Mental models and consistency

P.N. Johnson-Laird

In Bertram Gawronski and Fritz Strack (Eds.):

Cognitive Consistency: A Unifying Concept in Social Psychology, New York: Guilford Press, 2012.

[30 manuscript pages total, including title page, text body, references, notes, tables, and figures.]

Author's address

Professor P.N. Johnson-Laird

Department of Psychology

Princeton University

New Jersey 08540

USA

Email: phil@princeton.edu

Fax: ++1 609 258 1113

Introduction

Lady Bertram in Jane Austen's novel, *Mansfield Park*, has a dog called 'Pug'. In Chapters 7 and 8 of the novel, Pug is male. But, by Chapter 33, Lady Bertram is promising Fanny, the novel's protagonist, a puppy from Pug's next litter. Robinson Crusoe strips naked and swims out to his wrecked ship, where he stuffs his pockets with 'biskit'. And Emma Bovary has eyes that change color from one chapter to another in Flaubert's masterpiece. Novelists are often inconsistent. Their admirers are irked when close readers point out these flaws, and concur with Emerson: *A foolish consistency is the hobgoblin of little minds*. Indeed, these examples suggest that inconsistencies are a trivial matter of no great consequences for rationality: one assertion clashes with another. They seem no more serious than inconsistent desires: *Nec tecum possum vivere nec sine te* (I can live neither with you nor without you), as Martial wrote of his mistress nearly two thousand years ago. Such desires are part of the human condition, but again they are not a threat to rationality. The present author therefore declares that this chapter will be inconsistent too.

The chapter begins with more significant consequences of inconsistencies in logic and in life. Our ability to detect them, however, shows that we have some deductive competence, because there is a direct link between making a deduction and evaluating a set of assertions as inconsistent. The chapter accordingly proposes a theory of how we detect inconsistencies, and it reports several corroboratory lines of research. It then turns to the question of what happens after we have detected an inconsistency. It surveys what philosophers have had to say on this matter – in essence, that we should make only minimal changes to our beliefs in the face of an inconsistency, e.g., Miss Austen should have made Pug a bitch from the outset of her novel, because this change calls merely for altering two pronouns. In my view, however, philosophers

and psychologists have put far too much emphasis on the revision of beliefs. As the chapter argues, the real task is to resolve inconsistencies, and their resolution calls for the creation of explanations. One side effect is that explanations then imply that certain prior beliefs are false. Manifest inconsistencies can be useful, but those that go unnoticed are often a cause of catastrophe. The chapter concludes with this phenomenon.

Inconsistency in logic

In orthodox logic, inconsistency is disastrous. Bertrand Russell discovered in 1901 that the then current formulation of set theory yielded an inconsistency. Because set theory lies at the foundations of arithmetic, the inconsistency was devastating, and forced logicians to reformulate the theory. The consequences of a contradiction in orthodox logic are indeed explosive. They allow one to prove any conclusion whatsoever. Given, say, the premise: *The dog is male and the dog is not male*, a simple proof yields the conclusion, say, that *God exists*. The proof hinges on the fact that an inference of the form:

The dog is male.

Therefore, the dog is a male or God exists, or both.

is valid. That is, if the premise is true, then the conclusion must be true too (even though logically-untrained individuals find the inference too strange to be acceptable). The final step is to use the other constituent of the premise:

The dog is not male

and the preceding disjunction to infer validly:

Therefore, God exists.

A skeptic asked Russell whether it was true that anything whatsoever follows from a contradiction. Russell replied that it was. So, the skeptic demanded, prove that I am the Pope from 1 + 1 = 1. Russell said: you're one and the Pope is one, but one plus one equals one, and so you and the Pope are one.

The neglect of inconsistencies in daily life is a recipe for catastrophe. The engineers in charge of the experiment at the Chernobyl nuclear power plant believed that the reactor was still intact several hours after the explosion, even though firemen came to the control room carrying pieces of graphite that could have come only from inside the reactor. This inability of the engineers to recognize the inconsistency delayed them in alerting the authorities in Moscow to the disaster. And this delay exacerbated the consequences of the disaster (see Medvedev, 1990).

Inconsistencies, alas, are not mere conflicts between one proposition and another. For instance, consider these three assertions:

The reactor isn't dangerous if and only if it is intact.

If it is intact then all its graphite is inside it.

The reactor isn't dangerous and some of its graphite isn't inside it.

They cannot all be true at the same time, and so they are inconsistent. Yet, any two out of the three assertions *are* consistent. This principle applies in general: a set of *n* assertions can be inconsistent even though any *n-1* of them yields a consistent set. A corollary is that the evaluation of a set of assertions as consistent or inconsistent is computationally intractable (Cook, 1971). For instance, consider a set of assertions based on 100 propositions, which can each be true or false, such as: *the reactor is intact*. In principle, the evaluation of consistency may call for an examination of all 2^{100} possibilities. The number may not seem so large. But, in fact, it is more than a million times a million times a million times a million times a million, i.e.,

1,000,000[5]. So, even if you could examine each possibility in a millionth of a second, it would still take much longer than the universe has existed to examine them all. Granted that our beliefs together depend on many more than 100 propositions, we may well, like the White Queen in *Alice in Wonderland*, believe six impossible things before breakfast. We need to keep our beliefs segregated into separate topics so that we can check their consistency for one topic at a time (Klein, 1998).

How do we evaluate consistency?

