KI - Künstliche Intelligenz
https://doi.org/10.1007/s13218-019-00599-w

SURVEY

Why Machines Don't (yet) Reason Like People

Sangeet Khemlani[1] · **P. N. Johnson-Laird**[2,3]

Received: 15 February 2019 / Accepted: 19 June 2019

Abstract
AI has never come to grips with how human beings reason in daily life. Many automated theorem-proving technologies exist, but they cannot serve as a foundation for automated reasoning systems. In this paper, we trace their limitations back to two historical developments in AI: the motivation to establish automated theorem-provers for systems of mathematical logic, and the formulation of nonmonotonic systems of reasoning. We then describe why human reasoning cannot be simulated by current machine reasoning or deep learning methodologies. People can generate inferences on their own instead of just evaluating them. They use strategies and fallible shortcuts when they reason. The discovery of an inconsistency does not result in an explosion of inferences—instead, it often prompts reasoners to abandon a premise. And the connectives they use in natural language have different meanings than those in classical logic. Only recently have cognitive scientists begun to implement automated reasoning systems that reflect these human patterns of reasoning. A key constraint of these recent implementations is that they compute, not proofs or truth values, but possibilities.

Keywords Reasoning · Mental models · Cognitive models

1 Introduction

The great commercial success of deep learning has pushed studies of reasoning in artificial intelligence into the background. Perhaps as a consequence, Alexa, Siri, and other such "home helpers" have a conspicuous inability to reason, which in turn may be because AI has never come to grips with how human beings reason in daily life. Many automated theorem-provers that implement reasoning procedures exist online, but they remain highly specialized technologies. One of their main uses is as an aid to mathematicians, who seek simpler proofs (see [38, 72] for the notion of simplicity in proofs—a problem that goes back to one of Hilbert's problems for mathematicians). Their other uses include the verification of computer programs and the design of computer chips. The field has its own journal: the *Journal of Automated Reasoning*. There have even been efforts to use machine learning to assist theorem-provers [20].

Deep learning systems were inspired by human learners, but they do not learn concepts and categories the same way humans do, and so they can be fooled in trivial ways known as "adversarial attacks" [52, 54, 66]. Nevertheless, they have achieved success and pervasiveness despite such failings, because they can discover latent patterns of relevance—they do so by analyzing vast amounts of data generated by humans. Theorem-provers likewise diverge from human reasoning, and yet they rarely enter our everyday technologies. Why not? The central thesis of this paper is that theorem-provers—and other automated reasoning systems—cannot simulate human reasoning abilities. Any machine reasoning system built to interact with humans needs to understand how people think and reason [12]. But theorem-provers have no way of discovering human-level reasoning competence on their own, and they implement principles that yield counterintuitive patterns of reasoning. In what follows, we trace their limitations back to two historical developments in AI: the motivation to establish automated theorem-provers for systems of mathematical logic (see, e.g., [3]), and the

✉ Sangeet Khemlani
sunny.khemlani@nrl.navy.mil; skhemlani@gmail.com

P. N. Johnson-Laird
phil@princeton.edu

[1] US Naval Research Laboratory, Navy Center for Applied Research in Artificial Intelligence, Washington, DC 20375, USA

[2] Department of Psychology, Princeton University, Princeton, NJ 08540, USA

[3] Department of Psychology, New York University, New York, NY 10003, USA

Published online: 04 July 2019

formulation of nonmonotonic systems of reasoning. We then describe human reasoning patterns that no machine reasoning system—or deep learning system, for that matter—can mimic. Finally, we present the functions that any future automated reasoning system should perform in order to approach human-level thinking abilities.

2 Automated theorem-proving

Automated theorem-proving aimed originally to find logical proofs in an efficient way. A central constraint of such systems is that they are built on the foundation of mathematical logic. Most theorem-proving programs take as input the logical form of an argument, and then search for a derivation of the conclusion using formal rules of inference and perhaps axioms to capture general knowledge. In the early days, some theorem-provers aimed to simulate human performance [53]. But, most others were exercises in implementing a logical system [64] or intended to help users to discover novel mathematical proofs [71], and so they were not intended to embody human patterns of reasoning. For instance, theorem-provers operate by searching for derivations of given conclusions rather than generating conclusions of their own. The main differences among them are in the nature of the logical rules of inference on which they rely and in the methods they use to search for proofs.

The first theorem-provers relied on mathematical logic, as did contemporary psychological theories of reasoning. The latter uses rules of inference from logic as formulated in a "natural deduction" system [5, 63]. The idea of natural deduction was due to the logician Gentzen [9], and intended to provide a more "natural" way of doing logic than axiomatic systems. So, it eschews axioms, and it relies on formal rules that can introduce each sentential connective into a proof, e.g.:

A.

Therefore, $A \vee B$, [disjunction introduction]

where '$\vee$' is a symbol for inclusive disjunction (A or B or both), and can eliminate it from a proof, e.g.:

$A \vee B$.

$\neg A$.

Therefore, B. [disjunction elimination]

where '$\neg$' denotes negation. Likewise, some AI systems adopted natural deduction too [2]. The challenge for those theorem-provers, of course, was to determine which natural deduction rules of inference to try next as they searched for proofs.

