Individuals become more logical without feedback

Marco Ragni (ragni@cs.uni-freiburg.de)
Cognitive Computation Lab, Technical Faculty, University of Freiburg, Germany

Nicolas Riesterer (riestern@cs.uni-freiburg.de)
Cognitive Computation Lab, Technical Faculty, University of Freiburg, Germany

Sangeet Khemlani (sangeet.khemlani@nrl.navy.mil)
US Naval Research Laboratory, Washington, DC 20375 USA

P. N. Johnson-Laird (phil@princeton.edu)
Princeton University, Princeton NJ 08540, USA
New York University, New York, NY 10003, USA

Abstract

Many theories of reasoning and many experiments presuppose that human ability is stable over time, and so people usually draw the same conclusion from the same premises. The assumption has hitherto had little or no empirical investigation. We therefore analyzed a study in which 20 participants drew their own conclusions to the 64 sorts of syllogisms on *two occasions* separated by roughly a week. We report the nature of the changes in the participants' conclusions including their spontaneous improvement in logical accuracy, and use a model-based program, mReasoner, to explain the results.

Keywords: cognitive stability; logical improvements; mental models; reasoning; syllogisms

Introduction

Logically naive individuals are able to make valid inferences. How they do so is highly controversial. Some theories postulate that they use formal rules of inference akin to those of logic (e.g., Rips, 1994); some theories postulate that they manipulate probabilities (e.g., Chater & Oaksford, 1999); and some theories postulate that they derive conclusions from mental models of the premises (e.g., Johnson-Laird, 2006). But, theories often take for granted that the inferential mechanism is stable, and so in the absence of variations in an inference, individuals usually draw the same conclusion to the same premises. Some studies have examined the development of reasoning strategies over sets of similar inferences (e.g., Bucciarelli & Johnson-Laird, 1999; Lane, Fletcher, & Fletcher, 1983; O'Brien & Overton, 1982; Van der Henst, Yang, & Johnson-Laird, 2002). But, our concern is how individuals cope with identical inferences when they encounter them for a second time. Theories of reasoning need to account for systematic changes in conclusions that cannot be attributed to external factors, such as changes in the premises, the instructions, or the framing of the inference; or else to internal factors, such as noise. We analyzed the stability of reasoning for Aristotelian syllogisms, such as:

(EA1) None of the artists is a beekeeper.
All the beekeepers are chemists.
What, if anything, follows?

Most people infer: *none of the artists is a chemist.* But the inference is invalid. In contrast, a weaker conclusion is valid: *some of the chemists are not artists*, which does not imply its converse.

In general, syllogisms have two premises and a conclusion, which each contain a single quantifier, e.g., "All the artists". Each of these assertions is in one of four *moods*, which we present here with their traditional abbreviations:

A: All the A are B.
I: Some of the A are B.
E: None of the A is a B.
O: Some of the A are not B.

There are four different figures, which are arrangements of the terms, A, B, and C, in the premises (numbered as in Khemlani and Johnson-Laird, 2012):

Figure 1	Figure 2	Figure 3	Figure 4
A-B	B-A	A-B	B-A
B-C	C-B	C-B	B-C

The example syllogism provided earlier contains premises that can be abbreviated as: Eab (none of the artists is a beekeeper) and Abc (all of the beekeepers are chemists). It is in the first figure, and so the syllogism is abbreviated as: EA1. A valid conclusion to EA1 can be abbreviated as: Oca (some of the chemists are not artists).

At least 12 psychological theories of syllogistic reasoning exist. A meta-analysis was feasible for seven of them – the authors of the remaining theories explained that their theories were unsuited to such an analysis (Khemlani & Johnson-Laird, 2012). The results showed that what best fit the data were the verbal-models theory (Polk & Newell, 1995) and the theory of illicit conversions of premises (Chapman & Chapman, 1959; Revlis, 1975). The remaining theories attained only a mediocre level of accuracy (e.g., Chater & Oaksford, 1999; Rips, 1994). But, the subsequent development of a computer program, mReasoner, based on the theory of mental models, outperformed all the theories, and had an overall accuracy of just under 90% in the meta-analysis (Khemlani & Johnson-Laird, 2013).