Some theorists are skeptical that logically-untrained individuals are capable of deductive reasoning: an Italian logician once accused the present author of lying because he had not made this point in print (see also Oaksford & Chater, 2007, for a more profound case against deductive competence). But, if untrained individuals are able to detect inconsistencies, they can make deductions. The two tasks may seem very different, but they are two sides of the same skill. This fact is exploited in one method of doing logic: to prove that a conclusion follows validly, you add its negation to the premises and show that the resulting set of propositions is inconsistent (see, e.g., Jeffrey, 1981). We can therefore establish that untrained individuals are deductively competent if we show that they can detect inconsistencies.

One way in which they might evaluate consistency depends on the use of tacit rules of inference – of the sort that are said to underlie the ability of logically-naïve individuals to reason (see, e.g., Rips, 1994; Braine & O'Brien, 1998). The general procedure for assessing consistency is the mirror-image of the method of logic described above. You choose any assertion in the set, and try to prove its negation from the remaining assertions. If you succeed, then the original set of assertions is inconsistent. If you fail after an exhaustive search, then the

original set is consistent. It follows that the detection of inconsistency should be easier than the detection of consistency: inconsistency depends on finding a single proof, whereas consistency depends on an exhaustive search of all possible proofs.

Another way in which individuals might evaluate consistency is based on an alternative theory of the psychology of reasoning – the theory of mental models. This theory postulates that reasoners envisage all the possibilities to which premises refer – relying on both meaning and general knowledge – and they treat as valid any conclusion that holds in all these models. The theory has been described many times (see, e.g., Johnson-Laird, 2006; Johnson-Laird & Byrne, 1991), and so an outline suffices here. Because each mental model represents a possibility, an immediate way for reasoners to evaluate the consistency of a set of assertions is to seek a model in which all the assertions hold. That is, the model *satisfies* the assertions, where to *satisfy* a set of assertions is to show that they have a single model and are therefore consistent. If reasoners find such a model, they evaluate the assertions as consistent; otherwise, they evaluate them as inconsistent. Contrary to approach based on rules of inference, the use of models predicts that the detection of consistency should be easier than the detection of inconsistency. Consistency depends on finding a single model, whereas inconsistency depends on an exhaustive search of all possible models.

A central assumption of the *model* theory, as it is known, is its principle of truth. The principle stipulates that mental models represent only what is possible according to assertions, and not what is impossible. The principle further stipulates that each mental model represents only those clauses in assertions that are true in the relevant possibility. Working memory, as Simon (1982, 1983) argued, is a major bottleneck in cognition. Hence, the advantage of the

principle of truth is that it imposes less of a load on working memory. As an example of the principle, consider the assertion:

The reactor isn't dangerous if, and only if, it is intact.

It refers to the two possibilities shown here on separate lines:

The reactor: not-dangerous intact

. . .

The first line denotes a mental model of the possibility in which the reactor is not dangerous and is intact. The second line – the ellipsis – denotes a mental model of the possibility in which it is false that the reactor isn't dangerous. Individuals do not make explicit this possibility unless the task demands it and they can easily flesh out their models in this way. Mental models are not words and phrases as here, but iconic representations of the world, in which each part of a model corresponds to a part of what it represents (for the notion of an icon, see Peirce, 1931–1958, Vol. 4). In contrast to mental models, the *fully explicit* models of an assertion represent both what is true and what is false. Hence, for the assertion above, they are as follows:

The reactor: not-dangerous intact [PRINTER, PLEASE ALIGN

dangerous not-intact VERTICALLY, AS SHOWN]

The assertion therefore refers to the same possibilities as the disjunction:

Either the reactor is intact or else it is dangerous, but not both.

This equivalence is not obvious for most of us, because we tend to rely on mental models rather than fully explicit models. The principle of truth yields parsimonious representations, but it also has a striking consequence to which we will return presently. And the model theory makes testable predictions about the evaluation of consistency.

An immediate prediction is that if the first model that individuals tend to construct satisfies a set of assertions then the evaluation task should be easier than when they have to search for an alternative model. As an example of the difference, consider the following set of assertions about what is on top of a table:

If there isn't an apple then there is a banana.

If there is a banana then there is a cherry.

There isn't an apple and there is a cherry.

The salient mental model of the first assertion is one that satisfies its two clauses:

not-apple banana

The second assertion holds in this model and updates it to:

not-apple banana cherry

The third assertion holds in this model, and so reasoners should respond that the assertions are consistent.

A contrasting set of assertions is as follows:

There is an apple or there is a banana.

There isn't a banana or there is a cherry.

There isn't an apple and there is a cherry.

Individuals are likely to start with a model of the first clause of the first assertion:

apple

They may continue to update it only to discover that the third assertion is not consistent with it. They now have to start over with an alternative model of the first assertion:

not-apple banana

This model refutes the first clause of the second assertion, and so its second clause holds:

not-apple banana cherry

The third assertion holds in this possibility, and so individuals should now respond that the three assertions can all be true at the same time.

An experiment compared various problems of the two sorts, and corroborated the prediction (Johnson-Laird, Legrenzi, Girotto, & Legrenzi, 2000). When the first model sufficed for an evaluation of consistency, correct responses occurred on 93% of trials; when the first model had to be replaced by an alternative, correct responses occurred on only 74% of trials. You might suppose that conditionals are easier to understand than disjunctions. It's true, but this factor was counterbalanced in the experiment. We tested 522 participants from among the best high-school graduates in Italy, and the difference between the two sorts of problems was robust.