One answer that seemed attractive was instead to use just a single rule of inference, the resolution rule, in one of its many variants [43]. The rule calls for all the premises and conclusion to be transformed into inclusive disjunctions, which is feasible in classical logic. For example, material implication, $A \rightarrow B$, in logic, which is analogous to *if A then B*, can be transformed into: $\neg A \vee B$, because both $A \rightarrow B$ and $\neg A \vee B$ are false if and only if A is true and B is false. Other connectives, such as material equivalence ('$\leftrightarrow$') can likewise be transformed into inclusive disjunctions. In its simplest form, the resolution rule is:

$A \vee B$.

$\neg A \vee C$.

Therefore, $B \vee C$.

Not surprisingly, naive individuals presented with two such disjunctive premises and asked what follows seldom draw the conclusion in the rule [24].

Inferences in classical logic concern relations between entire sentences, where each sentence can be true or false. The predicate calculus includes the sentential calculus, but also allows inferences that depend on the constituents of sentences. For example:

Any actuary is a statistician.
Jean is an actuary.
Therefore, Jean is a statistician.

The quantifier "any actuary" ranges over individual entities. The first-order predicate calculus concerns only quantification of entities, whereas the second-order predicate calculus allows quantifiers also to range over properties. The calculus treats the above quantified statement as:

$\forall x$(x is an actuary $\rightarrow$ x is a statistician)

where '$\forall x$' ranges over all values of x. Theorem-provers exist for most branches of logic—the sentential calculus, the predicate calculus, and modal logics, which concern 'possible' and 'necessary,' temporal relations, and deontic relations such as permissions and obligations. Most theorem-provers dealing with the predicate calculus are restricted to first-order logic, because second-order logic cannot have a complete proof procedure. But, quantifiers such as "more than half of the actuaries" cannot be expressed in first-order logic.

One standard way to cope with universal quantifiers in automated proofs (e.g. "any actuary is a statistician") is to delete the determiner ($\forall x$), and to transform the remaining formula, such as:

x is an actuary $\rightarrow$ x is a statistician

into its disjunctive equivalent:

x is a not an actuary $\vee$ x is a statistician.

The process of unification can set the value of a variable in one expression to its value in another expression that has the same predicate. So, the unification of these two formulas:

x is an actuary

Jean is an actuary

sets the value of x equal to Jean. (It is equivalent to the rule of universal instantiation in predicate logic [19].) The

resolution rule with unification applies to the following premises:

x is not an actuary $\vee$ x is a statistician.

Jean is not a statistician $\vee$ Jean is a probabilist.

It cancels out the one clause and its negation and unifies the two remaining clauses to yield:

Jean is not an actuary $\vee$ Jean is a probabilist.

It follows that:

Jean is an actuary $\rightarrow$ Jean is a probabilist.

A more complex treatment, which we spare readers, is needed for existential quantifiers, such as: "Some statisticians," which as the name suggests establish the existence of entities, whereas universal quantifiers, such as "any," do not.

The formal derivation of a deductive conclusion resembles the execution of a computer program, and the analogy was exploited in the development of programming languages, such as PROLOG, which compute in a way analogous to a proof that a statement follows from others in a fragment of the predicate calculus [39]. The programmer formulates a set of declarative statements that characterizes a problem and the constraints on its solution; computation consists in a search for a solution based on a form of resolution theorem-proving.

In logic, an influential proof procedure—the tree method [65], which is also known as the method of "semantic tableaux"—adopts a different approach to the formal procedures of resolution and unification. A valid inference yields a conclusion that is true given that the premises are true, and so the negation of its conclusion together with the premises yields an inconsistent set of sentences. An inconsistent set of sentences cannot refer to any situation, e.g., nothing can be both a square and not a square. So, the method proceeds as follows:

1. Make a list of the premises and the negation of the putative conclusion.
2. Search for a case in which they are all true, using rules for each connective.
3. If the search fails, then the conclusion follows validly from the premises, and if it succeeds, then it discovers a counterexample to the inference, i.e., a case in which the premises are true but the conclusion is false.

For an inference such as modus ponens:

$A \rightarrow B$

A

Therefore, B

the first step is to list the premises and the negation of the conclusion:

$A \rightarrow B$

A

$\neg B$

The rule for the material implication in the first premise allows the list to be expanded into an "inverted tree," because the connective is equivalent to: $\neg A \vee B$. Each alternative is added to the list as follows:

$$
\begin{array}{c}
A \rightarrow B \\
A \\
\neg B \\
\hline
\neg A \quad B
\end{array}
$$

The left-hand branch of the tree contains two contradictory elements: A and $\neg A$, so that case fails; and the right-hand branch of the tree contains two contradictory elements: $\neg B$ and B, so that case fails. So, there is no situation in which the premises and the negation of the conclusion are true: they are inconsistent. And, so, the original conclusion is a valid inference from the premises. The tree method extends naturally to quantifiers, to a lucid pedagogy [19], and to modal logics [11]. It also underlies methods in the representation of knowledge [14], and in automated theorem-provers [4, 40].

Propositional logic is computationally intractable, and so as the number of atomic propositions in inferences gets larger so the search for a proof calls for more memory and takes longer—until it is beyond the competence of any finite organism in the lifetime of the universe [8]. Hence, whether a theorem-prover is based on natural deduction, resolution, or trees, the search for a proof becomes intractable for problems with multiple premises. The same applies, of course, to human reasoners, who cease to be able to infer valid conclusions long before AI programs do. Earlier theorem-provers were often incomplete in that they were unable to find proofs for certain valid theorems [57]. The advent of sets of test problems and an annual competition has led to complete theorem-provers [67].

