PNAS RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES

Human verifications: Computable with truth values outside logic

Philip N. Johnson-Laird[a,b,1], Ruth M. J. Byrne[c], and Sangeet S. Khemlani[d]

Contributed by Philip N. Johnson-Laird; received June 21, 2023; accepted August 23, 2023; reviewed by Maya Bar-Hillel and Philipp Koralus

Cognitive scientists treat verification as a computation in which descriptions that match the relevant situation are true, but otherwise false. The claim is controversial: The logician Gödel and the physicist Penrose have argued that human verifications are not computable. In contrast, the theory of mental models treats verification as computable, but the two truth values of standard logics, *true* and *false*, as insufficient. Three online experiments (n = 208) examined participants' verifications of disjunctive assertions about a location of an individual or a journey, such as: 'You arrived at Exeter or Perth'. The results showed that their verifications depended on observation of a match with one of the locations but also on the status of other locations (Experiment 1). Likewise, when they reached one destination and the alternative one was impossible, their use of the truth value: *could be true and could be false* increased (Experiment 2). And, when they reached one destination and the only alternative one was possible, they used the truth value, *true and it couldn't have been false*, and when the alternative one was impossible, they used the truth value: *true but it could have been false* (Experiment 3). These truth values and those for falsity embody counterfactuals. We implemented a computer program that constructs models of disjunctions, represents possible destinations, and verifies the disjunctions using the truth values in our experiments. Whether an awareness of a verification's outcome is computable remains an open question.

computability | counterfactuals | mental models | logic | truth

Significance

Logic treats sentences as either true or false, and Gödel's proof that certain true sentences are unprovable in any consistent logic for elementary arithmetic has led some theorists to argue that verification is not computable. Our experimental results show that people use a richer set of truth values than the binary pair of logic, *true* and *false*, and that human verifications can rely both on matching a description with an observation and also on imagining what might have happened but did not. Our computer simulation of the theory delivered correct truth values. So, these verifications are computable, but no-one knows if awareness of their truth values, or of outcomes of other cognitive processes, is also computable.

In 1972, the late Sydney Brenner gave a talk at the Princeton Institute for Advanced Study in which he argued that biological processes are algorithmic (1). Afterward, the logician Kurt Gödel announced that the talk showed what he had long believed: 'Vitalism is correct'. The remark seemed like a grotesque misunderstanding; Brenner did not reply. Yet, the status of human verification is a repercussion of Gödel's famous incompleteness proof. It shows that a self-referential sentence asserting its own unprovability is indeed unprovable in any consistent formal system equivalent to an algorithm for elementary arithmetic (2). Some humans can grasp the truth of the self-referential sentence and even why it is true. So their ability to verify assertions seems to go beyond what algorithms, even biological ones, can do. That's why Gödel argued for vitalism. Others have drawn analogous conclusions, notably Roger Penrose, who argued from a similar basis that awareness of the results of verifications is not computable and calls for a new physics (3). Of course, most verifications are straightforward as psychologists know from many experiments (4–8) even including studies of brain activity (9–11). The main surprise was that individuals are faster to determine that an affirmative assertion is true rather than false whereas they are faster to determine that a negative assertion is false rather than true (12). An explanation of this interaction (5, 13–17) is that an affirmative assertion, 'It is the case that the car arrived at Perth,' is true in case the destination in the predicate matches the car's destination, and false if it does not. A negative assertion, 'It is not the case that the car arrived at Perth' is false in case its predicate matches the car's destination, but true if it does not. Since negative assertions are harder to understand than affirmative assertions, and mismatches are harder to process than matches, the observed interaction follows. All these studies are compatible with the treatment of verification in standard logics (see below). In what follows, we report results showing that humans verify assertions in ways outside these logics, but that are nonetheless algorithmic. So, too, could be their grasp of the truth of Gödel's self-referential sentence.

Standard logics are the sentential calculus, which concerns idealized counterparts of *not*, *and*, *or*, and *if* (18), and all logics that include this calculus, e.g., a countable infinity of modal logics which concern possibilities (19), and the system that Gödel used for his proof, which includes axioms for arithmetic (2). They have only two truth values, *true* and *false*. For example, consider this assertion, which concerns a car journey along one of two divergent roads:

Author affiliations: [a]Department of Psychology, Princeton University, Princeton, NJ 08544; [b]Department of Psychology, New York University, New York, NY 10003; [c]School of Psychology and Institute of Neuroscience, Trinity College Dublin, University of Dublin, Dublin 2, Ireland; and [d]Navy Center for Applied Research in Artificial Intelligence, US Naval Research Laboratory, Washington, DC 20375

Author contributions: P.N.J.-L., R.M.J.B., and S.S.K. designed research; R.M.J.B. and S.S.K. performed research; R.M.J.B. and S.S.K. analyzed data; P.N.J.-L. wrote the programs; and P.N.J.-L., R.M.J.B., and S.S.K. wrote the paper.

Reviewers: M.B.-H., The Hebrew University of Jerusalem; and P.K., Oxford University.

The authors declare no competing interest.

[1]To whom correspondence may be addressed. Email: phil@princeton.edu.

Published September 25, 2023.