One of the first studies to examine all 64 possible pairs of syllogistic premises tested 20 participants. They drew their

own conclusions to each syllogism, twice within the course of about a week. And they did not know that they would return to the laboratory to be retested until the experimenter invited them to do so. In the first session, none of the participants drew a valid conclusion to the example above – they all inferred either Eac, or Eca – but on their second test, three participants correctly inferred Oca.

Our aim in what follows is three-fold. First, we outline the theory of mental models that underlies the mReasoner program. Second, we analyze the results of the experiment on the stability of syllogistic reasoning. And, third, we show how the model theory and its mReasoner implementation elucidate changes in the conclusions that the participants inferred.

Models and the mReasoner program

The theory of mental models – the "model" theory for short – was first developed in order to explain syllogistic reasoning (Johnson-Laird, 1983). The modern version of the theory postulates that individuals use their linguistic knowledge to construct models of premises, and this account is implemented in the mReasoner program (Khemlani & Johnson-Laird, 2013).

mReasoner implements two inferential pipelines: an intuitive pipeline (called "System 1" after Stanovich (1999)), which builds an initial model from the premises, and a deliberative pipeline (called "System 2"), which can construct additional models and can revise its initial models.

To illustrate how the two systems operate, consider premises of the sort:

(EA1) None of the A is a B.
All the B are C.

The model theory's System 1, which yields intuitive inferences, constructs the following sort of model of these premises, where each row represents an individual, and '¬' represents negation:

A	¬ B	
A	¬ B	
	B	C
	B	C

The first two individuals in this model accordingly have the properties of A and not-B. The model yields the conclusion:

(Eac) None of the A is a C

or its converse:

(Eca) None of the C is an A.

But, as we noted earlier, neither of these conclusions is valid.

The theory's System 2, which handles deliberations, searches for counterexamples to putative conclusions using various ways to modify models. It finds the following alternative to the model above by adding properties to it:

A	¬ B	C
A	¬ B	C
	B	C
	B	C

Both premises remain true in this model, but the only conclusion that holds for both models is:

Some of the C are not A.

This conclusion is valid. The inference should be difficult in comparison to syllogisms that yield initial conclusions that are not susceptible to counterexamples, e.g.:

(IA1) Some of the A are B.
All the B are C.
What follows?

The initial model of these premises yields the conclusion:

Some of the A are C.

And there is no counterexample to this conclusion. Perhaps not surprisingly, even seven year-old children can draw valid conclusions from such syllogistic premises.

In general, syllogistic premises can always be represented in a single model. If this model yields a valid conclusion, then the inference is easy. Of the 64 pairs of syllogistic premises, 27 yield valid conclusions, but only 10 of them yield valid conclusions from their initial models. The remainder call for a search for counterexamples, and hence are more difficult to reason about – they take longer, and people make systematic errors when they draw conclusions from them.

The implementation of the theory in the mReasoner[1] program introduces four parameters governing its performance:

1. The λ parameter controls the size of an initial mental model, i.e., the maximum number of entities it represents. It bases this number on a sample drawn from a Poisson distribution of parameter λ.

2. An ε parameter (from 0 to 1) is the probability of choosing instances for the model from all the possibilities as opposed to those that are typical, e.g., for *All the A are B* a typical instance is an *A* that is a *B*, whereas the complete possibilities include a *B* that is $\neg A$.

3. A σ parameter (from 0 to 1) is the probability of searching for a counterexample (i.e., engaging System 2).

4. If the program searches for a counterexample, then the parameter ω is the probability that the system weakens the conclusion, as opposed to responding that no valid conclusion follows. Only weakening can lead to a valid conclusion for certain syllogisms (as in the initial example above).

We show presently how the program fits the results of an experiment, to whose description we now turn.