As logicians and computer scientists began to implement automated reasoning systems, they discovered that those systems were brittle and often impractical. The reason was simple: orthodox logic can operate sensibly when information is static, i.e., when the same set of assumptions exists from one moment to the next. Real life, however, affords no such stasis: new information often overturns old assumptions. The scientists who recognized this disparity between logic and life developed methods that grant additional flexibility to systems of logic.

3 Nonmonotonic reasoning

One of the pioneers of AI, the late Marvin Minsky wrote:

> 'Logic' is the word we use for certain ways to chain ideas. But I doubt that pure deductive logic plays much of a role in ordinary thinking [51].

Others too have long been pessimistic about the prospects of artificial intelligence based on formal logic [48]. Part of this pessimism comes from the pertinent discovery in AI of the need for "nonmonotonic" reasoning—the idea that individuals draw a tentative conclusion that they may withdraw later. To use a once hackneyed example:

> Fido is a dog.
> Dogs have four legs.
> (So, Fido has four legs.)
> But, Fido lost at least one leg in an accident.
> So, Fido does not have four legs.

Classical logic is monotonic in that no subsequent premise can lead to the withdrawal of a valid conclusion: further premises can only add to the set of valid conclusions. In daily life, people abandon such conclusions in the face of definitive evidence to the contrary, and so everyday reasoning cannot be based on classical logic. One defense of logic is to argue that the premise, "Dogs have four legs," is false if it is interpreted to mean that all dogs have four legs. One should assume instead:

> If a dog is born intact, has not been in any accidents, has not had a leg amputated, or lost one as result of accident or illness, and so on ..., then it has four legs.

This solution is problematic, however, because no guarantee can exist that all the relevant conditions have been taken into account. In contrast, people often make inferences in daily life without knowing the complete information. It is useful to assume that dogs have four legs, humans have two legs, fish have no legs. The inferences that these claims support usually yield true conclusions.

An alternative stratagem first developed by Minsky [50] is the idea that conclusions can be withdrawn in the light of subsequent information. The concept of a MAMMAL includes a variable with a value denoting NUMBER-OF-LEGS, and, if there is no information to the contrary, then by default this value is equal to four. Dogs are mammals, and so they inherit the default value from this class inclusion. The idea is analogous to Wittgenstein's notion of a criterion [70]. He argued that many concepts have no essential conditions, but depend instead on criteria. The criteria for doghood includes having four legs. But, criteria are neither necessary conditions for doghood—a particular dog might be three-legged—nor inductions from observation, because one could not count the number of legs on dogs until one had some way of identifying dogs. Criteria are fixed by our concepts.

Various attempts were made in AI to incorporate default values in hierarchies of concepts, and Touretzsky formulated a semantic theory for such systems [69]. The attempts in part led to "object-oriented" programming languages, such as Java and Common Lisp. They allow for a hierarchy of class inclusions, such as TERRIER → DOG → MAMMAL → ANIMAL, in which default values for variables can be set up at any level in the hierarchy, and overruled by specific values lower in the hierarchy. A new class can be a child of one or more existing classes in a "tangled" hierarchy, e.g., a terrier is also a member of the hierarchy: TERRIER → DOG → PET → ANIMAL, and the new class inherits variables and default values from its parents [27].

AI systems of nonmonotonic reasoning began with attempts to model "common sense" reasoning: individuals draw a conclusion on the basis of incomplete information, and so later they may withdraw the conclusion [10, 45]. An intelligent robot has the same problem, and so there was a rapid development of AI systems of nonmonotonic reasoning. Some of these systems originally took a "proof theoretic" approach to the problem [62], invoking formal rules of inference in a default logic, such as:

> For any x, if x is a dog and it is consistent to assume that x barks, then one can infer that x barks.

Other systems captured the same idea with a clause to the effect that it is not disprovable that x barks [49]. Still other systems took a semantic or "model theoretic" approach to the problem. Circumscription minimizes the number of entities to which a predicate applies. It is then feasible to restrict those entities that are abnormal, such as dogs that do not have four legs [46]. The various approaches to nonmonotonic reasoning continue to be extended to yield novel advances [6, 18, 44]. Indeed, the foundations of human reasoning, as we will argue, are nonmonotonic, and so nonmonotonic logics appear well-poised to express "common sense" patterns of thinking. But they, too, fail to explain much of human reasoning [59].

What makes human reasoning so difficult for theorem-provers and nonmonotonic logics to capture? The answer may be that such systems are not constrained by data on human reasoning: they developed independently from discoveries of how people think. The lack of such a constraint may have resulted in highly creative formal frameworks for possible ways of computing inferences. But these frameworks cannot progress towards the flexibility of human-level reasoning capacities without a clear picture of what is unique about human reasoning.

4 What separates human from machine reasoning

A salient difference between the two sorts of reasoning are their respective starting points. Machine reasoning typically starts with the logical form of premises, whereas if human reasoning starts with verbal premises, they are in everyday language. A logic has a simple unambiguous grammar, with rules of the sort:

sentence → sentence connective sentence

and a simple lexicon that specifies, for example:

∨ → connective

So, it is simple to parse an input such as:

$A \vee B$

as a well-formed sentence. Indeed, such well-formed sentences in logic are unambiguous. In contrast, assertions in natural language are often ambiguous, and thereby create a major headache for theorists trying to establish their logical form. But, consider the following conditional assertion:

> If she plays a musical instrument then she doesn't play a harp.

