The consistency of durative relations

Laura Kelly and Sangeet Khemlani
{laura.kelly.ctr, sangeet.khemlani}@nrl.navy.mil
Navy Center for Applied Research in Artificial Intelligence
US Naval Research Laboratory, Washington, DC 20375 USA

Abstract

Few experiments have examined how people reason about durative relations, e.g., "during". Such relations pose challenges to present theories of reasoning, but many researchers argue that people simulate a mental timeline when they think about sequences of events. A recent theory posits that to mentally simulate durative relations, reasoners do not represent all of the time points across which an event might endure. Instead, they construct discrete tokens that stand in place of the beginnings and endings of those events. The theory predicts that when reasoners need to build multiple simulations to solve a reasoning problem, they should be more prone to error. To test the theory, an experiment provided participants with sets of premises describing durative relations; they assessed whether the sets were consistent or inconsistent. The results of the experiment validated the theory's prediction. We conclude by situating the study in recent work on temporal thinking.

Keywords: events, temporal reasoning, durative relations, mental models, consistency

Introduction

A police officer stopped a driver on the suspicion of drunk driving near Vero Beach, FL. As the officer began to speak to the driver, he noticed an open bottle of Jim Beam on the passenger's seat. The driver explained to the officer that he had not, in fact, been drinking *while* driving – because he only drank when the car was stopped at traffic lights. He was arrested after failing a field sobriety test (Simmons, 2018).

In daily life, people use temporal relations such as "while" and "during" to convey information about events that endure across more than one point in time. Consider the function of the temporal preposition "during" in the following examples:

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

The statements each describe a punctate event, i.e., a single point in time (e.g., the breakdown, the graduation), that occurred in the context of a period that extends across multiple time points (e.g., the road trip, the Progressive Era). The sentential connective "while" can yield similar interpretations, as in the examples in (2):

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

The examples show how syntax can change the way events are interpreted. For instance, (2a) seems to suggests that the neighbors fought for longer than the man slept, whereas (2b) seems to convey the opposite. Perhaps the two statements are compatible with one another, as in the situation in which the man started sleeping right as the fight began and woke up when the fight ended.

Researchers in artificial intelligence have developed many systems of temporal logic to cope with reasoning about durative events (e.g., Allen, 1983, 1991; Freksa, 1992). Temporal logics often stipulate relations between intervals of time. The logics were designed to describe durative events as they occur in the world – they were not developed to capture how humans think about time. Hence, many temporal logics posit relations that don't map onto prepositions or connectives in English. For instance, Allen's (1983) system includes the following types of relation that connect event A with event B:

```
AAAA          A starts B.
BBBBBBBB

  AAAA        A finishes B.
BBBBBBBB

AAAABBBB      A meets B.
```

The repetitions of the letters are used to depict how one event endures across multiple points in time. The descriptions of the relations in natural language can be quite complex, e.g., you might describe the *starts* relation as: "Event A and event B began simultaneously, but event A ended before event B did." Hence, while the relation is primitive in Allen's calculus, it depends on the composition of several different concepts in natural language: beginnings, endings, and the preposition "before." Despite the disparity between language and logic (see Knauff, 1999), researchers have built a wide variety of tools in artificial intelligence designed to explain what kinds of inferences can be drawn from the way relations between intervals interact (for reviews, see Fischer, Gabbay, & Vila, 2005; Goranko, Montanari, & Sciavicco, 2004).

In contrast to the computational analyses of temporal reasoning, few studies have examined how people reason about durative relations such as "while" and "during." Many studies have examined temporal relations such as "before" and "after" (Clark, 1971; Münte, Schiltz, & Kutas, 1998; Zhang et al., 2012), but durative temporal relations appear to be more complex – children comprehend and produce "while" after they understand the meanings of "before" and "after" (Keller-Cohen, 1981; Silva, 1991; Winskel, 2003). Previous work by Schaeken and colleagues investigated how adults reason about "while" (Schaeken, Johnson-Laird, & d'Ydewalle, 1996) using premises of the form *X happened while Y happened.* However, reasoners could draw inferences

from such relations without considering the durative nature of "while", i.e., the problems in Schaeken et al.'s (1996) studies implied that the two events both started and ended at the same time. Nevertheless, their work revealed two central patterns of temporal reasoning: first, reasoners appear to simulate a mental timeline of events when they reason about time (Bonato, Zorzi, & Umiltà, 2012; Casasanto & Boroditzky, 2008). Second, some temporal reasoning problems are easy, and some are difficult: people are more prone to error and they take longer to complete certain temporal reasoning problems (Baguley & Payne, 2000; Schaeken & Johnson-Laird, 2000; Vandierendonck & De Vooght, 1997).

