**SUPPLEMENTARY MATERIALS**

 **A summary of the details of each of the 16 theories of the selection task.**

 Theories of the selection task are based on meanings, on formal rules of logic, on content-specific rules, on heuristics, on probabilities, and on neural models. Readers who want a quick survey should consult Table 5 in the paper itself. This section aims to provide enough detail for readers to judge each theory for themselves. But, we end each account, where relevant, with our own criticisms. References are only to those not listed in the paper itself.

**Theories** **based on meaning**

*1. Defective truth tables.*Wason’s initial approach to his task was to contrast his analysis of conditionals with the “material” conditional of logic (Wason, 1966, 1968). Consider the following hypothesis bearing in mind that all the cards have a letter on one side and a number on the other side:

 If there’s an A on one side of a card then there’s a 2 on the other side.

In predicate logic, its analysis can be paraphrased as follows:

 For any x, if x is a card with an A on one side then x is card with a 2 on the other side.

This assertion is true provided that there is no card with an *A* on one side and a number other than 2 on its other side (see the paper for the paradox of confirmation that this analysis yields). All other cards are consistent with the hypothesis.

 Wason’s analysis was different. Reasoners select p, he says, to find out whether the hypothesis is made true by the occurrence of q. They select q, if they do, to find out whether the hypothesis is made true by the occurrence of p. They don’t select $\overbar{p}$, because it is irrelevant to the truth of the hypothesis: its occurrence assigns no truth value to the conditional, so “irrelevant” is the third truth value. And they do not select $\overbar{q}$, because they fail to infer it from $\overbar{p}$. The grounds are that they are biased, through a long learning process, to expect a relation of truth to hold between sentences and states of affairs. Not surprisingly, for an initial foray, the theory is incomplete. It does not account for falsifying selections.

 *2. The Insight theory.* Figure SM1 presents the flow-diagram for the original algorithm (see Figure 2 of Johnson-Laird & Wason, 1970b). The model theory maintains the same functionality, but replaces the “truth tables” of logic with mental models for intuitions (no insight) and fully explicit models (partial or full insight). The functionality represented in the figure is described in the paper.



*. Figure SM1*. The flow-diagram for the original algorithm in Johnson-Laird & Wason, 1970b (see their Figure 2).

**Theories based on logical reasoning**

1. *PSYCOP.* The most thorough-going defense of formal logic and the predicate

calculus as the basis of human reasoning is due to Rips (1994). Because of the controversies over effects of content, Rips (1994, p. 319) had misgivings about the selection task’s suitability as an object of theory. He therefore offered an account only of the abstract task (Rips, 1994, p. 179 et seq.). He assumed that its performance depends, not on the meanings of conditionals, but on formal inferences, which his PSYCOP system can draw from the hypothesis and a description of each card. If an inference can be made, then the card must be selected, because when its conclusion does not hold the hypothesis is false. The respective inferences are as follows, which we have prefaced with their traditional names (and abbreviations):

 Modus ponens (MP): If p then q. p. Therefore, q. Ergo, p should be selected.

 Affirmation of the consequent (AC): If p then q. q. Nothing follows.

 Denial of the antecedent (DA): If p then q. not-p. Nothing follows.

 Modus tollens (MT): If p then q. not-q. Nothing follows, at least at first.

Rips’s PSYCOP model is able to deduce modus ponens, because it can apply its rule for this inference in the absence of a given conclusion. But, its rule of modus tollens for the fourth inference applies only if there is a given conclusion. The denial of the antecedent (DA) and the affirmation of the consequent (AC) are valid only for biconditionals (*if and only if p then q*). Hence, the reason that some individuals select the *q* card is that they assume the converse of the given conditional, namely, *if q then p*, from which given *q* they infer *p.* However, some other principle of selecting the cards that match those referred to in the hypothesis may also be in play (Beattie & Baron, 1988; Rips, 1994, p. 408, n. 9).

 Some individuals may make the modus tollens (MT) inference because they imagine what could be on the back of the $\overbar{q} $card (p or $\overbar{p}$). Hence, they generate their own potential conclusion and thereby can use the rule of inference. This step is more likely when individuals are required to enumerate the possible values on the back of the cards (Smalley, 1974). A similar improvement occurs when the hypothesis concerns, not two separable entities such as a letter and a number, but a single integrated entity, such as a red triangle (Wason & Green, 1984).

 The striking effects of content seem contrary to the use of formal rules of inference. But, Rips suggests that realistic contents remind individuals of relevant violations, or trigger the use of modal operators, such as those pertaining to permission. As he writes (p. 183), “The cover stories that experimenters tell to set up the contentful versions of the selection task may invite subjects to rely on background information, and this will, in turn, enable them to make inferences that go beyond what can be derived in the more pristine form of the task.” He even speculates that deontic contents might be handled within PSYCOP using rules of inference for deontic operators (p. 323; cf. Osherson, 1976).

 PSYCOP explains the diversity of selections with abstract hypotheses and the rarity of falsifications, but it cannot explain the partial insight that occurs in the repeated selection task, or the effect of negative *then*-clauses on selections. Rips wrote (p.c. 12/27/16) that he has never tried to fit PSYCOP to experimental results, “because the task itself seemed too limited and too subject to minor wording changes.”