It looks as though its logical form is:

she plays a musical instrument → ¬ she plays a harp

The material implication here allows one to prove an inference of the form known as modus tollens:

$A \rightarrow \neg B$

B

Therefore, $\neg A$

But, such an inference is catastrophically wrong in the present case:

> If she plays a musical instrument then she doesn't play a harp. ($A \rightarrow \neg B$)
> She plays a harp. (B, i.e., $\neg\neg B$)
> Therefore, she doesn't play a musical instrument. ($\neg A$)

How can such an error be avoided? Only by taking into account that harps are musical instruments. A theorem-prover might be programmed to retrieve any pertinent knowledge to a premise, and to add it as a further premise:

> If she plays a harp then she plays a music instrument.

So, now the set of premises is:

$A \rightarrow \neg B$

B

$B \rightarrow A$

Alas, because logic is monotonic, it is still possible to prove the modus tollens inference. A better strategy is to use knowledge to interpret the original conditional premise by constraining the set of finite alternatives—i.e., the possibilities—to which it refers:

> She plays an instrument ∧ ¬ She plays a harp
> ¬ She plays an instrument ∧ ¬ She plays a harp

where '∧' is the symbol for conjunction. It is now obvious that the further premise:

> She plays a harp

contradicts the conditional premise. If we follow this approach, we no longer need logical form. We can reason on the basis of possibilities.

At best, theorem-provers and nonmonotonic logics advocate a narrow view of human reasoning. For instance, neither sort of approach focuses on how inferences are generated: theorem-provers generate proofs, but they cannot decide on what to prove. Indeed, they cannot operate unless they are provided with a conclusion—a theorem—to evaluate. And they are designed to generate proofs for any valid conclusion, where "validity" refers to a conclusion that is true in every case that the premises are true [19]. But not every valid conclusion is interesting or useful. In fact, most aren't; any set of premises, including the empty set, yields a limitless number of valid conclusions. For example, theorem-provers can easily find proofs of the following conclusions:

> A.
> Therefore, A or not-A.
> Therefore, A or not-A or B.
> Therefore, A or not-A or B or not-B.

The pattern reveals a cascading infinitude of "vapid" deductions [26]. People eschew such vapidity. They can generate their own conclusions from scratch without any guidance apart from the premises. The conclusions they generate are systematic: they refrain from generating conclusions that are mere repetitions of the premises, or those throw semantic information away, e.g., by adding disjunctive alternatives to possibilities supposed by the premises [23], because such conclusions are uninformative. And they exhibit biases in the way they generate their own conclusions: they generate some conclusions faster and more often than others [22, 61], they adopt various reasoning strategies [16], and they fall prey to systematic reasoning errors [33]. These patterns reflect cognitive trade-offs that people must make in order to lower the computational costs of engaging in reasoning. Hence, people are discerning in the inferences they draw. Indeed, reasoners often refrain from drawing a conclusion from a set of premises: they spontaneously respond that "nothing follows" when faced with certain reasoning problems [35].

Perhaps a more subtle distinction between human and machine reasoning systems is that humans do not "compartmentalize" different sorts of reasoning—they have no trouble making inferences about different causal, spatiotemporal, and quantificational relations all at once. Consider this inference:

> Anyone standing to the right of Talia before 5pm makes her nervous.
> Aly is standing next to Talia at 4pm.
> Is it possible that Talia is nervous?
> Is it necessary that Talia's nervous?

Few people should have difficulty inferring that it's possible but not necessary that Talia is nervous. The inference is easy for people, but challenging for machines. The reason is because logical calculi typically describe valid inferences that pertain to a particular domain of reasoning—the calculi used to reason about time can operate on wholly separate assumptions from those used to reason about space. Several systems of logic would need to be stitched together to describe the spatiotemporal, causal, quantificational, and modal relations described in the premises above, and it is unclear whether any contemporary machine reasoning system could explain why the inferences are trivial.

In theorem-provers based on first-order logic, any conclusion whatsoever follows from a contradiction. Hence, when a theorem-prover can derive a contradiction from a set of premises, it can also be used to prove any arbitrary theorem. The detection of an inconsistent set of premises does not curtail the operations of the theorem-prover. Human intuitions diverge. Reasoners tend not to draw any decisive conclusion from a contradiction. In line with nonmonotonic reasoning systems, people often withdraw one of the premises, and inferences in daily life are accordingly defeasible, i.e., they are subject to be overturned. People are also capable of a different strategy for reasoning their way out of inconsistencies: when they possess relevant background knowledge about the topics that led to an inconsistency, they can construct an explanation of why the conflict arose in the first place [25, 34]. And those explanations make it faster and easier to know which premises to withdraw [28, 35].

The biggest single difference between theorem-provers and human reasoning is that the former are based on logic, whereas the latter is not. As we show below, the meaning of logical connectives differs from the meaning of their counterparts in logic. Connectives in natural language have meanings that can be overridden, i.e., they hold in default of information to the contrary. As a consequence, reasoning in daily life is always defeasible. What is valid in logic is in some cases invalid in everyday reasoning, and what is invalid in logic is sometimes valid in everyday reasoning. Perhaps, the simplest illustration of the difference between logic and life are the following pair of contrasting inferences:

> It is possible that Trump is in NY or he is in DC.
> Therefore, it is possible that he is in NY.

In daily life, we make such inferences, but they are invalid in all normal modal logics, e.g., those based on "System K" [60].

> It is possible that Trump is in NY.
> Therefore, it is possible that he is in NY or that he is in DC.

In daily life, we reject such inferences, but they are valid in all normal modal logics.