Though no studies have examined how people reason about durations, many have focused on people's ability to estimate the durations of experienced or anticipated events (Zakay & Block, 1997). In typical tasks, people make estimations in minutes and hours or by using more qualitative boundaries. The research has shown that people overestimate short time periods and underestimate longer ones (Lejeune & Wearden, 2009), a robust pattern known as Vierordt's law. Gennari and Wang (2019) showed that these estimation biases are correlated with the relative amount of represented information per timepoint. People "compress" representations to avoid maintaining a representation of all timepoints over which an event transpires (Faber & Gennari, 2015, p. 157). The lesson for researchers interested in temporal reasoning is that some event representations can be compressed into a single timepoint, and reasoners can construe them as punctate events. Other event representations may resist such compression by requiring reasoners to maintain information about durations, i.e., information that spans two or more timepoints. Of course, even punctate events have some duration, but their duration is irrelevant to how people make inferences from them.

One recent account by Khemlani, Harrison, and Trafton (2015a) sought to explain how reasoners construct a mental timeline to represent durative relations such as "while" and "during" by specifying how time representations can be compressed. The account builds on previous theories of temporal reasoning that assume people build mental simulations that consist of discrete tokens to reason about time (Schaeken & Johnson-Laird, 2000; Schaeken et al., 1996). But Khemlani et al.'s account extends beyond previous research to make predictions about how people carry out different temporal reasoning tasks, such as reasoning about what is necessary, reasoning about what is possible, and assessing the consistency of a set of assertions (Khemlani, Lotstein, Trafton, & Johnson-Laird, 2015b).

In this paper, we spell out the central principles of Khemlani et al.'s (2015a) account of durative reasoning and use it to derive predictions about whether certain reasoning problems should be easy or difficult. We describe a preregistered experiment that tested these predictions. We conclude by describing limitations of the study and why durative inferences pose unique challenges for investigators.

Mental models of durative relations

Khemlani et al.'s (2015a) account of durative reasoning is based on the idea that humans build discrete mental simulations of possibilities – mental models – when they reason (Johnson-Laird, 2006; Johnson-Laird, Girotto, & Legrenzi, 2004). The model theory applies to relational reasoning across several different domains (Goodwin & Johnson-Laird, 2005), including reasoning about space (Ragni & Knauff, 2013; Jahn, Knauff, & Johnson-Laird, 2007), time (Schaeken et al., 1996; Schaeken & Johnson-Laird, 2000), consistency (Jahn, Johnson-Laird, & Knauff, 2004; Johnson-Laird, Girotto, & Legrenzi, 2004), and kinematics (Khemlani, Mackiewicz, Bucciarelli, & Johnson-Laird, 2013). The theory rests on three fundamental assumptions:

- **Models are iconic.** Mental models are discrete, iconic representations of possibilities. Iconicity constrains models so that their structure reflects the structure of what they represent (see Peirce, 1931-1958, Vol. 4). In the case of two or more events, models should be structured to reflect the events' chronology, i.e., the way in which those events unfolded. Since models are discrete, they cannot directly represent how long one event took relative to another. The restriction allows reasoners to efficiently compress temporal models to uniformly represent events that endure across vastly different timescales, such as seconds or decades.
- **Intuition vs. deliberation.** Reasoners rely on two primary processes of inference: an intuitive construction process and a deliberative revision process. The intuitive construction process rapidly builds and scans an initial, preferred mental model (Jahn et al., 2007). The process is subject to various heuristics and biases, and so reasoners who engage just the initial process are prone to make systematic errors (Khemlani & Johnson-Laird, 2017). A slower deliberative process can revise the initial models to search for alternative models and counterexamples to validate and correct any conclusions inferred by the intuitive process.
- **More models, more difficulty.** A final assumption of the theory is that each model that a reasoner builds demands cognitive resources to maintain. Hence, reasoners tend to rely on their preferred models most of the time. If a reasoning problem can be solved successfully from the preferred model, it should be easy: reasoners should be faster and their responses should be more accurate. If, however, a problem demands that reasoners engage in deliberation, they should be slower and less accurate.

