mReactr: A computational theory of deductive reasoning

Sangeet Khemlani and J. Gregory Trafton
khemlani@aic.nrl.navy.mil, trafton@itd.nrl.navy.mil

Navy Center for Applied Research in Artificial Intelligence
Naval Research Laboratory, Washington, DC 20375 USA

Abstract

The mReactr system is a computational implementation of the mental model theory of reasoning (Johnson-Laird, 1983) that is embedded within the ACT-R cognitive architecture (Anderson, 1990). We show how the memory-handling mechanisms of the architecture can be leveraged to store and handle discrete representations of possibilities, i.e., mental models, efficiently. Namely, the iconic representation of a mental model can be distributed, in which each component of a model is represented by a "chunk" in ACT-R's declarative memory. Those chunks can be merged to create *minimal* mental models, i.e., reduced representations that do not contain redundant information. Minimal models can then be modified and inspected rapidly.

We describe three separate versions of the mReactr software that minimize models at different stages of the system's inferential processes. Only one of the versions provides an acceptable model of data from an immediate inference task. The resulting system suggests that reasoners minimize mental models only when they initiate deliberative mental processes such as a search for alternative models.

Keywords: reasoning, mental models, immediate inferences, mReactr, ACT-R

Introduction

People regularly make complex deductive inferences. For instance, if you know that *none of the lawyers in the room are men,* you might refrain from asking any of the men in the room for legal advice. If so, you have made an "immediate" inference from a single premise:

1. None of the lawyers are men.
   Therefore, none of the men are lawyers.

The inference is *valid* because its conclusion must be true given that its premise is true (Jeffrey, 1981, p. 1). You likely followed up the deductive inference above with an inductive inference:

2. None of the men are lawyers.
   Therefore, they do not possess legal knowledge.

The second inference is inductive – the conclusion is not necessary given the truth of the premise.

How do reasoners make deductive and inductive inferences like the ones above? One prominent answer is that they construct mental simulations of the things they already know or believe. They then manipulate those simulations to obtain information they did not have at the outset. The idea that reasoning depends on building simulations, or mental models, is the fundamental intuition behind the mental model theory of reasoning (Johnson-Laird, 1983). In the present paper, we outline the theory and address one of its major limitations, namely its inability to explain how models are stored and manipulated in memory. We describe a computational implementation of the theory that is embedded within the ACT-R cognitive architecture (Anderson, Bothell, Byrne, Douglass, Lebiere, & Qin, 2004), and we show how the memory-handling mechanisms of the architecture can be leveraged to store and handle mental models efficiently.

Reasoning and mental models

The "model" theory of reasoning proposes that when individuals comprehend discourse, they construct mental models of the possibilities consistent with the meaning of the discourse (Johnson-Laird, 2006). The theory depends on three main principles: 1) Individuals use a representation of the meaning of a premise and their knowledge to construct mental models of the various possibilities to which the premises refer. 2) The structure of a model corresponds to the structure of what it represents (see Peirce, 1931-1958, Vol. 4), and so mental models are iconic insofar as possible. 3) The more models a reasoner has to keep in mind, the harder an inference is. On a model-based account, a conclusion is *necessary* if it holds in all the models of the premises and *possible* if it holds in at least one model of the premises.

mReasoner (Khemlani, Lotstein, & Johnson-Laird, under review) is a unified computational implementation of the mental model theory of reasoning. It implements two interoperating systems for reasoning:

a) An intuitive system (system 1) for building an initial mental model and drawing rapid inferences from that model

b) A deliberative system (system 2) for more powerful recursive processes that search for alternative models. This system can manipulate and update the initial model created in system 1, and it can modify conclusions

The system is akin to dual-process models of reasoning (see, e.g., Evans, 2003, 2007, 2008; Johnson-Laird, 1983, Ch. 6; Kahneman, 2011; Sloman, 1996; Stanovich, 1999; Verschueren, Schaeken, & d'Ydewalle, 2005). Below, we describe the various processes that each system implements.