Researchers in AI continue to build and deploy systems based on formal logic, but recent developments in the last two decades have pushed the community to embrace probabilistic computation [41]. The central motivation was to build systems that take into account uncertainty in human reasoning, which can be computed using probability distributions. For instance, we might assume, all things being equal, that the uncertainty of Trump being in NY in the preceding example can be treated as a probabilistic statement, e.g., $\mathcal{P}$ (Trump is in NY) ≈ 0.50. The strength of such systems is that they can learn in a "rational" way: they can leverage Bayesian inference to update probabilities depending on the strength of new information. This connection to rational computation prompted many cognitive scientists to argue that human reasoning is best modeled using probabilities [1, 7, 55, 58]. Moreover, studies suggest that people treat the probability of a natural language conditional, e.g.,

> The probability that *if Trump is in NY, he's in Trump Tower*

as equivalent to a conditional probability, e.g.,

> $\mathcal{P}$ (Trump is in NY | Trump is in New York).

The result—known as the equation—forms the backbone of many probabilistic theories of reasoning. But because probability theory is an extension of logic, such theories inherit many of the same limitations exhibited by systems based on formal logic: for instance, they treat the vapid deductions above as valid; they have difficulty inferring that "nothing follows"; and they do not explain why some inferences are easy and why some are difficult, or why reasoners fall prey to systematic fallacies. Recent approaches attempt to resolve some of the limitations by integrating possibilities with probabilistic computation [15, 30, 68].

In sum, human reasoning diverges from contemporary machine reasoning in many ways. Humans generate inferences as well as evaluate them. They use strategies and fallible shortcuts when reasoning. They can reason about many different domains at once. The discovery of an inconsistency does not result in an explosion of inferences—instead, it often prompts reasoners to build explanations that help them know which premise to abandon. And the connectives people use in natural language—words such as "if" and "and" and "or"—have different meanings than those in classical logic and in probabilistic extensions of logic. Cognitive scientists have begun to implement automated systems that explain these patterns of human reasoning. A key feature of these implementations is

that they compute, not truth values or probability distributions, but possibilities.

5 Case studies in computing with possibilities

In this section, we describe recent computational simulations of human deductive reasoning. These automated reasoning systems are based on the notion that to reason about the world, people mentally simulate real, hypothetical, or imaginary situations [21, 22]—mental models. Like model theory in logic, mental models concern the semantics that make certain statements true or false. But, mental model theory postulates that all reasoning concerns possibilities, so even inferences from premises that make no reference to them, and that have a logical treatment in the classical sentential calculus, refer to possibilities.

A series of inferences can illustrate why reasoning about possibilities is fundamental. Recent studies show that people have no difficulty making the following inferences:

Either it rained or it is hot, or both.
Therefore, it's possible that it rained.
Therefore, it's possible that it's hot.
Therefore, it's possible that it rained and it's hot.

Indeed, the inferences seem obvious—and yet each inference above is invalid in all of the normal systems of modal logic [17, 60]. For instance, consider a scenario in which it is impossible that it rained but it is hot. The disjunctive premise is true, but the first conclusion above is false. Hence, the inference is invalid. Analogous arguments hold for the two remaining conclusions. And yet reasoners readily infer such modal conclusions from nonmodal premises.

In general, the theory of mental models postulates that all compound assertions refer to possibilities, which hold in default of information to the contrary. So, they interpret the first premise as referring to these default possibilities and one impossibity:

```
possible(rained)
possible(hot)
possible(rained & hot)
impossible(didn't-rain & not-hot)
```

Together, these four cases are in a conjunction, because possibilities can be conjoined even when their constituent propositions are incompatible with one another, e.g., it possible that it is raining and it is possible that it is not raining. Reasoners make all four of the inferences above [17, 60].

This treatment explains our earlier contrasting pair of inferences. The inference:

It is possible that Trump is in NY or he is in DC.
Therefore, it is possible that he is in NY.

is valid because the premise yields a conjunction of possibilities, and one of the corresponds to the conclusion. In contrast, the inference:

It is possible that Trump is in NY.
Therefore, it is possible that he is in NY or that he is in DC.

is invalid, because the premise does not imply the possibility that Trump is in DC, which is one of the possibilities in the conjunction to which the conclusion refers. Another consequence is that inferences from exclusive disjunctions to inclusive disjunctions are valid in logic, e.g.:

Either it rained or it is hot, but not both.
Therefore, either it rained or it is hot, or both.

Readers new to how theorem-provers work should consider the inference bizarre—the conclusion describes a situation explicitly ruled out by the premise. But those familiar with theorem-provers will recognize that theorem-provers cannot treat the inference as anything but valid! After all, any situation that renders the premise true will render the conclusion true as well. The antidote is to compute with conjunctions of possibilities instead of truth-values [31]. The premise denies the possibility in which both clauses hold ("...but not both"), but the conclusion directly refers to that possibility ("...or both"), and so the mismatch should cause reasoners to reject the inference. A recent computational implementation interprets sentential connectives into corresponding possibilities, and it closely mimics human reasoning (see Fig. 1).