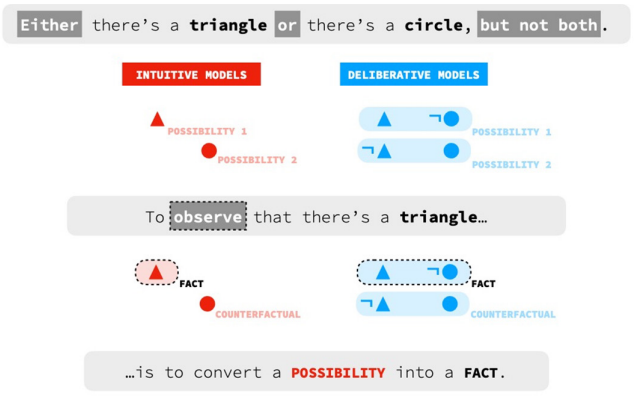

Fig. 1. Intuitive (red) and deliberative (blue) mental models of an exclusive disjunction. Each set of models represents a conjunction of possibilities, which each hold in default of knowledge to the contrary. Intuitive models, which are constructed rapidly without working memory for results of intermediate computations, represent only those clauses in the disjunction that are true in a possibility. Deliberative models, which can access working memory, represent in addition what is false, using negation (symbolized as '¬') to do so. The effect of the observation of a possibility is to change it into a fact, and to change the other possibility into a counterfactual possibility: One asserted to have been once possible but that did not happen (25).

You arrived at Exeter or at Perth.

Granted you cannot have arrived at both cities, the disjunction is 'exclusive': It is true if you arrived at one of the two cities; otherwise it is false.

The theory of mental models—the model theory, for short—has had a long development (20–23) leading to an alternative account of truth values. When individuals understand an assertion, they have in their minds, not truth values, but mental models of the possibilities to which the assertion refers. The models have the same structure, insofar as they can, as the situations that they represent. Each model is of a possibility holding in default of information to the contrary, though at least one possibility must hold for the assertion to be true (24–26). Fig. 1 presents a simple exclusive disjunction, and illustrates the difference between models that intuition and more thoughtful deliberation yield. They are simulated in a computational version of a 'dual system' conjecture due to Wason (27–30). In the model theory, intuitions have no access to the results of intermediate computations, whereas deliberations have free access to a working memory for them (the Lisp source code of the program, mSentential, is at https://www.modeltheory.org/models/). The effect of verifying one of the possibilities of the *Exeter or Perth* disjunction is that it becomes a fact, and the alternative possibility becomes a counterfactual (Fig. 1), which if it is true refers to a situation that was once possible but that did not occur (25, 31–33).

In standard modal logics, the meanings of possibilities are almost always treated in terms of 'possible worlds'. The assertion, 'you may have arrived at Perth,' is true if 'you arrived at Perth' is true in at least one relevant possible world, otherwise the assertion is false (34). Each possible world determines the truth values of all assertions about worlds to which it is relevant (aka 'accessible'), and so it is too vast for a human brain to contain (35, 36). In the model theory, the brain builds a small intuitive model of your arrival at Perth, and allows for an alternative model in which you do not arrive there. A disjunctive assertion is intuitively true if one of its models holds in (a model of) the situation. You arrived at Perth, and so the disjunction is true. Deliberation can verify counterfactual possibilities. It starts with a model of the facts of the journey, undoes the car at its destination, and simulates its counterfactual journey down the other road. If the road is open, the car arrives at Exeter, and the counterfactual possibility is true. If the road is closed, the car cannot arrive there, and the counterfactual possibility is false. Individuals might make imaginative simulations in which, say, someone removes a barrier blocking the road. But, according to the theory, the simulation of counterfactuals makes no unnecessary changes in undoing the car at its destination (ref. 31; see also ref. 37).

In sum, when one clause of a disjunction is true and the other refers to a counterfactual possibility that is true, the disjunction is certain to be true; likewise, if both clauses are false, it is certain to be false. But, when one clause is true and the other clause is false, certainty about the disjunction's truth value should decrease. How this uncertainty manifests itself depends on the verification procedure. Our studies examined three procedures, which introduced truth values outside standard logic.

Results

Experiment 1. This experiment (n = 48) established the impact of facts and their counterfactual alternatives on the verification of rival pairs of disjunctions, e.g.:

John says: Bill is in Dublin or London.
Mary says: No, Bill is in Belfast or Paris.
In fact, Bill is in Dublin, but otherwise he would have been in Paris.
Who is right? 1) *Mary.* 2) *John.* 3) *Both of them.* 4) *Neither of them.*

As the model theory predicts, when the fact verified a clause in a disjunction, participants judged its speaker to be right, i.e., John in the example above. But, the counterfactual in the description of the outcome—'otherwise he would have been in Paris' in the example—also affected their verifications. In all and only those cases in which the counterfactual predicted a different judgment from the fact, it led to a small but reliable number of judgments that the speaker who referred to this counterfactual possibility was right (Fig. 2). In all 14 problems, the fact predicted the most frequent judgment (Binomial test, prior probability of 0.25, $P < 0.25^{14}$). Yet, 92% of judgments matched the fact when the counterfactual predicted the same judgment, whereas only 77% of judgments matched the fact when the counterfactual predicted a different judgment (Wilcoxon test, $z = 3.3$, $P < 0.0005$, $r = 0.48$). The only unpredicted responses were judgments that neither speaker was right when the counterfactuals matched no clauses in their disjunctions.

