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

Position Paper: Knowledge-Based Mechanisms for Deception

Scott E. Fahlman
Carnegie Mellon University, Language Technologies Institute
sef@cs.cmu.edu

Abstract

In an earlier paper, I described in some detail how a system based on symbolic knowledge representation and reasoning could model and reason about an act of deception encountered in a children's story. This short position paper extends that earlier work, adding new analysis and discussion about the nature of deception, the desirability of building deceptive AI systems, and the computational mechanisms necessary for deceiving others and for recognizing their attempts to deceive us.

Background

At the 2011 AAAI Fall Symposium Series, in the workshop on Advances in Cognitive Systems, I presented a paper (Fahlman 2011) entitled "Using Scone's multiple-context mechanism to emulate human-like reasoning". That paper (available online) presented a brief overview of the Scone knowledge-base system that we are developing at Carnegie Mellon University, with special emphasis on Scone's *multiple-context mechanism* – a facility that allows Scone to create multiple distinct-but-overlapping world-models within a single knowledge-base.

The 2011 paper goes on to describe how this multiple-context capability can be used to model various aspects of human-like reasoning: the ability to reason about hypotheses and counter-factuals; the ability to describe and reason about what is true at a specific time; how an action or event can change the state of the world from a before-state to an after-state; the ability to reason about changes that occur spontaneously, without a specific action taking place (e.g. ice melting); and the ability to model the mental states (belief-states, knowledge-states, goal-states) of another agent, whether human or robot.

This last capability is illustrated using the example of a scene in a children's story. The third little pig is safe in his house, the wolf is outside, the pig knows that, and the wolf knows that the pig knows. So the wolf tries to *deceive* the pig into believing that he has gone away. In one ending of the story, the deception is successful: the pig opens the door, and is eaten. In the alternative ending, the pig "sees through" the deception and remains safely inside the house.

The analysis in this 2011 paper, and the knowledge-based mechanisms described there, seem quite relevant to the current 2015 symposium on "Deceptive and Counter-Deceptive Machines". Rather than repeating or trying to summarize the material in the earlier paper, I will just include it by reference. This short position paper extends that earlier work, adding new analysis and discussion about the nature of deception, the desirability of building deceptive AI systems, and the computational mechanisms necessary for deceiving others and for recognizing and countering their attempts to deceive us.

Can Deception Ever Be A Good Thing ?

Yes, of course. Deception has been around for as long as there has been complex life on earth.

- The earliest form of deception was camouflage: organisms evolved coloration that blends into the background. The deceptive message here is simple: "I'm a rock (or a weed or some pebbles) – not your lunch and not a lurking predator."
- Some animals, including flounders and many cephalopods, have the ability to *adaptively* change their color and texture as needed to blend into various backgrounds.
- Some animals can puff themselves up to a larger apparent size to scare away predators or to attract mates.
- In the realm of behavior, rather than appearance, animals make use of deceptive movements in combat or pursuit situations. For example, an antelope being pursued by a faster predator can sometimes escape by feinting in one direction, then turning sharply in the other. The pursuer is deceived into overshooting the prey, and it takes time

to recover. The feints are probably reflexive – not consciously planned.

In every case, these deceptive acts or attributes confer some survival advantage on the deceiver – that's why these capabilities evolve and why they persist in the population. The deceiver's advantage usually comes at the expense of the "victim", so there is also a selective advantage in being able to identify or "see through" some deceptions.

For this workshop, we will focus on *behavioral* deception at the cognitive level – that is, deception that is deliberately planned and executed. While some animals may be capable of this, we will focus on deception between adversaries who both possess human-like cognitive ability.

At first glance, it might seem crazy to build a capability for deception into our robot servants. Do we want an autonomous car to lie to us about how it was dented or why it was out driving around on its own between midnight and 4am? Probably not. There is (or should be) no adversarial relationship here, so it is best for everyone if the robots are truthful.

But as soon as these machines start acting in an *adversarial environmen*t – combat, diplomacy, a game of some kind, or even a negotiation over price – deception suddenly becomes an important tool for achieving the machine's (or rather, its owner's) goals. In the case of a military drone, the ability to confuse an adversary about its status and intentions may be essential both for completion of the mission and for the drone's own survival.