Recent advances show how automated reasoning systems based in possibilities can explain human reasoning patterns across several different domains, including reasoning about spatial and temporal relations [32, 61], counterfactual relations [31], probabilities [36], kinematics [37], and quantifiers [29, 30]. A primary constraint imposed on each of these systems is that the possibilities they build are *iconic*, i.e., the structures of those possibilities reflect the structures of the situations they represent [56]. Hence, when a possibility-based system for reasoning about quantifiers interprets the assertion, "all the actuaries are statisticians", it recognizes that the assertion refers to a set of individual entities, and so it builds a possibility consisting of a small set of tokens that represent those entities, as depicted in this diagram;

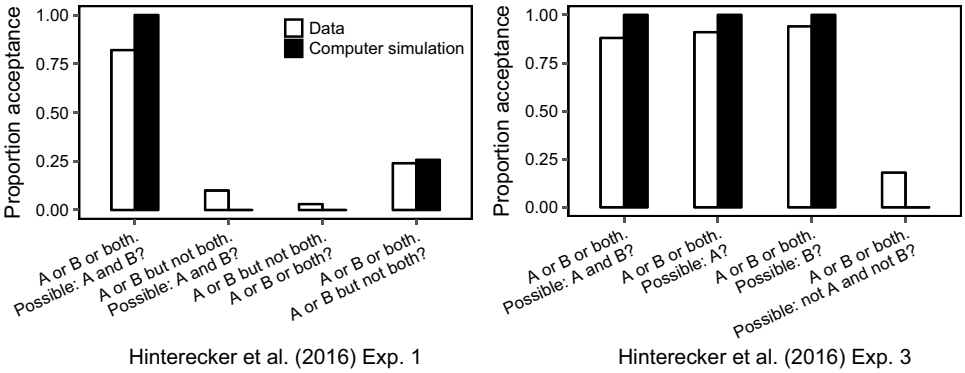

Fig. 1 Patterns of human data (white bars) from two experiments conducted by Hinterecker et al. [17] for different inferences that relate modal conclusions to non-modal premises, along with the results generated by an automated reasoning system (black bars) that is based on computations of possibilities

```
actuary     statistician
actuary     statistician
actuary     statistician
```

The number of tokens is derived stochastically. The structures of these resulting representations can be modified in various ways, e.g., tokens can be added, removed, or swapped to yield a different possibility consistent with the quantifier:

```
actuary     statistician
actuary     statistician
actuary     statistician
            statistician
```

The structures of these possibilities can then be "scanned" to yield inferences. For instance, in the first possibility, all of the statisticians are actuaries. In the second possibility, some of the statisticians are not actuaries. In general, the kinds of inferences that reasoners can draw are emergent properties of the iconic structures of the possibilities [13]. This method of drawing conclusions from iconic possibilities deviates from the way theorem provers operate (see above). The fundamental computational processes of inference concern how possibilities are constructed, how they're scanned, and how they're revised [29]. In principle, a computational system capable of efficiently exploring the entire space of possibilities consistent with various assertions can achieve perfect reasoning ability. In practice, however, reasoners make mistakes [33], and the account above explains them: possibilities impose a computational processing cost on humans and machines alike. If a reasoning problem can only be solved by considering multiple possibilities, reasoners are likely to err; but if the initial possibility suffices to yield a correct answer, people should be adept at delivering it.

In sum, the construction and manipulation of finite, iconic representations of possibilities appears to be a promising foundation for automated reasoning systems. When provided with unlimited processing resources, those systems can in principle achieve performance approaching logical inference. But when those resources are restricted to reasoning from just one possibility, they can be used to mimic the frailties of human reasoning, e.g., systematic errors and biased conclusions. Because an infinite set of possibilities can be captured by finite alternatives, and because conclusions emerge from their structures, representations based on possibilities yield many strengths of human inference for "free." For instance, they can be used to infer modal conclusions from non-modal premises; they can be restricted to yield only relevant and informative conclusions; and they can infer that "nothing follows" from a set of premises.

6 Conclusions

Why should any automated reasoning system be constrained by human reasoning patterns? After all, a major goal of artificial intelligence is to leverage technologies to exceed human abilities. Perhaps it is better that machines don't reason the way people do, because by ignoring human reasoning, machines can avoid the costly mistakes people are apt to make. At first blush, this perspective seems sensible. It has permitted researchers to build rich systems in a variety of domains that can solve novel problems, and the automated theorem-proving technologies that emerged from these explorations have helped generate unique solutions to long-standing puzzles (see, e.g., [47]). But, it has a devastating consequence: it overlooks three vast differences between reasoning based on logical form and reasoning in daily life.

First, logical form is not transparent in utterances in natural language: its recovery depends on grasping the meaning of assertions, but meaning suffices for reasoning, so logical form is superfluous. Second, validity in a logic depends on the conditions in which assertions are true; the dependence on truth values yields counterintuitive and vapid inferences. Third, reasoning about possibilities appears fundamental to human thinking. Logical systems known as "modal" logics

were designed to make deductions about possibilities, and there exists a countable infinity of them: they differ in the meanings that they assign to "possible" (and to "necessary"). But typical systems of modal logic treat as invalid certain inferences that humans consider obvious and uncontroversial. In daily life, the meanings of the word "possible" fall into three main categories: alethic modals, as in, "The conclusion follows as a possibility from the premises", deontic modals, as in, "It is not permissible for you to do that", and epistemic modals, as in "It is not possible for you to do that." Deontic modals raise severe problems for modal logics, because of inferences, such as the following one, that illustrate the well-known paradox of "free choice":

You can have the soup or the salad.

Therefore, you can have the soup.

Epistemic modals do not conform to any of the normal modal logics, and are akin to non-numerical subjective probabilities [42, 60]. The model theory accordingly treats compound assertions, such as disjunctions, as referring to conjunctions of possibilities, and the result explains the discrepancies we have illustrated between validity in human reasoning and validity in modal logics.