Experiment 2. In both the present experiment (n = 78) and the next one, the participants on each trial saw a picture of the end of one of four sorts of journey (Fig. 3 *A*–*D*). Their car had taken one of two diverging roads to arrive either at a city, or else at a barrier that had prevented it from getting there, and the road to the alternative city was likewise open or else a barrier blocked it. Their task was to verify a single disjunctive assertion, such as: *You arrived at Exeter or at Perth*, in relation to the picture. They chose whichever of three truth values was appropriate: *false*, *possibly true and possibly false*, and *true*. The intermediate truth value is outside standard logic, and we assigned a rank order of truthfulness scores from falsity to truth of: 0, 0.5, and 1. The distributions of the participants' choices (Fig. 4) showed a highly reliable trend in their truthfulness scores over the four sorts of journey (Page's trend test, $z = 11.71$, $P < 0.1^{7}$), e.g., the percentages of *True* evaluations over the four sorts of journey (A, B, C, D) were 79%, 70%, 10%, and 0%, and the trend for *False*, was opposite, though not wholly independent: 8%, 14%, 72%, and 99%. The overall trend corroborated the prediction that when the fact and the counterfactual yield opposite truth values for a disjunction, a small but reliable number of evaluations switched from the

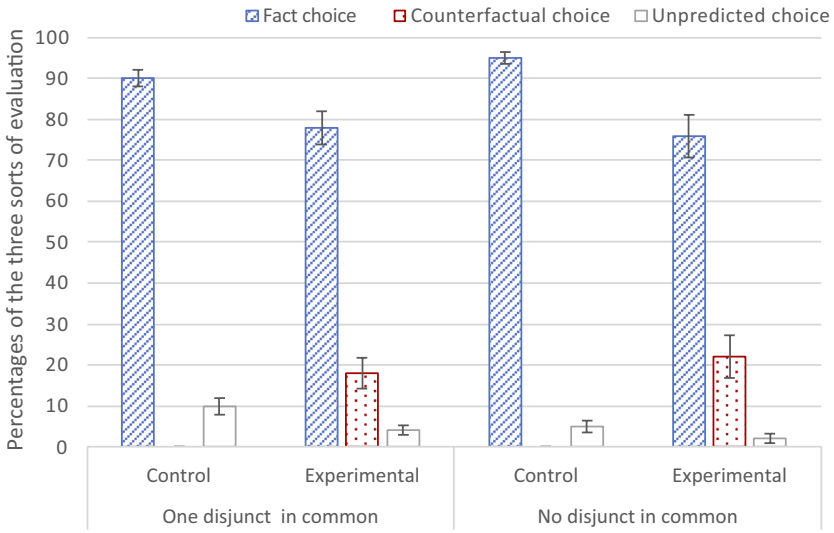

Fig. 2. Experiment 1 (n = 48) in which each problem had an outcome describing a fact and a counterfactual possibility, such as: 'In fact, Bill is in Dublin, but otherwise he would have been in Paris'. Two speakers asserted rival disjunctions, and the figure presents the percentages of three sorts of participants' evaluations of which speaker was right in asserting one of the rival disjunctions: 1) Evaluations that the speaker whose disjunction referred to the actual location of a person was right (blue bars), 2) Evaluations that the speaker whose disjunction referred to the counterfactual location of a person was right (red bars), and 3) Unpredicted errors that neither speaker was right when the counterfactuals matched no clauses in their disjunctions. In the control trials, the fact and the counterfactual led to the same choice, and in experimental cases they led to different choices. Participants had two trials for each of eight pairs of disjunctions with one disjunct in common, and six pairs of disjunctions with no disjunct in common. Error bars are SEM.

fact's truth value to either the intermediate truth value or to the opposite truth value. As Fig. 4 also shows, the percentages for the intermediate truth value tended to be larger when the fact and the counterfactual had different truth values (17%) than when they had the same truth values (7%; Wilcoxon test, $z = 3.25$, $P < 0.0006$, Cliff's $\delta = 0.13$). Experiment 2 therefore showed that counterfactuals biased evaluations of truth values even when nothing cued them—not even the set of truth values from which participants chose their options. We infer that the participants made a spontaneous verification of counterfactuals.

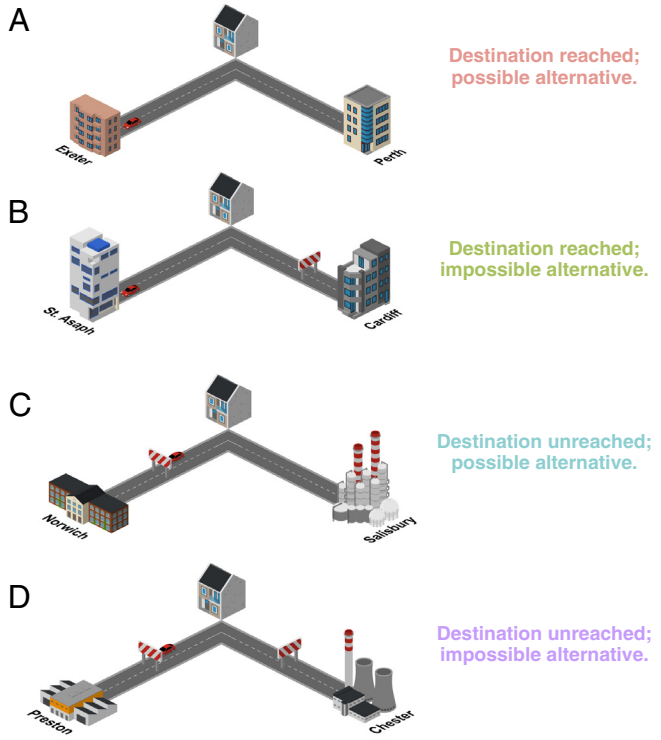

Fig. 3. The pictures of four sorts of journey used in Experiments 2 and 3. The small red car arrived at a city, or else a barrier made its journey impossible, and the other city was a possible destination, or else a barrier made it impossible. The descriptions in color (not shown to the participants) summarize the four sorts of journey.