The experimental data

We outline the original Experiment 2 in Johnson-Laird and Steedman (1978). The participants were 20 undergraduates from Columbia University who were tested individually in the experiment (carried out under the auspices of the late Professor Janellen Huttenlocher). They received all 64

[1] Source code at `http://mentalmodels.princeton.edu/models/`

pairs of syllogistic premises in different random orders, and had to create their own spontaneous conclusions to each of them. Every participant performed this task twice about a week apart, but they had no inkling of the second test until they were called for it. The contents of the premises were one name of an occupation and two names of vocations, e.g.: storekeeper, bowler, gourmet. They were timed and they were told to be both accurate and as quick as possible. They were also told that their answers should be based only on what could be deduced with absolute certainty from the premises, and that they should restrict their answers to one of the four moods interrelating the two terms occurring in separate premises, or else they should state that no such valid conclusion (NVC) followed from the premises. The principal result was the corroboration of the difficulty of syllogisms depending on the number of models that an inference requires and on the relation between the order of the terms in a valid conclusion and the figure of the premises. We now turn to an analysis of the changes from the first test to the second.

The stability of human reasoning

By far the most important result was that the participants improved in their reasoning from one test to the next. They spontaneously increased from 58% to 68% logically correct responses from the first to the second test (19 out of the 20 participants improved, Binomial test, $p < .0001$), and the percentages were very similar both for valid syllogisms and for NVC syllogisms. Because the participants had not known that they would be tested twice, the improvement must reflect the experience that they gained over the task of syllogistic reasoning. Figure 1 presents the number of changes that each of the 20 participants made from one test to the other. The mean was 27.5 changes, which was significant in comparison to no changes (Wilcoxon test, $p < .0001$). As the figure shows, the participants varied from 15 changes (in 64 syllogisms) to 39 changes with a $SD = 7.25$. The participants differed reliably one from another in how many changes they made (a resampling test of the observed SD had $p < .005$). So, not only is human reasoning unstable, but it also tends to improve, and its instability varies reliably from one person to another.

Table 1 presents the frequencies of transitions in the moods of the participants' conclusions. These transitions show that changes to conclusions were highly sensitive to the polarity of conclusions, i.e., whether they were affirmative or negative: 95.9% of changes were to the same polarity and only 4.1% of changes were to a different polarity (Wilcoxon test, $z = 7.11$, $p < .001$). The changes in polarity were cases in which I conclusions changed to O or E conclusions.

The 64 syllogisms differed in the number of changes to conclusions that occurred from the first test to the second test, and the difficulty of an inference according to the mReasoner program predicted the tendency for changes to occur. Syllogisms that support a valid conclusion from their initial model yielded only 52% changes from the first test to the second, whereas those that support a valid conclusion only from a search for counterexamples yielded 79% changes from the first test to the second (Wilcoxon test, $z = 3.88$, $p < .00001$).

Table 1: The frequencies of the transitions of the mood of the conclusions in the first test to the mood of the conclusions in the second test, where A denotes "All the _ are _", I denotes "Some of the _ are _", E denotes "None of the _ is a _", O denotes "Some of the _ are not _," NVC denotes "No valid conclusion", and the last column denotes miscellaneous errors.

		Week 2					
		A	I	E	O	NVC	Misc.
Week 1	A	36	10	0	0	2	0
	I	6	123	8	13	51	1
	E	0	0	100	43	37	2
	O	0	0	21	151	89	2
	NVC	3	25	25	90	407	10
	Misc.	1	3	5	6	10	0

Method and procedure

To simulate the participants performance on Week 1 and Week 2, mReasoner generated simulated datasets for every possible combination of quantized settings of its four parameters. For each unique parameter setting, the system generated a dataset in which it carried out 64 syllogisms 100 times. The parameter settings were quantized to span their ranges as follows:

- λ (size): 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0
- ε (canonicality): .0, .2, .4, .6, .8, 1.0
- σ (counterexample search): .0, .2, .4, .6, .8, 1.0
- ω (weakening conclusions): .0, .2, .4, .6, .8, 1.0

Hence, the system generated $7 \times 6 \times 6 \times 6 = 1512$ separate simulated datasets. A grid search located the best fitting parameter settings for Week 1 and the best fitting parameter settings for Week 2. mReasoner was set to these parameter settings, and then used to generate two synthetic datasets of 1000 simulated participants, one per week. Those datasets were analyzed against the data from their corresponding weeks to assess a fit of the computer model's performance against the data.

