Deceptive and Counter-Deceptive Machines
Papers from the AAAI 2015 Fall Symposium

# Analogical Abduction and Prediction: Their Impact on Deception

**Kenneth D. Forbus**
Qualitative Reasoning Group, Northwestern University
2133 Sheridan Road, Evanston, IL, 60208

### Abstract
To deceive involves corrupting the predictions or explanations of others. A deeper understanding of how this works thus requires modeling how human abduction and prediction operate. This paper proposes that most human abduction and prediction are carried out via analogy, over experience and generalizations constructed from experience. I take experience to include cultural products, such as stories. How analogical reasoning and learning can be used to make predictions and explanations is outlined, along with both the advantages of this approach and the technical questions that it raises. Concrete examples involving deception and counter-deception are used to explore these ideas further.

## Introduction

Deception can be viewed as an attempt to inject errors into the abductive and/or predictive processes of others**.** This makes understanding how human abduction and prediction work an important component of understanding deception and counter-deception. Most models of abduction and prediction rely on rules and other forms of logically quantified knowledge (e.g. Hobbs, 2004; Ovchinnikova 2012; Meadows et al. 2014). This paper makes a different proposal: That most human abduction and prediction is performed via analogical reasoning, fueled by the accumulation of experience and the refinement of that experience via analogical generalization into more rule-like representations (Gentner & Medina, 1998). If correct, this hypothesis has several implications for reasoning about deception, both in terms of how to deceive and how deception might be detected.

We start by briefly reviewing relevant background on analogical processing, and then discuss how analogy can be used for abduction and prediction. Particular features of analogical processing, such as the importance of surface similarity in retrieval, will be examined in terms of how they may impact deception. Finally, several technical issues are raised that need to be addressed within this approach.

## Background

We use Gentner's (1983) structure-mapping theory of analogy and similarity as a foundation. Analogy and similarity are viewed as comparisons involving structured, relational representations. (Given the central role of explanation and causation in social reasoning in general and in deception in particular, it is hard to see how any feature-vector account can provide much leverage.) Three operations are viewed as fundamental, and each builds upon the previous operations. We discuss each, and our computational models of them, in turn.

Matching takes as input two descriptions (*base* and *target*) and constructs one or more *mappings* between them. Mappings describe what goes with what, i.e. how entities and statements in one description align with another, via a set of *correspondences*. Mappings also include a *similarity score*, indicating the structural quality of the match. Finally, mappings include *candidate inferences*, suggestions of how structure in one description can be mapped to the other, based on the correspondences in the mapping. Candidate inferences can go in both directions (i.e., from base to target and from target to base), and are surmises, rather than being deductively valid. Our computational model of this process is the Structure-Mapping Engine (SME; Falkenhainer et al. 1989), which operates in polynomial time by using a greedy matching algorithm (Forbus & Oblinger, 1990; Forbus et al. under review).

Retrieval takes as input a *probe* description and a *case library*. Both the probe and the contents of the case library are again structured representations. The case library can in principle be enormous, e.g. one's entire episodic memory, although the theory is currently agnostic on this point. However, there is ample evidence that retrieval tends to be sensitive to surface similarity, although relational information is involved as well, and people prize structural remindings more than surface remindings (Gentner et al. 1993). Our computational model is MAC/FAC (Forbus et al. 1995), which captures the dissociation between the surface and structural properties via a two-stage process, designed for scalability. The first stage (MAC) uses a redundant feature vector representation for the probe and each case in the case library that is automatically computed from the structured representation. These *content vectors* are designed so that the dot product of two of them provides an estimate of how large a representation might be built via SME for the two corresponding structural descriptions. The content vector for the probe is matched against all of the content vectors for the cases in the case library, conceptually in parallel. The top three remindings are passed to the second stage (FAC). FAC uses SME to compare the structured descriptions from the case library elements returned via MAC against the probe, with again up to three remindings being returned if they are sufficiently close to the best.