The principle of truth makes a surprising prediction, which we discovered by accident in the output of a program that I wrote to simulate the theory (Johnson-Laird & Savary, 1999). In some cases, mental models turn out to be wrong, i.e., they do not correspond to the correct possibilities, which are captured only in fully explicit models. This discrepancy predicts the occurrence of illusions of consistency: individuals should evaluate certain sets of assertions as consistent when, in fact, they are inconsistent. Likewise, there should be illusions of inconsistency in which individuals evaluate certain other sets of assertions as inconsistent when, in fact, they are consistent. An example of an assertion that should mislead people is an exclusive disjunction of two conditional assertions about what is on top of a table:

If there's an apple then there's a banana, or else if there's a cherry then there's a banana.

You might want to think about what the possibilities are given this assertion. Most people think of those that correspond to the assertion's mental models:

apple banana [PRINTER, PLEASE ALIGN VERTICALLY AS SHOWN]

cherry banana

In contrast, the fully explicit models of the assertion take into account that when one conditional is true, the other conditional is false. And individuals take the falsity of a conditional, such as *if there's an apple then there's a banana* to mean that *there's an apple without a banana* (e.g., Barres & Johnson-Laird, 2003). When the first conditional is true, the second conditional is false and so *there is a cherry and not a banana*, which is compatible the truth of the first conditional in case there isn't an apple and isn't a banana. Conversely, when the second conditional is true, the first conditional is false, and so the fully explicit models of the assertion are:

apple not-banana not-cherry [PRINTER, PLEASE ALIGN VERTICALLY AS

not-apple not-banana cherry SHOWN]

The discrepancies between the mental models and the fully explicit models, which are correct, predict various illusions and various control problems that should yield correct responses. Hence, each of the following assertions if paired with the preceding disjunction elicits a different sort of evaluation:

There is an apple and a banana: an illusion of consistency, because it is consistent with the first of the mental models of the assertion above but with neither of its fully explicit models.

There is an apple and not a cherry: a control problem for consistency, because it is consistent with the first of the assertion's mental models and the first of its fully explicit models.

There is an apple and not a banana: an illusion of inconsistency, because it is not consistent with any of the mental models, but consistent with the first of the fully explicit models.

There is not an apple and not a cherry: a control problem for inconsistency because it is not consistent with any or the mental models or any of the fully explicit models.

An experiment examined four different sets of such problems, and yielded the following overall percentages of correct evaluations (Johnson-Laird, Girotto, & Legrenzi, 2000):

Illusions of consistency: 9%

Controls for consistency: 96%

Illusions of inconsistency: 20%

Controls for inconsistency: 86%

We tested 129 participants, and 128 of them were more accurate with the control problems than with the illusory problems, and there was one tie (Binomial test, $p = 0.5^{128}$; this statistic is the only one cited in the present paper, because it is the author's most significant result ever, occurring by chance less than one time in $1{,}000{,}000^6$).

Such a level of significance prompts critics to assert that perhaps the only illusions are in the minds of the experimenters, who have an erroneous view of the assertions. One potential error is in their account of conditionals. The criticism seems specious, because the illusions depend solely on the well-established phenomenon that individuals take a conditional, *if A then B*, to be false in the case of *A and not-B* (e.g., Barres & Johnson-Laird, 2003). But, in order to generalize the phenomenon, a more recent study corroborated similar illusions based solely on disjunctions (Byrne, Lotstein, & Johnson-Laird, 2010).

Critics may suppose instead that the root of the problem is the disjunction, 'or else', which perhaps naïve individuals interpret as an inclusive disjunction, which allows both of its clauses to be true. This hypothesis, however, fails to account for the illusions of inconsistency, which remain illusions whether the disjunctions are interpreted as exclusive or inclusive. And a further study also ruled out this criticism (see Legrenzi, Girotto, & Johnson-Laird, 2003). The task was to consider descriptions of objects, and if all of the assertions in a description could be

true to write down a direct description of the object. Otherwise, the participants had to write down that no such object was possible. Here is a typical trial:

Only one of the following two propositions is true:

The tray is heavy or elegant, or both.

The tray is elegant and portable.

The following proposition is definitely true:

The tray is not elegant but portable.

The initial pair of assertions – of which only one can be true – yield the mental models of the tray:

heavy [PRINTER PLEASE ALIGN VERTICALLY

elegant AS SHOWN]

heavy elegant

elegant portable

The third assertion is not consistent with any of these mental models, and so the participants should respond that the object is impossible, and 90% of them did so. However, fully explicit models take into account that when the first assertion is true, the second assertion is false, and vice versa. So, there is a tray that is possible, and it has this description:

heavy not-elegant portable

The description holds when the first assertion is true and the second assertion is false, and it satisfies the third assertion. Once again, the percentages of correct responses in the experiment corroborated the model theory's predictions:

Illusions of consistency: 8%

Controls for consistency: 93%

Illusions of inconsistency: 13%

Controls for inconsistency: 96%

Illusions of consistency and inconsistency occur for a variety of other sorts of assertion (see, e.g., Johnson-Laird, Girotto, & Legrenzi, 2004). They are a litmus test for the use of mental models in evaluating consistency, because the theory and its computer implementation predict the illusions from the principle of truth. At present, no alternative theories can account for them. Yet, performance with the control problems is very good, and shows that individuals do have the competence to evaluate consistency. Hence, they are capable of deductive reasoning.