 *4. Relevance theory.* Sperber and Wilson’s (1995) relevance theory motivated Sperber et al.’s (1995) empirical study, which we described in the text. The theory drew its inspiration from Grice’s (1975) pragmatics, but departs from it in innovative ways. Relevance theory rests on two principles. First, human cognition is geared to maximizing the relevance of any utterance that it processes. Second, every utterance conveys a presumption of its own optimal relevance to those to whom it is addressed. The greater the cognitive effort to process an utterance, the lower its relevance, but the greater the cognitive effect of its processing, the greater the relevance of the utterance. And an utterance has optimal relevance to those to whom it is addressed provided that it is relevant enough to be worth processing, and it is as relevant as is compatible with the speaker’s abilities and preferences (Sperber, p.c., 1-26-17).

An experimental setting yields only a low expectation of relevance, but participants in the selection task understand that it calls for them to select evidence potentially relevant to finding out whether the hypothesis is true or false, or, in deontic versions of the task, to finding out whether the cards denote cases that conform to, or contravene, the deontic principle.  Their understanding of the task’s description provides them with intuitions about the relevance of the cards, which in turn determine their selections.
 The theory postulates that individuals reason logically about the hypothesis (cf. Rips’s PSYCOP theory above).  A conditional, *if p then q*, yields three main inferences: an inference from *p* to *q*, which should yield the selection of p; an inference to *p and q*, which should yield the selection of pq; and an inference denying the counterexample to the hypothesis, *not (p and not-q)*, which should yield the selection of p$\overbar{q}$.  One way to elicit a falsifying selection is to make the hypothesis in the task itself a denial, *if p then not q*. Participants should treat the hypothesis as denying the conjunction of *p and q*, and so they should select pq. Likewise, deontic versions of the task lead to the selection of the cards, p$\overbar{q}$, as potential contraveners of the principle, *if p then q*.   As the theory implies, two factors are pertinent to inferences: the effort required to make them, and the cognitive effects they have. In general, it is harder to interpret the hypothesis *if p then q* as the denial of the conjunction, *p and not-q,*than to make the other two inferences. Hence, the falsifying selection is less likely to occur in a standard selection task. According to the theory, a successful recipe for increasing this selection should be to make the counterexample to the hypothesis both relevant to the task and easy to envisage (see the account of their experiments in the main text, and also an analogous study of the deontic selection task (Girotto, Kemmelmeir, Van der Henst, & Sperber, 2001).

 From our standpoint, it is odd that a theory that hinges on understanding meanings calls for deductive reasoning in the selection task (see the General Discussion in the paper). If a person has to decide whether an assertion is true or false, they should be able to rely on the meaning of the assertion. Theorists who offer no account of the meanings of sentential connectives often suppose that the formal rules of inference that apply to them can serve the same purpose as semantics (e.g., Rips, 1994). So, when Sperber et al. (1995) argue that the selection task sheds no light on how people reason, we concur: deductions from the hypothesis are not needed to make a selection. Indeed, the notion of making deductions from its hypothesis as opposed to thinking about its meaning stands in need of an independent motivation. The theory has never been implemented in a computational model (Sperber, p.c., 1-26-2017), and it would be difficult to formulate an algorithmic account of relevance. But, its use of counterexamples overlaps with the model theory (see the main text).

 *5. Multiple interpretations.* Stenning and van Lambalgen (2008) argued that the picture of logic current in psychology and cognitive science is completely mistaken (p. 5), and that until recently nearly everyone from Wason onwards was wrong about the selection task (p. 47). Their own theory, as they concede, is complicated. It is based on their unshakeable devotion to logic of various sorts, to the rationality of human beings, and to introspections as evidence. They engage in Socratic dialogs with participants making selections, and use the resulting introspective protocols to delineate different sorts of interpretation that underlie selections. They combine these putative interpretations with experiments designed to follow up their consequences.

 The end result of their analysis is, by their account, a partial list of the parameters that affect interpretation. One effect, for instance, is that a participant makes a deontic interpretation. In which case, the participant has few further interpretational problems: the correct cards to select contravene the principle, *if p then q*, and are p$\overbar{q}$. The main global parameter for the abstract task is whether or not the conditional hypothesis is interpreted as allowing exceptions. If it is “brittle” and does not tolerate them, then participants may think that one selection affects others, e.g., if p were to falsify the hypothesis, it would be pointless to make any more selections. But, if the hypothesis is “robust” and tolerates exceptions, then there is a conflict with the idea that just four cards can determine its truth or falsity, and it becomes problematic to distinguish between exceptions and counterexamples. (In our view, the original selection task has an untestable hypothesis, because it applies only to the four cards on the table.) Participants may then override the instructions and interpret the hypothesis as referring to a population from which the four cards are merely a sample. Other factors matter, such as whether participants think that the hypothesis implies its converse, and whether they assume that the hypothesis couldn’t be false because otherwise the experimenter would be dishonest.

