Proceedings of the Ninth Annual Conference on Advances in Cognitive Systems Submitted 9/2021; published 11/2021

# A Computational Perspective on Some Cognitive Illusions

**Kenneth D. Forbus** FORBUS@NORTHWESTERN.EDU
Qualitative Reasoning Group, Northwestern University, 2233 Tech Drive, Evanston, IL, 60208 USA

## Abstract

Reasoning is arguably at the heart of cognitive systems. Human reasoning, while still outperforming AI reasoners in many ways, has some well-explored limitations, called *cognitive illusions* in the psychological literature. This paper provides an initial theoretical analysis of several cognitive illusions in computational terms. The basic phenomenon is outlined and three illusions are summarized. An abstract model for reasoning systems is described to provide a program-independent way to characterize reasoning. This model is then used to propose explanations for the three illusions, including a novel psychological prediction. It also examines potential ways the reasoning model could be extended to either model human reasoning more closely, or to build cognitive systems that better complement weaknesses in human reasoning.

## 1. Introduction

While humans are likely the smartest creatures on the planet, our reasoning is far from error-free. Cognitive illusions (also called cognitive biases) have been extensively catalogued in the psychological literature. Like optical illusions providing insights into how vision works, such cognitive illusions can be used to provide insights into how human reasoning works. Considerable research has explored the role of heuristics in human reasoning, ranging from the seminal work of Tversky and Kahneman (1974) to Gigerenzer's (2007) fast-and-frugal methods. These heuristics are built upon the particulars of our cognitive capabilities, such as how our memories work. A classic example is the availability heuristic (Tversky & Kahneman, 1974), where we assess the probability of something by how easy it is to retrieve it. Such heuristics sacrifice soundness and completeness in favor of efficiency, enabling us to reason flexibly and easily under tight resource constraints. Given that, in many real situations, information is incomplete and actions must be taken in a timely manner, these trade-offs are often reasonable. However, the same illusions operate in situations where more careful analysis is required, and plague even trained professional intelligence analysts (Heuer, 1999).

The psychological mechanisms that give rise to these cognitive illusions are only partially understood. But given the growing need to create AI reasoning systems that can work with people doing professional reasoning, it makes sense to also look at these illusions from a computational perspective. There are three reasons for this:

1. If AI reasoning systems can be constructed that are immune to these cognitive illusions, that could make them more useful by complementing human capabilities.

2. Today's AI reasoning systems are superhuman for particular narrow types of reasoning (e.g. SAT solving, model checking), but still lack the flexibility of human reasoning and the ability to use reasoning with natural modalities to help frame problems. A better understanding of human cognitive heuristics could lead to principles for creating more flexible AI reasoners.
3. Given that human professional reasoning often involves tasks that require highly expressive representations (e.g. reasoning about knowledge, belief, and contexts), AI reasoning algorithms will invariably have their own trade-offs in terms of soundness, completeness, and efficiency. Understanding what new cognitive illusions they might be susceptible to would help us improve them and design practices to minimize negative impacts.

This essay begins with a brief introduction to cognitive illusions. It is far from comprehensive: the number of cognitive illusions that have been identified in the cognitive psychology literature is large[1]. However, their distinctions are in terms of externals, rather than underlying mechanisms. A computational analysis of underlying mechanisms might ultimately provide a more concise account of them. Next an abstract AI reasoning model is described, introducing some relevant mechanisms. These mechanisms will form the basis for the proposed explanations and solutions to the particular cognitive illusions examined here. Some conclusions and suggestions for future work wrap up the essay.

## 2. Three Key Cognitive Illusions

We focus on three illusions here because they have been identified as problems for intelligence analysts and implicated in the negative impacts of misinformation (aka "fake news"). This section provides a concise summary of them, to set the stage.

### 2.1 Confirmation bias

Confirmation bias is an umbrella term referring to the tendency for people to prefer supporting beliefs that they already have. This includes failing to seek evidence against such beliefs and discounting negative evidence when it is found. Examples from recent history include conspiracy theories about COVID-19 and about the 2020 US Presidential election. Several possible reasons have been suggested for confirmation bias. One component may be due to the nature of retrieval from long-term memory. Priming is the phenomena where what has been recently considered affects the probability of what is retrieved subsequently. Hence an explanation for an action is more likely to retrieve other memories relevant to that explanation, rather than alternative explanations for that action. If competing explanations are retrieved, they can be discounted by focusing on differences between that situation and the current situation, rendering them poorer matches.

An interesting computational model for confirmation bias has been based on ACT-R (Lebiere et al. 2013; Thomson et al. 2014). They explored how mechanisms of that architecture could lead to confirmation bias, as well as other biases, namely anchoring, representativeness, and

[1] For example, the Wikipedia article "List of cognitive biases" includes 200 items, although many of them are variations of more general types, and 44 concern properties of memory recall. (Retrieved 7/1/21)

probability-matching. Confirmation bias was explained in two ways, the use of blended retrievals of instances and the effect of those utility estimates on subsequent information gathering efforts. The model was successfully matched against human performance on a sensemaking task. However, the model only uses a simple attribute-based model of instances, and it is not clear that it can be extended to handle the kinds of relational structure that apply more generally in reasoning tasks, e.g. explanations and causal models.