The revision of beliefs

When a fact is inconsistent with our beliefs what should we do? A standard philosophical answer is that we should revise our beliefs so that they no longer conflict with the fact. In fact, people sometimes perseverate with beliefs even though they cannot all be true (see, e.g., Baron, 2008). Logic, alas, provides no guide to how we should revise our beliefs in the face of an inconsistency, but a primal view about this matter goes back to William James (1907, p. 59): '[The new fact] preserves the older stock of truths with a minimum of modification, stretching them just enough to make them admit the novelty'. This minimalist view has many defenders (e.g., de Kleer, 1986; Harman, 1986; Gärdenfors, 1992). A separate principle governs those beliefs that we should abandon: 'when it comes to choosing between candidates for removal, the least entrenched ought to be given up' (Fuhrmann, 1997, p. 24). Entrenchment itself is likely to depend on the evidence for a proposition; and various systems of 'truth maintenance' in artificial intelligence aim to keep track of the propositions that support a belief (e.g., de Kleer, 1986). But, the entrenchment of a proposition is also likely to depend on the

degree to which it coheres with other beliefs (e.g., Gärdenfors, 1992; Thagard, 2000). Psychological studies have corroborated the role of entrenchment in the revision of beliefs. In one study, participants evaluated assertions such as:

All vertebrates have a backbone.

This amoeba does not have a backbone.

and the 'fact':

This amoeba is a vertebrate.

The participants tended to hold on to the generalization, which is common knowledge, and to give up the specific proposition, which is not (Revlis, Lipkin, & Hayes, 1971). Other investigators have reported analogous effects of knowledge and experience (e.g., Dieussaert, Schaeken, De Neys, & d'Ydewalle, 2000; Politzer & Carles, 2001; Markovits & Schmeltzer, 2007).

The model theory predicts that another factor should affect the revision of beliefs. If the facts are inconsistent with a mental model of a proposition, then, other things being equal, individuals should abandon this proposition. This principle is pertinent when all the propositions are equally believable, and the fact does not conflict with any individual belief, but instead is inconsistent with the set as a whole. Here is an illustration of this *mismatch* principle:

Speaker A asserts: If the President owns a villa then he owns a yacht.

Speaker B asserts: He owns a villa.

But, you know as a matter of fact: He does not own a yacht.

The fact mismatches one of the mental models of what the president owns according to speaker A's conditional assertion:

villa yacht

. . . [PRINTER, PLEASE CENTER ELLIPSIS BETWEEN 2 WORDS ABOVE]

where the ellipsis represents other possibilities in which it is false that the President owns a villa. So, individuals should give up the conditional, because of the mismatch between the fact and this model. Of course, the two assertions are not truly inconsistent, because the conditional is consistent with the President not owning a villa and not owning a yacht. But, this possibility is not represented explicitly in the mental models of the conditional. A contrasting example is as follows:

Speaker A asserts: If the President owns a villa then he owns a yacht.

Speaker B asserts: He doesn't own a yacht.

But, you know as a matter of fact: He owns a villa.

In this case, the fact matches the salient model of the conditional (shown above), but it is not represented in the model of speaker B's assertion, and so it's this assertion that should be abandoned. These are indeed the most frequent revisions for specific examples of this sort (Elio & Pelletier, 1997; Revlin, Cate, & Rouss, 2001; Hasson & Johnson-Laird, 2003).

In a replication, we also examined more complex assertions to ensure that the participants were not using a syntactic strategy of matching, or mismatching, clauses in the assertions (Girotto, Johnson-Laird, Legrenzi, & Sonino, 2000). Here is an example of one sort of trial:

Speaker A asserts: If the President owns a villa and a pool then he owns either a plane or else a yacht.

Speaker B asserts: The President owns a villa and a pool.

But, you know as a matter of fact: He owns a plane and a yacht.

The fact conflicts with the mental models of the conditional, and so reasoners should reject it. A contrasting trial was of this sort:

Speaker A asserts: Either the President owns a villa and a pool or else if he owns a plane then he owns a yacht.

Speaker B asserts: The President owns a villa and a pool.

But, you know as matter of fact: He owns a plane and a yacht.

The fact matches one model of Speaker A's assertion, and so reasoners should reject Speaker B's assertion. The majority of responses corroborated the mismatch principle.

The explanation of inconsistencies

To illustrate an inconsistency in daily life, consider an actual example. A friend and I were waiting outside a small café in Provence. We knew two things. Our other friends had gone to get the car to pick us up. And if they had gone to get the car, then they'd be back in ten minutes. Ten minutes went by with no sign of them, and then another ten minutes. There was a conflict between a consequence of our beliefs, namely, that they would be back in ten minutes, and the facts of the matter. Our immediate thought was: what's happened to them? We needed an explanation that would resolve the inconsistency, and various possibilities occurred to us – they'd gone to a bar instead, they'd gotten lost, they'd run into a complicated one-way system and had to make an excursion round the town, or they'd had difficulty in starting the car. We could dismiss most of the possibilities, but the last one seemed likely, because the car had been difficult to start on a previous occasion. We reasoned that we'd better not walk to the car park to see what had happened in case they had gotten the car to start and were coming on some alternate route to pick us up. Indeed, they arrived not long afterwards, and had needed a tow to

get the car started. The point of this anecdote is that not once did we say to ourselves: which of our beliefs should we abandon? Instead, we thought of a plausible explanation – they'd had difficulty in starting the car, which had the side-effect of refuting our belief that if they'd gone to get the car then they'd be back in ten minutes. The direct effect of our reasoning was our decision to sit tight and to wait for them.

