could also mean that the two events began together and ended together. Only a few empirical studies have examined durative relations, which, they suggest, are more complex than the mere order of events in time. For example, children comprehend and produce the word "while" only after they have mastered the words "before" and "after" (Winskel, 2003; Silva, 1991; Keller-Cohen, 1981).

Individuals can make inferences based on "while" using premises of the form "X happened while Y happened" (Schaeken et al., 1996). This and other studies showed that they appear to simulate a mental timeline of events (Bonato, Zorzi, & Umiltà, 2012; Casasanto & Boroditsky, 2008; Gentner, 2001). The studies also established that some temporal reasoning problems are easy and some are difficult—people take longer to make them and are more prone to errors (Baguley & Payne, 2000; Schaeken & Johnson-Laird, 2000; Vandierendonck & De Vooght, 1997). Temporal logics and event calculi are neither intended to explain these errors, nor can they.

Psychologists have also examined how people perceive the durations of experienced or anticipated events (Zakay & Block, 1997). In typical tasks, participants estimate how long a particular event took in minutes, in hours, or in some qualitative measure. They tend to overestimate short periods and underestimate longer ones (Lejeune & Wearden, 2009). This robust pattern is known as Vierordt's law after the German physiologist who discovered it. The two biases correlate with the amount of information that individuals represent per unit of time (Wang & Gennari, 2019). They compress information to avoid representing all the time points over which an event takes place (Faber & Gennari, 2015). Indeed, people can conceive of a long and complex sequence of processes as though they were a single punctate event, for example, "Hillary and Tenzing climbed Everest in 1953." They are liable to make such compressions when nothing of importance happened during an interval of time. However, particular tasks or demands may inhibit compressions. In addition, very brief events that would otherwise seem punctate, such as a blink of an eyelid, can be conceived as a sequence in slow motion—an object looms in front of a person's eye and triggers a reflex, which in turn closes the eyelid.

Khemlani, Harrison, and Trafton (2015) explained how reasoners construct a mental timeline from descriptions that used "while" and "during." The idea of the compression of sequences to form punctate events was one that they added to a previous account (Schaeken & Johnson-Laird, 2000; Schaeken, 1996). The theory predicted how individuals cope with durations to make modal inferences about what is necessary and possible and to assess the consistency of descriptions (Khemlani, Lotstein, Trafton, & Johnson-Laird, 2015). This study investigated two further consequences of this theory, one negative and one positive. The negative consequence is that human reasoners do not rely on any formal temporal system akin to an event calculus. The positive consequence is that they rely instead on the contents of descriptions to construct mental models of temporal relations. To test these claims, participants in experiments had to assess whether or not a set of statements in a temporal description could all be true at the same time, that is, whether the statements were consistent with one another, which in logic means that they were "satisfiable." The advantage of this task for our purposes is the striking contrast between the predictions of formal calculi and the predictions of the theory of mental models. We return to this contrast after we explain how models of durations work.

Mental Models of Relative Durations

Khemlani, Harrison, et al.'s (2015) account of reasoning about durations is based on the idea that language, memory, and imagination rely on mental simulations of possibilities, that is, mental models (Johnson-Laird, 2006; Johnson-Laird, Girotto, & Legrenzi, 2004). Not all cognitive processes require the use of mental models; for example, a musician can improvise a tune without them. However, the theory of mental models—the model theory, for short—proposes that inferences depend on the construction, inspection, and revision of models. They explain systematic patterns of reasoning about spatial relations (Ragni & Knauff, 2013; Jahn, Knauff, & Johnson-Laird, 2007; Jahn, Johnson-Laird, & Knauff, 2004) and abstract relations (Khemlani, Wasylyshyn, Briggs, & Bello, 2018; Goodwin & Johnson-Laird, 2005). Khemlani, Harrison, et al. (2015) developed a computational theory that shows that mental models can also explain systematic patterns of reasoning about time. The theory rests on three main assumptions, which are discussed hereinafter.

Models of Time Are Iconic

Models are iconic representations of events in relation to one another; that is, the structure of a model corresponds to the structure of what it represents (see Peirce, 1931–1958, Vol. 4). Models can also include noniconic elements, such as the symbol for negation (Khemlani, Orenes, & Johnson-Laird, 2012). There are two types of iconic models

of events. The first type uses space to represent the temporal position of events in a sequence. For instance, the following statements

3. The clouds gathered after the sunrise.
   The rain began after the clouds gathered.

can be represented in a static model with a temporal axis. In the following model, the axis runs from left to right to represent the chronology:

sunrise clouds rain

Each word stands for a mental simulation of an event. In this case, the events can be visualized. However, mental imagery often incorporates irrelevant details, and so it can impede reasoning about relations (Knauff & May, 2006; Knauff & Johnson-Laird, 2002). A model containing a single mental token representing each event often suffices for many inferences. Because the model is iconic, it can be scanned to yield different conclusions that emerge from its structure. Scanning the model of (3) shows that it supports the conclusion that the rain occurred after the sunrise. The conclusion is not explicit in the statements but is instead an emergent consequence of the model. Temporal models therefore imply a potential chronology, that is, the order in which events occurred.

The second type of iconic representation is kinematic in that a sequence of models unfolds in time to represent a sequence of events. It uses time itself to represent time, although not necessarily at the same rate—a simulation can occur at a faster or slower rate than the real events. People can use such models to reason. For example, people can understand informal algorithms that describe loops of operations, as in "while there are dishes in the sink: select a dish, wash it, and then rest it to dry." Experiments called for participants to use descriptions of informal algorithms to infer the order of train cars on a set of tracks. Reasoners spontaneously envisaged the temporal sequence of moves of the cars, relying on a kinematic simulation (Khemlani, Mackiewicz, Bucciarelli, & Johnson-Laird, 2013).

Khemlani, Harrison, et al. (2015) showed that models can be used to represent and to reason about relations between durations and punctate events. For inferences about durations, models often need to demarcate the starts and ends of events. There are various ways in which such representations can be constructed, and they need to take into account that some states have no clear beginning or ending, for example, "Pat loves Viv," whereas other habitual events occur as a series, for example, "He climbs mountains." The following temporal relation

4. The meeting happened during the sale.

can be represented as a minimal iconic model in which time runs from left to right, as in the following diagram, where the square brackets denote the start and end of an event:

```
[        sale        ]
   [meeting]
```

The inclusion of one set of brackets within another represents the temporal inclusion of one event during another. So, the diagram shows that the meeting occurred during the sale. For clarity, events can be represented in separate temporal streams as above. A computational model implementing the theory (Khemlani, Harrison, et al., 2015) uses an analogous representation in which lists represent the starts and ends of events, as in the following: $\text{sale}_{\text{START}}$ $\text{meeting}_{\text{START}}$ $\text{meeting}_{\text{END}}$ sale_{END}. These sorts of model are uniform, efficient, and computable until the number of indeterminacies between events becomes intractable. They can represent events over different timescales from nanoseconds to millennia, and they can represent definite, indefinite, or unknown durations, as in Example (4). Yet, because of the additional processes required to track the starts and ends of events, reasoning about durations should be more difficult than reasoning about punctate events, and when possible, reasoners should tend to collapse durations into punctate events. As we show below, they also use other strategies.