Experiment 3. Experiment 3 (n = 82) used the same setup as the previous experiment, but a set of four truth values using counterfactuals. We assigned truthfulness scores from falsity to truth as follows:

0 *False and it couldn't have been true.*
1 *False but it could have been true.*
2 *True but it could have been false.*
3 *True and it couldn't have been false.*

These truth values are outside standard logics, because counterfactuality is built into their meanings. The distributions of the participants' verifications (Fig. 5) showed a highly reliable trend in truthfulness scores (Page's trend test, $z = 13.4$, $P < 0.1^7$), e.g., the percentages of *True and it couldn't have been false* evaluations over the four sorts of journey (a through d in Fig. 3) were 77%, 35%, 10%, and 0%, and the trend for *False and it couldn't have been true*, was opposite, though not wholly independent: 1%, 10%, 18%, and 95%.

Computer Programs for Verification. Robots equipped with perceptual, motor, and linguistic organs could carry out verifications (38). A computer program without access to such organs can formulate its own descriptions for a domain that it simulates, and carry out an algorithm that determines whether each description is true or false for any member of the domain. If the domain has a counterpart in the real world, then its output can be helpful to its human users provided that they can check that its algorithm is correct. We refer to such programs as carrying out a *verification* algorithm. Computer programs containing such algorithms exist. We implemented a verification algorithm in a program simulating the model theory of how individuals reason, which verifies various sorts of assertion. Here is a simple example using the disjunction:

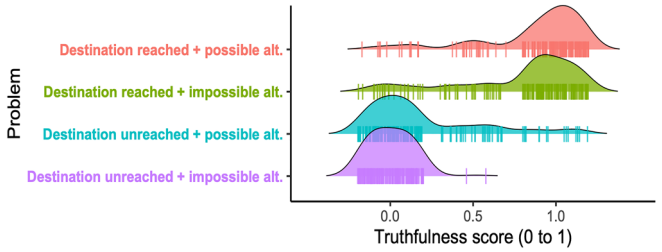

Fig. 4. The distributions of the disjunctions' mean ranks of truthfulness scores in Experiment 2 (n = 78) for four sorts of journey summarized on the *y* axis where 'alt' abbreviates 'alternative destination' (see Fig. 3 *A*–*D* for these journeys). The participants selected one of three truth values for each journey shown here with the truthfulness scores, which the participants did not see: 0 *False*; 0.5 *Possibly true and possibly false*; and 1 *True*. The curves plot the numbers of participants at each mean score both as areas under a curve, and as vertical bars representing participants according to their mean scores on the *x* axis (with a small random perturbation to prevent superpositions in the graphs).

You arrived at Preston or at Chester.

The verification algorithm constructs two intuitive models of your possible destinations according to the disjunction. Each model is shown here on a separate line, with the name of its possible destination:

Preston

Chester

The algorithm then builds a model of the town that you drove to, and any other destinations relevant to the disjunction. In our experiments, there is only one alternative destination, and so a typical model of the towns is akin to a map showing the end of your journey:

(You/ Preston) (/ Chester)

where the slashes denote barriers that prevent you reaching the destination and its alternative. So this model represents the journey in Fig. 3*D* above. You set out for Preston but could not reach it because of the barrier, and you could not have reached Chester either because of its barrier. The program compares the two sets of models—for the disjunction and for your journey—and yields the truth value:

False and it couldn't have been true.

The program copes with a wider variety of verifications than occurred in Experiment 3: it deals with those in Experiment 2, and with verifications concerning possible destinations out of a larger set, yielding truth values, such as it could be true and those

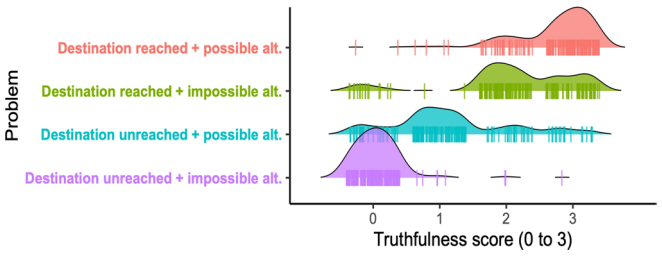

Fig. 5. The distributions of the disjunctions' mean ranks of truthfulness scores in Experiment 3 (n = 78) for four sorts of journey summarized on the *y* axis where 'alt' abbreviates 'alternative destination' (see Fig. 3 *A*–*D* for these journeys). The participants selected one of four truth values for each journey shown here with the truthfulness scores, which the participants did not see: 0 *False and it couldn't have been true*; 1 *False but it could have been true;* 2 *True but it could have been false;* and 3 *True and it couldn't have been false*. The curves plot the numbers of participants at each mean score both as areas under a curve, and as vertical bars representing participants on the *x* axis according to their mean scores (with a small random perturbation to prevent superpositions in the graphs).

referring to the probability of truth. Its Lisp code is at https://www.modeltheory.org/models/.