Individuals are adept in creating explanations that resolve inconsistencies. In an unpublished study, Tony Anderson (of the University of Strathclyde) and I gave participants four different scenarios that each implied an initial conclusion, e.g.:

> Tom's enemy Jerry was stabbed to death in a cinema during the afternoon. Tom was on a train to Edinburgh when the murder took place.

The participants' initial response was that Tom was innocent. In one condition, the experimenter replied with information that was inconsistent with this proposition:

> That's not in fact true. Can you think of any other possibility?

He made this same response to all the participants' subsequent suggestions. In another condition, the experimenter told the participants:

> That's possible. Can you think of any other possibility?

And he made this same response to all the participants' subsequent suggestions.

The participants were able to create potential explanations in response to the experimenter's feedback. Their initial creation of alternative scenarios depended on simple spatio-temporal manipulations of the situation, e.g.:

> The cinema was on the train.

One participant similarly suggested:

> The train ran through the middle of the cinema, and the suspect had a long knife and

leaned out the window.

The next suggestion tended to be by analogy with other crimes:

The suspect had an accomplice who committed the murder.

Some participants, especially engineering students, then proposed various sorts of action at a distance:

The suspect had suspended a knife in a block of ice over the victim.

The suspect put a spring-loaded knife in the victim's seat in the cinema.

The suspect used a radio-controlled robot.

A few participants thought of even more remote possibilities:

The suspect gave the victim a post-hypnotic suggestion to stab himself.

Sooner or later, all the participants ran out of ideas, but they created more explanations in response to the direct inconsistency that their previous idea was not true than in response to a request for another possibility. Likewise, there were reliable sequences, such as the one above, in the order in which individuals thought of potential explanations. This order suggests that they tended to use the same sorts of simulations in order to create their explanations.

To establish what makes a plausible explanation, we examined participants' spontaneous responses to problems based on inconsistencies, such as:

If someone pulled the trigger, then the gun fired.

Someone pulled the trigger. But the gun did not fire.

Why not?

In a preliminary study (Johnson-Laird et al., 2004), we asked the participants to list all the explanations that they could imagine for a series of such scenarios, and they responded with examples, such as:

A prudent person unloaded the pistol and there were no bullets in the chamber.

This explanation refutes the conditional assertion, but not in a minimal way. Each of 20 scenarios elicited a variety of different explanations, with a mean of 4.75 different explanations per problem. As the mismatch principle predicts, however, the vast majority of explanations amounted to refutations of the conditional proposition (90% of trials) rather than the categorical proposition. Only on 2% of trials were the participants unable to come up with an explanation.

A legitimate but puzzling question is: where do explanations come from? That is, by what processes do individuals create them? The explanations depend on knowledge, but in some cases they are novel – the individual who created them had never thought of them before. One answer to this question is that we all have knowledge of causal relations – models in long-term memory of causes and their effects, and we can use this knowledge to construct a causal chain that simulates a sequence of events. The optimal solution is indeed a chain consisting of a cause and an effect in which the effect resolves the inconsistency. I wrote a computer program that illustrates this process (see Johnson-Laird et al., 2004). For the example above, the program constructs a model of the possibility described in the first two assertions:

trigger pulled pistol fires

The fact that the pistol did not fire is inconsistent with this model. Yet, the conditional expresses a useful idealization, and so the program treats it as the basis for these mental models:

trigger pulled not(pistol fires) [the facts]

trigger pulled pistol fires [counterfactual possibilities]

. . . [PRINTER, ALIGN AS SHOWN]

In a knowledge-base, the program has fully explicit models of various ways in which a pistol may fail to fire, i.e., disabling conditions such as, if the pistol doesn't have any bullets in it, or if

its safety catch is on. The model of the facts above triggers one such model corresponding, say, to the first of these cases, and the relevant model modulates the facts to create the following possibility:

not(bullets in pistol) trigger pulled not(pistol fires)

The new proposition in this model, not(bullets in pistol), can in turn trigger a causal antecedent from another set of models in the knowledge base representing a cause for the absence of bullets in a pistol, e.g., if a person empties the bullets from the pistol. In this way, the program can construct a novel causal chain. The resulting possibility explains the inconsistency: a person emptied the pistol and so it had no bullets. And the counterfactual possibilities yield the claim: if the person hadn't emptied the pistol then it would have had bullets, and it would have fired. The fact that the pistol did not fire has been used to create an explanation from knowledge, which in turn refutes the generalization and transforms it into a counterfactual claim (see Byrne, 2005).

One prediction from this account is that explanations consisting of a cause and an effect that resolves the inconsistency should be rated as very probable, and indeed as more probable than the cause alone, the effect alone, or a refutation of the categorical assertion in the problem, e.g., the person didn't pull the trigger hard enough. In a experiment, the participants assessed the probabilities of putative explanations resolving inconsistencies based on those from the previous study (see Johnson-Laird et al., 2004). For each of 20 scenarios, they assessed the probabilities of the following sorts of explanation and two additional foils. They had to rank order the explanations from the most probable to the least probable:

1. Cause and effect: A prudent person had unloaded the gun and there were no bullets in the chamber.

2. Cause alone: A prudent person had unloaded the gun.

3. Effect alone: There were no bullets in the chamber.

4. Rejection of the categorical proposition: The person didn't really pull the trigger.

5. Non-causal conjunction: The gun was heavy, and there were no bullets in the chamber.

The cause and effect is a conjunction, so the non-causal conjunction was included as a control.