### 2.2 Mirroring

Mirroring refers to the tendency of using one's own beliefs, values, and motivations when reasoning about someone else. In cognitive development, the use of analogy between self and others (e.g. Meltzoff's (2005) "like me" hypothesis) has been proposed as a valuable means of bootstrapping knowledge, as infants learn from those around them. Similarity is only a good guide when the actors are similar in relevant respects, of course, hence for reasoning about other cultures, analogical inferences must be more carefully scrutinized. Such inferences may or may not be conscious. If unconscious, it may be an example of *attribute-substitution* (Kahneman & Shane, 2002), where to estimate a property that is unknown (e.g. the goal of an opponent) one substitutes an easier to compute property (e.g. one's own goals).

Several attempts have been made to model mirroring computationally. Yalcin & DiPaola (2018) argue that modeling empathy requires mirroring at multiple levels, but they do not propose a mechanism by which mirroring occurs. Similarly, several papers have examined computational models of mirror neurons and tissues involving them but have not led to computational systems that can actually do mirroring (Thill et al. 2013). An ACT-R model of like me simulation has been used for human-robot interaction experiments (Kennedy et al. 2009; Hiatt et al. 2011), but only for predicting properties of other agents in an embodied environment, not the kind of conceptual reasoning we are focusing on here.

### 2.3 Misinformation effects

The social impacts of misinformation, such as conspiracy theories in politics and in anti-vaccination campaigns, are proving quite significant. Misinformation is a tough problem because, even when people learn that a piece of news is incorrect, people often continue to use it in reasoning (Johnson & Seifert, 1994). When exposed to incorrect information, that incorrect information can override their prior knowledge (Rapp et al. 2020). For example, participants read stories where deliberately false incidental information was injected, e.g. "Someone, maybe an actor named Oswald, has killed Lincoln." Later, when asked who killed Lincoln, some answered "Oswald", while others gave the correct answer "Booth", but did so more slowly (Gerrig & Prentice, 1991; Rapp & Salovich 2018). Put another way, these studies suggest that recent information can overcome people's knowledge of facts that they have known for years. This suggests that people are poor at tracking dependencies in their reasoning, enabling information from different contexts to "leak" into the current situation. A heavily studied aspect of misinformation effects is the continued influence effect (e.g. Lewandowsky et al. 2012), where retractions of misinformation are ineffective, in that the retracted information is still used in reasoning. Evidence suggests that the timing of corrections does not matter (e.g. immediately versus two days later), and that belief in the mistaken information can return post correction (Rich

& Zaragoza, 2020). We lump these phenomena together because their underlying source appears to be the processes involved in fact storage, retrieval, and verification.

Psychological studies of these phenomena are relatively recent, and there do not seem to be computational models of these effects yet.

## 3. An Abstract Reasoning Model

To explore how cognitive illusions might arise in AI systems, we need a model of reasoning systems that we can use to compare against properties of human reasoning. This section describes an abstracted version of the FIRE reasoning engine (Forbus et al. 2010) to play that role. It captures the functional capabilities of FIRE while suppressing irrelevant properties and engineering details. We choose FIRE because it is designed to model key aspects of human reasoning, including the ability to work with the highly expressive symbolic representations of knowledge needed to capture human conceptual structure, and a heavy reliance on analogical reasoning and learning (Gentner 2003). Some aspects of it, e.g. the analogical processing models, are indeed psychological models and have been tested against a variety of phenomena and used to make novel predictions (e.g. Forbus et al. 2016). Other aspects are not, as noted below. I point out similarities and differences between how it operates and human reasoning, to the extent that we know them at this point.

### 3.1 Knowledge Representations

There is ample evidence that human knowledge includes structured, relational representations (Gentner & Maravilla 2018). These include both propositional statements describing particular states of affairs and logically quantified knowledge to represent rules and other general statements. We assume a higher-order logical representation, capable of using predicates as constants, in order to express type-level axioms and metaknowledge and to handle modal statements (i.e. explicit statements about knowledge and belief). The kinds of information such knowledge encodes includes event schemas, causal laws, qualitative models, and action models. We further assume that there is a hierarchical arrangement of concepts, providing the multiple layers of description needed to reason with partial information and to make broad, robust generalizations (e.g. that animals require food to survive). Schemas are implemented via neoDavidsonian conventions, i.e. schemas are represented via a bundle of assertions involving an explicit entity representing the event or situation of the schema, with role relations connecting the parts to the event.

An important aspect of representation often overlooked are representations for context. We build on the Cyc notion of *microtheories* (Guha 1991). A microtheory is a collection of statements taken together as a unit. There is an inheritance relationship between microtheories, which enables contexts for a particular task to be dynamically constructed. Typically the contents of a microtheory are internally consistent, and alternate contradicting perspectives are represented in terms of distinct microtheories. For example, Newtonian, relativistic, and quantum laws of physics would be stored in distinct microtheories. The ability to explicitly refer to contexts is important for reasoning with them, e.g. in solving a physics problem, there are rules of thumb that suggest which set of laws might be useful, and criteria for determining when one has made an

incorrect choice[2]. Similarly, when an analyst is considering alternative explanations, these explanations can be worked up in distinct microtheories, which can then be compared and contrasted to guide further elaboration and decision-making.