One of our earlier programs, mAbducer, creates its own algorithms for making combinatorial rearrangements of the order of items, i.e., cars in a train (39), e.g., to reverse their order, to carry out a Faro shuffle of their order, and so on, regardless of the number of cars. The user gives it two input–output examples of a target rearrangement of different lengths of trains, and it infers the recursive loops of actions required for their algorithms. It constructs them in Lisp, which it then translates into simple English. To ensure that a program is correct, the program verifies it for trains of any tractable length. Fig. 6 illustrates the program with an example of its verification algorithm that checks its program to reverse the order of the cars in a train of any length. It also illustrates the parallel between this verification algorithm and one for Gödel's theorem (2).

Discussion

Standard logics, which we defined at the outset, are a product of human thinking (21, 40). They have only two truth values, *true* and *false*, which occur in a 'metalanguage' to formulate the semantics of the main language for proofs (18). No such segregation occurs in natural languages, and so assertions can have inconsistent meanings, such as the well-known 'liar' paradox:

This sentence is false.

If this self-referential sentence is true then it follows from the sentence itself that it is false, and vice versa. Such inconsistencies in a standard logic are catastrophic: any conclusion whatsoever follows validly from them. Hence, the logician Tarski argued that natural languages should not be used for science (41). We define three main sorts of truth value that they contain: the *elementary* values of standard logics, 'true' and 'false'; *modal* values, such as 'possibly true and possibly false'; and *counterfactual* values, such as 'true and it could not have been false'. All three sorts can occur in probabilistic assertions, such as 'probably true and could not have been false'. They can have explicit numerical values in numerate cultures (refs. 42 and 43). Our participants coped well with instances of all three sorts of truth value, including those that are not in standard logics. They based their verifications on observable facts, but they evaluated counterfactuals (Experiment 1), and they did so even when nothing cued their use (Experiment 2). Many counterfactuals cannot be verified, e.g., 'If the Viennese were three-legged they would march in waltz time'. The model theory postulates that those that are verifiable elicit a mental simulation in three steps:

Step 1 starts with a model of the facts—you arrived at Exeter (Fig. 3*A*).

Step 2 modifies this model to accommodate the counterfactual possibility—it removes you from your destination at Exeter.

Step 3 tries to simulate the counterfactual—you journey to the alternative destination of Perth. If it is possible, the counterfactual is true; if it is impossible the counterfactual is false. So, the truth or falsity of a counterfactual can enter into the truth value of assertions.

When an inference flouts standard logics, one defense of them appeals to pragmatics, such as the conventions of discourse (refs. 44–46, cf. ref. 47). But pragmatics can hardly introduce a new sort of truth value into a logic, because it would call for additions to its grammar and semantics. Likewise, one of the counterfactual truth values in Experiment 3 referred to truths that are necessary—they could not be false (see also ref. 32) contrary to an influential view (48) that they are justified only if they depend on logic, as in the tautology: 'Either she arrived at Exeter or she did not'. Of course nonstandard logics may contain nonstandard truth