 Few of these interpretations apply to the deontic task in which items, not the deontic principle, are evaluated: the items, such as cards describing individuals, are selected independently from one another, the principle doesn’t imply its converse, and the truthfulness of the experimenter is not at issue. But, when a hypothesis has to be evaluated as true or false, many factors affect interpretation. The theory offers no algorithm for the mental processes underlying interpretations or for those that lead from interpretations to selections. The authors instead tabulate some of the possible interpretations that can lead to each of the canonical selections. For example, if participants think there is no point in selecting any further card because p shows that the hypothesis is false, then that is the only card that they select. The theory aims to account for how each particular individual interprets the task on a particular occasion, and, as a result, makes a selection for a particular and perhaps unique reason. The set of factors determining interpretation interact, and the differences among individuals are large, even when they make the same selection.

 The project’s scope is almost boundless; its authors’ persistence is admirable and unsparing; their theory is vast. Yet, it seems mistaken. It is wrong that the selections in the deontic task are independent of one another (see Table 3 in the main text). It is wrong that conditionals in the deontic task never call for the selection of more than two cards (see Politzer & Nguyen-Xuan, 1992, for selections of all four cards). It is wrong that the falsifying selection in the abstract task is a mistake because it depends on treating the conditional hypothesis as a material conditional. Other meanings of conditionals imply that the falsifying selection is correct (see the model theory in the main text, which also demands that reasoners establish that instances of the hypothesis also exist). The theory is informal, and its implementation in a computer program would not be easy. Indeed, its methods may be incompatible with the exercise (though cf. Fugard & Stenning, 2013).

**Reasoning with content-specific rules or modules**

*6. Pragmatic reasoning schemas.* A formal rule of inference is one that contains only logical terms, variables, and punctuation. Any rule that introduces other terms is content-specific, though it may be so only in a minimal way (see, e.g., Hewitt, 1971; Kolodner, 1993). Cheng and Holyoak (1985) used such rules, which they referred to as “pragmatic reasoning schemas,” to explain the selection task (see also Cheng, Holyoak, Nisbett & Olivier, 1986). “Schema” refers to a set of rules supposedly induced (by unknown processes) from everyday experience with causation, permission, and obligation. A schema for the deontic selection task concerns permission, and consists in four rules that are heuristics:

 1. If the action is to be taken, then the precondition must be satisfied.

 2. If the action is not to be taken, then the precondition need not be satisfied.

 3. If the precondition is satisfied, then the action may be taken.

 4. If the precondition is not satisfied, then the action must not be taken.

They depend on context to elicit them. For example, the deontic conditional in the selection task

 If a person is drinking beer then the person must be over 19

elicits the permission schema. Rule 4 implies that if a person is not over 19 then the person must not drink beer, which leads to the selection of the card, not over 19, as a potential counterexample. Other schemas are for causation and for obligation, and the theory postulates that reasoning is typically based, not on logic, but on schemas. So, failure in the selection task implies failure to elicit a relevant schema.

 The schema theory makes predictions beyond the selection task, and experiments have corroborated them (Cheng & Holyoak, 1985). And Cheng, Holyoak, Nisbett, and Oliver (1986) showed that training on the schema for obligation enhanced performance on the selection task, whereas training with formal rules of inference was ineffective. Instruction about the nature of obligations improved performance for various conditionals. These theorists allowed that logical rules of inference could co-exist with schemas, but, that if so, schemas should take priority, and that individuals should fall back either on logic in some cases or on purely superficial strategies, such as matching, in other cases (Cheng et al., 1986).

 The schema above contains conditional assertions, but the theory says nothing about the meanings of conditionals, or about the possibility that these meanings rather than schemas underlie the selection task. The rules in the schema also contain modal auxiliary verbs, such as “may” and “must”, which are ambiguous between a deontic interpretation, as in: “you may smoke,” and an epistemic one, as in “you may get cancer”. A plausible interpretation of rules 1 and 2 in the schema above can be paraphrased in a biconditional, and likewise rules 3 and 4 can be paraphrased in another biconditional, and so the schema reduces to two assertions:

 If and only if the action is to be taken, then the precondition must be satisfied.

 If and only if the precondition is satisfied, then the action may be taken.

Indeed, the whole schema seems equivalent to a single deontic biconditional:

 The precondition is obligatory if and only if the action is permissible.

Cheng and Holyoak might rightly object that such a formulation no longer yields the appropriate selections. But, this objection reminds us that the theory offers no account of how the selection task triggers the appropriate rule in the appropriate schema, or of how this rule leads to a selection. The theory is informal rather than algorithmic (Holyoak, p.c. 1/2/17).

*7. Innate modules.*The paper in which Cosmides (1989) lays out her theory of innate modules has been cited more often than any other on the selection task, even Wason’s original publication of the problem. She makes the bold conjecture that evolutionary adaption has equipped the mind with a set of domain-specific modules for reasoning. Their existence, she claims, can be corroborated in their enhancing effect on reasoning about contents that concern them. An obvious method to test this conjecture is the selection task; and experiments showed the predicted enhancement for “checking for cheaters”, that is, as Cosmides says, for checking whether others are violating the cost-benefit relations of a deontic principle. The theory posits other innate modules, but it is does not list them. And it is a theory that gives an account of *what* one putative module computes, but not of *how* it computes it. So, how such modules work is unknown.