There may be an urgent need to reconcile the differences between machine and human reasoning: today's best interactive AI and robotic systems cannot reason. The popularity and prevalence of machine learning in everyday household technologies reveals that advances can be made without the need to engage in collaborative reasoning. Nevertheless, many sorts of scientist aim to develop machines that can form rich, meaningful interactions. Interactive technologies cannot function successfully without understanding how and why humans make certain inferences. Such systems need to infer that a human driving a car is, say, pressed for time, and to tailor their interactions to prevent distracting the driver. They should infer that a human lacks a critical piece of information, and provide it in a lucid format. In some cases, they should suggest a counterintuitive course of action, and provide a well-reasoned justification for it. To achieve these goals, cognitive scientists can learn from researchers in AI to design computational theories that are explicit, efficient, and principled. And researchers in AI can learn from cognitive scientists how individuals reason—both their systematic errors and their striking explanatory ability.

References

1. Baratgin J, Douven I, Evans J, Oaksford M, Over D, Politzer G (2015) The new paradigm and mental models. Trends Cogni Sci 19(10):547–548
2. Bledsoe W (1977) Non-resolution theorem proving. Artif Intell 9:1–35
3. Bonacina MP (1999) A taxonomy of theorem-proving strategies. In: Artificial intelligence today. Springer, Berlin, Heidelberg, pp 43–84
4. Bonacina MP, Furbach U, Sofronie-Stokkermans V (2015) On first-order model-based reasoning. In: Martí-Oliet N, Ölveczky P, Talcott C (eds) Logic, rewriting, and concurrency. Springer, Berlin
5. Braine MDS (1978) On the relation between the natural logic of reasoning and standard logic. Psychol Rev 85:1–21
6. Brewka G, Dix J, Konolige K (1997) Nonmonotonic reasoning: an overview, vol 73. CSLI publications, Stanford
7. Elqayam S, Over DE (2013) New paradigm psychology of reasoning: an introduction to the special issue edited by elqayam, bonnefon, and over. Think Reason 19(3–4):249–265
8. Garey MR, Johnson D (2002) Computers and intractability. W.H. Freeman, New York
9. Gentzen G (1969) Investigations into logical deduction. The collected papers of Gerhard Gentzen, pp 68–131
10. Ginsberg ML (1994) AI and nonmonotonic reasoning. In: Handbook of logic in artificial intelligence and logic programming, vol 3. Oxford University Press, Inc., pp. 1–33
11. Girle R (2009) Modal logics and philosophy. Routledge, Abingdon
12. Goodrich MA, Schultz AC (2008) Human-robot interaction: a survey. Found Trendsin Hum Comput Interact 1(3):203–275
13. Goodwin GP, Johnson-Laird P (2005) Reasoning about relations. Psychol Rev 112(2):468
14. Halpern JY, Vardi M (1991) Model checking vs. theorem proving: a manifesto. Artif Intell Math Theory Comput 212:151–176
15. Hattori M (2016) Probabilistic representation in syllogistic reasoning: a theory to integrate mental models and heuristics. Cognition 157:296–320
16. Van der Henst JB, Yang Y, Johnson-Laird PN (2002) Strategies in sentential reasoning. Cogn Sci 26(4):425–468
17. Hinterecker T, Knauff M, Johnson-Laird P (2016) Modality, probability, and mental models. J Exp Psychol 42(10):1606
18. Jalal S (2015) Non-monotonic reasoning: mimicking human thought process through argumentation. University of California, Davis
19. Jeffrey R (1981) Formal logic: its scope and limits, 2nd edn. McGraw-Hill, New York City
20. Jiang Y, Papapanagiotou P, Fleuriot J (2018) Machine learning for inductive theorem proving. In: International conference on artificial intelligence and symbolic computation. Springer, Cham, pp 87–103
21. Johnson-Laird PN (1983) Mental models: towards a cognitive science of language, inference, and consciousness. Harvard University Press, Cambridge
22. Johnson-Laird PN (2006) How we reason. Oxford University Press, Oxford
23. Johnson-Laird PN, Byrne R (1991) Deduction: essays in cognitive psychology. Laurence Erlbaum Associates, Mahwah
24. Johnson-Laird PN, Byrne RM, Schaeken W (1992) Propositional reasoning by model. Psychol Revi 99(3):418
25. Johnson-Laird PN, Girotto V, Legrenzi P (2004) Reasoning from inconsistency to consistency. Psychol Rev 111(3):640
26. Johnson-Laird PN, Khemlani SS, Goodwin GP (2015) Logic, probability, and human reasoning. Trends Cogn Sci 19(4):201–214
27. Keene S (1989) Object-oriented programming in Common LISP: a progammer's guide to CLOS. Addison-Wesley, Boston
28. Khemlani S, Johnson-Laird P (2013) Cognitive changes from explanations. J Cogn Psychol 25(2):139–146
29. Khemlani S, Johnson-Laird PN (2013) The processes of inference. Argum Comput 4(1):4–20
30. Khemlani S, Lotstein M, Trafton JG, Johnson-Laird P (2015) Immediate inferences from quantified assertions. Q J Exp Psychol 68(10):2073–2096