The intuitive system

Model building. The system builds an initial model from the meaning of a premise, and it updates that initial model if additional premises occur. The system begins by building a model with a small, arbitrary set of individuals. For example, the model of *some of the artists are bohemians* is built by first constructing a set of artists:

artist
artist
artist
artist

In the diagram above, each row represents an individual with the property of being an artist, and so the model as a whole represents a finite number of individuals. Mental models are representations of real individuals, not letters or words, which we use here for convenience. The meaning of the assertion *some of the artists are bohemians* provides instructions for the system to add additional properties, namely the property of being bohemian. The model is updated accordingly:

artist	bohemian
artist	bohemian
artist	
artist	
	bohemian

The model therefore represents a set of individuals, some of whom are both artists and bohemians, some of whom are just artists, and one who is just a bohemian. Once a premise has been represented, the system can assess whether the given conclusion is true in the initial model.

Assessing initial conclusions. When reasoners have to assess a given conclusion, the system inspects the initial model to verify that the given conclusion holds or does not hold. For instance, suppose that reasoners are asked to decide whether it is possible that *all bohemians are artists* given the previous premise. From the model above, the system initially responds in the negative, i.e., the putative conclusion is impossible. The process is simple, and the response is rapid. However, it is incorrect: the system's ability to assess and generate initial conclusions is fallible. For instance, one can indeed show that *all of the bohemians are artists* is possible. To revise its initial conclusion, the system needs to find an alternative model in which both the premise and conclusion hold. We turn to the second system to explain how such a model is found.

The deliberative system

Searching for alternative models. In the preceding section, we focused on how the system assesses conclusions based on an initial model. However, the conclusions it draws can be invalid. System 2 attempts to revise initial conclusions by searching for alternative models. To do so, it uses three separate operations: *adding* properties to individuals, *breaking* one individual into two separate individuals, and *moving* properties from individual to another (see Khemlani, Lotstein, & Johnson-Laird, under review). The operations correspond to those that naïve participants spontaneously adopt when they reason about syllogisms (as evidenced by their manipulations of external models, see Bucciarelli & Johnson-Laird, 1999). Consider our example above. After an individual represents the initial model and provides an initial conclusion that is false, it can modify that conclusion by adding properties to the initial model. If the system can successfully create a model in which *some of the artists are bohemians* and *all of the artists are bohemians* are both true, then it can conclude that it is possible, but not necessary that all of the artists are bohemians. By adding properties, the system finds such a model:

artist	bohemian
artist	bohemian
artist	bohemian
artist	bohemian
artist	bohemian

The new model, which contains individuals who are all artists and bohemians at the same time, refutes the conclusion that it is impossible that *all the bohemians are artists*. However, the search for alternative models places a considerable tax on working memory. Until now, the limitations of the model theory have prevented it from characterizing the cost of holding models in memory.

Limitations of the model theory

The model theory and its unified implementation explain many aspects of how people make inferences. The theory provides an explanation of how discourse is mapped to high-level representations. It accounts for why some reasoning problems are hard and others are easy (Khemlani & Johnson-Laird, 2012). It provides working algorithms for how individuals assess whether a given conclusion is possible, necessary, or consistent with a given set of premises. And the model theory as a whole can explain deductive, inductive, and abductive inferences (Johnson-Laird, 2006). As such, it represents a unified theory of reasoning.

The theory is limited by design, however, in that most of its predictions are qualitative. For instance, it can explain that an inference that requires a reasoner to hold one model in working memory should be easier than an inference that requires three models in memory, but it cannot explain or predict the degree of the difficulty. Is the former inference twice as easy or thrice as easy as the latter? And how long should each inference take? The computational model is silent on these matters, because it specifies only those algorithms that are pertinent to how individuals make inferences. It ignores other aspects of cognition, such as how models are stored in working memory and how they are retrieved. To overcome these limitations, we

implemented the theory in the ACT-R cognitive architecture, and we describe the resulting hybrid system below. The framework, which we call *mReactr* (mReasoner + ACT-R), imbues the model theory with a more robust account of how models are represented and manipulated. It also stands as a novel application of the ACT-R system, which has had only limited success in accounting for behavior on high-level deductive tasks (e.g., Emond, 2003, and Ragni, Fangmeier, & Brüssow, 2010).