Cosmides is right that different contents in the selection task can elicit different levels of performance. But, many ways exist to explain these differences without the need to postulate innate modules specializing in different topics. Discoveries since Cosmides published her paper challenge her theory, e.g., the large effects of the relevance and accessibility of counterexamples on falsifying selections (Sperber et al., 1995). Its purview is narrow; and granted her proposed module, it is paradoxical that cheats and liars so often prosper in daily life. She claimed that the only hypotheses in the selection task that yield robust and replicable effects are those with contents pertaining to the social contract; critics have pointed out counterexamples (e.g., Cheng & Holyoak, 1989). Other theorists have postulated an additional innate mechanism: the ability to detect genuine altruism (Brown & Moore, 2000). The notion of innate modules has prompted much thought, and much controversy. It remains a provocative and founding exercise in evolutionary psychology, but innate origins are more persuasive for such matters as stereopsis or emotions, that have obvious antecedents in the evolutionary ancestors of humans. The present theory takes for granted that verification of sentences depends on reasoning, and has nothing to say about many of the robust phenomena that occur in the selection task.

**Theories based on heuristics**

*8. The stochastic model.* Evans’s (1977) earliest theory of the selection task postulated that the task should be modeled stochastically. It makes three assumptions:

1. The selections of cards are independent of one another.
2. The probability of selecting a card depends on two tendencies: one is to rely on logic,

and the other is to rely on matching, i.e., to select a card that matches a card referred to in the hypothesis under test.

3. Individuals either reason or match, and so:

 Probability(reasoning) = αLogical-tendency + (1 - α)Matching-tendency

where “α” is the probability of reasoning as opposed to matching. It is fixed for a particular experiment, and tends to be high for everyday generalizations, and so people are more likely to make falsifying selections. There are two other probabilistic parameters, but Evans makes several simplifying assumptions in order to model Evans and Lynch’s (1973) results on the effects of negatives in hypotheses (see the account in the text). The tendency to reason has two parameters: αa for the *if*-clause, and αc for the *then*-clause. The parameter for selecting a matching card has only two values: a match occurs (Rm) or it does not occur (R$\overbar{m}$). So, given a hypothesis, *if p then q*, the probabilities of selecting each of the four cards are as follows:

The card The formula

 p αa + (1 - αa) . Rm

 $\overbar{p} $ (1 - αa) . R$\overbar{m}$

 q (1 - αc) . Rm

 $\overbar{q}$ αc + (1 - αc) . R$\overbar{m}$

The fits with the actual probabilities for selecting each of the four cards are reasonable. Of course, as Evans concedes, the model offers no account of the mental processes that enable individuals to use logic or matching (van Duyne, 1973). Evans recently suggested that this theory might be appropriate to fit to data (p.c. 12/16/16). But, it assumes that the choices of cards in a selection are independent, and, as we show in the main text, this assumption is wrong. And so it cannot model the frequencies of canonical selections. We consider below an alternative theory due to Evans, the heuristic-analytic theory.

*9. Matching and verifying.*Krauth’s (1982) starting point was Evans’s preceding stochastic model, but in analyzing its implementation, he decided that it would fit results better if it also embodied a tendency to verify the hypothesis. So, the model has three states: a tendency to verify the hypothesis, a tendency to falsify the hypothesis, and a tendency to match the hypothesis. If the tendency to verify has a probability of 0, and falsification is equivalent to the tendency to rely on logic, then Krauth’s model is equivalent to the stochastic model. Not all responses, he points out, can be explained by verifying, falsifying, or matching, e.g., the selection of $\overbar{p}$ for a hypothesis such as *if p then q*, which does occur occasionally. He assumes that these “non-logical” tendencies occur in all three states. The theory accordingly has five parameters:

pA: the probability of falsifying the *if*-clause of the hypothesis.

pC: the probability of falsifying the *then*-clause of the hypothesis.

 pV: the probability of verifying.

pM: the probability of matching.

pG: the probability of a non-logical response.

So, the probability of selecting p given the hypothesis *if p then q* is equal to the value of the formula:

pA + pV + pM. (1 – pA – pV)

and, of course, the probability of selecting $\overbar{p}$ in this case = pG. The resulting model yields a better fit than the stochastic model for the data from Evans and Lynch (1973), Manktelow and Evans (1979), and Krauth’s own studies. Like the stochastic model, however, the theory presupposes that the selections of the four cards are independent of one another. It cannot predict the four canonical selections are the ones that occur more often than chance.

*10. The heuristic-analytic theory.* Evans (1984, 1989) formulated a second theory in which heuristics determine which cards are deemed relevant and may be followed up by analysis. The two heuristics are “matching” in which individuals attend to those cards that are referred to in the *then*-clause of the hypothesis, whether or not the clauses are affirmative or negative, and the “if-heuristic” in which individuals attend to true instances of the *if*-clause. These biases explain why so many participants make the falsifying selection, pq, for the hypothesis, *if p then not q*, which we described in the main paper. Analytical reasoning is an option that could occur after the use of the heuristics, but Evans (1984, p. 457) wrote: “I suggest that card selections do not reflect *any* process of reasoning, in the sense of analytic processing, and are due entirely to heuristic processes”.