31. Khemlani SS, Byrne RM, Johnson-Laird PN (2018) Facts and possibilities: a model-based theory of sentential reasoning. Cogn Sci 42(6):1887–1924
32. Khemlani SS, Harrison AM, Trafton JG (2015) Episodes, events, and models. Front Hum Neurosci 9:590
33. Khemlani SS, Johnson-Laird P (2017) Illusions in reasoning. Minds Mach 27(1):11–35
34. Khemlani SS, Johnson-Laird PN (2011) The need to explain. Q J Exp Psychol 64(11):2276–2288
35. Khemlani SS, Johnson-Laird PN (2012) Hidden conflicts: explanations make inconsistencies harder to detect. Acta Psychol 139(3):486–491
36. Khemlani SS, Lotstein M, Johnson-Laird PN (2015) Naive probability: model-based estimates of unique events. Cogn Sci 39(6):1216–1258
37. Khemlani SS, Mackiewicz R, Bucciarelli M, Johnson-Laird PN (2013) Kinematic mental simulations in abduction and deduction. In: proceedings of the national academy of sciences, p. 201316275
38. Kinyon M (2019) Proof simplification and automated theorem proving. Philos Trans R Soc A 377(2140):20180034
39. Kowalski R (2011) Computational logic and human thinking. Cambridge University Press, Cambridge
40. Kowalski R, Hayes PJ (1983) Semantic trees in automatic theorem-proving. In: Automation of Reasoning. Springer, Berlin, Heidelberg, pp. 217–232
41. Lake BM, Ullman TD, Tenenbaum JB, Gershman SJ (2017) Building machines that learn and think like people. Behav Brain Sci 40
42. Lassiter D (2017) Graded modality: Qualitative and quantitative perspectives. Oxford University Press, Oxford
43. Loveland DW (2016) Automated Theorem Proving: a logical basis. Elsevier, Amsterdam
44. Marek VW, Truszczynski M (2013) Nonmonotonic logic: context-dependent reasoning. Springer, Berlin
45. McCarthy J (1960) Programs with common sense. In: Proceedings of the teddington conference on the mechanization of thought processes. H.M. Stationery Office
46. McCarthy J (1986) Applications of circumscription to formalizing common sense knowledge. Artif Intell 28:89–116
47. McCune W (1997) Solution of the Robbins problem. J Autom Reason 19(3):263–276
48. McDermott D (1987) A critique of pure reason. Comput Intell 3(1):151–160
49. McDermott D, Doyle J (1980) Non-monotonic logic i. Artif intell 13(1–2):41–72
50. Minsky M (1975) Frame-system theory. In: proceedings of the 1975 workshop on theoretical issues in natural language processing, Association for Computational Linguistics, pp. 104–116.
51. Minsky M (1985) The Society of Mind. Simon and Schuster, New York City
52. Moosavi-Dezfooli SM, Fawzi A, Frossard P (2016) Deepfool: a simple and accurate method to fool deep neural networks. In: proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2574–2582
53. Newell A, Shaw JC, Simon HA (1963) Empirical explorations with the logic theory machine. In: Feigenbaum E, Feldman J (eds) Computers and Thought. McGraw-Hill, New York City
54. Nguyen A, Yosinski J, Clune J (2015) Deep neural networks are easily fooled: high confidence predictions for unrecognizable images. In: proceedings of the IEEE conference on computer vision and pattern recognition, pp. 427–436
55. Oaksford M, Chater N (2007) Bayesian rationality: the probabilistic approach to human reasoning. Oxford University Press, Oxford
56. Peirce CS (1931–1958) Collected papers of Charles Sanders Peirce. In: Hartshorne C, Weiss P, Burks A (eds) vol 8. Harvard University Press, Cambridge, MA
57. Pelletier FJ (1986) Seventy-five problems for testing automatic theorem provers. J Autom Reason 2:191–216
58. Pfeifer N (2013) The new psychology of reasoning: a mental probability logical perspective. Think Reason 19(3–4):329–345. https://doi.org/10.1080/13546783.2013.838189
59. Ragni M, Eichhorn C, Bock T, Kern-Isberner G, Tse APP (2017) Formal nonmonotonic theories and properties of human defeasible reasoning. Minds Mach 27(1):79–117
60. Ragni M, Johnson-Laird P (2018) Reasoning about possibilities: human reasoning violates all normal modal logics. In: proceedings of the 40th annual conference of the Cognitive Science Society
61. Ragni M, Knauff M (2013) A theory and a computational model of spatial reasoning with preferred mental models. Psychol Rev 120(3):561
62. Reiter R (1980) A logic for default reasoning. Artif Intell 12:81–132
63. Rips L (2019) Cognitive processes in propositional reasoning. Psychol Rev 1:90
64. Robinson JA (1979) Logic: form and function. Edinburgh University Press, Edinburgh
65. Smullyan RR (2012) First-order logic, vol 43. Springer, Berlin
66. Su J, Vargas DV, Sakurai K (2019) One pixel attack for fooling deep neural networks. IEEE Transactions on Evolutionary Computation. https://doi.org/10.1109/TEVC.2019.2890858
67. Sutcliffe G (2015) The 9th IJCAR automated theorem proving system competition–CASC-J9. AI Communications, (Preprint), pp 1–13
68. Tessler M, Goodman N (2014) Some arguments are probably valid: syllogistic reasoning as communication. In: Proceedings of the annual meeting of the cognitive science society (vol. 36 No. 36)
69. Touretzky D (1986) The mathematics of inheritance systems. Morgan Kaufmann, Burlington
70. Wittgenstein L (1953) Philosophical investigations. Macmillan, London
71. Wos L (1988) Automated reasoning: 33 basic research problems. Prentice-Hall, Upper Saddle River
72. Wos L, Pieper GW (2003) Automated reasoning and the discovery of missing and elegant proofs Automated reasoning and the discovery of missing and elegant proofs. Rinton Press, Princeton