Insert Table 1 about here

Table 1 shows the overall rank orders of the probabilities for each of the 20 scenarios. As the table shows, the results corroborated the model theory: the participants' tended to rank the cause-and-effect explanations as the most probable. Hence, individuals do not always accommodate a new fact with a minimal change to their existing beliefs. The acceptance of a conjunction calls for a greater change than the acceptance of just one of its constituent propositions. The rankings are thus instances of the 'conjunction' fallacy in which a conjunction is judged as more probable than its constituents (Tversky & Kahneman, 1983). As these authors argued, such judgments are often based on a heuristic of 'representativeness': if a property is representative of membership of a set, we judge entities with that property as likely to be members of the set. In the present case, a causal chain is more representative of an explanation than either of its constituents. Other recent studies have borne out these effects (e.g., Walsh & Johnson-Laird, 2009). And participants also rate explanations as more probable than minimal revisions to either the generalizations in scenarios or the categorical assertions (Khemlani & Johnson-Laird, 2010). The most plausible explanation is not always minimal.

The importance of detecting inconsistencies

Inconsistencies do have their uses. One effective rebuttal of an invalid argument is to show that it leads to an inconsistency, and logically-untrained individuals do sometimes use this

method (Bucciarelli & Johnson-Laird, 1999; Johnson-Laird & Hasson, 2003). A brain-imaging study showed that it elicits activation in the right frontal pole of prefrontal cortex, whereas simple inferences do not (Kroger et al., 2009). The use of inconsistencies can also be informal. For example, Relativists argue that the principles of reasoning differ depending on culture, and one culture's views about inference are as good as any other culture's view (e.g., Barnes & Bloor, 1982). In contrast, Rationalists argue that certain principles of reasoning are universal (e.g., Hollis, 1970). Relativists must allow that Rationalists are right in their culture, but since Rationalists make a universal claim, their conclusion is inconsistent with Relativism. So, Relativism leads to its own rebuttal. The usefulness of such manifest inconsistencies in argument contrasts with the dangers of unnoticed inconsistencies.

Perhaps the single biggest error in human thinking is to overlook a possibility that is inconsistent with what we tacitly believe. The 'Darwin awards' are monuments to such errors. As their founder, Wendy Northcutt (2000) has observed, they are posthumously awarded to those who have 'improved the gene pool by eliminating themselves from the human race in astonishingly stupid ways'. What one learns from the citations is the ubiquitous failure to think of a possibility, e.g., what may happen – depending on where you stand – if you give an elephant an enema. Likewise, a common human error in causing disasters is the failure to think of a possibility, as at Chernobyl and in many collisions at sea. As Perrow (1984, p. 230) pointed out, '... despite the increasingly sophisticated equipment, captains still inexplicably turn at the last minute and ram each other. … they built perfectly reasonable mental models of the world, which work almost all the time, but occasionally turn out to be almost an inversion of what really exists.'

Accidents occur, as Wagenaar and Groeneweg (1987) wrote, because people fail to consider their possibility. A tragic but illustrative case is *The Herald of Free Enterprise* disaster. On March 6th 1987, a ship of this name was in the port of Zebrugge in Belgium. It was a 'roll on roll off' car ferry in which the cars drive down a ramp through doors in the bow, and exit in the reverse way. The cars had driven into the ferry, but the assistant bosun whose job it was to close the bow doors was asleep in his bunk. The bosun saw that the doors were open, but it was not his job to report the matter. The company required the Chief Officer to be on the bridge 15 minutes before departure, and the company had a policy in which vessels had to sail on time – a policy that was urgent because of a delay earlier that day. The Master did not ask for any report about the doors, and the company had a policy of 'negative' reporting in which only untoward events were reported. At 6.05pm, the *Herald* set sail with the bow doors open. Once it left the harbor, the sea poured in, the vessel capsized, and amid scenes of great heroism 188 individuals drowned. The subsequent inquiry blamed everyone from the director of the company down to the assistant bosun. It also faulted the policy of negative reporting. A prophylactic that would prevent such errors is a checklist, just as checklists have proved so effective in hospitals (Gawande, 2009). But, how can we minimize the failure to consider inferential possibilities?

One technique is the so-called 'model' method developed by Bell (1999). Consider this problem:

The convertor or the signal system, or both, are working.

If the signal system isn't working then the G force is excessive.

Hovering is possible only if the signal system is working.

The propellers are functioning normally, and the G force is not excessive.

Is hovering possible even if the convertor isn't working?

The answer is not obvious, but a simple procedure of listing the possibilities, makes the problem much easier. The first assertion is consistent with three possibilities for what is working:

Convertor		[PRINTER: PLEASE ALIGN VERTICALLY AS
	Signal	SHOWN HERE, AND LIKEWISE BELOW.]
Convertor	Signal	

The subsequent assertions add further information, and eliminate possibilities, to yield two possibilities:

	Signal	Hovering	Propellers
Convertor	Signal	Hovering	Propellers

Hovering is therefore possible even if the convertor isn't working. The use of the procedure with a set of problems of this sort increased the accuracy of the participants' conclusions from 65% to 95%, and speeded up their correct responses from a mean of 25 s to a mean of 15 s (Bell, 1999). Once individuals have learned the model method, it is even effective if they can no longer write down the possibilities, but have to imagine them instead.