The theory was revised (Evans, 1998), and Evans wrote: “In spite of appearances, I believe that for most participants, the analytic system is actively engaged on this task” (Evans, 2006, p. 388). And he takes the analytic system to carry out reasoning, which for most individuals selects cards that could verify the hypothesis, and only highly intelligent individuals consider whether cards could falsify the hypothesis.

The theory gives no account of the mechanisms underlying its heuristics or analyses, and, as Evans recognizes, in themselves they seem merely to re-describe results. Why, for example, should matching occur for the *then*-clauses of conditionals but not for their *if*-clauses? Which heuristics apply, say, to selections for a disjunctive hypotheses? Why does the theory postulate that the analytic process is based on reasoning as opposed to knowledge of the meaning of the hypotheses? The answers to these questions are not obvious. The theory applies to various sorts of reasoning and decision-making. It is informal, and Evans has never implemented it in a computer program. However, the next theory is similar and it has been programmed.

*11. The inference-guessing model.* Klauer et al. (2007) proposed a set of related models, including one that postulates that individuals either reason or else use heuristics such as guessing. This model has 10 parameters, which are each the probability that one sort of process occurs rather than another, and so each parameter ranges in value from 0 through 1. The model’s first parameter, *a*, is the probability that the inferential component governs the selection as opposed to the guessing component. The guessing component makes independent selections of the four cards according to four parameters that are the respective probabilities of choosing each card independently as a result of guessing or some other factor such as matching bias.

The inferential component does not use the meanings of conditionals to make selections, but instead the four basic inferences from conditionals (MP, DA, MT, and AC, see our earlier account of Rips’s PSYCOP model). The theory is neutral about the mechanism for reasoning, which, its authors write, could depend on rules, models, or suppositions. Which inferences occur depend on five parameters. The first parameter, *c*, is the probability that the rule, *if p then q*, is interpreted as a conditional as opposed to a biconditional (see Figure 2 in the main text of our paper). The second parameter, *d*, is the probability that the inference is forwards from the *if*-clause (MP or DA) as opposed to backwards from the *then*-clause (MT or AC). Directionality depends at least in part on the formulation of the conditional, e.g., *p only if q* invites more backward inferences than *if p then q*. The third parameter, *x*, affects only the biconditional interpretation. It is the probability that the interpretation is “bidirectional”, *if p then q & if q then p*, as opposed to a “case distinction”, *if p then q & if not-p then not-q*. With the bidirectional interpretation, the distinction between forward and backward inferences does not apply because both are made, but with a case distinction interpretation, the distinction still applies. The fourth parameter, *s*, is the probability that an inference from a conditional or a bidirectional biconditional is a sufficient one as opposed to a necessary one. Normally, *p* is judged sufficient to infer *q* from *if p then q*, but sometimes *p* is judged necessary to infer *q,* as when the conditional is interpreted as stating an enabling condition akin to *only if p then q*. A forward sufficient inference is MP, whereas a forward necessary inference is DA; and backward sufficient inference is AC, whereas a backward necessary inference is MT. The fifth and final parameter, *i*, is the probability that inferences are made only about the visible sides of cards – cards are “irreversible” – as opposed to the invisible sides of cards too, i.e., individuals can envisage possible items on them.

In sum, individuals make selections either from cues about the four individual cards or

else from reasoning about the rule. Cues include guessing, matching, and even perhaps an optimal gain in information. Reasoning yields one or two inferences for either a conditional or biconditional interpretation. Figure 2 (in the main text of our paper) presents a multinomial tree summarizing the possible selections that reasoning yields. For example, the topmost branch of the tree is taken when individuals make a forward and sufficient inference from a conditional interpretation of the rule, taking only the visible side of the card into account. They make the selection, p, because they infer that q should be on its other side. The top branch from the biconditional interpretation yields the selection, pq: individuals infer the selection, q, in addition because a backwards sufficient inference implies that p should be on its other side. The falsifying selection, p$\overbar{q}$, occurs either when individuals make a forward and sufficient inference from a conditional interpretation based on the invisible side of a card, or when individuals make a backward necessary inference (i.e., MT) from a conditional interpretation based on the invisible side of the card.

 As Figure 2 (in the main text) shows, the inferential component yields 11 out of the 16 possible selections. The others can be made from the guessing component of the model. Klauer et al. (2007) fitted the theory to their own results. It provided a better fit of selections of individual cards or pairs of cards than Evans’s stochastic model (see above), his heuristic-analytic model (see above), or a version of the optimal data selection model that yields the probabilities of selecting individual cards (see below for Hattori’s, 2002, version).