The process of generalization involves incrementally assimilating new examples into an ongoing analogical model consisting of generalizations and outlier examples. Our computational model, SAGE (McLure et al. 2015) organizes analogical models via *generalization contexts*, each representing a concept. For example, spatial terms such as prepositions would each be represented via a separate generalization context. Given an example of a concept, one or more prior examples or generalizations are retrieved from the generalization context for that concept via MAC/FAC. If the similarity score is sufficiently high, then if the retrieved item was a generalization, the new example is assimilated into this generalization, and if the retrieved item was an example, then a new generalization is created by assimilating them. The assimilation process assigns a probability to each statement in the correspondences found via SME between the two items based on the frequency to which there is a statement that aligns in the examples that went into the generalization. For example, if 99 out of 100 swans seen were white and one was black, the probability for the color of a swan being white would be 0.99 and the probability for its color being black is 0.01. Entities that are not identical are replaced with generalized entities, which have only those properties that come from the statements in the description. For example, if Chicago and Rome were aligned within a generalization, a new entity would be created to replace them in all of the facts in the generalization. While more abstract, these new entities are still not logical variables. They share whatever attributes held for the entities that they came from (e.g. being a city, in this case). Moreover, two variables can be bound to the same value via unification, whereas two entities cannot be mapped to the same value via structure-mapping, because that would violate the 1:1 constraint.

These models are both compatible with existing psychological evidence and have been used to make novel predictions that have been borne out in laboratory studies. Moreover, they have been engineered to be components in performance-oriented systems, which enables larger-scale experiments than have been done with other cognitive models of analogy. Thus they form a solid basis for exploring the roles of analogy in abduction and prediction.

## Abduction and Prediction

Abduction and prediction are knowledge-intensive forms of reasoning. But what kinds of knowledge? The most common answer is some form of first-principles knowledge, i.e. logically quantified statements that can be instantiated on a wide range of situations. These can take multiple forms, e.g. forward chaining rules are commonly used for predictions, while backward chaining rules are commonly used for abduction. Special forms of such statements are often proposed to facilitate particular kinds of reasoning, e.g. causal laws (Strasser & Aldo, 2015) for reasoning about actions, or model fragments (Falkenhainer & Forbus, 1991) for reasoning about continuous phenomena. Such representations, by themselves, have two key problems:

1. Combinatorial explosions. The number of abductive proofs can skyrocket if the set of predicates that can be assumed is not tightly controlled. Similarly, forward chaining over even qualitative states is exponential in most domains. Tactics like beam search can help, but the fine-grained nature of such reasoning makes for a daunting challenge.
2. Logically quantified models are better at capturing what is possible than what is typical. For example, it is logically possible for flipped coins to land on their edges and for all of the air molecules in a room to suddenly rush to a corner, leaving its occupants gasping for breath. But neither of these are contingencies that we think of when flipping a coin nor walking into a room.

Another kind of knowledge, missing from such accounts, is what might be deemed experiential knowledge. Various attempts have been made to model this via extra layers that add probabilities (qualitative or quantitative) to logical statements. But that begs the question of where such probabilities come from.

People have vast amounts of experience, gleaned by interacting with the world and with their culture. Since our concern here is social reasoning and deception, we focus on cultural experience. (Hamsters and house cats sense, act, and learn from the physical world, and some birds make tools, but we are the only creatures that build schools.) By cultural experience I include conversations with other people and assimilating information about what people can do from cultural products (e.g. reading books and newspapers or watching TV and movies). While some statements in these products are generic, interpretable as logically quantified statements, most of them are about particulars. And this is where analogy comes in.

As Falkenhainer (1990) pointed out, the candidate inferences produced by a mapping can be deductive or abductive. If the candidate inference consists of an instantiation of a logically valid axiom and the mapping completely pins down the statements in it, then the inference provides a deduction. For example,

Base: `(HouseCat Nero)`
`(implies (HouseCat Nero)`
`(huntsMice Nero))`
Target: `(HouseCat Archie)`
Mapping: `Nero` ↔ `Archie`
`(HouseCat Nero)` ↔ `(HouseCat Archie)`

Assuming that all house cats hunt mice, then one may deductively conclude that Archie, too, hunts mice. On the other hand, if we know only

Target: `(huntsMice Archie)`

with the same correspondences, the best we can do is infer that one possible reason why Archie hunts mice is that Archie is a house cat. Archie could be a terrier or an exterminator just as easily, knowing this little about the situation. This means that we can use analogy both as a means of deduction and abduction. What is interesting about this mechanism is that, if one of the cases has a rich, elaborate explanation, the process of applying that explanation is handled by the matching process, all at once. It is like walking through a space of inferences in seven-league boots, rather than on tip-toes. This potentially makes analogical inference more efficient.