mReactr: Mental models in memory

The ACT-R cognitive architecture is a modular computational theory of human cognition (Anderson et al., 2004). It is a collection of interoperating modules that store and retrieve information relevant to a particular task. The central control system, called the *procedural* module, directs the way the system accesses capacity-limited buffers. The system also contains a *declarative* module for storing knowledge of facts and procedures. Facts are stored in structures called *chunks*, and procedures are represented by *productions*, i.e., condition-action pairings. The productions direct the procedural model to monitor the buffers for the existence of certain sorts of chunks, and if a chunk appears in a buffer in the manner that a production expects, the relevant action will be initiated. Each chunk has an associated level of activation. If the chunk's activation is low, ACT-R will take longer to retrieve it, but if it is high, it will be retrieved quickly. Accordingly, the system automatically calculates the time it takes to trigger productions, modify goals, retrieve chunks, and clear buffers.

The architecture efficiently manages chunks in declarative memory. In particular, if it detects that two chunks are identical in every respect, it merges those chunks into one chunk. The merged chunk will then have a higher activation than either individual chunk. This "chunk-merging" feature of the system is particularly important for how mental models are handled.

The mReactr system is an implementation of mental model theory in ACT-R. The system can build initial models and assess putative conclusions (system 1) and likewise it can modify those models to search for alternative models (system 2). It stores models in declarative memory by assigning each individual to a separate chunk. Thus, the system will store the model of *all the artists are bohemians* as five separate chunks:

artist	bohemian	(chunk 1)
artist	bohemian	(chunk 2)
artist		(chunk 3)
artist		(chunk 4)
	bohemian	(chunk 5)

The system therefore represents the model in a distributed fashion, as a collection of chunks with similar properties. However, several of the separate chunks are identical to one another, and so ACT-R will try to merge those chunks automatically, to produce just a condensed version of the model:

artist	bohemian	(chunk 1')
artist		(chunk 3')
	bohemian	(chunk 5')

By merging the chunks, the underlying architecture automatically produces a *minimal* mental model, i.e., a model that only retains information about the different *types* of individuals. The process of minimizing mental models is not something that is built into mental model theory as yet; the basic mechanisms of memory management within ACT-R provide a way to efficiently store and retrieve models. But, is there any evidence that reasoners minimize models? And if so, do they minimize models at the outset, or at a later stage of processing? To answer both of these questions, we compared mReactr's accuracy and latency predictions against data from a recent reasoning experiment.

An assessment of the model

We assessed whether the mReactr system could model that data from a recent study on so-called "immediate" deductive inferences akin to our introductory example above (1). Psychologists have investigated immediate inferences for many years (e.g., Begg & Harris, 1982; Newstead & Griggs, 1983; Wilkins, 1928), but have yet to resolve how logically untrained individuals make them. The inferences are based on singly-quantified assertions in four different *moods* of the premise:

All the Xs are Ys
Some of the Xs are Ys
None of the Xs are Ys
Some of the Xs are not Ys

and 8 different sorts of conclusion (4 moods by 2 *figures*, i.e., arrangements of terms 'X' and 'Y'). Therefore, there are 32 possible immediate inference problems based on these premises. A typical problem looks like this:

Suppose that some of the students are Virginians.
Is it possible that all of the Virginians are students?

Immediate inferences were chosen because the model theory and mReactr distinguish between the relative difficulties of three sorts of immediate inference: a) zero-model inferences, b) one-model inferences, and c) multiple model inferences.

Zero-model inferences are those in which the conclusion is identical to the premise, and so individuals needn't even build a model to be able to solve the problem. For instance, consider the following problem:

All the aldermen are barters.
Is it possible that all the aldermen are barters?

Reasoners should realize that the answer is true immediately; however, they should nevertheless need to extract the meanings from the assertions, and they need to establish a set of subgoals in order to infer a conclusion.

One-model inferences are those in which the conclusion holds in the initial model of the premise, and so individuals can rapidly determine that an assertion is possible. For example:

All the aldermen are barters.
Is it possible that some of the barters are aldermen?