The most salient problem with the theory is that its algorithm cannot account for the canonical selections or their statistical reliability. It fails to generate the selection pq$\overbar{q},$ and yet

its inferential component predicts five selections that occur so rarely – below chance in the 228

experiments – that they should be attributed to the guessing component, i.e., the selections: $\overbar{p}, \overbar{q}, \overbar{p}\overbar{q},$ p$\overbar{p}, and $q$\overbar{q}.$ It offers no obvious explanation for other phenomena, such as the dependence of falsifying selections on task, contents, instructions, and framing; or why negative *then*-clauses increase falsifying selections, whereas negative *if*-clauses merely increase entropy. Likewise, it offers no explanation for why individuals focus on the cards to which the rules refer, for why they can have zero, partial, or complete insight into selecting falsifying instances in the repeated version of the task, for why they refer to “truth” in commenting on verifying selections but “falsity” in commenting on falsifying selections, or for why intellectual competence predicts falsifying selections (see the main text for other criticisms). Nonetheless, the model is impressive, because it yields a better fit than the heuristic-analytic theory and the optimal data selection theory. We therefore fit a simplified version of the theory to our test-bed of experimental results (see main text).

**Probabilistic theories**

*12. Probable utility.*Kirby’s (1994) theory of probable utility has precursors in both Cosmides’s (1989) use of costs and benefits in her account (see above), and Manktelow and Over’s (1991) requirement that mental models need to represent utilities in order to account for deontic thinking. Kirby, however, argued that the selection task depends on decision-making and therefore on the subjective expected utility of selections. A participant might realize, for example, that a card could falsify the hypothesis, but refrain from selecting it on the grounds that such an outcome is unlikely (low subjective probability) or unimportant (low utility). So, in line with decision theory, the theory assumes that individuals should select a card if the expected utility of doing so is greater than that of not doing so. There are four possible outcomes from selecting an individual card, which can be categorized as in signal detection theory:

a *hit* selects a card that falsifies the hypothesis,

a *miss* fails to select such a card,

a *false alarm* selects a card that does not falsify the hypothesis,

a *correct rejection* fails to select such a card.

Expected utility depends on the product of probability and utility, and the theory uses the following criterion for selecting a card, where P denotes a subjective probability:

 P(falsifying the hypothesis | card) > Utility(correct rejection) - Utility(false alarm)

 P(not falsifying the hypothesis | card) Utility(hit) - Utility(miss)

The ratio on the left side states the “posterior odds” of a falsification given the selection of a

card. On the assumption that the utilities in the ratio on the right side remain constant, the greater the posterior odds, the more likely a card should be selected. Consider the *if p then q* hypothesis:

If a card has a vowel on one side then it has an even number on the other side.

The set corresponding to p has 5 members, and the set corresponding to $\overbar{p}$ has 21 members; and so the posterior odds of falsifying on the selection of $\overbar{q}$ equals 5/21. However, if utilities are constant and we increase the number of members in the p set relative to the number of members in the $\overbar{p}$ set, e.g., we change the p set to consonants, the posterior odds of falsifying on the selection of $\overbar{q}$ will increase, and so the participants should be more likely to select it. Kirby’s experiments corroborated this prediction, but the effect was on the selection of $\overbar{q}$, not on the selection of p$\overbar{q}$. The theory is a miss for dependent canonical selections.

*13. Optimal data selection, Version 1.*Oaksford and Chater (1994)’s two versions of their theory are briefly described in the main paper, but here we spell them out in detail. They made the radical proposal that the selections of p, pq, and p$\overbar{q}$, were all rational, whether the hypothesis was abstract or an everyday generalization, because for almost all participants the task concerns, not the truth or falsity of the hypothesis, but the statistical dependency that the theorists claim it expresses. Over the years, they clarified their theory (Oaksford & Chater, 1996) and made one major revision to it (Oaksford & Chater, 2003, 2007, Ch. 6), but its constant assumption is that the optimal cards to select are those “that lead to the greatest expected reduction in uncertainty” (Oaksford & Chater, 1996, p. 610).

Uncertainty is measured using Shannon’s (1948) entropy (or informativeness), which we described in the main paper. Here, we abbreviate as I in order to distinguish from hypotheses, H. Given n mutually exclusive hypotheses, the uncertainty of any one of them, *H*i, fits the equation:

 I(Hi) = - P(Hi) log2 P(Hi).

Shannon’s measure was devised for the information conveyed by a signal as a function of its probability, but here it concerns hypotheses: the more improbable a hypothesis, the more information that its corroboration conveys.
 If you were to turn over a card, or to carry out the repeated selection task, you would get information about a hypothesis that would enable you to update its probability according to Bayes’s theorem. The conditional probability of a hypothesis given the data is P(Hi | D), which Bayes’s theorem shows to be:

 P(Hi | D) = $\frac{P\left(Hi\right) P(Hi) }{\sum\_{j=1}^{n}P\left(Hj\right) P(Hj)}$

The gain in information about the *ith* hypothesis from turning a card and receiving information *D* is:

Igain = I(Hi | D) – I(Hi)

Because participants select a card without knowing what is on the other side, they have to calculate the expected gain in information, taking into account the two sorts of possibility on the other side of the card:

 Expected (Igain) = Expected [I(Hi | D) - I(Hi)]

 To illustrate the two statistical hypotheses that a conditional is supposed to concern, we consider the example:

 If a bird is a raven then it is black.