Importantly, analogical inference can still be used when the candidate inference is not a logically correct axiom. The qualification problem is well-known, i.e. there are an indefinite number of preconditions that might need to hold for an action to have its intended outcome. The results must be taken as surmises, but there are some heuristics which can be used to help evaluate them. For example, if the two examples are very similar, e.g. house cat to house cat, or even house cat to puma, then the inference is more likely to hold. Similarly, if the two examples are very different, e.g. house cat and dolphin, the inference is less likely to hold (Heit & Rubenstein, 1994).

How can we arrange memories so that the analogical generalizations produced are useful for abduction and prediction? A promising candidate organization scheme is to carve up experiences (personal or from cultural narratives) into *causal triple cases* that include causally connected statements, e.g.

(cause <*ante* > <*conse*>)
<*description of* <*ante*>>
<*description of* <*conse*>>

where <*ante*>, <*conse*> are situations or events, represented via explicit tokens in the ontology (e.g. situations can include qualitative states, configurations holding for some piece of space-time – I do not mean situation calculus here, but the Cyc ontology meaning of situation). Such cases would be stored twice, once in a generalization context for explanations and once in a generalization context for predictions. That is, when storing in the context for explanations, MAC/FAC would be given a requirement that a match be found for <*conse*>. This provides relative frequency information as to possible causes for <*conse*>. Similarly, when a new case is added to the context for predictions, MAC/FAC would be required to provide a match for <*ante*> in any retrieval, thereby constructing generalizations which provide relative frequency estimates for possible consequences of <*ante*>. Figure 1 illustrates, where Cg is a generalization of <*conse*>s and Ag is a generalization of <*ante*>s.

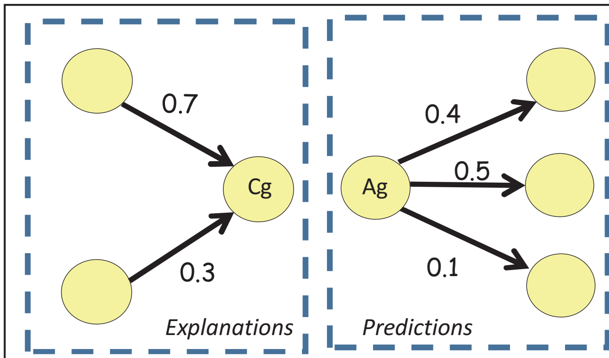

*Figure 1: How generalization contexts might be used for facilitating analogical abduction and prediction.*

I think this account is a promising start as a model for analogy's role in human abduction and prediction. We constantly see behaviors unfolding, and hear (or construct) explanations for why they occur. Those specific explanations can be done much more straightforwardly than postulating logically quantified rules that will work in all potential situations. If the conclusions are about a reason for something, or an event that happened in the past, we can consider it to be an abductive inference. If the conclusions are about something that might happen in the future, we can use that as a prediction.

This account has several implications for deception and counter-deception. First, I assume that people track occurrences of deception, since it is useful to understand when particular actors, categories of actors (e.g. telemarketers), and situations (e.g. playing poker, political campaigns) are likely to involve deception. There are multiple ways that this could be accomplished in this framework. The simplest is to mark some relationships in the stored cases as lies, so that they will be retrieved as such, and probabilities as to whether someone is lying could come out of the matching process. When an explanation as to the reason for lying is available, including that as well as a causal connection for the lie would help in providing evidence for future cases where deception might be at work. In addition to lying, tracking someone's accuracy is also useful, as well as properties such as whether they tend to check their sources. Modeling these epistemic properties of someone's social interactions could also be facilitated via the frequency information kept in analogical generalizations.