Reasoners have to construct the meanings of the assertions, use them to build a model, and evaluate the truth of the conclusion in the model.

When the conclusion fails to hold in the initial model, but does hold in an alternative to it, then participants have to search for that alternative model. We refer to such problems as multiple-model inferences. For instance:

All of the aldermen are barters.
Is it possible that some of the barters are not aldermen?

For multiple-model inferences, mReactr predicts that reasoners extract the meaning of the assertion and build an initial model, but their initial model suggests an erroneous evaluation of whether or not the conclusion is possible. To obtain a correct evaluation, reasoners have to modify their initial model to produce an alternative model. The theory therefore predicts that zero-model inferences should be easier than one-model inferences, and one-model inferences should be easier than multiple-model inferences. Likewise, mReactr provides precise latency predictions for how long zero-, one-, and multiple-model inferences should take.

We used mReactr to simulate an experiment conducted by Khemlani, Lotstein, & Johnson-Laird (in revision). In the study, the participants carried out all 32 problems (4 sorts of premise x 8 sorts of conclusion), and they responded "yes" or "no" to a conclusion about a possible conclusion to each problem. The contents of the problems were based on nouns referring to common occupations. The instructions stated that the task was to respond to questions about a series of assertions concerning what was possible given the truth of the assertion.

Simulation

Our goals in simulating immediate inference data were two-fold. First, we sought to test the fidelity of the mReactr system as an instantiation of the model theory. We restricted our simulation to valid immediate inferences, i.e., 22 of the 32 problems. The theory distinguishes between three sorts of problem, and so mReactr should reflect the same distinction. A failure of the computational model to capture those data indicates a poor implementation of the model theory. We retained all of the default values of the ACT-R architecture, except we increased the architecture's default tracking ability so that it could track 10 individual chunks (i.e., the `:declarative-num-finsts` parameter).

Second, we attempted to examine whether mReactr could fit the data better when it actively engaged in minimizing models by merging chunks. We created three separate versions of mReactr:

1) no chunk-merging version
2) system 1 chunk-merging version
3) system 2 chunk-merging version

In the *no chunk-merging* version, chunks were kept separate and ACT-R's automated chunk-merging capability was disabled. In the *system 1 chunk-merging* version, chunks were merged before the system engaged in any inferential processes. And in the *system 2 chunk-merging* version, chunks were kept separate in the initial model. They were merged only when mReactr initiated a search for alternative models. The best performing version of the theory can help establish whether and when models should be minimized.

Results and discussion

The results of the experiment corroborated the theory's predictions of difficulty (Khemlani et al., in revision), and they yielded the following trend: reasoners were 98% correct for zero-model problems, 85% correct for one-model problems, and 71% correct for multiple-model problems (Page's trend test, $L = 340.0$, $z = 3.88$, $p < .0001$). Immediate inferences are relatively easy to deduce, nevertheless participants exhibit predictable patterns of difficulty. The mean latencies also corroborated the predicted trend: 4.30 s for zero-model problems, 5.17 s for one-model problems, and 5.41 s for multiple-model problems (Page's trend test, $L = 336.0$, $z = 3.33$, $p < .0005$).

Figure 1 illustrates the empirical latencies and the predicted latencies from the different versions of mReactr. As the figure shows, the system yielded the closest match to

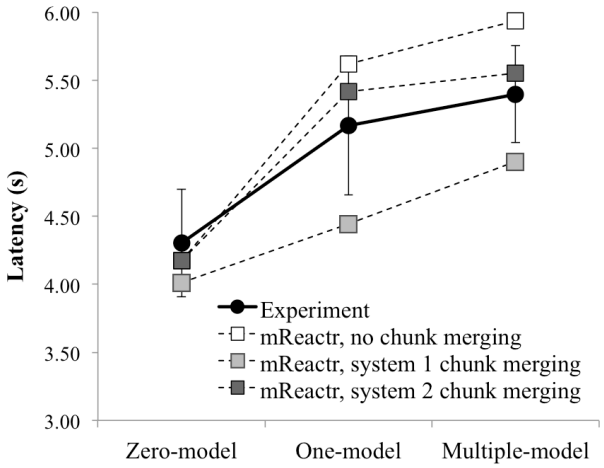

Figure 1: Participants' mean latencies (in s) to solve zero-, one-, and multiple-model problems, and the latencies predicted by the three separate versions of mReactr.