The dependence hypothesis, HD, is that a bird’s being black depends on it being a raven; the

independence hypothesis, HI, is that the two properties are statistically independent of one another. As this contrast implies, Oaksford and Chater reject a deterministic meaning for conditionals, which they take to be the material conditional, and instead treat its meaning as probabilistic (see Adams, 1998). And so a conditional tolerates exceptions, i.e., cases in which a raven (p) is not black ($\overbar{q}$), and the probability of such an exception, ε, equals P($\overbar{q}$ | p). Table SM1 shows the contingency tables for the two hypotheses. The row and column marginals are identical for the two contingency tables. Reasoners seek evidence that provides the most information to discriminate between the two hypotheses. Given a particular card, they compute the expected gain in information from its selection.

Table SM1

*The contingency tables for the dependence and independence hypotheses in Oaksford & Chater’s theory of the selection task for the hypothesis, if p then q, where ε denotes P(*$\overbar{q}$ *| p), and each variable in a cell entry denotes its own probability, so p stands for P(p), and so on (after Oaksford & Wakefield, 2003)*

|  |  |
| --- | --- |
|  Dependence hypothesis: HD |  Independence hypothesis: HI |
|  | Q | $$\overbar{q}$$ | Q | $$\overbar{q}$$ |
| p | p(1 - ε) | ε | Pq | p(1 – q) |
| $$\overbar{p}$$ | q - p(1 - ε) | (1 - q)pε | (1 - p)q | (1 - p)(1 - q) |

 In order to model the results of experiments that do not introduce explicit probabilities, Oaksford and Chater assume that for most conditionals in daily life, such as the one about ravens, the four categories have the following rank order of increasing frequencies:

 p: ravens < q: black things < $\overbar{q}$: non-black things < $\overbar{ p}$: non-ravens.

This order mirrors the decline in the potential informativeness of a card: the smaller its probability, the more informative it is according to Shannon’s metric. Oaksford and Chater (1994, p. 613) showed in a meta-analysis that this order corresponds to the order that tends to occur in 34 experiments using the abstract selection task.

 Oaksford and Chater (1994) extended Pollard’s (1985) analysis of dependence, using his method of Fisher-Yates tests (see above), to show that selections are dependent. In order to model dependence, they assumed that every card, even $\overbar{ p}$, has some probability of being chosen even if it is not informative, and so they added a constant .1 to all expected gains in information. This factor is not a rational one: it is akin to guessing or noise in the system. They then assumed that selections are competitive. They scaled the new expected gain of each card by dividing it by the overall mean value for the four cards. Participants choose cards as a monotonic function of these values. To model the dependency of selections, they carried out the preceding procedure computing expected gain for small increments in P(p) and P(q), but assuming various constraints between them in order to capture the rarity assumption, and the assumption that P(q) is only marginally bigger than P(p). For each computation, they then calculated the rank order correlation between each of the six possible pairs of card (p and $\overbar{ p}$, p and q, etc.). They compared the polarity of these correlations with those in the data, and the agreement in whether the polarity was positive or negative was reliable, e.g., there is a positive correlation in the predictions between p and q, and there was also a positive correlation between them in the data.

 When a conditional hypothesis contains a negative, as in “If there’s an A on one side of a card then there is not a 3 on the other side,” there is a change in the size of the set to which the *then*-clause refers. The clause “there is not a 3” refers to a much larger set than does “there is a 3”, and so individuals should be more likely to select the 3 card, because it yields a greater expected gain in information. One experimental study concluded that both matching and set size had effects (Yama, 2001). Another study made a more explicit manipulation of probabilities, and corroborated set size (Oaksford & Moussakowski, 2004). Neither of the preceding studies manipulated negation in the *if*-clause of the hypothesis. But, Oaksford and Chater (1994) had reported that such negations cause surprising difficulties in comprehension, and so the correlations are weaker. Indeed, these negations reduce the frequency of verifying selections, and increase the entropy of selections. Yet, their effects on expected information gain should be straightforward. Otherwise, optimal data selection accounts for the rank order of the frequencies of selecting the four cards, and the polarity of the correlations between the 6 possible pairs of cards. The theory relies on three parameters: the prior probability of the independence model, P(p), and P(q). We point out three drawbacks to the theory in the main text of our paper.

 *14. Optimal information gain, version 2.* The previous theory did not make quantitative predictions until Hattori (2002) developed his version of it. He used independent estimates of the subjective probabilities of p and of q to calculate the expected information gain and its scale over the four cards. He derived a “logistic selection tendency” function that transforms this scale into numerical predictions of the independent frequencies with which participants select each card. The function depends on two “semi-fixed” parameters (Hattori, 2002, p. 1252). Hattori’s empirical studies led him to several major conclusions. The subjective probabilities of pand q in the abstract selection task are usually roughly equal. When they deviate from equality, results from the selection task tend not to support his model. The way in which experiments present probabilistic information is important: it is better to use “natural sampling” (Gigerenzer & Hoffrage, 1995) than merely to present direct statements of probabilities. Experiments that emphasize probabilities affect performance in the selection task. Participants differ in their strategies, some are sensitive to probabilities and their results fit his model, but others are not – in fact, the majority of participants in his studies. Oaksford and Chater (2003, 2007, p. 172) have endorsed this revision to their theory (see also Oaksford & Wakefield, 2003, who adopted naturally sampling in an experiment). There is a cost, however. As Hattori (2002, p. 1262) points out, the new model implies that the selections of individual cards are independent of one another. So, the theory of optimal data selection has changed from making rank order predictions about the frequencies of the four cards, and binary predictions about positive or negative correlations between pairs of cards, to making independent numerical predictions about the probabilities of the four cards.