The computational model captured participants' performance on the two weeks well. For both weeks, the responses generated by mReasoner were highly correlated with the individual participants' performance ($r = .84$ for Week 1; $r = .85$ for Week 2).

The model was further applied to simulate the 20 individual participants' performance on each week. Table 3 presents the optimal settings of the parameters to model each participants conclusions in Week 1 and in Week 2.

Table 2: The settings of the four parameters in mReasoner to fit three subgroups of individual reasoners

	Settings of the four parameters to model the three sorts of individual reasoner			
Subgroup of Ss	λ: Size of model	ε: Less typical entities	σ: search for counterexample	ω: weakening conclusion
Intuitive	2.0	0.0	0.4	0.6
Intuitive + deliberative	3.0	0.6	0.8	0.6
Deliberative	2.0	0.0	1.0	0.8

The fit of the model theory

Most theories of reasoning have been developed to account for data from a group of participants, and so not all theories can be adapted to predict the inferences that an individual makes. A previous study of the participants in our test experiment could be explained using mReasoner (Khemlani & Johnson-Laird, 2016). A preliminary cluster analysis of the participants yielded three main subsets of them: individuals who relied on intuition, on a mixture of intuition and deliberation, or on deliberation. Table 2 presents the parameter settings for mReasoner that best modeled these three groups. The settings make sense in characterizing the three groups performance, and provided a good fit with their conclusions (with values of *r* equal to .74, .82, and .9, respectively).

Table 3 presents the optimal settings of the parameters to model each participants conclusions in the first test and conclusions in the second test. mReasoner fits 60% of the participants' inferences in the first test and 65% of their inferences in the second test. The best-fitting parameter values for the two tests were reliably correlated (Kendall's coefficient of concordance, $W = .87$, $p < .03$). So, if mReasoner predicts an individual reasoners performance in one test, then it does so for the other test.

The tendency for changes to be in the same polarity follows at once from mReasoners procedures for weakening putative conclusions. Likewise, according to mReasoner, the principal reasons for an improvement in accuracy is that reasoners have become more likely to search for a counterexample and more likely, when they find one, to weaken their conclusion. These predictions are corroborated in the increase in the value of σ, the probability of a search for counterexamples, from .66 in modeling the first test to .75 in modeling the second test (Wilcoxon test, $z = 2.76$, $p < .003$, one tail), and, given such a search, an increase in ω, the probability of weakening the conclusion, from .52 in modeling the first test to .7 in modeling the second test (Wilcoxon test, $z = 2.45$, $p < .01$, one tail). Neither of the other two parameters changed their values reliably from one test to the next.

General discussion

A common but tacit assumption is that the human reasoning system is deterministic but noisy. The conclusions that an individual reasoner draws therefore tend to be stable apart from some slight "jitter". As a consequence, theories of reasoning often neither allow for alternative inferences nor provide an inbuilt mechanism for improving the accuracy of inferences. In contrast, the model theory derives from a theoretical tradition embracing non-determinism from its outset. In one of the first algorithmic accounts of high-level cognition – an explanation of how individuals select evidence to test hypotheses – the algorithm allowed for intuitive selections, deliberative selections, and those that combined elements of both (see Johnson-Laird and Wason (1970); and Ragni, Kola, and Johnson-Laird (in press) for a meta-analysis of this theory's fit to the results of over 200 experiments). Likewise, the model theorys account of syllogistic reasoning postulates that individuals use intuition, a mixture of intuition and deliberation, or pure deliberation. These two systems are implemented in the mReasoner program.