From the standpoint of planning deception, this account provides some computational rationale for several well-known heuristics for deception:

- Use excuses that will be highly salient to the victim. Since the evaluation of the plausibility of the deceptive alternative will involve analogical retrieval, picking highly salient experiences for the victim means the deception is more likely to resonate and be undetected.
- Avoid excuses that have become stereotyped. A soon to be obsolete classic is "The dog ate my homework", for instance. Once an excuse can be mapped to such a stereotype ("the printer ate my homework"), it starts being associated with being an excuse rather than a reason.
- Use excuses that are hard to check. If the kind of information that it would take to invalidate it is hard to obtain, as estimated by causal triple cases pertaining to information-gathering, it is a better excuse.

Some examples of this are considered in the next section.

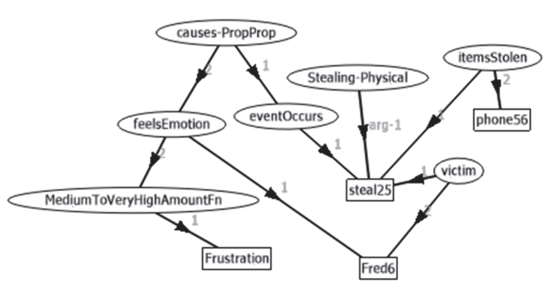


*Figure 2: Formal representation using Cyc ontology.*

## Some Examples

Considering the following very simple story:

"Fred was angry because his phone was stolen."

Figure 2 shows a translation of this into predicate calculus. Now suppose we are given the first sentence of another story:

"George was furious."

Comparing the representation for this sentence with the prior sentence, SME produces the following candidate inferences:

```
(victim (skolem steal25) George7)
(causes-PropProp
 (eventOccurs (skolem steal25))
 (feelsEmotion George7 (HighAmountFn Anger))
```

That is, the reason why George is furious might be that he too was the victim of a stealing event. The term `(skolem steal25)` is an *analogy skolem*, a new individual whose existence is hypothesized because of the projected relational structure. If one tentatively accepts the existence of this event, and extends the mapping with the requirement that the new event in the target map to `steal25`, then another skolem will be introduced, in this case `(skolem phone56)`, to represent the item that was stolen. This gives us another expectation, that something like a phone was stolen from him.

One of the tasks in analogical reasoning and learning is resolving such skolems. There are several techniques that have been used, including postulating a new category of individual in the domain theory (Falkenhainer, 1990), projecting the base domain individual into the target (e.g. Klenk & Forbus, 2009, good for within-domain analogies), and using a constraint solver over the union of constraints implied by the mapping (e.g. Forbus et al. 2003, good for spatial entities). Context often provides answers: If the next sentence were

"His laptop was gone."

then we could conjecture that the thing like the cell phone is the laptop. Moreover, if we know that something being gone is one of the consequences of it being stolen, we can project the act of stealing, which provides a possible explanation of why the laptop is gone.

Let us consider how a clever thief might use their experience to change their modus operandi to reduce their likelihood of being caught. A victim who doesn't notice a theft can't report it. How might they detect the theft of a laptop? There are several possible ways: They might find their bag lighter than it would be otherwise, or squish flatter than it normally does. They might try to get it out, in order to use it. All of these possibilities could be predicted based on experience, as anyone who has noticed a missing device, for whatever reason, knows. Removing the detectable differences would change the world sufficiently that the causal antecedents (as encoded in the mapped relational structure) would be incorrect and thus the prediction would not hold.

Going to an extreme, replacing the victim's laptop with a functioning machine of the exact same make and model would minimize the differences and thus delay detection the most. (Such reasoning, I assume, is based on comparative analysis via qualitative models (Weld, 1990), rather than having, say, numerical probabilities for detection as a function of quantitative differences.) Unless the thief's goal is the information on the laptop rather than the physical device itself, this is an extreme expense and probably not worth it. Replacing it with something roughly similar: A glossy magazine, in the case of modern lightweight machines, or a phone book, for a turn-of-the-century model, might increase risk of earlier detection but would be a lot more profitable. How might that tradeoff be evaluated? If you keep track of how often your fellow laptop thieves are arrested and what their preferred techniques are, you can build up (again via analogical generalization) an estimate of the likelihood of being caught for various techniques.