We illustrate how the three principles apply to temporal reasoning by contrasting how the model theory treats punctate and durative events. Consider the premises in (3):

3. The meeting happened before the sale.
   The sale happened after the conference.
   The meeting happened before the conference.

The durations of the events in (3) are irrelevant, and so the premises can be represented as punctate events. The mental model representing the premises in (3) can be depicted in the following diagram:

meeting conference sale

The diagram shows an arrangement of events in which time moves from left to right. Only one such arrangement is possible for (3). The model is parsimonious; it can be used to infer many different relations that are not made explicit in the premises:

4a. The conference happened after the meeting.
b. The sale happened after the meeting.
c. The conference happened before the sale.

In contrast, the premises in (5) concern a durative relation:

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

The description is consistent with the following model:

$sale_{START}$ $sale_{END}$
meeting conference

The model represents the durative aspect of the sale as two separate tokens (following Khemlani et al., 2015a): one token marks the sale's beginning and the other marks its end. And the premises in (5) are consistent with at least one other model:

$sale_{START}$ $sale_{END}$
meeting conference

Hence, the premises in (3) are consistent with just one model, while the premises in (5) are consistent with multiple models.

In general, the model theory predicts that people should be less accurate when reasoning about descriptions consistent with multiple models than about those consistent with one model. No other theory of reasoning makes an analogous claim (Khemlani, 2018; Knauff, 1999, p. 286 et seq.). We next describe an experiment that tested and corroborated the prediction.

Experiment

To test whether participants make more errors when reasoning about problems that elicit multiple models, our experiment presented them with one- and multiple-model descriptions of events that consisted of premises that described temporal relations. Their task was to evaluate the consistency of the premises by assessing whether all of them can be true at the same time. Previous studies used similar problems, but they examined how participants deductively inferred relations between two specified events (Schaeken et al., 1996). In daily life, reasoners are seldom provided such constraints, and so our experiment used a task that does not provide participants with any restriction on which premises to consider. The approach also has the advantage of using the same question across all problems, and so it uses a uniform task to test participants' durative deductions.

To balance out participants' responses, half the problems were consistent and half were inconsistent. The theory predicts that people should be more accurate in assessing the consistency of one-model problems than multiple-model problems.

Method

Participants. 50 participants completed the experiment for monetary compensation ($2 and a 10% chance of a $10 bonus) through Amazon Mechanical Turk. All of the participants were native English speakers, and all but 6 had taken one or fewer courses in introductory logic. 5 participants were excluded from the analysis, either because of excessive and inappropriate keypresses, or else because the participant produced irrelevant debriefing responses. The analyses reported below are based on the remaining 45 participants (21 female, mean age = 35.0).

Preregistration and data-availability. The predicted effects were pre-registered through the Open Science Framework platform (https://osf.io/q45mw). The same link makes the data from the study available.

Task and design. Participants carried out 16 different problems. Each problem comprised 3 premises that describe how 3 different events relate to one another. They were asked to judge whether the 3 premises could all be true at the same time. Half the problems concerned descriptions that were designed to yield one-model after the first 2 premises and the other half yielded multiple models after the first 2 premises. And half the problems used a 3rd premise that was consistent with the previous premises, while the rest used a 3rd premise that was inconsistent with the previous premises.

The first premise of each problem was of the form: *X happened during Y*. Hence, the following is an example of a problem designed to yield one model:

6a. X happened during Y. Y_{START} X Y_{END}
b. Y happened before Z. Y_{START} X Y_{END} Z
c. X happened before Z. **Y_{START} X Y_{END} Z**

A compressed model of events is provided next to each premise to show how Khemlani et al.'s (2015a) system would update the represention after interpreting new information. The bolded text shows how the final model would look. The problem presents a consistent description of events, since all three premises can be true at the same time.

In contrast, the set of premises in (7):

7a. X happened during Y. Y_{START} X Y_{END}
b. Z happened before X. Z Y_{START} X Y_{END} (i)
Y_{START} Z X Y_{END} (ii)
c. Y happened during Z. NO MODEL POSSIBLE

corresponds to a multiple-model problem, because the 2nd premise is consistent with at least two different situations: one in which Z happened before Y started (i), and another in which Z happened before X and they both happened during Y (ii). But neither of those possibilities are consistent with the third premise, therefore (7) is an inconsistent multiple-model problem.