 The optimal data selection theory has provoked a variety of reactions. It has elicited criticisms that it is mistaken in it normative assumptions (Evans & Over, 1996), in its Bayesian presuppositions (Laming, 1996), in its descriptions of results (Evans & Over, 1996), and in the adequacy as a theory (Almor & Sloman, 1996). Its authors replied to these criticisms (Oaksford & Chater, 1996). Their theory is a brilliant integration of Bayes theorem and Shannon’s entropy, and it has addressed most of the phenomena of the selection task. Its recent assumption that the four cards are selected independently is not a central tenet of their theory but one made for a useful index of fit (Oaksford, p.c., 1-20-2017). One of the advantages of the theory is that it is a special case of a general Bayesian approach to cognition (see, e.g., Oaksford & Hall, 2016). One of its proponents has recently argued that brains need not represent or calculate probabilities at all, and are poorly adapted to do so (Sanborn & Chater, 2016). Estimates instead call for the sampling of information. And the decision of whether or not to select a card, likewise depends on sampling the possible outcomes of turning a card, and on sampling their informativeness (Chater, p.c., 1-22-2017). We point out the drawbacks of this theory in the main text.

**Neural models**

 *15. Parallel distributed processing.* Leighton and Dawson (2001) devised three separate parallel distributed networks to carry out the selection task. The networks contained 4 input units – one for each card, 16 output units, and the number of hidden units depended on the selection. A standard backwards propagation of error algorithm trained three such networks. The one for selecting p required 3 hidden units, but those for selecting pq and p$\overbar{q}$ were able to learn only with 8 hidden units. The principal challenge was to model the selections rather than to formulate a full-fledged account of the frequencies of selections. The account is not intended to be complete, but, contrary to the complexity of the networks, selections of pq are more frequent than selections of p.

 *16. Models of real neurons.* Eliasmith (2005) met the challenge of representing complex structures, such as conditional hypotheses, in a system modeling real neurons that compute distributed vectors. His model consists of ten interconnected populations of neurons in a total population of around 20,000. It mimics performance in with abstract and with everyday hypotheses. It compares the resulting vectors with ones encoding their typical selections. Hence, the network learns to infer the common verifying selection, pq, for the abstract task, and the rarer falsifying selection, p$\overbar{q}, $for the everyday generalization. Like the parallel-distributed model, it is important as a demonstration that such models can produce contrasting selections. But, it is not intended to be a complete account of the selection task.

**References (not in the paper itself)**

Adams, E. W. (1998). *A primer of probability logic*. Stanford, CA: Center for the Study of

 Language and Information.

Cheng, P. W., Holyoak, K. J., Nisbett, R. E., & Oliver, L. M. (1986). Pragmatic versus syntactic

 approaches to training deductive reasoning. *Cognitive Psychology*, *18*, 293-328.

Fugard, A. J. B., & Stenning, K. (2013). Statistical models as cognitive models

 of individual differences in reasoning. *Argument & Computation*, *4*, 89-102.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without

 instruction: Frequency formats. *Psychological Review*, *102*, 684-704.

Grice, P. (1975). Logic and conversation. In Cole, P. & Morgan, J. (Eds) *Syntax and Semantics*

 *Vol. 3: Speech Acts*. Pp. 41-58. New York: Academic Press.

Hewitt, C. (1971). Procedural embedding of knowledge in Planner. *International Joint*

 *Conference on Artificial Intelligence*, *1*, 167-184.

Kolodner, J. (1993). *Case-Based reasoning*. San Mateo: Morgan Kaufmann.

Oaksford, M., & Moussakowski, M. (2004). Negations and natural sampling in data selection:

 Ecological versus heuristic explanations of matching bias. *Memory & cognition*, *32*, 570-

 581.

Osherson, D. N. (1976). *Logical abilities in children, Vol. 4*. Hillsdale, NJ: Erlbaum.

Smalley, N. S. (1974). Evaluating a rule against possible instances. *British Journal of*

 *Psychology*, *65*, 293-304A.

van Duyne, P. C. (1973). A short note on Evans' criticism of reasoning experiments and his

 matching response hypothesis. *Cognition,* *2*, 239-242.

Wason, P. C., & Green, D. W. (1984). Reasoning and mental representation. *Quarterly Journal*

 *of Experimental Psychology*, *36*, 597-610.

Yama, H. (2001). Matching versus optimal data selection in the Wason selection task. *Thinking*

 *& Reasoning*, *7*, 295-311.