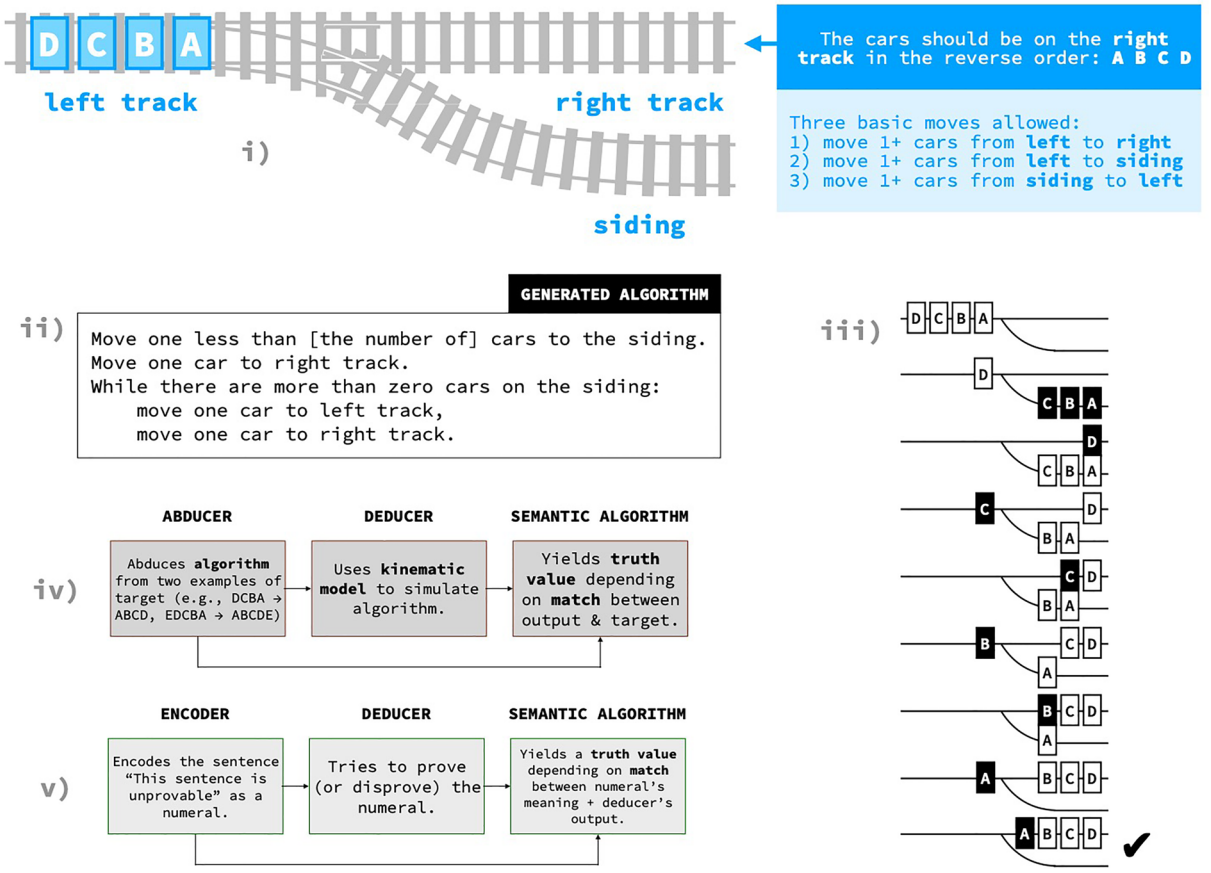

Fig. 6. Two programs with verification algorithms. (*i*) The initial state of the railway, an illustrative target rearrangement, and the three legal moves. (*ii*) The algorithm that the mAbducer program creates to rearrange the cars in trains, which it translates into simple English. (*iii*) The effects of the rearrangement program on the order of the cars, and its verification as true, on which both the left end of the track and the siding function as stack-like working memories (*iv*) The structure of the algorithm for verifying algorithms that the rearrangement program creates (39). (*v*) The structure of a verification algorithm for Gödel's first incompleteness theorem (2).

values, cf. intuitionistic logics, specialized modal logics, paraconsistent logics. The chatbot Bard uses the modal truth value, 'possible'. We have searched and not found as yet any system that uses counterfactual truth values. In any case, our first conclusion is that standard logics cannot underlie everyday discourse, which makes free use of complex truth values that are outside them.

Not all humans can verify assertions; and not all assertions can be verified. And no finite observations, such as our participants' and programs' performances, can establish that all human verifications are computable. The immediate crux, however, is the computability of human verifications of the truth of Gödel's self-referential sentence. It is akin to the liar paradox though not paradoxical:

This sentence is unprovable.

And it is unprovable in a system of formal proofs for the elementary arithmetic of natural numbers—a system that has independent proofs of its consistency. Readers of this article are likely to judge it to be true, though the formal system for proofs cannot make this evaluation. Four factors may make the judgment of its truth appear uncomputable. First, it cannot occur in a formal system. But, as we have illustrated, verification algorithms can be computable, and in principle they can use perceptible inputs. Even a great mathematician has distinguished between 'official' formal proofs and informal ones that rely on meanings (49). Second, the circularity of self-reference may seem uncomputable. In fact, many forms of self-reference are virtuous, not least the definition in a computer program of a function that calls itself. Third, the incompleteness proof is so complex that many people are unable to understand it, and so it may be beyond the power of a theorem-proving program. This factor is irrelevant. You can verify 'Clicking this icon stops the computation' without understanding why it does so. Likewise, you can verify Gödel's sentence without understanding its proof. Fourth, Gödel's self-referential sentence is encoded in a vast natural number, and you cannot carry out the computation to decode its meaning. Tractability aside, a computer program can make this computation. But, again, it is irrelevant. You verify the meaning of the sentence, not its 'official' translation. A verification algorithm should be able to do so too (Fig. 6). Our second conclusion is therefore that the human ability to verify Gödel's self-referential sentence may not depend on vitalism or on a process that is not algorithmic. Are computers aware of what they have accomplished in verifying a description? No. This qualification applies to any of their programs—from calculating a payroll to simulating gravitational waves. Awareness relies at least on access to a model of oneself (21, 50), but whether other of its components are computable remains an open question.

The main sources of factual truths—observations and witnesses—are fallible, as are inferences from the information they provide. So, too, are AI systems such as GPT-4, which do not verify their assertions. If our thesis is correct, their devisers cannot ignore this task on the grounds that no algorithm can make verifications. Human observations concern external physical and social situations, and internal matters such as bodily feelings and mental states. Yet, as we have shown, verification can consider counterfactual possibilities. Truth can therefore depend on something that did not happen and cannot be observed—the viability of the road not taken.

Materials and Methods

The experiments received prior approval from the ethics committee of Trinity College Dublin and the Naval Research Laboratory. The participants carried out the experiments on-line, they gave their consent to take part in the experiment, and the instructions made clear that they could withdraw at any point. The materials and raw data for all three experiments are accessible at https://osf.io/2wtc6/. The preregistration for the experiments, the analysis scripts, materials, and experimental code are also accessible there. We used SPSS and the R statistical analysis software to compute all the (nonparametric) statistical tests.