Two Cognitive Systems Exist: Intuition and Deliberation

The model theory postulates that reasoners rely on two systems of inference: an intuitive process that builds a single mental model and a deliberative process that considers alternative models, if any, to the initial intuitive one (see, e.g., Khemlani & Johnson-Laird, 2013). The late Peter Wason was the first to propose two systems of reasoning (Wason & Johnson-Laird, 1970), and they have remained a core component of the model theory (see, e.g., Johnson-Laird, 1983, Chap. 6). The two modes of thinking have been adopted in many other dual-process frameworks (e.g., Pennycook, Fugelsang, & Koehler, 2015; Kahneman, 2011; Stanovich & West, 2000). Recent evidence corroborates differential brain networks for the two processes: Intuitions often recruit long-term memories, whereas deliberations recruit cognitive control and working memory mechanisms (Williams, Kappen, Hassall, Wright, & Krigolson, 2019).

According to the model theory, intuitions underlie the rapid construction of an initial mental model (Johnson-Laird, Khemlani, & Goodwin, 2015). The process is subject to various heuristics (Jahn et al., 2007), and so reasoners who engage in only the intuitive process are prone to make systematic errors (Khemlani & Johnson-Laird, 2017). To validate an initial conclusion based on an intuitive model, a slower deliberative process can revise the initial model. It can yield "alternative models," that is, models in which the premises remain true. An alternative model can invalidate an initial conclusion—it is a model of a "counterexample" in which the premises are true but the conclusion is false. A conclusion is "necessary" if it holds in all models of the premises, "probable" if it holds in most models of the premises, "possible" if it holds in at least one model of the premises, and "impossible" if it holds in no models of the premises. Likewise, a set of

premises is "consistent" if there is at least one model in which all the premises are true, and it is "inconsistent" if there is no such model.

A working memory for intermediate results, or initial models, is the heart of computational power (see, e.g., Johnson-Laird, 1983, Chap. 1). Human working memory is limited in capacity, and deliberation requires a model to be held there while a search is made for an alternative. So, reasoners tend to rely on their initial models. If such a model yields a correct conclusion, its inference should be easy: It should be faster and more accurate. If, however, the initial model yields an incorrect conclusion, then deliberation is needed to arrive at a correct one, and reasoners should be slower and less accurate (see, e.g., Khemlani, 2018; Knauff, 1999; Johnson-Laird & Bryne, 1991, p. 124).

Models Should Underlie the Assessment of Consistency

The model theory explains in principle how individuals could assess the consistency of a set of statements, that is, whether or not they can all be true at the same time. Consider, for example, this set of statements that concerns a duration:

5. The meeting happened during the sale.
   The conference happened before the sale.
   The conference happened before the meeting.

The first two statements together yield the following model:

```
                    [ meeting ]
[ conference ]  [        sale        ]
```

The third statement holds in this model too, and so the set of statements is consistent. In contrast, consider this set of statements:

6. The meeting happened during the sale.
   The conference happened before the meeting.
   The conference happened during the sale.

The first two statements hold in the model above, but the third statement is inconsistent with it, which suggests that the set of statements is inconsistent. Deliberation, however, can yield an alternative model of the first two statements:

```
          [conference]  [meeting]
[                sale                ]
```

This model accommodates the third statement, and so the set is, in fact, consistent. Hence, reasoners should be more likely to make an error with (6) than with (5).

When the first two statements are consistent with more than one temporal model, the theory postulates that people introduce representations of events into a model in the same order in which the premises mention these events. They also enter representations in a way that aims to avoid having to rearrange those events that the model already represents (see Ragni & Knauff, 2013, who reported an analogous heuristic for spatial reasoning).

Another two illustrations concern inconsistent sets of statements. Consider this problem:

7. The meeting happened during the conference.
   The sale happened before the conference.
   The meeting happened before the sale.

The first two statements yield only one model:

```
                  [ meeting ]
[ sale ]  [ conference ]
```

This model is inconsistent with the third statement, and no alternative model exists, and so the set of statements is inconsistent too. They yield the "null" model, which corresponds to an inconsistency. Now, consider these statements:

8. The meeting happened during the conference.
   The sale happened before the meeting.
   The conference happened during the sale.

The first two statements yield at least two different models:

```
[ sale ]  [meeting]
          [ conference ]
```

and

```
[ sale ]  [ meeting ]
[        conference        ]
```

There are yet other possibilities; for example, the sale can overlap the beginning of the conference but end before the start of the meeting. However, no model of the first two statements can accommodate the third, and so the set of statements is again inconsistent. The detection of these inconsistencies should be quite straightforward, because any model of the first two statements is inconsistent with the third statement. Such inconsistencies often prompt reasoners to try to explain how the conflict arose (Khemlani & Johnson-Laird, 2011, 2012). For more complex descriptions, reasoners may disregard inconsistent information to build a coherent model (Otero & Kintsch, 1992).

The Crucial Predictions

The model theory predicts that reasoners tend to rely on intuitions rather than deliberations. Hence, for consistent sets of statements, people should be correct more often for sets yielding only one model than for sets yielding multiple models; for inconsistent sets of statements, however, no reason exists for them to differ in accuracy between problems whose first two premises yield one model or multiple models, because whichever model they construct, the third premise will be inconsistent with it. Event calculi and other theories based on axioms and rules of inference diverge from these predictions. These systems enable reasoners to derive valid conclusions from premises. They have only one general procedure that can yield a correct assessment of whether or not a set of statements is consistent. If there is a proof that the negation of a statement in

the set follows from the other statements in the set, then the set is inconsistent; otherwise, if there is no such proof, then the set is consistent. Granted that it is easier to find a proof than to make an exhaustive search that fails to find one, this approach predicts that correct assessments of inconsistency should be easier than those of consistency. Because proofs do not use models, these theories have no grounds to predict that consistent problems with only one model should be easier to assess than those with multiple models. The two accounts therefore make opposing predictions. In particular, if the model theory is correct, then one-model consistent problems should yield more accurate responses than multiple-model consistent problems, whereas if the logical theories are correct, then the two sorts of problem should not differ reliably. If the logical theories are correct, then inconsistent problems should yield more accurate responses than consistent problems, but if the model theory is correct, then there should be no reliable difference between them. We carried out three preregistered experiments to test these predictions.

