Eric Naeseth
THEA 209

Human Logic: Fallacies and Cheaters

Human beings aren't really logical creatures. This might seem like a categorically false statement; after all, human beings employ the most cognitive faculties and have access to the most reasoning ability in the entire animal kingdom, so how can we be illogical? We may have the most reasoning ability, but our reasoning doesn't necessarily follow the rules of logic. Logic can be thought of as a special, "perfect" kind of reasoning which is conducted using strict principles of validity. It progresses from the premises to the conclusion using only a set of fixed, well-defined rules to derive new statements from previous statements. The key is not the rules themselves but the fact that the rules are *truth-preserving*; that is, if the previous statements was true, then the new statement must be true. What's interesting is that we don't *always* deviate from these rules of formal logic. There are a few situations in which we appear to use a different reasoning process that follows formal logic. Though some researchers have proposed that this asymmetry is necessarily due to separate, domain-specific reasoning modules, context and relevance alone are fully capable of explaining this phenomenon.

Human intuitive reasoning was perhaps most famously tested in Wason (1966)'s card-flipping experiment, commonly referred to as the selection task. It has been repeated many times by other researchers with variations to test specific claims, but the basic principle is the same. Subjects are shown four double-sided cards which are laid flat on a table, with only one side visible to the subject, and are asked to consider a conditional sentence of the form $P \rightarrow Q$ (read as "if P, then Q", where P and Q are independ-

ent clauses). The prompt sentences are of course actually given in normal English (with variations, which, we will see, are important), but that "if" clause is always present in some form. The sides of the cards that are visible to the subjects show statements corresponding to P, ¬P ("not P"), Q, and ¬Q. If the visible side of the card has a statement corresponding to the antecedent printed on it, the hidden side of the card has the consequent, and vice-versa. The subjects are asked to turn over the cards that would show if the rule were true.

In formal logic, conditional statements ($P \rightarrow Q$) are false *only* when the antecedent (P) is true, but the consequent (Q) is false. (If P is false, it doesn't matter whether or not Q is true or false because the statement only makes a claim about the case where P is true.) Thus, only two cards need to be turned over: the P card, to ensure that the back reads Q; and the ¬Q card, to ensure that the back reads ¬P. However, depending on how the question is phrased, the cards which a majority of subjects actually do turn over varies (Cosmides/Tooby 1992). A simple example with an abstract rule is: "If there is a vowel on one side of the card, then there is an odd number on the other". The subject is presented with cards showing 'A', 'B', '1', and '2'. When the rule is stated as such (in the form of a simple description), with these abstract cases, most subjects turn over the card with the A on it, representing P; and the card with the 1 on it, representing Q, even though even if there were a consonant on the other side, it would not actually invalidate the rule (Wason 1966).

However, change the phrasing of the rule to that of a social obligation, and the responses change dramatically. Say the rule is presented instead as, "If a person is drinking alcohol, he must be over 21 years old", and the cards are "Beer", "Coke", "22 years", and "16 years". Overwhelmingly, subjects agree with formal logic in their choices and pick

the beer card and the underage card (GRIGGS/COX 1982). One might point out that this distinction doesn't follow good scientific practice because it changes two aspects of the study at once: the concreteness of the properties and the addition of a deontological aspect to the rule. However, varying just the concreteness produces no change in subject responses (PINKER 1997; p. 336).

Even more interesting results are obtained when an element of role-play is introduced (FIDDICK/COSMIDES/TOOBY 2000). All employees of a certain business sign an employment contract which contains the following clause: "If an employee works on the weekend, then that employee gets a day off". Subjects are presented with cards labeled "Worked on the weekend", "Worked only during the week", "Got a day off", "Did not get a day off". Subjects asked to verify the rule that take the position of the employee pick the cards suggested by first-order logic: "Worked on the weekend" and "Did not get a day off". But, those asked to choose from the employer's point of view pick the exact *opposite* cards: "Worked only during the week" and "Got a day off". What happened?

Both the social responsibility and role-play factors change the nature of the condition the subjects are asked to check. The social responsibility factor turns the question into the search for violators, and the role-play specifies which violation the subject will be looking for. For, when a person who has entered into contract is looking to verify whether or not part of it is being upheld, that person will be looking to make sure the other party is following; the person isn't interested in verifying the whole contract, just the relevant end. So, the employees are interested in verifying $\forall x$ [WorkedWeekend(x) → DayOff(x)], while the employer is interested in verifying $\forall x$ [¬WorkedWeekend(x) → ¬Day-

$\mathrm{Off}(x)]$. Given this, subjects in *both* roles correctly choose the cards that will verify their internal translations of the proposition under formal propositional logic.