Conclusions

Logically-untrained individuals are deductively competent because they can evaluate the consistency of sets of assertions. The way they do so is to search for a mental model that satisfies all the assertions. The evidence for this hypothesis is that those problems for which the first model suffices are easier than those for which a search for an alternative model is necessary. It is also corroborated by the occurrence of illusions of consistency and of inconsistency. Once individuals have detected an inconsistency, they ought to revise the propositions that yield it. In daily life, however, the main task is to explain how the inconsistency arose. Individuals are able

to do so, and their explanations tend to be based on causal chains, contrary to the doctrine, going back to William James, that they should revise their beliefs in a minimal way. Manifest inconsistencies have their uses. Individuals with no training in logic can refute an argument by showing that it yields an inconsistency – a form of argument known as reductio ad absurdum. They can likewise refute an inference by finding a counterexample to it, i.e., a possibility that is consistent with the premises but not with the conclusion. The single biggest error in thinking, however, is likely to be the failure to envisage a possibility. Errors of this sort often contribute to disasters. The chance of such an error can be reduced by the use of checklists and especially those in which we list possibilities.

Envoi: Was the chapter inconsistent? If you ignore my prophecy that it would be inconsistent, then the remainder of the chapter was either consistent or inconsistent. If the remainder was consistent then it was inconsistent with my prophecy; if the remainder was not consistent then my prophecy was correct. Either way, the chapter was inconsistent. But, to make doubly sure, the remainder *was* inconsistent: I wrote that the chapter would present the result of just one statistical test. In fact, it presented a further twenty (see Table 1).

Acknowledgements

This research was supported in part by a grant from the National Science Foundation SES 0844851 to study deductive and probabilistic reasoning. The author thanks Tony Anderson, Patricia Barres, Victoria Bell, Monica Bucciarelli, Ruth Byrne, Vittorio Girotto, Geoffrey Goodwin, Uri Hasson, Sangeet Khemlani, Paolo and Maria Legrenzi, Max Lotstein, and Fabien Savary, for their help in carrying out this research. They are not responsible for any other inconsistencies in the chapter.

References

Barnes, B. and Bloor, D. (1982) Relativism, rationalism, and the sociology of knowledge. In Hollis, M., and Lukes, S. (Eds.) *Rationality and relativism*. Oxford: Basil Blackwell.

Baron, J. (2008). *Thinking and Deciding*. 4th Edition. New York: Cambridge University Press.

Barres, P., & Johnson-Laird, P.N. (2003). On imagining what is true (and what is false). *Thinking & Reasoning*, 9, 1-42.

Bell, V. (1999). The model method. Unpublished Ph.D. thesis, Princeton University.

Braine, M.D.S., & O'Brien, D.P., Eds. (1998). *Mental Logic*. Mahwah, NJ: Lawrence Erlbaum Associates.

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

Byrne, R.M.J. (2005). *The Rational Imagination: How People Create Alternatives to Reality*. Cambridge, MA: MIT.

Byrne, R.M.J., Lotstein, M., & Johnson-Laird, P.N. (2010). Disjunctions: A theory of meaning, pragmatics, and inference. Under submission.

Cook, S.A. (1971). The complexity of theorem proving procedures. *Proceedings of the Third Annual Association of Computing Machinery Symposium on the Theory of Computing*, 151-58.

de Kleer, J. (1986). An assumption-based TMS. *Artificial Intelligence*, 28, 127-162.

Dieussaert, K., Schaeken, W., De Neys, W., & d'Ydewalle, G. (2000). Initial belief state as predictor of belief revision. *Current Psychology of Cognition*, 19, 277-288.

Elio, R., & Pelletier, F. J. (1997). Belief change as propositional update. *Cognitive Science*, 21,

419-460.

Fuhrmann, A. (1997). *An Essay on Contraction*. Stanford, CA: CSLI Publications.

Gärdenfors, P. (1992). Belief revision: An introduction. In Gärdenfors, P. (Ed.) *Belief revision*. (pp. 1-28). Cambridge: Cambridge University Press.

Gawande, A. (2009). *The Checklist Manifesto*. New York: Holt.

Girotto, V., Johnson-Laird, P.N., Legrenzi, P., and Sonino, M. (2000) Reasoning to consistency: How people resolve logical inconsistencies. In Garcia-Madruga, J., Carriedo, M, and Gonzalez-Labra, M. J. (Eds.) *Mental Models in Reasoning*. Madrid: UNED. Pp. 83-97.

Harman, G. (1986). *Change in View: Principles of Reasoning*. Cambridge, MA: MIT Press, Bradford Book.

Hasson, U., & Johnson-Laird, P. N. (2003). Why believability cannot explain belief revision. In R. Alterman & D. Kirsh (Eds.), *Proceedingsof the 25th Annual Conference of the Cognitive Science Society* (pp. 516-521). Mahwah, NJ: Erlbaum.

Hollis, M. (1970) Reason and ritual. In Wilson, B. R., (Ed.) *Rationality*. Oxford: Basil Blackwell. pp. 221-239.

James, W. (1907). *Pragmatism: A new name for some old ways of thinking*. New York: Longmans, Green.