The sixteen different problems used in the study are provided in the Appendix. The experiment implemented a 2 (problem type: one- vs. multiple-model) x 2 (consistent vs. inconsistent) fully repeated-measures design.

Materials. The temporal terms in each problem were replaced by descriptions of everyday events, e.g., X was replaced with "the meeting" and Y was replaced with "the snowstorm". The materials were drawn from 16 sets of 3 events. Each set was designed to describe events that endure at comparable timescales, so that any event in the set could take place during any other event, e.g.,

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. Events that elicit strong punctative interpretations, such as "the sneeze," were not used in the study, as they would yield peculiar and unbelievable descriptions, e.g., "The meeting happened during the sneeze." Likewise, events were chosen so that they did not bear causal relations to one another.

Each of the 16 materials was rotated over the designs for each participant. Therefore, across the experiment as a whole, each of the 16 material sets was applied to each of the 16 problems approximately the same number of times. For any given participant, once the materials were assigned to the problems, the order in which the problems appeared was randomized. The counterbalancing scheme eliminated the possibilities that order effects and carry-over effects could account for participants' responses.

Procedure. Participants interacted with the experiment by registering responses through keyboard presses. For each problem, the participants were asked to consider an initial premise, and then pressed the spacebar to reveal each of the remaining premises on the screen. Previously revealed premises remained on the screen whenever the experiment displayed the next premise. The sequential display sought to encourage participants to read the sentences in the order displayed. Once a participant revealed all three premises, a prompt would appear that said: "Can all three of these sentences be true at the same time?" The 'f' and 'j' keys were used to indicate "yes" and "no" responses, respectively. Before taking part in the experiment proper, they completed an example problem and were shown a schematic of how their fingers should be placed on the keyboard. After completing all 16 problems, the participants were asked four open response debriefing questions, which probed their intuitive definitions of "before" and "during" as well as their reasoning strategies.

Results and discussion

Figure 1 plots the proportion of participants' correct assessments of consistency as a function of whether the premises yielded one model or multiple models, and as a function of whether the problem they carried out was consistent or inconsistent. Participants were more accurate for one-model problems than multiple-model problems (78% vs. 69%; Wilcoxon test, $z = 3.02$, $p = .003$, Cliff's $\delta = .43$). The result corroborated the model theory's central prediction that reasoners should find it easier to reason about one-model problems than multiple-model problems. The difference between participants' accuracies did not reliably differ depending on whether the model was consistent or inconsistent (72% vs. 75%; Wilcoxon test, $z = 1.12$, $p = .27$, Cliff's $\delta = .17$). However, the interaction between the problem type (one- vs. multiple-model) and the consistency of the premises was reliable (Wilcoxon test, $z = 4.03$, $p < .0001$, Cliff's $\delta = .42$). The interaction is evident in Figure 1, which shows that consistent problems had a higher accuracy rate when the premises yielded one-model rather than multiple-models. There was little difference by model quantity for inconsistent problems.

To test whether the type of problem is robust to participant and item random effects, we fit a generalized logistic mixed model (GLMM) regression to the data. The fixed effects were the problem type (one- or multiple-model), the consistency of the problem, and their interaction. The random effects components included intercepts and random slopes for all 3 fixed effects by participant. Intercepts also controlled for the items (paired syntax and material sets) and for the pattern of temporal relations in the three premises, i.e., "during/during/during", "during/before/during", etc. The

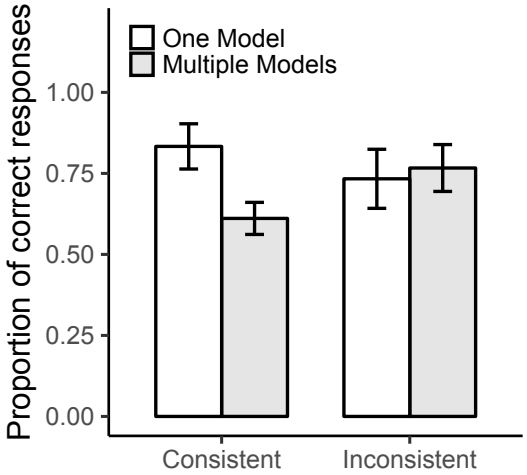


Figure 1. The proportion of correct responses in the experiment as a function of the type of problem (one- or multiple-model) and as a function of whether the premises was consistent or inconsistent. Error bars indicate 95% confidence intervals.