EXPERIMENT 1

To test participants' accuracy in assessing consistent one-model and multiple-model problems, and in assessing inconsistent and consistent problems, Experiment 1 manipulated relevant variables. It used descriptions of durations, which were either consistent or inconsistent, and the first two of the three statements in a description yielded either one model or multiple models. An example of a consistent problem with one model is the following:

The speech happened during the press coverage.
The press coverage happened before the fireworks.
The speech happened before the fireworks.
Can all three of these sentences be true at the same time?

The question is equivalent to asking participants to decide whether all three statements are consistent, but the word "consistent" often confuses naive participants, that is, those with no background in logic. The task had not been used before to study temporal reasoning—previous experiments restricted the number of relations reasoners had to consider by asking them to infer the relation between two given events (Schaeken, 1996). It also posed the same uniform question for all problems.

Methods

Participants

Fifty participants completed the experiment for monetary compensation ($2 and a 10% chance of a $10 bonus) on Amazon's Mechanical Turk (AMT). Five participants were excluded from the analysis—some because they made excessive and inappropriate key presses, and others because they provided debriefing responses that implied that they had misunderstood the problems. The analyses reported below are based on the remaining 45 participants (mean age = 35.0 years; 21 were female). All but one of these 45 participants were native English speakers, and only three had taken a course in introductory logic.

Preregistration and Data Availability

The preregistrations, data, and analysis scripts are available at osf.io/evprc/.

Task and Design

The participants acted as their own controls and carried out 16 different problems. Each problem had three premises that described the temporal relations among three different events and asked whether the premises could all be true at the same time. Four sorts of problems occurred equally often in the experiment. The first two premises could yield either one model or multiple models, and the third premise was either consistent with the first two premises or inconsistent with them, therefore yielding the null model representing contradictions.

The first premise of each problem was of the following form: X happened during Y. Hence, the following is an example of a problem designed to yield one model, which we present as it cumulates over the three premises:

9.	X happened during Y.	[Y [X]]
	Y happened before Z.	[Y [X]] [Z]
	X happened before Z.	[Y [X]] [Z]

The models of the events next to each premise show how Khemlani, Harrison, et al.'s (2015) system updates the representation after interpreting new information. The problem presents a consistent description of events, because all three premises can be true at the same time.

In contrast, the set of premises in (10)

10.	X happened during Y.	[Y [X]]	
	Z happened before X.	[Z] [Y [X]]	(i)
		[Y [Z] [X]]	(ii)
	Y happened during Z.	Null model	

corresponds to a multiple-model problem, because the second premise is consistent with at least two different situations: one in which Z happened before Y started (i) and another in which Z happened after Y started (ii). Neither of these possibilities is consistent with the third premise, and so (10) is an inconsistent problem with multiple models for the first two premises. Appendix A summarizes the 16 problems used in the study. For each participant, once the contents had been assigned to the problems, their order of presentation was random.

Materials

The variables in the schemas for each problem were replaced with everyday events, for example, "the meeting," "the snowstorm," and "the ceremony." The materials were

drawn from 16 sets of three events (see Appendix B). Each set was designed to describe events that last for indeterminate durations such that any event in a set could take place during any other event, for example,

The meeting happened during the snowstorm.
The snowstorm happened during the ceremony.
The meeting happened during the ceremony.
The snowstorm happened during the meeting.

and so on. Therefore, the sets did not group together an event that occurs in only a second or so, such as a sneeze, with an event that takes significantly longer, such as a concert. In addition, the events of each set did not have any obvious causal relations to one another. The 16 material sets were rotated over the 16 different problems. Therefore, across the experiment as a whole, each set of materials occurred about equally often with each of the problems.

Procedure

The instructions explained that the task was to judge whether or not sets of statements could all be true at the same time. The participants saw a schematic of how their fingers should be placed on the computer keyboard, and they carried out a simple practice problem. For each problem, the participants considered the initial premise and then pressed the spacebar to reveal each of the remaining premises in turn. To encourage participants to read the premises in the given order, the program required 1 sec to pass after it displayed each premise before a spacebar press could trigger the next premise. Each premise remained on the screen during the display of the subsequent premises. After a participant had revealed all three premises, the program displayed the question: "Can all three of these sentences be true at the same time?" The participant then pressed the "F" or "J" key to register a "yes" or a "no" response. After completing all 16 problems, the participants answered four open-response debriefing questions, which probed their intuitive definitions of "before" and "during" as well as their strategies for tackling the problems.

Analysis

The data for all the experiments were subjected to paired nonparametric Wilcoxon tests for each effect and to a generalized logistic mixed effects model (GLMM) regression (using the R package "lme4"; Bates, Mächler, Bolker, & Walker, 2015) that controlled for noise as a result of differences among participants, items, and temporal relations in each problem, for example, "during/during/during," "during/before/during," and so forth. We included this factor because several participants reported that they based their judgments on these patterns alone. The model gave estimates for the main effects of problem type (one- vs. multiple-model), consistency, and their interaction. We examined relevant simple effects using estimated marginal means (using the R package "emmeans"; Lenth, Singmann, Love, Buerkner, & Herve, 2019).

While we collected RT data in each experiment, some participants reported using strategies that, unbeknown to them, prevented their data from being interpretable: They described tapping the spacebar and waiting for all premises to be available before they started reading the premises. We therefore omit any report of RTs, but we include the data and the relevant analyses in the on-line supplement.

Results and Discussion

Figure 1 presents the proportion of participants' correct assessments of consistency depending on whether the first two premises yielded one model or multiple models and on whether the three premises were consistent or inconsistent. Participants made accurate judgments reliably more often than chance (30 of the 45 participants did so, five did not do so, and there were 10 ties; binomial test, $p < .0001$) and were more accurate for one-model problems than for multiple-model problems (78% vs. 69%; Wilcoxon test, $z = 3.02$, $p = .003$, Cliff's $\delta = 0.43$; GLMM, $\beta = 1.49$, $z = 2.49$, $p = .013$). The difference between participants' accuracies did not reliably differ depending on whether the problem was consistent or inconsistent (72% vs. 75%; Wilcoxon test, $z = 1.12$, $p = .266$, Cliff's $\delta = 0.17$; GLMM, $\beta = 0.50$, $z = 1.13$, $p = .257$). The results exhibited an interaction between the problem type (one- vs. multiple-model) and the consistency of the premises (Wilcoxon test, $z = 4.03$, $p < .0001$, Cliff's $\delta = 0.42$; GLMM, $\beta = -3.06$, $z = 3.24$, $p = .001$). The interaction reflected the participants' greater accuracy for consistent one-model problems than for consistent multiple-model problems (83% vs. 61%; Wilcoxon test, $z = 4.32$, $p < .0001$, Cliff's $\delta = 0.56$; GLMM, $\beta = -3.03$, $z = 3.22$,