Manipulation of Mental States

Fundamentally, this cognitive-level deception is a matter of modeling and manipulating your opponent's mental state. If we say that agent **D (**the "deceiver") takes action **A** to *deceive* agent **T** (the "target"), we mean, more or less, this:

- Initially, both **D**'s mental state and **T**' s mental state correctly represent the real world, or at least those aspects of the true world-model that are important in this situation.
- **D** performs action **A** – that is, **D** *says* or *does* something – that is intended to induce **T**' s to modify his mental state so that it no longer corresponds to reality.
- **T**'s new mental state is in some way advantageous to **D** – for example, it may cause **T** to open a door when **D**, a hungry wolf, is outside.
- If action **A** has the desired effect on **T**' s belief-state, then we say that the deception was *successful*; if **T** does not make the mental change that **D** desires, the deception was *unsuccessful* – in other words, **T** didn't "fall for" the trick.

Note that if **D** actually *believes* that the target mental state for **T** is correct in reality, then we don't call this "deception" – we call it "education" or "persuasion". The conventional moral strictures regarding this act are different from a deception, even if **D**'s belief is sincere but mistaken.

Note also that if action **A** is in the form of *saying something* rather than *doing something*, we call this action a "lie" – but only if the statements made by **D** are *literally and unambiguously untrue*. If they are just misleading, perhaps because **D** is presenting true statements very selectively, we generally don't call this "lying" or "fraud" – it's "persuasion" or "spinning". Again, the conventional moral and legal strictures are different. An actual lie, under some circumstances (e.g. testifying under oath) can be punished as perjury, but "misleading" generally is a lesser offense.

If the *motivation* for **D**'s description is benign or at least not harmful to **T** – perhaps **D** wants to spare **T** some pain or embarrassment, or wants to prevent **T** from doing something rash – we may call this a "good lie" or a "little white lie". Some people may view this as a moral transgression, but most people would forgive this as long as **D**'s motive is not to gain some advantage at **T**'s expense.

All of these variations are cousins in the hierarchy of action types. They all involve **D** trying to alter the mental state of **T**, but there are crucial differences in how society views these actions, depending on motive, method, and circumstances.

In the context of a game – *simulated* reality – the social rules are different. It is impossible to consistently do well at a multi-player game like poker or the board game Diplomacy without engaging in frequent and blatant deception. Even when playing for real money, deception is considered acceptable behavior – "all part of the game" – and it is not supposed to affect the level of trust between **D** and **T** in the real world, outside of the game context.

An interesting issue, beyond the scope of this paper, is what happens if the same agents interact *repeatedly* in an adversarial context, whether in a series of games, business dealings, or combat. In this case, deceptive behavior in one situation may be remembered and may affect behavior in later dealings. An agent may develop a *reputation* for honesty, general deceptiveness, or for preferring certain kinds of deceptive maneuvers such as bluffing. This reputation itself may become something that the agent tries to manipulate. This is an area where qualitative AI reasoning meets game theory – an interesting area for further research.

Required KRR Mechanisms for Deception

If deception is all about actions that affect mental states, and if we want to represent and reason about deception, we need a knowledge representation and reasoning (KRR) system that can easily represent a large number of mental states (distinct world models), keeping them all in memory at once without confusion, moving between them easily.

Scone's multiple-context model was designed specifically to deal with this kind of problem. Each mental state (knowledge-state, belief state, goal state, hypothesis, etc.) is represented by a distinct Scone context. The way these contexts operate is described in the 2011 paper, so I will not repeat that here. But I believe that the critical requirements for the knowledge base of a deceptive of counter-deceptive agent are as follows:

First, it must be a lightweight operation, both in memory and in processing time, to create a new context, so that we can afford to have many of them around at once. More specifically, it must be easy to create a context X that is "just like" existing context Y, except for some explicitly stated differences – both additions of new entities and statements that are only present in X, and cancellation of some things in Y that would otherwise be inherited by X.