Jeffrey, R. (1981). *Formal Logic: Its Scope and Limits*. 2nd Ed. New York: McGraw-Hill.

Johnson-Laird, P.N. (2006). *How We Reason*. New York: Oxford University Press.

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

Johnson-Laird, P.N., Girotto, V., & Legrenzi, P. (2004). Reasoning from inconsistency to

consistency. *Psychological Review*, 111, 640-661.

Johnson-Laird, P.N., & Hasson, U. (2003). Counterexamples in sentential reasoning. *Memory & Cognition*, 31, 1105-1113.

Johnson-Laird, P.N., Legrenzi, P., & Girotto, V. (2004). How we detect logical inconsistencies. *Current Directions in Psychological Science*, 13, 41-45.

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

Johnson-Laird, P.N., & Savary, F. (1999). Illusory inferences: A novel class of erroneous deductions. *Cognition*, 71, 191-229.

Khemlani, S., & Johnson-Laird, P.N. (2010). The need to explain. Under submission.

Klein, G. (1998). *Sources of Power: How People Make Decisions*. Cambridge, MA: MIT Press.

Kroger, J.K., Nystrom, L.E., Cohen, J.D., & Johnson-Laird, P.N. (2008). Distinct neural substrates for deductive and mathematical processing. *Brain Research*, 1243, 86-103.

Legrenzi, P., Girotto, V., & Johnson-Laird, P.N. (2003). Models of consistency. *Psychological Science*, 14, 131-137.

Markovits, H., & Schmeltzer, C. (2007). What makes people revise their beliefs following contradictory anecdotal evidence: The role of systemic variability and direct experience. *Cognitive Science*, 31, 535-547.

Medvedev, Z. A. (1990). *The Legacy of Chernobyl*. New York: W.W. Norton.

Northcutt, W. (2000). *The Darwin Awards: Evolution in Action*. New York: Penguin Putnam.

Oaksford, M., & Chater, N. (2007). *Bayesian Rationality*. Oxford University Press.

Peirce, C. S. (1931–1958). C. Hartshorne, P. Weiss & A. Burks (Eds.), *Collected papers*

of Charles Sanders Peirce, 8 Vols. Cambridge, MA: Harvard University Press.

Perrow, C. (1984). *Normal Accidents: Living with High-risk Technologies*. New York: Basic Books.

Politzer, G., & Carles, L. (2001). Belief revision and uncertain reasoning. *Thinking & Reasoning*, 7, 217-234.

Revlin, R., Cate, C. L., & Rouss, T. S. (2001). Reasoning counterfactually: Combining and rending. *Memory & Cognition*, 29, 1196–1208.

Revlis, R., Lipkin, S. G., & Hayes, J. R. (1971). The importance of universal quantifiers in a hypothetical reasoning task. *Journal of Verbal Learning and Verbal Behavior*, 10, 86–91.

Rips, L.J. (1994). *The Psychology of Proof*. Cambridge, MA: MIT Press.

Simon H.A. (1982). *Models of Bounded Rationality*, Vols 1 and 2. Cambridge, MA: MIT Press.

Simon, H.A. (1983). Search and reasoning in problem solving. *Artificial Intelligence*, 21, 7-29.

Thagard, P. (2000). *Coherence in thought and action*. Cambridge, MA: MIT Press.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, 90, 293-315.

Wagenaar, W.A., & Groeneweg, J. (1987). Accidents at sea: multiple causes and impossible consequences. *International Journal of Man-Machine Studies*, 27, 587-598.

Walsh, C.R., & Johnson-Laird, P.N. (2009). Changing your mind. *Memory & Cognition*, 37, 624-631.

Table 1: Participants' mean rank orders of the probabilities of five sorts of putative explanation of inconsistencies in 20 scenarios whose topics are shown below (Johnson-Laird et al., 2004), and the results of Page's L test on the predicted trend: *CE* denotes cause-and-effect, *C* denotes the cause alone, *E* denotes the effect alone, *R* denotes a refutation of the categorical assertion, and *NC* denotes a non-causal conjunction.

Scenarios	Page's L	z	p	Rank order of probability from highest to lowest
1. Tectonics	1059	7.1	<< .00001	CE E C R NC
2. Explosion	1069	7.5	<< .00001	CE C E R NC
3. Weather	988	3.9	< .00004	CE E C NC R
4. Melting	1062	7.2	<< .00001	CE C R E NC
5. Snake bite	1061	7.2	<< .00001	CE C R E NC
6. Diet	957	2.6	< .006	CE R E C NC
7. Indigestion	999	4.4	<< .00001	CE R E C NC
8. Aerobics	973	3.3	< . 0005	R CE C E NC
9. Car	1016	5.2	<< .00001	CE R C E NC
10. Reactor	969	3.1	< .002	CE R C R NC
11. Pistol	1035	6.0	<< .00001	E CE C R NC
12. Camera	1002	4.6	<< .00001	CE R C E NC
13. Forgetting	1049	6.7	<< .00001	CE C E R NC
14. Anger	1066	7.4	<< .00001	CE C E R NC
15. Liking	978	3.5	< .0003	CE NC E C R
16. Anxiety	1031	5.9	<< .00001	CE E R C NC
17. Politics	1083	8.2	<< .00001	CE C E R NC
18. Banks	998	4.4	<< .00001	CE C R NC E
19. Hotels	1027	5.7	<< .00001	CE C E R NC
20. Party	1037	6.1	<< .00001	CE C E R NC