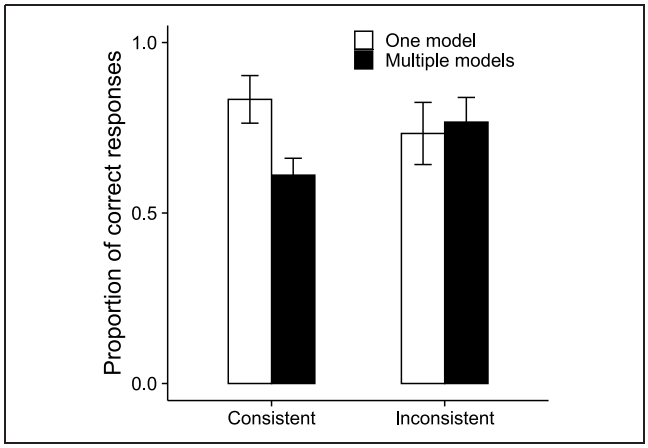

Figure 1. The proportions of correct responses in Experiment 1 (n = 45) depending on whether the initial two premises yielded one model or multiple models and on whether the three premises were consistent or inconsistent. Error bars indicate 95% confidence intervals.

p = .001), whereas inconsistent problems showed no marked difference between problem types.

Experiment 1 supported the predictions of the model theory, but as we noted above, some participants reported that they developed heuristics based on the patterns of relations across the three premises:

> "[A]bout halfway through [I] could skim…[such that] during-during before/after = no [and] during-during-during = yes"

These heuristics do not require reasoners to engage with the particular contents of the problem, and they could have led participants to make incorrect responses. All the problems in the experiment had a first premise in which one event occurred during another, and so there were only four possible combinations of relations that the other two premises could describe. We therefore carried out Experiment 2 as a replication in which each problem consisted of four premises interrelating four events. The design increased the number of possible relation patterns and made it less feasible for participants to develop these heuristics.

EXPERIMENT 2

Experiment 2 replicated Experiment 1 in that it manipulated whether temporal descriptions referred to one model or multiple models over the first three of their statements and whether the sets were consistent or inconsistent. It further sought to eliminate the use of the heuristics described above by using sets of four statements.

Methods

Participants

Fifty participants completed the experiment for compensation ($2) on AMT. One participant's results were excluded from the analysis, because of a mean accuracy of 2 *SD*s below the sample's mean accuracy. The analyses reported below were based on the remaining 49 participants (mean age = 36.6 years; 20 were female). All participants were native English speakers, and only two had taken a course in introductory logic.

Design

The same task was used as in Experiment 1, but the problems had four premises. For one-model problems, the first three premises were consistent with only one model, and for multiple-model problems, they were consistent with multiple models. Half the problems had a fourth premise that was consistent with the previous three, and half the problems had a fourth premise that was inconsistent with them. There were four instances of each sort of problem, and they were presented in a different random order to each participant. Appendix C summarizes the 16 problems.

Materials and Procedure

The experiment changed the sets of materials and the display of the problems. We added a new event to each set of materials to accommodate the increase to four premises; we shortened unnecessarily long names of events; for example, we used "wash cycle" instead of "dishwasher cycle." In addition, we eliminated those events that typically occur in a fixed sequence, for example, "earthquake," "light flicker," and "scream." Appendix B shows all the changes. Participants in the current experiment pressed the spacebar to view each premise including the first one. Otherwise, the procedure and analysis were the same as in Experiment 1.

Results and Discussion

Figure 2 presents the proportions of participants' correct assessments of consistency depending on whether the first three premises yielded one model or multiple models and on whether the problems were consistent or inconsistent. Participants made accurate judgments reliably more often than chance (36 of the 49 participants did so, four did not do so, and there were nine ties; binomial test, $p < .0001$). They were more accurate for one-model problems than multiple-model problems (78% vs. 54%; Wilcoxon test, $z = 5.34$, $p < .0001$, Cliff's $\delta = 0.66$; GLMM, $\beta = 1.27$, $z = 5.37$, $p < .0001$). Unlike in Experiment 1, participants were more accurate for inconsistent problems than for consistent problems (74% vs. 57%; Wilcoxon test, $z = 3.64$, $p < .001$, Cliff's $\delta = 0.45$; GLMM, $\beta = 1.20$, $z = 4.38$, $p < .0001$). However, this difference was attributable to an interaction between the type of problem (one- vs.

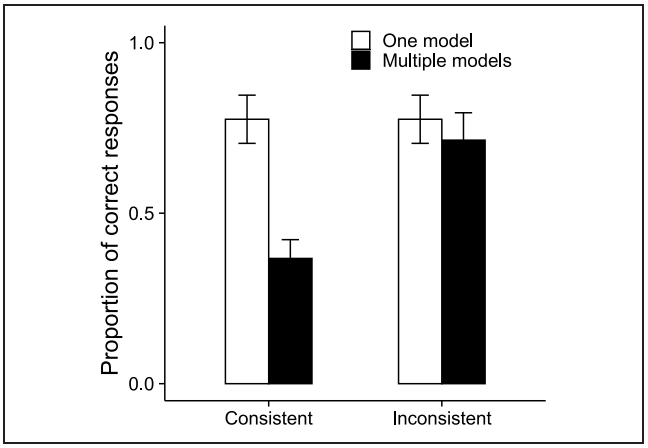

Figure 2. The proportions of correct responses in Experiment 2 (n = 49) depending on the type of problem (one- or multiple-model) and on whether the premises were consistent or inconsistent. Error bars indicate 95% confidence intervals.

multiple-model) and the consistency of the premises, which was reliable in the nonparametric analysis (Wilcoxon test, $z = 4.39$, $p < .0001$, Cliff's $\delta = 0.52$), but not in the logistic regression (GLMM, $\beta = -0.74$, $z = 1.32$, $p = .188$). As in Experiment 1, the interaction reflected the participants' greater accuracy for consistent one-model problems than for consistent multiple-model problems (76% vs. 37%; Wilcoxon test, $z = 5.35$, $p < .0001$, Cliff's $\delta = 0.76$; GLMM, $\beta = 1.65$, $z = 4.17$, $p < .0001$). Responses were also more accurate for inconsistent one-model problems than for inconsistent multiple-model problems, but the difference was much smaller (78% vs. 71%) and only reliable in the GLMM (Wilcoxon test, $z = 1.76$, $p = .078$, Cliff's $\delta = 0.12$; GLMM, $\beta = 0.90$, $z = 2.62$, $p < .01$).