### 3.2 First-Principles Reasoning

We assume that mechanisms for reasoning with rules containing variables are available. Rules need not be logically sound, although the ability to detect contradictions, perhaps via additional reasoning, is assumed. Moreover, rules may be used abductively, e.g. some antecedents might be marked as abducible, so that if they are not known, they can be assumed if needed to draw desirable conclusions (Hobbs et al. 1993). We note that first-principles reasoning can be combinatorially explosive, and hence tends to be done with resource bounds, which trades off completeness for efficiency. Often sets of rules can be treated as performing logical deduction, although non-monotonic predicates[3] extend inferential capabilities beyond this. We also note that logical rules are better at capturing what is possible or impossible, rather than what is typical. This is why probabilities are often used to guide abductive assumptions and why analogy is used as well as first-principles reasoning, as described below.

### 3.3 Memories

We assume a knowledge base that serves as a general storehouse, akin to long-term memory. Psychologists typically distinguish semantic memory, i.e. general knowledge, from episodic memory, i.e. knowledge of particular experiences. Semantic memory consists of general facts in the knowledge base. Episodic memories are encoded via microtheories in the knowledge base, e.g. what was gleaned from a reading a story, understanding a sketch, or solving a problem. Psychologists also distinguish declarative memories from procedural memory. In this model, procedural memory lies in the knowledge base's ability to store inference rules and task descriptions for plans as declarative representations, just like the other kinds of information in the knowledge base. Functionally, inference rules provide new conclusions while plans are used to construct sequences of behavior and take actions, including invoking reasoning.

A number of cognitive architectures focus on learning procedural knowledge via the accumulation and tuning of production rules (e.g. Anderson 2009; Laird 2012; Choi & Langley 2017). My conjecture is that skill learning has little to do with cognitive illusions, that is, the accumulation of declarative knowledge along with fixed reasoning mechanisms seem to suffice for explaining the illusions examined here.

A major difference between human long-term memory and FIRE's model is how retrieval works. Spontaneous remindings can happen in both, although in FIRE, this currently only happens via analogical retrieval, as described below. FIRE supports logical queries against its knowledge base, using a query pattern combined with a logical environment (i.e. a microtheory

[2] This provides a mechanism for implementing automatic model formulation as used in qualitative reasoning, which may be useful for some forms of social reasoning as well as physical reasoning (Forbus, 2019).

[3] For example, **`uninferredSentence`** is true exactly when the statement which is its argument cannot be proven within the current logical environment. With this predicate other nonmonotonic reasoning patterns, such as negation by failure, can be implemented.

and what it inherits from), and by default returns all matching answers. By contrast, human explicit retrievals tend to be small in number, even though the number of potential matches can be huge. Such tight bounds make sense for organisms that accumulate massive amounts of experience. SOAR and ACT-R have been used to model long-term memory retrieval using spreading activation, although obviously testing at human scale is currently beyond the state of the art.

Two other forms of memory typically used in psychological explanations are short-term memory and working memory. Short-term memory is the infamous 7 plus/minus 2 (Miller, 1955), which FIRE does not model at all. Working memory is harder to characterize. Ericsson & Kintsch (1995) provides evidence that it can actually be huge, and that for some kinds of information, a brief intervention in a laboratory experiment can increase a participant's capacity by a factor of 1,000, demonstrating that working memory involves expertise. FIRE's working memory is implemented via a reasoning system that incorporates a logic-based truth maintenance system (Forbus & de Kleer, 1993). Propositions are recorded with justifications in terms of other facts, using clauses. Thus any conclusion can be inspected and the underlying assumptions identified. When an assumption is retracted, all beliefs based on that assumption are also no longer believed, unless there is alternative support for them in the network of clauses that constitutes working memory. We note that the propositional reasoning algorithm used in the logic-based truth maintenance system may or may not be psychologically plausible, we treat it as an engineering approximation. Similarly, the degree to which people record dependencies during reasoning is an open question. The evidence cited below suggests that in people, dependency tracking is less than complete.

Both rule-based inference and analogical inference (see below) operate by adding facts and clauses to working memory. Moreover, specialized inference capacities (e.g. visual processing and natural language understanding) also write their results to this working memory, thereby providing a tight integration across all forms of reasoning. FIRE does not contain an automatic consolidation process to move material to long-term memory. Instead plans or procedures associated with particular systems select material to move, often based on writing working memory microtheories (in their entirety, or with bookkeeping facts filtered out) to the knowledge base. Dependency information from the working memory may or may not be stored during such consolidation, depending on the algorithm used.