Second, a context must function both as a container – a world model within which we can perform queries and inferences – and as a full-fledged entity in the knowledge base. This allows us to use the full expressive power of the knowledge-base system (KBS) to talk about an entire hypothesis or world-view. Within an agent's KBS, some context or set of contexts will be labeled as that agent's *current true beliefs*, some may be possible but uncertain beliefs, and some may be pure fiction.

This suggests that, as a minimum, we need default reasoning with explicit exceptions, and we need higher-order logic so that statements about statements are allowed, nested to any level. Scone's multiple-context mechanism is just a convenient way of re-packaging higher-order logic. It is hard to see how a system based on first-order logic or a less-expressive subset of FOL could get any real traction in dealing with deception and mental states.

Reasoning Strategies for Deception and Counter-Deception

Scone-like mechanisms for representation and fast simple inference may be necessary *enabling technolog*ies for complex deception or counter-deception, but of course this is not the whole story. Built upon this substrate of knowledge, making heavy use of it, are the higher-level reasoning mechanisms and strategies that actually plan and refine the deceptive words or actions, or that try to diagnose whether any deception is taking place.

What reasoning mechanisms and strategies are required? For now, let's focus on complex deceptions of the kind we see in detective stories. These illustrate the full range of mechanisms required. Simpler deceptions – a single lie about a single event – will often need many of these mechanisms as well.

Planning a complex deception – let's say a crime – requires that **D**'s mental machinery includes a planner – not necessarily an *optimal* planner, but one that is flexible enough to consider many alternatives and to polish the resulting plan until there is a reasonable chance that the plan will succeed and that the police will be deceived about what actually happened – who did it, how, why, and whether any crime was committed at all.

Such a planner requires a great deal of world-knowledge, both static (entities, properties, and relations), and recipes or plan-templates that can suggest an action sequence for accomplishing some goal. The planner also requires the ability to *simulate* (down to some level of detail) the results of executing the steps in a plan, in order to do as much debugging as possible before trying to execute the plan in the real world.

So, if **D** wants to poison someone, he must first have some knowledge of poisons: which ones are reliable, how long they take to act, how they are administered, how to obtain them without leaving a clear trail, and which poisons are easy for the police or the coroner to detect after the fact. It also helps if **D** knows which poisons have alternative uses that are both plausible and non-criminal, in case he is caught with the poison.

D may want to claim that he was at home at the time of the murder and that his car never left the garage. But if it's a rainy night and the car is in the garage, dripping wet, when the police arrive, this apparent contradiction may lead them to doubt **D**'s story. **D** may anticipate this and dry the car before the police arrive, or he might leave it outside in the rain, so that its wet state is not an anomaly.

So in addition to a good knowledge of poisons, **D**'s planning process must also have more mundane, common-sense knowledge about a vast range of topics – in this case, cars, garages (which have roofs), rain, things getting wet, how long things take to dry, that water on a car is more easily visible than, say, some unique kind of mud on the tires, and so on.

In this case, the police investigator is playing the role of **T**, the intended target of the deception. What reasoning mechanisms and strategies does **T** require in order to detect and understand the deception?

I would argue that the capabilities needed by **T** are almost the same as those needed by **D**: world-knowledge both general and domain-specific (e.g. poisons); a library of action-types and their effects; a library of plan recipes that **T** can try to fit to the case at hand, and a planner that can produce various alternative plans that **D** might have employed, or attempted to employ. That is, **T** must have the same planning capabilities as **D** (or better) in order to reconstruct what **D** may have been doing and why.

An additional capability needed by **T**, but perhaps not by **D**, is a *plan-recognition* capability: given a few observed steps in someone else's plan, what is the larger plan within which these steps appear, and is that plan plausible in the current context?

So, if **D** was seen putting on a ski mask on the night of the murder, and if it is rainy but not particularly cold outside, **D** probably is worried that he might be seen and wants to make sure that his face cannot be identified. So that is probably a part of **D**'s larger plan. There has been a lot of research in the robotics world on plan identification – this just takes it to a more abstract and much more complex domain.