	Model fits			
			Goodness of fit	
mReactr version	R^2	RMSE	D	p
a) By problem type				
No chunk-merging	.99	.40	.67	.60
System 1 chunk-merging	.94	.54	.67	.60
System 2 chunk-merging	.99	.18	.67	.60
b) By immediate inference				
No chunk-merging	.45	.70	.41	.05
System 1 chunk-merging	.23	.86	.50	.008
System 2 chunk-merging	.45	.57	.18	.86

Table 1: Model fits for the three versions of mReactr by problem type (zero-, one-, and multiple-model problems) and by the 22 valid immediate inferences. Note: a lack of significance for the Kolmogorov-Smirnov D statistic indicates a good fit.

the data when chunk-merging was initiated at a later stage of processing, i.e., the *system 2 chunk-merging* version (R^2 = .99, RMSE = .18). When chunk-merging was disabled in the *no chunk-merging* version, the system did well, but it took too long to search for alternative models, (R^2 = .99, RMSE = .40). In the *system 1 chunk-merging* version, mReactr performed faster than participants tend to perform, yielding a poorer fit of the data (R^2 = .94, RMSE = .54).

Across all three simulations, the system negatively correlated with participants' accuracy (r's < -.90, p's < .0001). Likewise, the simulations fit the latencies well. Table 1a gives the model fits for the three separate versions of the system across the three types of problems as a whole, as well as across the 22 different problems separately.

We ran a separate set of analyses to examine how the three versions of the system modeled the 22 valid immediate inferences separately (see Table 1b). This set of analyses would by definition yield poorer model fits as a result of the inherent variation between different problems, and so any significant correlation can be construed as support for the theory. The analysis replicated and elaborated upon the aggregated results. The system fit the data moderately well with chunk-merging turned off, but its RMSE was relatively high (R^2 = .45, RMSE = .70), and a Kolmogorov-Smirnov goodness of fit analysis indicated that the system exhibited reliably different distributional properties than that of the experiment (D = .41, p = .05). Likewise, the system provided a relatively poor fit of the data when models were minimized at the outset (R^2 = .23, RMSE = .86) and so mReactr produced results that came from a separate distribution (Kolmogorov-Smirnov test, D = .50, p = .008). Only when models were minimized before the system searched for alternative models did the system fit the data well (R^2 = .45, RMSE = .57), and the goodness-of-fit analysis indicated a close match between mReactr and the data (Kolmogorov-Smirnov test, D = .18, p = .86).

The results of the simulations showed that across all three version of mReactr, the system successfully implemented the model theory's predictions of difficulty, and it distinguished between zero-, one-, and multiple-model problems. However, the system performed best only when it initiated chunk-merging before it began a search for alternative models. The results have important implications for an overlooked process in the psychology of reasoning: representational minimization.

General Discussion

The computational theory, mReactr, is system implemented in the ACT-R cognitive architecture that simulates the construction of mental models in order to draw immediate inferences from singly-quantified premises. The cognitive architecture comes equipped with the ability to manage its declarative memory efficiently, namely by merging identical chunks. mReactr repurposes this chunk-merging functionality to produce minimal mental models at a particular stage of inference. At the outset, mReactr uses the same collection of iconic representations as is specified in the model theory. However, the full representation is ephemeral, and it lasts only until the system starts to modify the model. If and until the system initiates a search for alternative models, it minimizes the model. This process maps onto the psychological strategy of abstracting over the different sorts of individuals.

The theory predicts that individuals should be faster and more accurate when an inference can be drawn from an identity in the meanings of the assertions, i.e., when they do not need to consult a mental model. They should be next fastest and accurate when an inference can be drawn from the initial model constructed in system 1. And they should be slowest and least accurate when an inference can be drawn only from the discovery of an alternative model constructed in system 2. These rank-order predictions were borne out in the data from an experiment that tested all 22 valid inferences about possible conclusions in the set of 32 inferences.