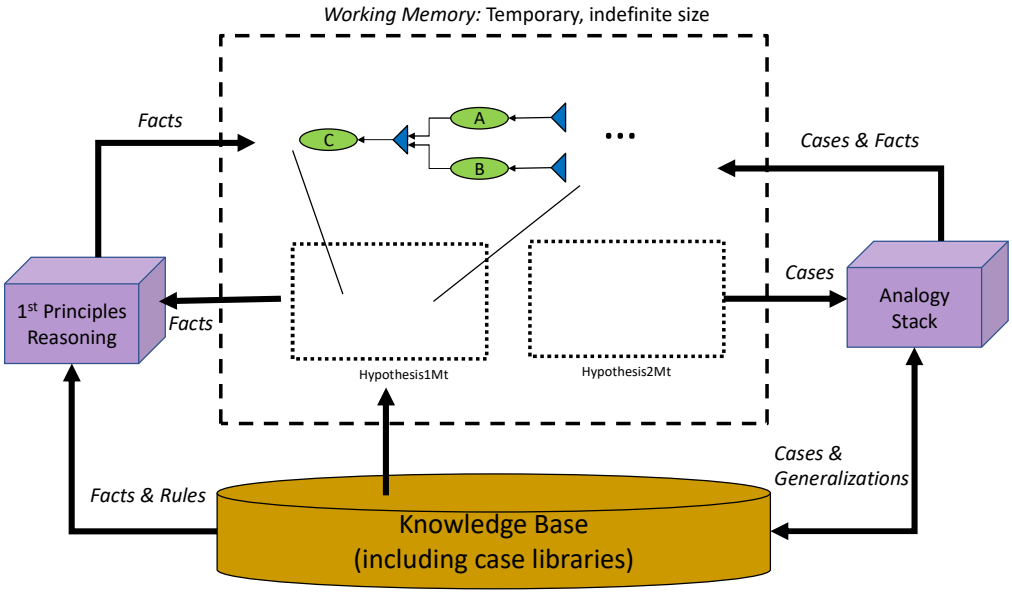


Figure 1: An abstract reasoning model. Cases and generalizations are microtheories.

Figure 1 illustrates the abstract reasoning architecture. The knowledge base is persistent, while working memory is temporary, cleared when a reasoning session is over. Like the knowledge base, the contents of working memory are structured in terms of microtheories. That is, every fact is in some microtheory. Here, two alternate hypotheses are being explored in separate microtheories (i.e. Hypothesis1Mt and Hypothesis2Mt). These microtheories inherit from other microtheories, such as a microtheory representing the shared background assumptions of the analytic problem being explored. This includes microtheories from the knowledge base, thereby providing the rules and other knowledge needed to reason with. The nodes A, B, and C are propositions, connected by a dependency structure that indicates C is believed on the basis of both A and B being believed. If either of those propositions lose support, then C would automatically lose support as well.

### 3.4 Analogy

The work of Gentner (2003) and others suggests that analogy is ubiquitous in human cognition. Functionally this makes a lot of sense, because people have vast amounts of experience from interacting with the world and with others in their culture. Analogical reasoning is powerful because it enables remembered experiences to be used directly via analogy in new situations, and supports learning more portable, transferable knowledge via analogical generalization (Forbus & Hinrichs, 2017). Analogy also supports one-shot learning, i.e. cases representing experience can be directly applied to new situations.

We have developed models of the key processes involved in analogical reasoning and learning based on Gentner's (1983) structure-mapping theory. These systems been used to both model a variety of human phenomena and to build performance-oriented AI systems. We consider this an *analogy stack* for cognitive architectures. It consists of

- Analogical matching (SME; Forbus et al. 2016)
- Similarity-based retrieval (MAC/FAC; Forbus et al. 1995)
- Generalization in long-term memory (SAGE; Kandaswamy & Forbus, 2012)
- Generalization in working memory (SageWM; Kandaswamy & Forbus, 2014)

Here we focus on the functional properties that are important for reasoning for each in turn.

*3.4.1 Matching*

The Structure-Mapping Engine (SME) compares structured, relational representations. Analogical inferences are constructed by projecting statements from the base into the target, based on correspondences found between other entities and statements, or vice-versa. Such analogical inferences can be deductive or abductive (Falkenhainer, 1990), based on the form of the relational knowledge projected. This is one solution to the qualification problem (McCarthy 1977), since analogical matching does not assume a complete set of preconditions in order to project facts between descriptions. This means, for example, that incomplete explanations can still be used to draw conclusions in new situations, as long as there is enough overlap between them so that the prior case is retrieved. As explained below, Rabkina et al. (2017)'s analogical theory of mind suggests a way that "like me" inferences in mirroring may be computed.

*3.4.2 Retrieval*

Analogical retrieval provides cases or generalizations to be used in reasoning. That is, given a situation in working memory, stored experiences (in *case libraries* that are part of the knowledge base) and/or generalizations constructed from experience (in *generalization pools,* see below) that are similar to it are retrieved, so that SME's candidate inference mechanism can be used to infer new facts about the situation. Analogical retrieval is sensitive to both surface properties (i.e. attributes and relationships among entities) and higher-order relations (e.g. relationships between statements), but since surface properties tend to be more easily encoded, they tend to dominate in retrieval from concrete examples. In people, the degree of similarity supports inference (e.g. Heit & Rubenstein, 1994), so being sensitive to similarity during retrieval is not unreasonable. Properties of analogical retrieval may be involved in both confirmation bias and misinformation effects, as described below.

*3.4.3 Generalization*

People learn incrementally and in a data-efficient manner from examples, leading to generalizations that can be used more broadly in reasoning. The Sequential Analogical Generalization Engine (SAGE; Kandaswamy & Forbus, 2012) models this. SAGE builds models of concepts, given an incremental stream of examples. Each concept is represented by a *generalization pool*. All generalization pools are part of the knowledge base. A generalization pool can contain both generalizations and outliers. Generalizations are constructed (or extended) when a new example being added retrieves a very similar item from the pool. If the item is an outlier, then a new generalization is formed, otherwise the example is added to the existing generalization. Every statement in a generalization has an associated probability, calculated directly by the frequency of which examples include a statement that aligns with it. Thus SAGE

provides a mechanism for constructing priors for probabilistic reasoning. SAGE has been used to automatically construct probabilistic rules and Bayes nets (Halstead & Forbus, 2005), as well as to model human conceptual change (Friedman & Forbus, 2009).