How does this take place? One view, that of the "cheater detection module", takes a rather Pinkerian view on things, although Pinker himself doesn't elaborate on it much in *How the Mind Works*. This view states that there is a specific module in the brain dedicated to the discovery of cheating, and that in studies like this, this module gets used to perform the analysis of the deontic questions instead of whatever generic logical faculties otherwise exist in the brain, explaining the different results. However, there are other explanations that don't require such a leap of thought.

Sperber and Wilson (1986) give an explanation that doesn't require a separate reasoning apparatus at all. Instead, it postulates a general reasoning apparatus that is context-sensitive; different contexts will cause different interpretations of the sentences to be used. After all, there are many, many details and nuances that can be transmitted by English languages that cannot be directly represented in formal logics. The resulting lack of ambiguity, of course, is largely the point of these artificial languages, but as anyone who has ever translated a complex English proposition into formal logic knows, very careful attention needs to be paid to all the implications of the original sentence.

It appears that the first thing the mind does when faced with a proposition such as the one in these experiments is to figure out if there really are relevant instances of the antecedent or the consequent in play as the start of its effort to find the most relevant context (SPERBER/WILSON 1986). So, the initial interpretation is not $\forall x\,[P(x) \rightarrow Q(x)]$ but rather $\exists x\,[P(x) \wedge Q(x)]$, a weaker claim. Further, if the subject interprets the domain as being merely the cards on the experimenter's table, (s)he might just suggest flipping over

the P and Q cards, to check if at least one is in fact $P \wedge Q$. Interestingly, when the initial proposition is of the form $P \rightarrow \neg Q$, and the same process happens, the subjects will test the cards representing P and $\neg Q$, which is correct propositional logic. The claim made by this theory is that, given a deontic selection task, the deontic part has cognitive effects and causes the rule to be interpreted as a prohibition (i.e. $\neg \exists x [P \rightarrow \neg Q]$), giving the same successful interpretation. It goes on to argue that, further, this correct logical interpretation can be chosen in non-social contexts (ATRAN 2001):

> For example, take the statement: "If a person wins a professional boxing match, then that person much be sober". The prediction is that subjects would pick the P card ("Wins Match") and the $\neg Q$ card ("Drunk") because information concerning a winning but drunk professional boxer more likely has cognitive effects than information concerning a successful sober boxer.

Frustratingly, the connection between subjects stopping with an existential claim during the card experiment and stopping at the same place during real-life reasoning is never elaborated on.

However, these frustrations pale in comparison to those from the explanation that Fodor offers in his *The Mind Doesn't Work That Way* (FODOR 2000; Appendix A). His logic goes something like this: instead of the deontic phrasing being internally interpreted as something like $\text{required}(P \rightarrow Q)$, it is actually interpreted as simply a categorical ban on Q. To show this, he submits a proof by contradiction (reproduced here with some cleanup):

i. Assume, for reductio, that "it's required that 'if P then Q'" is equivalent to $\mathsf{required}(P \rightarrow Q)$.
ii. Assume $\mathsf{required}(P \rightarrow Q) \wedge \neg Q$
iii. The inference scheme $[A \wedge \mathsf{required}(A \rightarrow B)] \rightarrow \mathsf{required}(B)$ is valid.
iv. Contraposition is valid in the scope of **required**.
v. $\neg Q \wedge \mathsf{required}(\neg Q \rightarrow \neg P) \rightarrow \mathsf{required}(\neg P)$

Then, he presents the following counterexample:

> Suppose everyone under 18 is obliged to drink coke. Then if Sam is under 18, he is prohibited from drinking whiskey. But *it does not follow* that if Sam is drinking whiskey, he is then obliged to be over 18. In fact Sam *can't* be obliged to be over 18 because he can't be obliged to do *anything* that he is unable to do. And with Sam, as with the rest of us, there's nothing much that he can do about how old he is (in, alas, either direction). I conclude that Authority cannot mandate the conditional (Sam drinks coke if he is under 18). The only course it can coherently pursue, having taken note of Sam's being under 18, is to mandate categorically that he drink coke.

This counterexample is intended to show that, in fact, (i) cannot be true, and instead that the requirements really take the mental form of just a categorical ban on Q, with that categorical ban simply only applying to P (Fodor 2000; p. 102). Assuming for the moment that those are actually distinct from each other (and, admittedly, it could be if it is simply a special form of a conditional that checks the consequent first and then backtracks to the antecedent to see if it was applicable), the supposed counterexample appears to be an invalid argument. Why is there this sudden shift from using "required" to using "obliged", and why is it suddenly able to compel physical objects to change? Yes, it is true that if you find Sam drinking whiskey, the law cannot compel him to become legal age on the spot, but that's not really a contradiction on (v). It simply states that if Sam is not 18 (true) and it's required that people who are under 18 not drink alcohol (also true),

that it's required that Sam not drink alcohol. It's the rule that allows the general regulation to be applied to specific individuals, and nothing more.