The difference in accuracy between one-model and multiple-model consistent problems was striking. For three of the four multiple-model consistent problems, the mean accuracy was less than 25%; that is, it was much lower than chance performance. One reason was evident in many of the participants' remarks in the debriefing questionnaires:

> "I didn't take into account that one event could last longer than another."
> "In my head 'during' became '='."
> "I thought if something happened during something else, the two were equivalent…."

This interpretation treats "the meeting happened during the conference" as though the meeting and the conference started and ended at the same time. Hence, we refer to it as the "equal duration" interpretation. It predicts erroneous judgments of inconsistency for the three consistent multiple-model problems mentioned above. In fact, 43% of all consistent multiple-model trials resulted in errors that could reflect the equal duration interpretation. Some participants in Experiment 1 had likewise reported using the same interpretation. Because the equal duration interpretation affected only consistent multiple-model problems, it is a confound. Experiment 3 eliminated this confound.

EXPERIMENT 3

Experiment 3 replicated the design of the previous experiment—it manipulated whether sets of four temporal statements referred to one model or to multiple models and whether they were consistent or inconsistent. Changes to the instructions were aimed at preventing the equal duration interpretation: They established that all the events in a problem had different durations.

Methods

Participants

Fifty participants completed the experiment for compensation ($2) on AMT. The results from four participants were excluded from analysis (for performance more than 2 *SD*s below the mean or for violations of the instructions). The analyses below are for the results of the remaining 46 participants (mean age = 37.0 years; 25 were female). All the participants were native English speakers, and only two had taken a course in introductory logic.

Design and Materials

They were the same as in Experiment 2.

Procedure

The experiment changed the instructions to prevent the participants from making an equal duration interpretation of the premises. They received an example of a problem paired with the instructions: "None of these events have the same duration" and "All of the events last for different lengths of time." They were then quizzed on their interpretation of the instructions. During the experiment itself, there was the following reminder underneath all the response options: "Remember: None of the events have the same duration." Otherwise, the procedure and analysis were the same as in Experiment 2.

Results and Discussion

Figure 3 presents the proportion of participants' correct assessments of consistency depending on whether the premises yielded one model or multiple models and on whether they were consistent or inconsistent. Participants made accurate judgments reliably more often than chance (28 of the 46 participants did so, four did not do so, and there were 14 ties; binomial test, $p < .0001$). Participants were more accurate for one-model problems than for

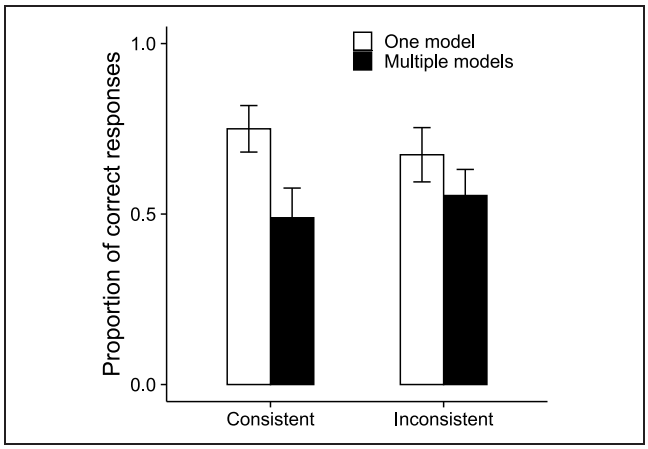

Figure 3. The proportions of correct responses in Experiment 3 ($n = 46$) depending on the type of problem (one- or multiple-model) and on whether the premises were consistent or inconsistent. Error bars indicate 95% confidence intervals.

multiple-model problems (71% vs. 52%; Wilcoxon test, $z = 4.88$, $p < .0001$, Cliff's $\delta = 0.51$; GLMM, $\beta = 1.01$, $z = 5.21$, $p < .0001$). There was no difference between participants' accuracies depending on whether the problem was consistent or inconsistent (62% vs. 61%). As Figure 3 shows, there was a trend toward an interaction between problem type (one- vs. multiple-model) and consistency, but it was not reliable. The results once again validated the primary effect predicted by the model theory: Participants were more accurate for consistent one-model problems than for consistent multiple-model problems (75% vs. 49%; Wilcoxon test, $z = 4.36$, $p < .0001$, Cliff's $\delta = 0.49$; GLMM, $\beta = 1.20$, $z = 3.76$, $p < .001$).

The postexperimental questionnaires showed that the instructions had blocked the equal duration interpretation: The participants acknowledged that the events need not have the same durations. For example, one participant noted that "during" means two events are "[h]appening at the same time but [that] doesn't necessarily mean they will start or end together." Participants nevertheless made more errors on consistent multiple-model problems than on consistent one-model problems.

GENERAL DISCUSSION

How do people mentally represent and reason about temporal relations, such as those expressed with "during" and "before"? Many logical frameworks describe ideal temporal reasoning (Fisher et al., 2005; Goranko et al., 2004), and a plausible candidate for everyday reasoning is an event calculus (e.g., Kowalski & Sergot, 1986). Such frameworks are compatible with theories of reasoning based on standard logic (see, e.g., Rips, 1994), which can invoke postulates that capture the logical properties of connectives, such as the transitivity of "during":

11. X happened during Y.
    Y happened during Z.
    Therefore, X happened during Z.

The drawbacks of such an approach are threefold.

First, the formal rules and axioms of standard logic do not state the conditions in which assertions are true or the conditions in which they are false. So, they provide no machinery for how an individual seeing a sequence of events can determine that a temporal description of them is true or else false.

Second, standard logic allows that infinitely many conclusions are provable from any set of premises, and it provides no principles guiding which conclusions are worth drawing. Rips (1994) therefore compensates for this problem: His theory focuses on the evaluation of given conclusions and curbs the power of rules, such as *A*—therefore, *A* or *B*—or *both*, which can be used to introduce an indefinite number of new statements. In contrast, iconicity in the model theory constrains the possibilities that people consider and the conclusions that they draw. When reasoners build a model of a set of premises, they aim to maintain semantic information and to draw a conclusion that is true in all models of the premises but that is not stated in an explicit premise (Johnson-Laird, 1983, pp. 37–40; for an early algorithm embodying these principles, see Johnson-Laird & Bryne, 1991, Chap. 9).

Third, the only general way to use formal rules to assess whether or not a set of statements is consistent is to try to prove the negation of one member of the set from the other members of the set. The existence of such a proof establishes the inconsistency of the set, whereas a failure to derive a proof after an exhaustive search establishes the consistency of the set. This procedure is cognitively implausible (Johnson-Laird, Legrenzi, Girotto, & Legrenzi, 2000). The model theory has a simple solution (Khemlani, Lotstein, et al., 2015): If and only if there is a model of all the premises, then they are consistent.