Experiment 1. The participants carried out verifications of 14 pairs of disjunctions. Eight pairs had a clause in common, and six pairs did not have a clause in common (Fig. 2). Each outcome described a fact and a counterfactual possibility, referring to both disjunctions, to one disjunction, or to neither disjunction. For half the problems, the counterfactual possibility led to the same conclusion as the fact (control problems); for the other half, the counterfactual led to a different conclusion from the fact (experimental problems). The initial 51 participants were recruited from the general public on the Prolific website and paid 2 pounds sterling for their participation in the experiment, and their ages ranged from 18 to 67 y. Their number was in accordance with a prior analysis of the power needed to detect a significant effect. We omitted the data of three of them from statistical analysis, because they had failed one or both online attention checks. The materials were exclusive disjunctions using well-known names of cities, which were assigned at random to the 14 pairs of disjunctions. The trials were presented in a different random order to each participant.

Experiment 2. Each trial started with a schematic picture of a destination of a journey (Fig. 3). The participants verified a disjunction such as: *You arrived at Exeter or Perth,* in relation to the picture. The experiment manipulated whether a picture showed that the car had arrived at a city or not, and whether the road to the alternative destination was open or not. The participants were told to judge the disjunctive description by choosing the best evaluation from three options: *false*, *possibly true and possibly false*, and *true*. The 95 participants were recruited on the Cloud Research platform and paid $2.50 USD for their participation in the experiment, and their ages ranged from 24 to 72 y. Their number was in accordance with a prior analysis of the power needed to detect a significant effect. We omitted 17 participants' data from statistical analysis, because they had failed an online attention check. But the predicted trend was also significant over the entire sample. Each participant was tested with the four problems in a different random order. An algorithm categorized participants on the three-point truthfulness scale (0, 0.5, 1), and introduced a small random perturbation to prevent overlaps in Fig. 4.

Experiment 3. The design and procedure were the same as those for the previous experiment except that there were four different truth values concerning counterfactual possibilities. The Cloud Research platform recruited 95 members of the general public, ranging in age from 22 to 69 y. We omitted 13 participants' data from statistical analysis, because they failed an online attention check, but the predicted trend was also significant over the entire sample.

Data, Materials, and Software Availability. Experimental results data have been deposited in OSF (https://osf.io/2wtc6/) (51).

ACKNOWLEDGMENTS. We thank Geoff Goodwin, Mark Keane, Cristina Quelhas, Marco Ragni, and Célia Rasga, for their help and advice. We are grateful to Maya Bar-Hillel and Philipp Koralus for their stringent critiques of an earlier draft. We also thank the members of the CONCATS seminar in the Department of Psychology at NYU for their comments on a presentation of some of these results in December 2021. The research was funded in part by a grant from the US Naval Research Laboratory.