Intuition calls only for the premises to be represented in a model, which can sometimes yield a valid conclusion. Those premises with this property yield easy deductions. Most syllogisms, however, lack this property. They call for a search for counterexamples. Individual who combine intuition with deliberation may use a counterexample to infer that nothing follows from the premises. In certain cases, however, syllogisms depend not only on the discovery of a counterexample, but also on weakening an initial conclusion to one that accommodates the alternative model. This exercise in deliberation can, in principle, yield a valid conclusion to any syllogism. The mReasoner program embodies these processes, but unknown factors determine an individuals level of performance.

Several questions remain open. No data are available about the order in which the participants dealt with the syllogisms in the two sessions in the Johnson-Laird and Steedman (1978) study. Likewise, we cannot be certain that none of the participants studied syllogistic reasoning between the two sessions. But, the fact that all but one of them improved in performance suggests that such interventions are unlikely to explain the improvement. It results from a shift from intuition to a greater reliance on deliberation, or from "credulous" to "skeptical" reasoning (Stenning & Cox, 2006). Current theories of reasoning tend to predict only general patterns of inference, and have nothing to say the performance of individuals. In contrast, the mReasoner program fitted the results of 20 individual reasoners, who drew their own conclusions to the 64 pairs of syllogistic premises in two separate sessions. They spontaneously changed their minds about many inferences, and in general improved in their ability: they were more likely to search for a counterexample the second time around.

Over the course of drawing conclusions to 64 syllogisms, individuals are likely to realize that a conclusion drawn from an initial, and intuitive, model, can be fallible. For example, given the premises that none of the gourmets is a beekeeper,

Table 3: The parameter settings of mReasoners best-fitting simulations for each participants performance in the first test and the second test.

SubjectID	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16	S17	S18	S19	S20
Fit (week1)	0.52	0.53	0.45	0.50	0.81	0.70	0.57	0.78	0.70	0.56	0.73	0.63	0.64	0.58	0.58	0.43	0.55	0.43	0.83	0.49
Fit (week2)	0.63	0.70	0.56	0.50	0.80	0.80	0.51	0.75	0.78	0.69	0.75	0.53	0.64	0.70	0.64	0.54	0.57	0.57	0.76	0.57
λ_1	2.00	2.00	3.00	2.00	2.00	2.00	3.50	2.00	2.00	3.00	3.00	2.00	3.50	4.00	2.00	3.00	2.00	2.00	2.00	2.00
λ_2	2.50	2.00	2.00	2.50	2.00	3.00	2.00	2.00	4.00	2.00	2.00	2.00	3.50	3.00	2.00	3.50	2.00	4.00	4.00	3.50
ε_1	0.40	0.00	0.60	0.40	0.00	0.00	0.00	1.00	0.00	0.60	0.40	0.20	1.00	0.40	0.40	0.80	0.20	0.00	0.20	0.00
ε_2	0.00	0.60	0.40	0.00	0.00	1.00	0.00	1.00	0.60	0.20	0.20	0.00	1.00	0.40	0.60	0.60	0.80	0.00	0.60	0.40
σ_1	0.60	0.60	0.40	0.40	1.00	0.80	0.60	0.80	0.80	0.80	0.80	0.60	0.60	0.60	0.60	0.40	0.60	0.40	1.00	0.80
σ_2	0.60	0.80	0.60	0.60	1.00	1.00	0.80	0.80	0.80	0.80	0.80	0.60	0.60	0.80	0.80	0.40	0.60	0.60	1.00	1.00
ω_1	0.60	0.60	0.60	1.00	0.80	0.40	0.20	0.40	0.60	0.60	0.80	0.60	0.60	0.20	0.60	1.00	0.20	0.00	0.40	0.20
ω_2	1.00	0.60	1.00	1.00	0.80	0.60	0.60	0.80	0.80	0.80	1.00	1.00	0.60	0.20	0.60	0.20	1.00	0.20	0.60	0.60