The system we describe is limited, however, and it can be improved to yield a more fine-grained processing account of the data. We suggest two separate ways of proceeding. One way to improve the fit of the system is to make the system sensitive to the direction in which it scans models. For instance, if reasoners read a particular premise, e.g., *all artists are bohemians*, they may be biased to scan the model in the opposite directions by considering bohemians before artists. This *figural* bias is widely documented in syllogistic reasoning (Khemlani & Johnson-Laird, 2012) and it is likely to make a difference when reasoning about immediate inferences as well.

Another way to improve the system's overall performance is to consider the process of model minimization as something that may or may not occur depending on strategy and individual differences (Bucciarelli & Johnson-Laird, 1999). Some reasoners may be more likely to minimize their models, and others might prefer to keep the full model representation in mind.

In sum, model minimization is an important way in which individuals can optimize the storage and retrieval of mental models. It is embodied in the computational system of deductive reasoning that we developed.

Acknowledgments

This research was funded by a National Research Council Research Associateship awarded to SK and ONR Grant #s N0001412WX30002 and N0001411WX20516 awarded to JGT. We are also grateful to Len Breslow, Magda Bugajska, Hua Gao, Tony Harrison, Cathy Haught, Laura Hiatt, Phil Johnson-Laird, Gorka Navarrete, Marco Ragni, and Tobias Sonntag for their helpful comments.

References

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y . (2004). An integrated theory of the mind. *Psychological Review 111,* 1036-1060.

Begg, I., & Harris, G. (1982). On the interpretation of syllogisms. *Journal of Verbal Learning and Verbal Behavior*, *21*, 595–620.

Bucciarelli, M., & Johnson-Laird, P. N. (1999). Strategies in syllogistic reasoning. *Cognitive Science, 23,* 247-303.

Emond, B. (2003). Cognitive representations and processes in syllogistic reasoning: existential graphs and cognitive modelling. *Psychologica*, 32, 311-340.

Evans, J.St.B.T. (2003). In two minds: Dual process accounts of reasoning. *Trends in Cognitive Sciences*, *7*, 454–459.

Evans, J. St. B. T. (2007). *Hypothetical thinking: Dual processes in reasoning and judgement* . Hove, UK: Psychology Press.

Evans, J. St. B. T. (2008). Dual-processing accounts of reasoning, judgment and social cognition. *Annual Review of Psychology, 59,* 255-278.

Jeffrey, R. (1981). *Formal logic: Its scope and limits* (2nd Ed). New York: McGraw-Hill.

Johnson-Laird. P.N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Cambridge: Cambridge University Press. Cambridge, MA: Harvard University Press.

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

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

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

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

Khemlani, S., Lotstein, M., & Johnson-Laird, P.N. (in revision). Immediate inferences from quantified assertions. Manuscript in revision.

Khemlani, S., Lotstein, M., & Johnson-Laird, P.N. (under review). A unified theory of syllogistic reasoning. Manuscript under submission.

Newstead, S.E., & Griggs, R.A. (1983). Drawing inferences from quantified statements: A study of the square of opposition. *Journal of Verbal Learning and Verbal Behavior*, *22*, 535–546.

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

Ragni, M., Fangmeier, T., & Brüssow, S. (2010). Deductive spatial reasoning: From neurological evidence to a cognitive model. In D. D. Salvucci & G. Gunzelmann (Eds.), *Proceedings of the 10th International Conference on Cognitive Modeling* (pp. 193-198). Philadelphia, PA: Drexel University.

Sloman, S.A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*, 3–22.

Stanovich, K.E. (1999). *Who is rational? Studies of individual differences in reasoning*. Mahwah, NJ: Erlbaum.

Verschueren, N., Schaeken, W., & d'Ydewalle, G. (2005). A dual-process specification of causal conditional reasoning. *Thinking & Reasoning*, *11*, 278–293.

Wilkins, M. C. (1928). The effect of changed material on the ability to do formal syllogistic reasoning. *Archives of Psychology*, *16*, No. 102.