## Open Questions

The account presented here raises a number of open questions. The four most obvious are:

1. *Granularity of memory*. Using a single generalization context for all explanations and predictions may or may not scale. On the other hand, there are reasons to build up analogical models (via generalization) of categories of events and/or situations independent of prediction and explanation, such as being able to recognize them. We have explored automatic introduction of hierarchal generalization contexts, to model human conceptual structure more closely (Liang & Forbus, 2014), so that may be a better way to proceed. However, current hierarchical clustering methods are batch-oriented, which is not psychologically plausible. Moreover, while events often correspond to verbs and thus are very well ontologized, situations are less so.
2. *Skolem resolution strategies*. Even within-domain analogies can involve more sophisticated strategies for resolving skolems than simply projecting the base entity. If the next sentence in the George example had been "He lost his laptop." then most readers would align a LosingSomething event with the Stealing-Physical event. This could involve either estimating the similarity of the two types of events to see if they are plausibly aligned, or decomposing them to see if the specific ways in which they are alike could explain what happened (i.e. they both involve loss of control over a piece of property).
3. *Estimation of probabilities for candidate inferences*. As noted above, SAGE constructs probabilities for every statement in a generalization. The probabilities are frequency information, i.e. how often does a statement in the cases that make up the generalization align with that statement in the generalization. It is straightforward to convert such probabilities into estimations of various explanations or predictions, by making a closed-world assumption over the set of causal transitions. However, it is less obvious how to combine those probabilities with the similarity of a retrieved generalization to the current situation to estimate the likelihood of particular predictions or explanations.
4. *Qualitative probabilities in cultural products*. Cultural products can provide a skewed view of human life: In IBM's PRISMATIC knowledge base, for example, "the most common actions that Pat Garrett was involved in are kill, shot, and capture; and the most common object of these actions is Billy the Kid." (Fan et al. 2012, p 5:1). That Pat Garrett far more frequently breathed, ate, and slept are simply not mentioned in stories, because they are uninteresting even while being vital to know, at least tacitly. The idea of tracking all of the mundane aspects of everyday life in order to estimate accurate probabilities of them seems unlikely. Instead, once we establish something as normal, frequent, and typical, we may stop accumulating information and simply mark them in some way that is recognized when combining information as being a normal default. Understanding these representations and how they are used is an interesting challenge.

## Other Related Work

Clark (Clark, 2011) has investigated building systems whose lies are more plausible because they coincide with well-known cognitive illusions in human reasoning. It would be interesting to see under what conditions these same illusions occur under analogical reasoning.

Bello & Guarini (2010) explore some of the representational and reasoning capabilities of mental simulation required for mindreading, an essential operation in deception. They treat the problem as a first-principles reasoning task, using Polyscheme's inheritance plus overrides mechanism. This paper has not addressed mind-reading, however it is interesting to note that analogical reasoning has been proposed as a means of performing the mental simulation involved in learning about others (e.g. Meltzoff 2007). This seems like a potentially useful approach to explore.

Heuristics appear to play a major role in human reasoning (Tversky & Kahneman, 1974; Gigerenzer, Todd, & ABC group, 1999). There seem to be useful connections between these ideas and analogical reasoning in two directions. Analogical retrieval seems a useful candidate for modeling some of the accessibility-based heuristics, and heuristics might be usefully employed in skolem resolution.

## Conclusions

Humans are creatures with limited capabilities and resources, operating with incomplete and uncertain information in complex environments. Shortcuts are always necessary. Analogical reasoning and learning over experience potentially provides a basis for rapidly reasoning and learning, without requiring complete or correct domain theories. To the extent that people are indeed using analogy heavily in everyday reasoning, those wishing to deceive, or to counter deception, should be aware of its psychological properties. This paper has provided an initial step in that direction.

Beyond this initial step, there are two things that need to be done. The first is further theoretical analyses, to better understand the issues involved in using analogical reasoning and learning at scale. The second are experimental investigations: What will it take to create systems that achieve human-like analogical reasoning at scale? This provides several daunting challenges, including creating systems by which such experiences can be automatically encoded with minimal human labor, both to reduce tailorability and for practicality.