temporal relation pattern was included because a handful of participants reported making judgments based on the pattern alone, and so we treated it as a relevant random effect factor beyond the individual items. The analysis revealed a reliable difference between one- and multiple-model problems when both participants and items were controlled for, $\beta = 2.02$, $z = 3.41$, $p = .0006$, and a reliable interaction between consistency and the problem type ($\beta = 2.35$, $z = 3.21$, $p = .0013$). The GLMM therefore confirmed the nonparametric analyses. However, it further revealed a reliable effect of consistency: participants were more accurate on inconsistent problems than consistent problems ($\beta = 1.82$, $z = 4.34$, $p < .0001$); the model theory did not predict whether there would be an effect of consistency. The results of the regression analysis accordingly support the predictions of the model theory, though the analysis suggests that future studies should examine a broader set of problems to generalize beyond the four in each condition of the present experiment.

General discussion

How do people mentally represent and reason about durative temporal relations, i.e., relations such as "while" and "during"? Such relations describe events that persist across multiple points in time, and many logical frameworks exist that describe ideal temporal reasoning patterns (Fischer et al., 2005; Goranko et al., 2004). But those frameworks do not explain how people represent durations, and so they cannot characterize the mental processes or the strategies people use when reasoning about time. A recent treatment of temporal reasoning explains how people mentally represent durations when they reason. It is based on the idea that people construct mental models, i.e., iconic mental simulations, to draw conclusions from premises or observations (Johnson-Laird, 2006). Models implement a mental timeline, which people use to reason about events (Schaeken et al., 1996). To mentally simulate durative relations, reasoners do not represent all of the time points across which an event might endure. Instead, they construct discrete tokens that stand in place of the beginnings and endings of durative events (Khemlani et al., 2015a).

One way to diagnose the model theory is to investigate how people assess the consistency of durative premises. In principle, the theory should make predictions about people's assessments of consistency. It posits that if they can construct a coherent model of the premises, then those premises are consistent – they can all be true at the same time. If they fail to build a model of the premises, however, then they should consider the premises to be inconsistent (Johnson-Laird et al., 2004). The theory has explained how people reason about the consistency of spatial relations (Jahn et al., 2004), and, given the strong relationship between temporal and spatial metaphors (Casasanto & Boroditsky, 2008; Gentner, 2001), it should also explain how people reason about the consistency of durative relations.

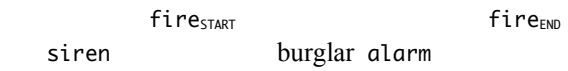
We report an experiment that suggests it does. The experiment presented participants with a description of events as in (8):

8a. The burglar alarm happened during the fire.
b. The siren happened before the burglar alarm.
c. The fire happened during the siren.

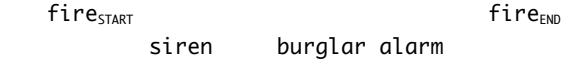
The description is inconsistent, and participants capably judged such descriptions to be inconsistent. In contrast, they had difficulty with descriptions of the following form:

9a. The burglar alarm happened during the fire.
b. The siren happened before the burglar alarm.
c. The siren happened during the fire.

Many reasoners incorrectly evaluated the three assertions as inconsistent. What explains the disparity? The model theory suggests that people rapidly constructed a simulation of the first two statements to yield a model akin to the following:

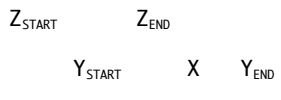


As the theory posits, models cost cognitive resources to maintain, and reasoners are reticent to alter the initial models of the premises they construct. The premise (9c) does not hold in the initial model above, and so reasoners who base their judgments of consistency on it should consider (9) inconsistent. Only reasoners who construct an alternative model of the premises should get the correct answer, as in this model:



In general, the model theory uniquely predicts that participants should find one-model problems easier to reason about than multiple-model problems, and the experiment corroborated the prediction.

One concern of the experiment is that the definition of "during" as an enclosure relation may have been overly restrictive. Consider example (7) above. If people take "during" to refer any two overlapping events instead of an enclosure relation, then the three premises could yield the following model:



Yet, a more permissive notion of "during" does not impact the central outcome of the experiment: if a problem is consistent under the restrictive construal of "during", then it is consistent under the permissive construal as well. Nevertheless, reasoners were least accurate on consistent multiple-model problems – a result that corroborates the theory on any construal of "during."