There is also a working memory version of SAGE, SageWM (Kandaswamy & Forbus, 2014), which has been used to model immediate generalization effects, e.g. learning during forced-choice tasks (Kandaswamy & Forbus, 2014). SageWM keeps a bounded set of examples, sorted by recency, which is used to help provide context. SageWM may be directly involved in misinformation effects that occur within a single experimental session, as described below.

## 4. Characterizing Some Cognitive Illusions Computationally

As noted above, we focus on three cognitive illusions that are particularly relevant for intelligence analysis and misinformation: Confirmation bias, Mirroring, and Misinformation Effects.

### 4.1 Confirmation Bias

In confirmation bias, once a hypothesis is formed, people tend to (1) gather and pay attention to evidence that supports it and (2) ignore evidence that contradicts it. My hypothesis is that confirmation bias in knowledge gathering is, at least in part, a cost paid for powerful human pattern-recognition capabilities. Consider a problem being analyzed, where spontaneous memory retrieval suggests a similar problem whose solution can be adapted to solve the current problem. For example, a physics problem might bring to mind a problem with a similar diagram, even though the principle used to solve the problem is entirely different (Chi et al. 1981), and hence it will turn out to be irrelevant. Such appearance matches are common in human retrieval. If that first retrieval remains in working memory, then its properties could become part of the probe used to find the next reminding. This will increase the likelihood that examples similar to the first retrieved are found, rather than seeking a more distinct alternative. In MAC/FAC, a retrieved 2nd example will likely be a literal similarity[4] to the first retrieval, because the candidate inferences from the first retrieval will be added to the probe. Hence any example sharing both the problem set-up and solution will be the most similar. Thus retrieving a different solution, perhaps better able to provide a relevant principle, will become harder. This suggest that AI systems using similarity-based retrieval will be susceptible to confirmation bias in information gathering, unless precautions are taken. How might this be overcome? Consider using Cyc-style microtheories in working memory as the probes for analogical retrieval. If the inferences from the first retrieval are added to a new microtheory, rather than the problem microtheory, then an unpolluted problem microtheory can again be used as a probe to retrieve the next-most similar example[5].

The reasoning model does not place values on prior beliefs, and thus as is could not exhibit selective attention about evidence for or against a cherished belief because it has none. I actually view this as a limitation, given the need to focus reasoning becomes more acute as the scale of knowledge and tasks increases. We plan to add automatic compiling of statistical metadata about

[4] That is, overlapping in both appearances and causal structure (Gentner, 1983).

[5] Obtaining alternate remindings is done by temporarily suppressing previously retrieved cases from the case library.

facts to FIRE, to enable using such metadata in knowledge refinement, akin to how it is used in SOAR (Laird 2012). For example, measuring accuracy (i.e. how often they lead to correct inferences) and utility (i.e. how often they are used) should be useful in detecting incorrect learned facts arising from misinformation. Another arena where confirmation bias in evidence evaluation arises is in education. Feltovich et al. (2001) has argued that one reason for the persistence of misconceptions is that learners often resort to mental shields that block evidence which would force them to change cherished beliefs. Friedman's TIMBER model of conceptual change uses preferences between explanation properties to capture aspects of this phenomenon (Friedman et al. 2018). TIMBER includes four dimensions of preference, namely specificity of information, whether or not an explanation is supported by instruction, whether it is compatible with prior knowledge, and whether an explanation uses constituents that are refinements of earlier knowledge. The hypothesis is that different people rank these dimensions of explanation evaluation differently. For example, by varying explanation preference rankings, TIMBER was able to account for 90% of student model transitions in a psychological experiment on students learning about circulatory systems (Friedman & Forbus, 2011). Using TIMBER-like mechanisms with preferences that are biased in favor of new information might be able to ameliorate confirmation bias in gathering and evaluating information.

### 4.2 Mirroring

In mirroring, a reasoner believes (often implicitly) that other actors think the way we do. As noted earlier, this can be viewed as a form of attribute substitution, i.e. when we don't have a model of another person (or culture), we substitute reasoning about ourselves/our own culture. Such like me reasoning is powerful for bootstrapping human social reasoning and theory of mind when it works. For example, Rabkina's Analogical Theory of Mind experiments demonstrate that our analogy stack can be used to model learning theory of mind inference from examples and from language (Rabkina et al. 2017, 2018), and can be used in AI systems to infer the goals of other agents (Rabkina et al. 2020). Thus mirroring can be a source of increased flexibility in reasoning, by drawing on a system's experience and/or lessons from stories to draw conclusions in circumstances where it does not have prior knowledge or sufficient rules to directly infer something.

Analogy can also lead one astray, of course. How would this occur in the reasoning model here? I assume that episodic memories include things that happened to the system itself, but also observations it makes about other people, and the contents of stories. If all these materials are placed in a single case library (or into a set of generalization pools differentiated by, for example, type of event) no matter who they occurred to, then analogical retrieval could include memories based on the actions and events affecting others, which could lead to mirroring. If, on the other hand, generalization pools are further differentiated by the actor, more refined models could be generated. Christmas traditions in the US, Germany, and Japan are rather different, for example. By choosing generalization pools built to model aspects of particular cultures, cross-culture contamination can be eliminated when desired. For example, MoralDM (Dehghani et al. 2008) used different libraries of cultural stories to express protected values for different cultures in moral decision-making. Since unions of case libraries (and generalization pools) are themselves

treated as case libraries (or generalization pools), retrievals can be sought more broadly when appropriate.