## Acknowledgements

I thank Tom Hinrichs, David Barbella, and Joe Blass for helpful suggestions. This research was supported by the Cognitive Science Program of the Office of Naval Research.

## References

Bello, P. & Guarini, M. (2010). Introspection and Mindreading as Mental Simulation. In *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society*.

Clark, M., (2011). Mendacity and Deception: Uses and Abuses of Common Ground. In *AAAI Fall Symposium on Building Representations of Common Ground with Intelligent Agents*.

Falkenhainer, B. (1990). A unified approach to explanation and theory formation. In Shrager, J. & Langley, P. (Eds.), *Computational Models of Scientific Discovery and Theory Formation*. San Mateo, CA: Morgan Kaufmann.

Falkenhainer, B., Forbus, K.D., & Gentner, D. (1989). The Structure-Mapping Engine: Algorithm and Examples. *Artificial Intelligence*, 41, 1-63.

Fan, J., Kalyanpur, A., Gondek, D., & Ferrucci, D. (2012). Automatic knowledge extraction from documents. *IBM Journal of Research and Development*, 56(3/4), pp 5:1-5:10.

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

Forbus, K. and Gentner, D. (1997). Qualitative mental models: Simulations or memories? Proceedings of the Eleventh International Workshop on Qualitative Reasoning, Cortona, Italy, June 3-6, pp. 97-104.

Forbus, K. D., Klenk, M., & Hinrichs, T. (2009). Companion cognitive systems: Design goals and lessons learned so far. IEEE Intelligent systems, 24(4), 36-46.

Forbus, K.D. & Oblinger, D. (1990). Making SME greedy and pragmatic. *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*. Hillsdale: Erlbaum.

Forbus, K., Usher, J., & Chapman, V. (2003). Qualitative Spatial Reasoning about Sketch Maps. *Proceedings of the Intelligence User Interfaces Conference*, 2003, January: Miami, Florida.

Forbus, K., Ferguson, R., Lovett, A., & Gentner, D. (under review). Extending SME to Handle Large-Scale Cognitive Modeling.

Gentner, D., 1983. Structure-mapping: A theoretical framework for analogy. Cognitive Science 7(2), 155-170.

Gentner, D., & Medina, J. (1998). Similarity and the development of rules. *Cognition*, 65, 263-297.

Gentner, D., Rattermann, M.J., & Forbus, K.D. (1993). The roles of similarity in transfer: Separating retrievability from inferential soundness. *Cognitive Psychology* 25, 524-575.

Gigerenzer, G., Todd, P. & the ABC Group (1999). *Simple Heuristics That Make Us Smart*. Oxford University Press.

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

Hobbs, J.R. 2004. Abduction in Natural Language Understanding. In Horn, L. and Ward, G. (eds.), *Handbook of Pragmatics*, pp. 724-741, Blackwell.

Klenk, M. and Forbus, K. (2009). Analogical Model Formulation for AP Physics Problems. *Artificial Intelligence*, 173(18), 1615-1638.

Liang, C. and Forbus, K. (2014). Constructing Hierarchical Concepts via Analogical Generalization. *Proceedings of the Cognitive Science Society*.

McLure, M.D., Friedman S.E. and Forbus, K.D. (2015). Extending Analogical Generalization with Near-Misses. *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, Austin, Texas.

Meadows, B., Langley, P., & Emery, M. (2014). An Abductive Approach to Understanding Social Interactions. *Advances in Cognitive Systems* 3, pp 87-106.

Meltzoff, A. (2007) The "like me" framework for recognizing and becoming an intentional agent. *Acta Psychologica*, 124(1), pp. 26-43.

Strasser, Christian and Antonelli, G. Aldo, Non-monotonic Logic. *The Stanford Encyclopedia of Philosophy* (Fall 2015 Edition), Edward N. Zalta (ed.).

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

Ovchinnikova, E. (2012) *Integration of World Knowledge for Natural Language Understanding*, Atlantis Press, Springer

Weld, D. (1990). *Theories of Comparative Analysis*. MIT Press.