Because the model theory distinguishes between two different systems, intuitive and deliberative, it follows that statements for which intuition (System 1) yields a correct model should be easier to assess as consistent than statements that call for deliberation (System 2) to build an alternative model. Consider, for instance, the following consistent problem from Experiment 1:

12. The meeting happened during the snowstorm.
    The ceremony happened before the snowstorm.
    The ceremony happened before the meeting.

The first two statements yield the following model, where the square brackets denote the start and end of events:

```
            [meeting]
[ceremony]  [ snowstorm ]
```

The third statement holds in the model, and there is no alternative model. So, this one-model problem should be easy. A more difficult problem is the following:

13. The meeting happened during the snowstorm.
    The ceremony happened before the meeting.
    The ceremony happened during the snowstorm.

The first two statements yield the same intuitive model as in (12) above, but the third statement is inconsistent with this model, and so many participants should incorrectly judge the set of statements as inconsistent. Those who deliberate may discover this alternative model of the statements:

```
[ceremony]  [meeting]
[          snowstorm          ]
```

So, the set of statements is, in fact, consistent, but it should be difficult to make this assessment. The model theory and logical rule theories therefore make divergent predictions:

> Models predict that one-model consistent problems should be easier than multiple-model consistent problems, whereas logic does not. Logic predicts that

inconsistent problems should be easier than consistent problems, whereas the model theory does not.

Our three experiments corroborated the model theory. Judgments of consistency were more accurate for descriptions with one model than for descriptions with multiple models, and these judgments were not reliably less accurate than those for inconsistent descriptions (Experiment 1). Some participants reported that they had developed a strategy in which they assessed only the relational terms in the three statements—they responded "yes" to three occurrences of "during"; otherwise, they responded "no." Reasoners do often spontaneously discover such strategies (see, e.g., Schaeken & Johnson-Laird, 2000). The use of four statements in descriptions prevented the participants from developing them. Yet, the difference between the two sorts of consistent descriptions still occurred: Those depending on one model yielded a greater accuracy than those depending on multiple models (Experiment 2). However, now, consistent descriptions yielded fewer accurate responses than inconsistent ones—a phenomenon that reflected the considerable difficulty of consistent descriptions yielding multiple models (see Figure 2). Some participants reported that they had assumed that, when one event occurred during another, both events started together and ended together. This interpretation might have explained their difficulty with consistent descriptions that had multiple models. When a new study used instructions and practice to prevent this unwarranted assumption, the main pattern of results was corroborated once again. One model led to more accurate judgments of consistency than multiple models, and no reliable difference in accuracy occurred between consistent and inconsistent descriptions (Experiment 3).

These results conflict with the hypothesis that individuals use some sort of logic or event calculus to assess the consistency of temporal descriptions. Likewise, reasoners' difficulty in deducing temporal relations correlates with the number of models they need to consider rather than the steps needed to prove those relations (Schaeken & Johnson-Laird, 2000). Defenders of logical calculi may counter that perhaps a different set of rules could predict the results. In contrast, these studies of consistency are not open to this defense—logical calculi have no direct procedure for assessing the consistency of a set of statements, and so they are bound to predict that inconsistent sets should be easier to assess than consistent ones. Our experiments show otherwise.

Concerns and Limitations

Some readers may disagree with our analysis of "during." Consider the following set of statements (from Example 8 above):

14. X happened during Y.
    Z happened before X.

They allow for the following model in which Z continues beyond the start of Y:

```
          [X]
      [   Y   ]
  [ Z ]
```

However, following Kamp (2017), an accurate description of the preceding model is as follows:

15. Part of Z happened during part of Y and ended before X.

We took (15) to be inconsistent with the third statement in (8)

16. Y happened during Z

because only part of Y happened during part of Z. Perhaps people are more liberal in their reading of "during." Even so, the possibility of these interpretations does not alter the impact of our results: Consistent problems should be consistent on any reasonable interpretation of "during." More liberal interpretations of "during" can explain only why people sometimes judge inconsistent descriptions to be consistent, but such judgments are contrary to our findings.

Our experiments were limited in at least two ways. First, they did not explore the causes of errors. Errors on consistent problems with multiple models could result from a failure to deliberate or from a failure of deliberation to yield an appropriate alternative model. If researchers can identify the cause of errors, then they may be able to develop interventions to improve reasoning. Second, our studies focused on the preposition "during" and, by design, ignored the many other ways in which to describe the co-occurrence of events. For example, the sentential connective "while" has a similar interpretation to "during," and both are in the 200 most frequent words in American English (Davies, 2008). As we noted earlier, "while" also has an important interpretation in informal descriptions of algorithms; for example, "While this condition holds, carry out the following operations," and participants use other similar expressions to describe loops of operations, such as "as long as" and "until" (see Khemlani et al., 2013). Future studies should examine how individuals interpret these different ways to express relative durations.

Conclusion

As in our opening example, everyone can envisage a man who drinks whiskey in his car during its stops at traffic lights, but not while it is in motion. That imaginative ability is the main postulate of the model theory. In addition, the theory has the great advantage over temporal logics and event calculi that models represent what speakers can perceive, imagine, and communicate. An accurate model is an immediate demonstration of the consistency of a description, and it explains the spontaneous conclusions that individuals draw. They capitalize on the iconic nature of models.

APPENDIX A: THE 16 SORTS OF PROBLEM USED IN EXPERIMENT 1

Number of Models	*Consistency*	*First Premise*	*Second Premise*	*Third Premise*
One model	Consistent	X happened during Y	Y happened before Z	X happened before Z
One model	Consistent	X happened during Y	Z happened during X	Z happened during Y
One model	Consistent	X happened during Y	Y happened during Z	X happened during Z
One model	Consistent	X happened during Y	Z happened before Y	Z happened before X
Multiple models	Consistent	X happened during Y	X happened during Z	Z happened during Y
Multiple models	Consistent	X happened during Y	Z happened during Y	Z happened during X
Multiple models	Consistent	X happened during Y	Z happened before X	Z happened during Y
Multiple models	Consistent	X happened during Y	Z happened during Y	X happened before Z
One model	Inconsistent	X happened during Y	Y happened before Z	Z happened during X
One model	Inconsistent	X happened during Y	Z happened during X	Z happened before Y
One model	Inconsistent	X happened during Y	Y happened during Z	X happened before Z
One model	Inconsistent	X happened during Y	Z happened before Y	X happened before Z
Multiple models	Inconsistent	X happened during Y	Z happened before X	Y happened during Z
Multiple models	Inconsistent	X happened during Y	X happened during Z	Z happened before Y
Multiple models	Inconsistent	X happened during Y	X happened during Z	Z happened before X
Multiple models	Inconsistent	X happened during Y	X happened before Z	Z happened before Y