### 4.3 Misinformation Effects

The crux of the problem with misinformation is that (a) people often do not detect that a story contains facts that they know to be incorrect and use those incorrect facts and (b) even when people learn that a piece of news is incorrect, they often continue to use it in reasoning. Why might that be? Consider the impact of inserting incorrect facts into a story (Hinze et al. 2014; Rapp & Salovich 2018). For example, "I named the boat after the mythical high civilization that sank into the sea, Pompeii." When, immediately after reading the story, participants were asked what mythical city fell into the sea, some answered "Pompeii" instead of "Atlantis". Others replied "Atlantis", but took longer to answer than participants who read versions of the story without the incorrect fact. How can such local information override facts that participants already knew?

I propose an explanation based on implicit analogical processing. While most research on analogy has assumed it is a conscious operation, in fact there is evidence that implicit analogies are commonly used in human cognition. These are not distant, cross-domain analogies, but rather literal similarities, i.e. within-domain analogies that are commonly used in reasoning. Consider Day & Gentner (2007), which found that participants used information from a previously read story to understand a new story, without any awareness that they had done so. This explanation assumes that all information in working memory is in one or more microtheories, just as information in the knowledge base is[6], and that implicit analogical retrieval is used to answer questions. Such implicit retrievals operate over SageWM as well as long term memory. There are three cases to explain:

1. Participants who gave the wrong answer: Their retrieval strategy is to stop after a reasonably matching answer is found, and they hit the representation of the story in SageWM and used that information.
2. Participants who gave the correct answer, but more slowly: Their retrieval strategy looks in parallel at SageWM and the KB, and when they get multiple answers, they use provenance information and statistical metadata to choose which answer to accept. The explanation preferences used in Friedman's TIMBER model of conceptual change (Friedman et al. 2018) could be adapted to choose which answer is preferred.
3. Participants who gave the correct answer, but with no speed difference: Their encoding strategy did more vetting during the story understanding process, and marked the relevant fact as incorrect, perhaps even including the correct fact with a relationship to the incorrect fact.

Psychologically, based on the findings of Day & Gentner (2007), this explanation predicts that (1) the intrusion of misinformation could occur for up to several days, and (2) intrusions will be more likely when the new stimuli are very similar to the original example. This would make an interesting experiment. While the experiment above looked at immediate effects, other

[6] There is some psychological evidence that the understanding of stories is "compartmentalized", as per Gerrig and Prentice (1991), which is compatible with our assumption of microtheory-based storage.

experiments have looked at retrieval over longer periods, e.g. two days (Rich & Zaragoza, 2020). Their use of the identical situation is the strongest similarity case, and hence compatible with implicit analogical processing, but if the explanation is implicit analogical processing, then such intrusions should also occur with similar but not identical situations. This is a prediction that seems worth investigating.

Computationally, the reasoning model presented earlier would need two modifications to fully capture these effects. The first concerns recency. SageWM does incorporate recency, in that the working memory generalization pools are temporally ordered, and so more recent acceptable retrievals will be found first. But SAGE in long-term memory, and MAC/FAC, do not incorporate recency. We think it unlikely that working memory stays intact for several days, so to explain the implicit analogy findings, extending MAC/FAC and SAGE to use recency seems necessary. Time is used in the reasoning model for a number of purposes, e.g. ascertaining when particular cached data is stale, but we currently do not use temporal discounting when evaluating retrieved facts, as ACT-R and SOAR do.

What about the role of dependency information? Seifert (2002) suggests one possible explanation is that while some of the incorrect information has been edited out, at least some of its consequences remain intact. Since many psychological experiments exploring these phenomena operate over a single session, this suggests that human dependency tracking in working memory is hit-or-miss. Similar results were found by Rich & Zaragoza (2016), who further note that implied misinformation is harder to correct than explicit misinformation, compatible with this hypothesis.

How can we build AI systems to avoid misinformation effects? The reasoning model already likely records more dependency information than people do, and this provides better provenance information. More vetting of incoming information is another strategy, i.e. detecting misinformation immediately. This of course would increase comprehension time, which could be problematic, especially in interactive dialogue.

What about misinformation that has made it into the knowledge base? There are two suggestions that might help. The first is to use memory consolidation methods that preserve dependencies, so that they are retrieved along with potential answers. This would enable additional scrutiny during evaluating retrieved answers. The second is based on a technique for regularizing knowledge proposed in TIMBER, namely retrieving other relevant cases when a qualitative domain theory changes, to re-analyze them in terms of the new information. This would place the additional vetting into an off-line rumination process, as in Forbus et al. (2007).

## 5. Conclusions and Future Work

Understanding reasoning in people and machines is a key goal of cognitive science, and improving our computational understanding of human reasoning should help us make better cognitive systems. This theoretical paper has looked at three important cognitive illusions, confirmation bias, mirroring, and misinformation effects in computational terms. An abstract reasoning model was described and used to suggest computational explanations for them, including a novel psychological prediction. These explanations also suggest ways to avoid such illusions in AI reasoning systems aimed at complementing human reasoning.