1. S. Brenner, Computers and the biological sciences (1972). https://albert.ias.edu/handle/20.500.12111/2684. Accessed 9 September 2023.
2. K. Gödel, "On formally undecidable propositions of Principia Mathematica and related systems I" (Trans., E. Mendelson, of original publication in 1931) in *The Undecidable*, M. Davis, Ed. (The Raven Press, 1965), pp. 5–38.
3. R. Penrose, *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics* (Oxford University Press, 2016).
4. P. C. Wason, Response to affirmative and negative binary statements. *Brit. J. Psychol.* **52**, 133–142 (1961).
5. H. H. Clark, W. G. Chase, On the process of comparing sentences against pictures. *Cognit. Psychol.* **3**, 472–517 (1972).
6. S. Khemlani, C. Wasylyshyn, G. Briggs, P. Bello, Mental models and omissive causation. *Mem. Cognit.* **46**, 1344–1359 (2018).
7. G. Agmon, Y. Loewenstein, Y. Grodzinsky, Negative sentences exhibit a sustained effect in delayed verification tasks. *J. Exp. Psychol. Learn.* **48**, 122–141 (2022).
8. N. Skovgaard-Olsen, P. Collins, K. C. Klaier, Possible worlds truth table task. *Cognition*, in press (2023).
9. E. D. Reichle, P. A. Carpenter, M. A. Just, The neural bases of strategy and skill in sentence–picture verification. *Cognit. Psychol.* **40**, 261–295 (2000).
10. J. Lüdtke, C. K. Friedrich, M. De Filippis, B. Kaup, Event-related potential correlates of negation in a sentence–picture verification paradigm. *J. Cognit. Neurosci.* **20**, 1355–1370 (2008).
11. H. S. Bremnes, J. Szymanik, G. Baggio, Computational complexity explains neural differences in quantifier verification. *Cognition* **223**, 105013 (2022).
12. P. C. Wason, S. Jones, Negatives: Denotation and connotation. *Brit. J. Psychol.* **54**, 299–307 (1963).
13. H. H. Clark, W. G. Chase, Perceptual coding strategies in the formation and verification of descriptions. *Mem. Cognit.* **2**, 101–111 (1974).
14. P. A. Carpenter, M. A. Just, Sentence comprehension: A psycholinguistic processing model of verification. *Psychol. Rev.* **82**, 45–73 (1975).
15. H. H. Clark, *Semantics and Comprehension* (De Gruyter Mouton, 2019).
16. R. Dale, N. D. Duran, The cognitive dynamics of negated sentence verification. *Cognit. Sci.* **35**, 983–996 (2011).
17. I.-A. Tan, N. Kugler-Etinger, Y. Grodzinsky, Do two negatives make a positive? Language and logic in language processing. *Lang. Cogn. Neurosci.* **38**, 1–16 (2023).
18. R. C. Jeffrey, *Formal Logic: Its Scope and Limits* (McGraw-Hill, ed. 2, 1981).
19. G. E. Hughes, M. J. Cresswell, *A New Introduction to Modal Logic* (Routledge, 1996).
20. K. Craik, *The Nature of Explanation* (Cambridge University Press, 1943).
21. P. N. Johnson-Laird, *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness* (Harvard University Press, 1983).
22. P. N. Johnson-Laird, R. M. J. Byrne, *Deduction* (Erlbaum, 1991).
23. R. M. J. Byrne, *The Rational Imagination: How People Create Alternatives to Reality* (MIT Press, 2005).
24. T. Hinterecker, M. Knauff, M. P. N. Johnson-Laird, Modality, probability, and mental models. *J. Exp. Psychol. Learn.* **42**, 1606–1620 (2016).
25. S. S. Khemlani, R. M. J. Byrne, P. N. Johnson-Laird, Facts and possibilities: A model-based theory of sentential reasoning. *Cognit. Sci.* **42**, 1887–1924 (2018).
26. P. N. Johnson-Laird, M. Ragni, Possibilities as the foundation of reasoning. *Cognition* **193**, 103950 (2019).
27. P. C. Wason, P. N. Johnson-Laird, A conflict between selecting and evaluating information in an inferential task. *Brit. J. Psychol.* **61**, 509–515 (1970).
28. J. St, B. T. Evans, Dual-process account of reasoning, judgement and social cognition. *Annu. Rev. Psychol.* **59**, 255–278 (2008).
29. M. Ragni, I. Kola, P. N. Johnson-Laird, On selecting evidence to test hypotheses. *Psychol. Bull.* **144**, 779–796 (2018).
30. D. Kahneman, *Thinking Fast and Slow* (Farrar, Strauss, Giroux, 2011).
31. R. M. J. Byrne, P. N. Johnson-Laird, If and or: Real and counterfactual possibilities in their truth and probability. *J. Exp. Psychol. Learn.* **46**, 760–780 (2019).
32. A. C. Quelhas, C. Rasga, P. N. Johnson-Laird, The analytic truth and falsity of disjunctions. *Cognit. Sci.* **43**, e12739 (2019).
33. O. Espino, R. M. J. Byrne, P. N. Johnson-Laird, Possibilities and the parallel meanings of factual and counterfactual conditionals. *Mem. Cognit.* **48**, 1263–1280 (2020).
34. S. Kripke, Semantical considerations on modal logic. *Z. Math. Logik.* **9**, 67–96 (1963).
35. B. H. Partee, "Semantics – mathematics or psychology?" in *Semantics from Different Points of View*, R. Bäuerle, U. Egli, A. von Stechow, Eds. (Springer-Verlag, 1979), pp. 311–360.
36. P. N. Johnson-Laird, "Formal semantics and the psychology of meaning" in *Processes, Beliefs and Questions*, S. Peters, E. Saarinen, Eds. (Reidel, 1982), pp. 1–68.
37. T. Gerstenberg, N. D. Goodman, D. A. Lagnado, J. B. Tenenbaum, A counterfactual simulation model of causal judgments for physical events. *Psychol. Rev.* **128**, 936–975 (2021).
38. M. Sridharan, B. Meadows, Towards a theory of explanations for human–robot collaboration. *Künstliche Intelligenz* **33**, 331–342 (2019).
39. S. S. Khemlani, R. Mackiewicz, M. Bucciarelli, P. N. Johnson-Laird, Kinematic mental simulations in abduction and deduction. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 16766–16771 (2013).
40. P. Koralus, *Reason and Enquiry* (Oxford University Press, 2023).
41. A. Tarski, The semantic conception of truth. *Philosoph. Phenomenol. Res.* **4**, 341–375 (1944).
42. D. E. Over, N. Cruz, "Probabilistic accounts of conditional reasoning" in *International Handbook of Thinking and Reasoning*, L. J. Ball, V. A. Thompson, Eds. (Psychology Press, 2018), pp. 434–450.
43. M. Lopéz-Astorga, M. Ragni, P. N. Johnson-Laird, The probability of conditionals: A review. *Psychonomic Bull. Rev.* **29**, 1–20 (2022).
44. H. P. Grice, *Studies in the Way of Words* (Harvard University Press, 1989).
45. S. C. Levinson, *Pragmatics* (Cambridge University Press, 1983).
46. D. Sperber, D. Wilson, *Relevance: Communication and Cognition* (Harvard University Press, 1986).
47. C. Rasga, A. C. Quelhas, P. N. Johnson-Laird, An explanation of or-deletions and other paradoxical disjunctive inferences. *J. Cognit. Psychol.* **34**, 1032–1051 (2022).
48. W. V. O. Quine, "Two dogmas of empiricism" in *From a Logical Point of View* (Harvard University Press, 1953), pp. 20–46.
49. G. H. Hardy, Mathematical proof. *Mind* **38**, 1–25 (1929).
50. P. N. Johnson-Laird, "How could consciousness arise from the computations of the brain?" in *Mind Waves*, C. Blakemore, Ed. Blackwell, 1987), pp. 247–257.
51. P. N. Johnson-Laird, S. Khemlani, R. M. J. Byrne, Human verifications: Computable with truth values outside logic. Open Science Framework (OSF). https://osf.io/2wtc6/. Deposited 6 January 2021.