APPENDIX B: THE 16 SETS OF CONTENTS IN EXPERIMENTS 1–3

Event 1	*Event 2*	*Event 3*	*Event 4*	*Replacements*
burglar alarm	fire	siren	*explosion*	
dishwasher cycle	shower	baking	*news report*	*wash cycle*
speech	press coverage	fireworks	*interview*	*monologue*
hike	conversation	hailstorm	*negotiation*	
commute	sunrise	podcast	*windstorm*	
pep talk	hiccups	applause	*intermission*	
car alarm	argument	prank	*sunset*	
sale	headache	shopping trip	*book signing*	*game*
manhunt	summer camp	cold spell	*contest*	*camping trip*, *heat wave*
scream	light flicker	earthquake	*thunderstorm*	*phone call*, *nap*, *movie*
caroling session	fog	sled ride	*tribute*	*rap battle*, *publicity stunt*
bus ride	sunset	traffic jam	*fight*	*eclipse*
rainstorm	heat wave	thunder	*celebration*	*workshop*, *farmer's market*
meeting	snowstorm	ceremony	*video conference*	
concert	beach party	volleyball tournament	*bonfire*	*tournament*
vacation	engagement	flu	*carnival*	*flood*, *outbreak*

The plain text events were used in all three experiments. The underlined events were used only in Experiment 1. The italicized events were only used in Experiments 2 and 3, including a fourth event for each set and replacements for italicized items.

APPENDIX C: THE 16 SORTS OF PROBLEM IN EXPERIMENTS 2 AND 3

Number of Models	*Consistency*	*First Premise*	*Second Premise*	*Third Premise*	*Fourth Premise*
One model	Consistent	W happened during X	X happened during Y	Y happened before Z	W happened before Z
One model	Consistent	W happened during X	Y happened during W	X happened before Z	Y happened during X
One model	Consistent	W happened during X	Y happened before X	Z happened during W	Y happened before Z
One model	Consistent	W happened during X	X happened before Y	Z happened during Y	W happened before Z
Multiple models	Consistent	W happened during X	Y happened during X	Z happened during Y	Z happened before W
Multiple models	Consistent	W happened during X	W happened during Y	W happened before Z	Y happened during X
Multiple models	Consistent	W happened during X	Y happened before W	Z happened before Y	Y happened during X
Multiple models	Consistent	W happened during X	W happened before Y	Z happened during X	Y happened during Z
One model	Inconsistent	W happened during X	X happened during Y	Y happened during Z	Z happened before W
One model	Inconsistent	W happened during X	X happened before Y	Y happened before Z	W happened during Z
One model	Inconsistent	W happened during X	Y happened before X	Z happened during Y	Z happened during W
One model	Inconsistent	W happened during X	Y happened during W	Z happened before X	W happened before Z
Multiple models	Inconsistent	W happened during X	W happened before Y	X happened during Z	Y happened before Z
Multiple models	Inconsistent	W happened during X	Y happened before W	Z happened during Y	Z happened during W
Multiple models	Inconsistent	W happened during X	W happened during Y	Y happened before Z	Z happened before X
Multiple models	Inconsistent	W happened during X	X happened during Y	Z happened before W	Y happened during Z

Acknowledgments

This research was performed while the first author held an NRC Research Associateship award at the U.S. Naval Research Laboratory. It was also supported by a grant from the Office of Naval Research to the second author. We are grateful to Kalyan Gupta, Danielle Paterno, and Kevin Zish at the Knexus Research Corporation for their help in conducting the experiments. Finally, we thank Bill Adams, Gordon Briggs, Monica Bucciarelli, Hillary Harner, Tony Harrison, Laura Hiatt, Joanna Korman, Andrew Lovett, Marco Ragni, Mark Steedman, and Greg Trafton for their advice and comments.

Reprint requests should be sent to Laura Jane Kelly, U.S. Naval Research Laboratory, Navy Center for Applied Research in Artificial Intelligence, 4555 Overlook Ave. SW, Washington, DC 20375, or via e-mail: laura.kelly.ctr@nrl.navy.mil.

REFERENCES

Allen, J. F. (1983). Maintaining knowledge about temporal intervals. *Communications of the ACM*, *26*, 832–843.

Allen, J. F. (1991). Time and time again: The many ways to represent time. *International Journal of Intelligent Systems*, *6*, 341–355.

Baguley, T., & Payne, S. J. (2000). Long-term memory for spatial and temporal mental models includes construction processes and model structure. *Quarterly Journal of Experimental Psychology, Section A*, *53*, 479–512.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48.

Bonato, M., Zorzi, M., & Umiltà, C. (2012). When time is space: Evidence for a mental time line. *Neuroscience & Biobehavioral Reviews*, *36*, 2257–2273.

Casasanto, D., & Boroditsky, L. (2008). Time in the mind: Using space to think about time. *Cognition*, *106*, 579–593.

Clark, E. V. (1971). On the acquisition of the meaning of *before* and *after*. *Journal of Verbal Learning and Verbal Behavior*, *10*, 266–275.

Davies, M. (2008). The corpus of contemporary American English (COCA): 560 million words, 1990–present. Retrieved from corpus.byu.edu/coca/.

Faber, M., & Gennari, S. P. (2015). Representing time in language and memory: The role of similarity structure. *Acta Psychologica*, *156*, 156–161.

Fisher, M. D., Gabbay, D. M., & Vila, L. (2005). *Handbook of temporal reasoning in artificial intelligence*. Amsterdam: Elsevier.

Freksa, C. (1992). Temporal reasoning based on semi-intervals. *Artificial Intelligence*, *54*, 199–227.

Gentner, D. (2001). Spatial metaphors in temporal reasoning. In M. Gattis (Ed.), *Spatial schemas and abstract thought* (pp. 203–222). Cambridge, MA: MIT Press.

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

Goranko, V., Montanari, A., & Sciavicco, G. (2004). A road map of interval temporal logics and duration calculi. *Journal of Applied Non-Classical Logics*, *14*, 9–54.

Jahn, G., Johnson-Laird, P. N., & Knauff, M. (2004). Reasoning about consistency with spatial mental models: Hidden and obvious indeterminacy in spatial descriptions. In C. Freksa, M. Knauff, B. Krieg-Brückner, B. Nebel, & T. Barkovsky (Eds.), *Spatial cognition IV: Reasoning, action, interaction* (pp. 165–180). Berlin, Germany: Springer.

Jahn, G., Knauff, M., & Johnson-Laird, P. N. (2007). Preferred mental models in reasoning about spatial relations. *Memory & Cognition*, *35*, 2075–2087.

Johnson-Laird, P. N. (1983). *Mental models*. Cambridge, United Kingdom: Cambridge University Press.