and all the beekeepers are French, they may at first infer that none of the gourmets is French. But, on reflection, the conclusion may strike them as implausible, and so they examine more carefully whether it follows. Indeed, they may then discover that the premises are compatible with some, or even all, of the gourmets being French. What continues to hold, however, is that some of the French are not gourmets. The manipulation of models with the aim of refuting conclusions could, in principle, lead to insights into the need to search for counterexamples, and the need to examine whether a weaker conclusion follows from such a counterexample (Ragni, Khemlani, & Johnson-Laird, 2014). In sum, human reasoning is not stable. It does not necessarily draw the same conclusion to the same premises, but instead can spontaneously make more accurate inferences. Of course, the degree to which people improve varies reliably from one person to another, just as the ability to reason does.

Acknowledgments

This research has been supported in part by a Heisenberg grant to the first author (RA 1934/3-1 and RA 1934/4-1) and is supported by a project in the SPP "New Frameworks of Rationality" (RA 1934/2-1).

References

Bucciarelli, M., & Johnson-Laird, P. N. (1999). Strategies in syllogistic reasoning. *Cognitive Science: A Multidisciplinary Journal*, *23*(3), 247-303.

Chapman, L. J., & Chapman, J. P. (1959). Atmosphere effect re-examined. *Journal of Experimental Psychology*, *58*(3), 220–226.

Chater, N., & Oaksford, M. (1999). The Probability Heuristics Model of Syllogistic Reasoning. *Cognitive psychology*, *258*, 191–258.

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

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

Johnson-Laird, P. N., & Steedman, M. (1978). The psychology of syllogisms. *Cognitive psychology*, *10*(1), 64–99.

Johnson-Laird, P. N., & Wason, P. C. (1970). A theoretical analysis of insight into a reasoning task. *Cognitive Psychology*, *1*(2), 134–148.

Khemlani, S., & Johnson-Laird, P. N. (2012). Theories of the syllogism: A meta-analysis. *Psychological bulletin*, *138*(3), 427–57.

Khemlani, S., & Johnson-Laird, P. N. (2013). The processes of inference. *Argument & Computation*, *4*(1), 4–20.

Khemlani, S., & Johnson-Laird, P. N. (2016). How people differ in syllogistic reasoning. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society.*

Lane, D. S., Fletcher, D. N., & Fletcher, H. J. (1983). Improving conditional syllogism performance of young normal and gifted students with discovery and rule instruction. *Journal of Educational Psychology*, *75*(3), 441.

O'Brien, D. P., & Overton, W. F. (1982). Conditional reasoning and the competence-performance issue: A developmental analysis of a training task. *Journal of Experimental Child Psychology*, *34*(2), 274 - 290.

Polk, T. A., & Newell, A. (1995). Deduction as verbal reasoning. *Psychological Review*, *102*(3), 533–566.

Ragni, M., Khemlani, S., & Johnson-Laird, P. (2014). The evaluation of the consistency of quantified assertions. *Memory & Cognition*, *42*(1), 1–14.

Ragni, M., Kola, I., & Johnson-Laird, P. N. (in press). On selecting evidence to test hypotheses: a theory of selection tasks. *Psychological Bulletin*.

Revlis, R. (1975). Two models of syllogistic reasoning: Feature selection and conversion. *Journal of Verbal Learning and Verbal Behavior*, *14*(2), 180–195.

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

Stanovich, K. E. (1999). *Who is rational? studies of individual differences in reasoning*. Psychology Press.

Stenning, K., & Cox, R. (2006). Reconnecting interpretation to reasoning through individual differences. *Quarterly Journal of Experimental Psychology*, *59*(8), 1454–1483.

Van der Henst, J.-B., Yang, Y., & Johnson-Laird, P. N. (2002). Strategies in sentential reasoning. *Cognitive Science*, *26*(4), 425–468.