There are many avenues for future work. The first is developing one or more datasets that can be used for experiments. We are working on generating such an open-license dataset, both for replicability and to encourage reasoning research. The second is to expand the catalog of cognitive illusions examined, to develop computational descriptions that perhaps might provide more structure and order to the phenomenon. The third is based on the hypothesis that cognitive illusions arise from trade-offs. AI reasoners designed to overcome human cognitive illusions are still going to be designed based on trade-offs. What new cognitive illusions will they be subject to? This is an interesting and important question, in order to make cognitive systems whose conclusions we can trust.

## Acknowledgements

I thank David Rapp, Tom Hinrichs, and Dedre Gentner for helpful suggestions. This research was sponsored by the Air Force Office of Scientific Research, Grant #FA9550-20-1-0091.

## References

Anderson, J. R. (2009) *How can the Human Mind Occur in the Physical Universe?* Oxford University Press.

Chi, M., T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, 5, 121-152.

Choi, D., & Langley, P. (2017). Evolution of the ICARUS Cognitive Architecture. *Cognitive Systems Research*, 48:25-38.

Day, S. & Gentner, D. (2007). Nonintentional analogical inference in text comprehension. *Memory and Cognition*, 35, 39-49.

Dehghani, M., Tomai, E., Forbus, K., Klenk, M. (2008). An Integrated Reasoning Approach to Moral Decision-Making. *Proceedings of AAAI 2008*. Chicago, IL.

Ericsson, K. & Kintsch, W. (1995). Long-term working memory. *Psychological Review*, 102(2).

Falkenhainer, B. (1990). A unified approach to explanation and theory formation. In J. Shrager and P. Langley (Eds.), *Computational models of scientific discovery and theory formation* (pp. 157-196). Morgan Kaufmann Publishers.

Feltovich, P., Coulson, R., & Spiro, R. (2001). Learners' (mis)understanding of important and difficult concepts: a challenge to smart machines in education. In K. Forbus & P. Feltovich (Eds.), *Smart Machines in Education*, MIT Press, pp. 349-375.

Forbus, K. (2019). *Qualitative Representations: How People Reason and Learn about the Continuous World*. MIT Press.

Forbus, K. and de Kleer, J. (1993). *Building Problem Solvers*, MIT Press.

Forbus, K., Gentner, D., and Law, K. (1995). MAC/FAC: A model of similarity-based retrieval. *Cognitive Science*, 19, 141-205.

Forbus, K.D. & Hinrichs, T. (2017) Analogy and Qualitative Representations in the Companion Cognitive Architecture. *AI Magazine*.

Forbus, K., Riesbeck, C., Birnbaum, L., Livingston, K., Sharma, A., and Ureel, L. (2007). Integrating Natural Language, Knowledge Representation and Reasoning, and Analogical Processing to Learn by Reading. *Proceedings of AAAI-07*: Vancouver, BC.

Forbus, K., Hinrichs, T., de Kleer, J., and Usher, J. (2010). FIRE: Infrastructure for Experience-based Systems with Common Sense. *AAAI Fall Symposium on Commonsense Knowledge*, Arlington, VA

Forbus, K. D., Ferguson, R. W., Lovett, A., and Gentner, D. (2016). Extending SME to handle large-scale cognitive modeling. *Cognitive Science*, DOI: 10.1111/cogs.12377, pp 1-50.

Friedman, S. and Forbus, K. (2009). Learning Naïve Physics Models and Misconceptions. *Proceedings of the 31st Annual Conference of the Cognitive Science Society*. Amsterdam, Netherlands.

Friedman, S. E. and Forbus, K. D. (2011). Repairing Incorrect Knowledge with Model Formulation and Metareasoning. *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*. Barcelona, Spain.

Friedman, S., Forbus, K., & Sherin, B. (2018). Representing, Running, and Revising Mental Models: A Computational Model. *Cognitive Science*, 1110-1145. DOI:10.1111/cogs.12574.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7, 155-170.

Gentner, D. (2003). Why we're so smart. In D. Gentner and S. Goldin-Meadow (Eds.), *Language in mind: Advances in the study of language and thought* (pp.195-235). Cambridge, MA: MIT Press.

Gentner, D. & Maravilla, F. (2018). Analogical reasoning. L. J. Ball & V. A. Thompson (eds.) *International Handbook of Thinking & Reasoning* (pp. 186-203). NY, NY: Psychology Press.

Gerrig, R. & Prentice, D. (1991). The representation of fictional information. *Psychological Science*, 2(5):336-340.

Gigerenzer, G. (2007). *Gut Feelings: The Intelligence of the Unconscious*. Penguin Books.

Guha, R.V.: Contexts: a formalization and some applications. Technical Report STAN-CS-91-1399, Stanford CS Dept., Stanford, CA (1991)

Halstead, D. and Forbus, K. (2005). Transforming between Propositions and Features: Bridging the Gap. *Proceedings of AAAI-2005*. Pittsburgh, PA.

Heit, E. & Rubinstein, J. (1994). Similarity and property effects in inductive reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. 20:411-422.