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

Johnson-Laird, P. N., & Bryne, R. M. J. (1991). *Deduction*. Hillsdale, NJ: Erlbaum.

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. S., & Goodwin, G. P. (2015). Logic, probability, and human reasoning. *Trends in Cognitive Sciences*, *19*, 201–214.

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

Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.

Kamp, H. (2017). Events, discourse representations and temporal reference. *Semantics & Pragmatics*, *10*, 1–68.

Keller-Cohen, D. (1981). Elicited imitation in lexical development: Evidence from a study of temporal reference. *Journal of Psycholinguistic Research*, *10*, 273–288.

Khemlani, S. S. (2018). Reasoning. In S. Thompson-Schill (Ed.), *Stevens' handbook of experimental psychology and cognitive neuroscience* (pp. 385–427). Hoboken, NJ: Wiley.

Khemlani, S. S., Harrison, A. M., & Trafton, J. G. (2015). Episodes, events, and models. *Frontiers in Human Neuroscience*, *9*, 590.

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

Khemlani, S. S., & Johnson-Laird, P. N. (2012). Theories of the syllogism: A meta-analysis. *Psychological Bulletin*, *138*, 427–457.

Khemlani, S. S., & Johnson-Laird, P. N. (2013). The processes of inference. *Argument & Computation*, *4*, 4–20.

Khemlani, S. S., & Johnson-Laird, P. N. (2017). Illusions in reasoning. *Minds and Machines*, *27*, 11–35.

Khemlani, S. S., Lotstein, M., Trafton, J. G., & Johnson-Laird, P. N. (2015). Immediate inferences from quantified assertions. *Quarterly Journal of Experimental Psychology*, *68*, 2073–2096.

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

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

Khemlani, S. S., Wasylyshyn, C., Briggs, G., & Bello, P. (2018). Mental models and omissive causation. *Memory & Cognition*, *46*, 1344–1359.

Knauff, M. (1999). The cognitive adequacy of Allen's interval calculus for qualitative spatial representation and reasoning. *Spatial Cognition and Computation*, *1*, 261–290.

Knauff, M., & Johnson-Laird, P. N. (2002). Visual imagery can impede reasoning. *Memory & Cognition*, *30*, 363–371.

Knauff, M., & May, E. (2006). Mental imagery, reasoning, and blindness. *Quarterly Journal of Experimental Psychology*, *59*, 161–177.

Kowalski, R., & Sergot, M. (1986). A logical-based calculus of events. *New Generation Computing*, *4*, 67–95.

Lejeune, H., & Wearden, J. H. (2009). Vierordt's *The experimental study of the time sense* (1868) and its legacy. *European Journal of Cognitive Psychology*, *21*, 941–960.

Lenth, R., Singmann, H., Love, J., Buerkner, P., & Herve, M. (2019). emmeans: Estimated marginal means, aka least-squares means. Retrieved from cran.r-project.org/package=emmeans.

Miller, G. A., & Johnson-Laird, P. N. (1976). *Language and perception*. Cambridge, MA: Belknap Press of Harvard University Press.

Moens, M., & Steedman, M. (1988). Temporal ontology and temporal reference. *Computational Linguistics*, *14*, 15–28.

Münte, T. F., Schiltz, K., & Kutas, M. (1998). When temporal terms belie conceptual order. *Nature*, *395*, 71–73.

Øhrstrøm, P., & Hasle, P. F. V. (1995). *Temporal logic: From ancient ideas to artificial intelligence*. Dodrecht, The Netherlands: Kluwer.

Otero, J., & Kintsch, W. (1992). Failures to detect contradictions in a text: What readers believe versus what they read. *Psychological Science*, *3*, 229–235.

Peirce, C. S. (1931–1958). *Collected papers of Charles Sanders Peirce. 8 vols*. C. Hartshorne, P. Weiss, & A. W. Burks (Eds.). Cambridge, MA: Harvard University Press.

Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2015). What makes us think? A three-stage dual-process model of analytic engagement. *Cognitive Psychology*, *80*, 34–72.

Prior, A. N. (1967). *Past, present, and future*. Oxford, United Kingdom: Oxford University Press.

Ragni, M., & Knauff, M. (2013). A theory and a computational model of spatial reasoning with preferred mental models. *Psychological Review*, *120*, 561–588.

Reichenbach, H. (1947). *Elements of symbolic logic*. New York: Free Press.

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

Schaeken, W. (1996). Tense, aspect, and temporal reasoning. *Thinking & Reasoning*, *2*, 309–327.

Schaeken, W., & Johnson-Laird, P. N. (2000). Strategies in temporal reasoning. *Thinking & Reasoning*, *6*, 193–219.

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

Silva, M. N. (1991). Simultaneity in children's narratives: The case of *when*, *while*, and *as*. *Journal of Child Language*, *18*, 641–662.

Simmons, R. (2018). Florida man tells cops he wasn't drinking and driving—He was only drinking Jim Beam at stop signs, traffic lights. *Orlando Sentinel*. Retrieved from www.orlandosentinel.com/opinion/audience/roger-simmons/os-ae-florida-man-drinking-and-driving-20180711-story.html.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, *23*, 645–665.

Steedman, M. (2019). Form-independent meaning representation for eventualities. In R. Truswell (Ed.), *The Oxford handbook of event structure* (pp. 605–623). Oxford, United Kingdom: Oxford University Press.

Vandierendonck, A., & De Vooght, G. (1997). Working memory constraints on linear reasoning with spatial and temporal contents. *Quarterly Journal of Experimental Psychology, Section A*, *50*, 803–820.

Vendler, Z. (1967). *Linguistics in philosophy*. Ithaca, NY: Cornell University Press.

Wang, Y., & Gennari, S. P. (2019). How language and event recall can shape memory for time. *Cognitive Psychology*, *108*, 1–21.
Wason, P. C., & Johnson-Laird, P. N. (1970). A conflict between selecting and evaluating information in an inferential task. *British Journal of Psychology*, *61*, 509–515.
Williams, C. C., Kappen, M., Hassall, C. D., Wright, B., & Krigolson, O. E. (2019). Thinking theta and alpha: Mechanisms of intuitive and analytical reasoning. *Neuroimage*, *189*, 574–580.
Winskel, H. (2003). The acquisition of temporal event sequencing: A cross-linguistic study using an elicited imitation task. *First Language*, *23*, 65–95.
Ye, Z., Kutas, M., St. George, M., Sereno, M. I., Ling, F., Münte, T. F., et al. (2012). Rearranging the world: Neural network supporting the processing of temporal connectives. *Neuroimage*, *59*, 3662–3667.
Zakay, D., & Block, R. A. (1997). Temporal cognition. *Current Directions in Psychological Science*, *6*, 12–16.