For both the deceiver **D** and the counter-deceiver **T**, there is a requirement for *adversarial reasoning* – the kind that we see in chess programs: "If I make move M1, your best counter-move is … after much computation … M2, and then my best move on this branch is M3" and so on. It is important for **D** to consider the actions **T** might take to gather additional evidence. Of course, chess is a game of perfect information and has a very small, bounded set of options at each level. Real-world adversarial reasoning is much more complex, so considering *every* possibility is impossible. These agents don't have to be perfect – just better than their adversary.

A skilled deceiver **D** will first perhaps sketch his plan, then try to simulate how that plan will look to **T**. If there are incongruities that **T** is likely to notice, weakening the credibility of the deception, then **D** can try to modify the original plan. So if **D** anticipates that it may be raining and that **T** will notice the wet car, **D** can modify the plan to include drying off the car. Ah, but **T** might obtain a search warrant and find the wet towels, so **D** further modifies the plan to get rid of the towels, somewhere off the premises, before **T** arrives (but **D** can't use his car to do this). And so it goes, each agent trying to consider all the likely possibilities and each trying to anticipate the other's actions and thought processes.

Conclusions

None of the reasoning here is straightforward logical deduction. **D** is trying to construct a deception that, while not guaranteed to work, is *unlikely* to lead **T** to conclude that a deception is taking place. **T** is looking for the most-likely or most-plausible interpretation for what is going on. In a criminal investigation, **T** may be looking for "proof" of **D**'s deception, but this is not a mathematical "proof" – it is just enough to persuade a jury "beyond a reasonable doubt" that **T**'s version of the story is the only plausible one, and that **D**'s alternative version is not truthful. So the reasoning described here is much more like abductive reasoning (searching for the most likely cause for some observed state of the world) than like a formal deductive proof.

This also may involve a form of case-based reasoning. The elements of a case may remind **T** of a specific case that he has worked on before, or perhaps just heard about. (For this to happen, the knowledge base must be good at approximate, partial matching.) That old case may suggest a framework for what is really going on – a template that can be modified to fit the case at hand. These stored cases may be kept in more-or-less raw form, or they may have undergone some pre-processing to generalize them and create meta-information about them. Either way, a wealth of stored experiences can be valuable.

I should emphasize that this analysis of the reasoning strategies is all speculative, as of now. In the Scone research group, we have discussed these things, but we have not yet actually tried to build any deceptive AI systems. As argued in this paper, there is a lot of machinery required to plan and execute complex deceptions, but the striking thing is how little of this machinery is needed *only* for deceptive and counter-deceptive reasoning.

There is great overlap between the mechanisms I have described here and the mechanisms we are building for *co-operative, truthful* human-robot interaction in domains such as vehicle maintenance. In both cases we need a KB with a multi-context mechanism, static domain knowledge, a library of recipes (plan templates) for accomplishing various goals in the domain, a planner that can make use of these recipes, and a (mostly qualitative) simulator against which to test our plans before we try them in the real world. (Once we start executing the plan it may be difficult or impossible to back up and undo some actions already executed.) In cooperative human-robot scenarios, we need the capability to model the goal-states and belief-states of the other agents in order to predict what these agents will do, and we need the capability to recognize their intentions from a few observed actions. We may occasionally need to intervene in the mental state of other agents (i.e. tell them something) in order to prevent wasted effort.

So it would appear that same substrate of fundamental representation and reasoning mechanisms are needed for real-world problem-solving, for multi-agent cooperation on tasks, and for deception and counter-deception. Only the adversarial reasoning and a few of the higher-level strategies seem to be unique to the deception domain.

Acknowledgments:

The work reported here was supported in part by the U.S. Office of Naval Research, under grant N000141310224. Any opinions, findings, conclusions, or recommendations expressed here are those of the author, and do not necessarily reflect the views of ONR or the U.S. government.

References

Fahlman, S. E. 2011. Using Scone's multiple-context mechanism to emulate human-like reasoning, Proceedings of the AAAI Fall Symposium on Advances in Cognitive Systems. http://www.cs.cmu.edu/~sef/scone/publications/ACS-2011.pdf