Heuer, R. (1999) *Psychology of Intelligence Analysis.* Center for the Study of Intelligence.

https://www.cia.gov/resources/csi/books-monographs/psychology-of-intelligence-analysis-2/

Hiatt, L., Harrison, A., & Trafton, G. (2011) Accommodating Human Variability in Human-Robot Teams through Theory of Mind. *Proceedings IJCAI* 2011.

Hinze, S.R., Slaten, D.G., Horton, W.S., Jenkins, R., & Rapp, D.N. (2014). Pilgrims sailing the Titanic: Plausibility effects on memory for facts and errors. *Memory & Cognition*, 42, 305-324.

Hobbs, J., Stickel, M., Appelt, D. & Martin, P. (1993). Interpretation as abduction. *Artificial Intelligence*, 63:69-142.

Johnson, H., & Seifert, C. (1994). Sources of the Continued Influence Effect: When Misinformation in Memory Affects Later Inferences. *Journal of Experimental Psychology:Learning Memory & Cognition*, 24, 1483-1494.

Kahneman, D. & Shane, F. (2002) Representativeness Revisited: Attribute Substitution in Intuitive Judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.) *Heuristics and Biases: The Psychology of Intuitive Judgment*. Cambridge University Press.

Kandaswamy, S. and Forbus, K. (2012). Modeling Learning of Relational Abstractions via Structural Alignment. *Proceedings of CogSci 2012*. Sapporo, Japan

Kandaswamy, S., Forbus, K., & Gentner, D. (2014) Modeling Learning via Progressive Alignment using Interim Generalizations. *Proceedings of CogSci 2014*.

Kennedy, W., Bugajska, M., Harison, A., & Trafton, G. (2009). "Like-Me" Simulation as an Effective and Cognitively Plausible Basis for Social Robots. *Int.J.Soc.Robot*, 1:181-194, doi:10.1007/s12369-009-0014-6

Laird, J. (2012) *The SOAR Cognitive Architecture*. MIT Press.

Lebiere, C., Pirolli, P., Thomson, R., Paik, J., Rutledge-Taylor, M., Staszewski, J., & Anderson, J. (2013) A functional model of sensemaking in a neurocognitive architecture. *Computational Intelligence and Neuroscience*, https://doi.org/10.1155/2013/921695

Lewandowsky, S., Ecker, U., Seifert, C., Schwarz, N. & Cook, J. (2012). Misinformation and its correction: Continued Influence and Successful Debiasing. *Psychological Science in the Public Interest*, 12(3) 106-131. Doi:10.177/1529100612451018

McCarthy, J. (1977). Epistemological problems of artificial intelligence, *Proc. IJCAI-77*, Cambridge, MA, pp. 1038–1044

Meltzoff, A. (2005) Imitation and Other Minds: The "Like Me" Hypothesis. In S. Hurley & N. Chater (Eds.), *Perspectives on Imitation: From Neuroscience to Social Science*. Vol. 2, pp. 55-77. MIT Press, Cambridge, MA.

Miller, G. (1955). The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. *Psychological Review*, Vol. 101, No. 2, 343-352.

Rabkina, I., McFate, C., Forbus, K. D., & Hoyos, C. (2017). Towards a Computational Analogical Theory of Mind. In *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 2949-2954)

Rabkina,I., McFate, C., & Forbus, K. D. (2018). Bootstrapping From Language in the Analogical Theory of Mind Model. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society*.

Rabkina, I., Kathnaraju, P., Roberts, M., Wilson, J., Forbus, K., & Hiatt, L. (2020). Recognizing the Goals of Uninspectable Agents.. In *Proceedings of AAAI Workshop on Plan, Activity, & Intent Recognition (PAIR)*. New York, NY.

Rapp, D., Donovan, A., & Salovich, N. (2020). Assessing and Modifying Knowledge: Facts versus Constellations. In *Handbook of Learning from Multiple Representations and Perspectives*, Routledge.

Rapp, D. & Salovich, N. (2018). Can't we just disregard fake news? The consequences of exposure to inaccurate information. *Policy Insights from the Behavioral and Brain Sciences*, 5, 232-239.

Rich, P. & Zaragoza, M. (2016). The Continued Influence of Implied and Explicitly Stated Misinformation in News Reports. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. 42(1):62-74.

Rich, P. & Zaragoza, M. (2020) Correcting Misinformation in News Stories: An investigation of Correction Timing and Correction Durability. *Journal of Applied Research in Memory and Cognition*, 9:310-322.

Seifert, C. (2002). The continued influence of misinformation in memory: What makes a correction effective? *Psychology of Learning and Motivation*, 41:265-292. DOI:10.1016/S0079-7421(02)80009-3

Thill, S., Caligiore, D., Borghi, A., Ziemke, T., & Baldassarre, G. (2013) Theories and computational models of affordances and mirror systems: An integrative review. *Neuroscience & Biobehavioral Reviews,* 37:3, pp. 491-521

Thomson, R., Lebiere, C., Anderson, J., & Staszewski, J. (2014) A general instance-based learning framework for studying intuitive decision-making in a cognitive architecture. *Journal of Applied Research in Memory and Cognition*.

Tversky, A., & Kahneman, D. (1974) Judgment under Uncertainty: Heuristics and Biases. *Science*, 185(4157):1124-1131.

Yalcin, O., & DiPaola, S. (2018) A computational model of empathy for interactive agents. *Biologically Inspired Cognitive Architectures*, 26:20-25.