There are at least three limitations of the experiment we report. First, a small set of participants self-reported that they adopted reasoning strategies based on rapidly assessing the relations in the premises. It is not clear to what extent

participants' strategies attenuated or enhanced the difference in performance on one-model and multiple-model problems, but reasoners can spontaneously discover strategies when reasoning about punctate events (Schaeken & Johnson-Laird, 2000), and so future studies should investigate what kinds of strategies participants develop, and how those strategies promote or inhibit the construction of models. Second, the current design did not explore the nature of participants' errors. It could be that participants attempted to consider alternative models of the premises and failed; or it could be that participants chose not to consider alternative models in the first place. Future studies should explore why multiple-model problems yield systematic errors. Finally, only a limited number of problems could be designed for the study given that they described three relations among three events: hence, the study examined only the small number of configurations possible for three events. Future studies should explore an expanded set of problems. Indeed, the language used to describe durational events goes beyond the preposition "during". The connective "while" has a similar meaning, and both words are in the top 200 most frequent words in American English (Davies, 2008). Other words, e.g., "when", can sometimes be used to situate durative events, and the various ways people describe and discuss events, durative and punctate, can provide insight into how people represent and reason about time.

Temporal reasoning is an essential process that underlies how humans conceptualize time (Hoerl & McCormack, 2019; Kelly, Prabhakar, & Khemlani, 2019). Reasoners routinely make inferences about durations in order to carry out time-dependent tasks, such as picking a friend up at the airport. The model theory provides an explanation of the mental representations people build and processes people use when they think and reason about temporal sequences.

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 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, Phil Johnson-Laird, Joanna Korman, and Greg Trafton for their advice and comments.

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*, 53A, 479-512.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

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

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

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

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

Dierckx, V., Vandierendonck, A., Liefhooge, B., & Christiaens, E. (2004). Plugging a tooth before anaesthetising the patient? The influence of people's beliefs on reasoning about the temporal order of actions. *Thinking & Reasoning*, 10, 371-404.

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

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

Fischer, M., Gabbay, D., & Vila, L. (2004). *Handbook of Temporal Reasoning in Artificial Intelligence.* Elsevier.

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.

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.

Hoerl, C., & McCormack, T. (2019). Thinking in and about time: A dual systems perspective on temporal cognition. Manuscript in press at *Behavioral and Brain Sciences*.

Hothorn, T., Hornik, K., van de Wiel, M. A., Zeileis A. (2008). Implementing a Class of Permutation Tests: The coin Package. Journal of Statistical Software 28(8), 1-23. URL http://www.jstatsoft.org/v28/i08/.

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. Barkowsky (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.

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

Johnson-Laird, P. N., Girotto, V., & Legrenzi, P. (2004). Reasoning from inconsistency to consistency. *Psychological Review*, 11, 640.

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

Kelly, L., Prabhakar, J., & Khemlani, S. (2019). Updating and reasoning: Different processes, different models, different functions. Commentary in press at *Behavioral and Brain Sciences*.

Khemlani, S. (2018). Reasoning. In S. Thompson-Schill (Ed.), *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience*. Wiley & Sons.

Khemlani, S., Harrison, A. M., & Trafton, J. G. (2015). Episodes, events, and models. *Frontiers in Human Neuroscience*, 9, 1-13.

Khemlani, 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*, 110, 16766–16771.

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

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.

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

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

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

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.

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: https://www.orlandosentinel.com/opinion/audience/roger-simmons/os-ae-florida-man-drinking-and-driving-20180711-story.html

Silva, M. (1991). Simultaneity in children's narratives: the case of *when, while,* and *as. Journal of Child Language,* 18, 641-62.

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

Wang, Y., & Gennari, S. P. (2019). How language and event recall can shape memory for time. *Cognitive psychology*, *108*, 1-21.

Winskel, H. (2003). The acquisition of temporal event sequencing: a cross-linguistic study using an elicited imitation task. *First Language*, 23.

Zakay, D., & Block, R. A. (1997). Temporal cognition. *Current Directions in Psychological Science*, *6*, 12-16.

Zheng, Y. et al. (2012). Rearranging the world: Neural network supporting the processing of temporal connectives. *NeuroImage,* 59, 3662-3667.

Appendix. The 16 problems used in the experiment.

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