There is a word, though, for what happens in that situation with regard to rule (v). It's called a *violation*, and it could be this special sense of "contradicting" the requirement that triggers the emotion associated with being cheated. But if requirements don't use the special backtracking form that Fodor proposed, why are people better at checking the contrapositive when they are told it's a requirement? As Atran (2001) points out, human communication carries the assumption that things that are said are relevant to the listener. Saying that something is a requirement implies that people under that requirement would otherwise have a fairly strong incentive to not obey it, so people think to check the potential offenders, but when this clue is not present, it's easy to overlook.

There is a twist to this, though. Fiddick et al. (2000) predicted and then confirmed via experiments that, when presented with the conditional, "If in a hazardous situation that is costly to fitness, then take the benefit of precaution", that a majority of people's responses would correspond to formal logic. Here there is no explicit reference to a requirement or regulation; nothing that obvious is there to suggest to the subject to check the contrapositive. Fiddick et al. suggest that, especially when an element of role-play is added, relevance theory cannot reliably predict response patterns (ibid.), but instead these conditions must be handled by another module: the risk avoidance module.

They provide an example. Take the scenario of a primitive tribe being studied by an anthropologist, who, out of sympathy, provides rubber gloves to the tribe members involved in making poison darts. The condition is: "If you make poison darts, then you may use the rubber gloves". Subjects playing the roles of anthropologists who are trying

to ensure that only tribesmen involved in poison dart production are using the rubber gloves, make the ostensibly illogical choice of ¬P and Q. On the other hand, subjects playing the role of anthropologists who are trying to ensure that tribesmen who are making the darts are using protection concur with formal logic and choose P and ¬Q. These results, say Fiddick et al., are incompatible with relevancy theory, for the rule should not be able to produce both logical and illogical results depending on the perspective.

Unfortunately for this argument, relevance is in the eye of the beholder. The condition has an implicit reciprocity: "If you are using the rubber gloves, you [should] make poison darts", which effectively turns the conditional into a biconditional: ∀x [Risk(x) ↔ Benefit(x)] (ATRAN 2001). The "risk" interpretation looks for a violation of the left-to-right direction: ¬∃x [Risk(x) ∧ ¬Benefit(x)]; while the "privilege" interpretation looks for a violation in the opposite direction: ¬∃x [Benefit(x) ∧ ¬Risk(x)]. The different interpretations *can* be explained by relevancy alone, without resorting to adding any special detectors to the mind (though it is worth pointing out that this does not *rule out* detectors or a hybrid approach).

What's interesting is that even people who have been trained in formal logic frequently make these mistakes. This points to a built-in reasoning system common to all humans that cannot be so easily overridden through learning. As we have seen, however, people disagree about the form of its architecture, namely whether there are domain-specific "modules" that can take over for a general reasoning system, whether only a general reasoning system that is simply sensitive to context and relevant information, or, I assume, whether all reasoning takes place in domain-specific modules. While I do

believe that the sum of the evidence presented here points to the second interpretation, there is not enough to categorically disqualify the other possibilities. We just don't know enough to say for sure how the mind works.

References

Pinker, S. (1997) *How The Mind Works.* Norton: New York.

Fodor, J. (2000) *The Mind Doesn't Work That Way*. MIT Press: Cambridge, MA.

Pinker, S. (2005) *So, How* Does *the Mind Work?*. Mind & Language 20:1-24.

Atran, S. (2001) "A Cheater-Detection Module?". Evolution and Cognition 7 no. 2:1-6.

Wason, P. (1966) "Reasoning" in *New Horizons in Psychology*. Penguin Books: London, pp. 135-151.

Cosmides, L./Tooby, J. (1992) "Cognitive Adaptations for Social Exchange" in *The Adapted Mind: Evolutionary Psychology and the Generation of Culture.* Oxford University Press: New York, pp. 163-228.

Fiddick, L./Cosmides, L./Tooby, J. (2000) "No Interpretation Without Representation". Cognition 75:1-79.

Griggs, R./Cox, J. (1982) "The Elusive Thematic-Materials Effect in the Wason Selection Task". British Journal of Psychology 73:407-420.

Sperber, D./Wilson, D. (1986) "Relevance". Blackwell: Oxford.