(How) Can AI Bots Lie?

A Formal Perspective on the Art of Persuasion

Tathagata Chakraborti
IBM Research AI
Cambridge MA 02142 USA
tchakra2@ibm.com

Subbarao Kambhampati
Arizona State University
Tempe AZ 85281 USA
rao@asu.edu

Abstract

Recent work on explanation generation (Chakraborti et al. 2017) for decision-making problems has viewed the explanation process as one of *model reconciliation* where an AI agent brings the human mental model (of its capabilities, beliefs and goals) to the same page with regards to a task at hand. This formulation succinctly captures many possible types of explanations, as well as explicitly addresses the various properties – e.g. the social aspects, contrastiveness and selectiveness – of explanations (Miller 2018) studied in social sciences among human-human interactions. However, it turns out that the same process can be hijacked into producing "alternative explanations" – i.e. explanations that are not true but still satisfy all the properties of a proper explanation. In previous work (Chakraborti and Kambhampati 2019), we have looked at how such explanations may be perceived by the human in the loop, and alluded to one possible way of generating them. In this paper, we go into more details of this curious feature of the model reconciliation process and discuss similar implications to the overall notion of explainable decision-making.

The Model Reconciliation Process

One of the root causes[1] for the need of an explanation is that of model differences between the human and the AI agent. This is because, even if an agent makes the best decisions possible given its model, they may appear to be suboptimal or *inexplicable* if the human has a different mental model of its capabilities, beliefs and goals. Thus, it follows that the explanation process, whereby the AI agent justifies its behavior to the human in the loop, is one of model reconciliation.

The Model Reconciliation Process $\langle M^R, M_h^R, \pi \rangle$ takes in the agent model M^R, the human mental model of it M_h^R, and the agent decision π which is optimal in M^R as inputs and produces a model $\bar{M}_h^R$ where π is also optimal.

- **An Explanation** ϵ is the model difference $\bar{M}_h^R \Delta M_h^R$.

Thus, by setting the mental model $\bar{M}_h^R \leftarrow M_h^R + \epsilon$ (through means of some form of interaction / communication), the human cannot come up with a better *foil* or decision $\hat{\pi}$, and hence we say that the original decision π has been *explained*. This is referred to as the **contrastive property** of an explanation. This property is also the basis of persuasion since the human, given this information, cannot come up with any other alternative to what was done.

So how do we compute this model update? It turns out that there are several possibilities (Chakraborti et al. 2017), many of which have the contrastive property.

Minimal Explanations These minimize the size of an explanation and ensure that the human cannot find a better foil using the fewest number of model updates. These are referred to as *minimally complete explanations* or MCEs.

$$\epsilon_{MCE} = \arg\min \bar{M}_h^R \Delta M_h^R$$

Monotonic Explanations It turns out that MCEs can become invalid on updating the mental model further, while explaining a later decision. *Minimally monotonic explanations* or MMEs, on the other hand, maintain the notion of minimality as before but also ensure that the given decision π never becomes invalid with further explanations.

$$\epsilon_{MME} = \arg\min \bar{M}_h^R \Delta M_h^R \text{ such that}$$
$$\text{any } M^R \setminus \bar{M}_h^R + \epsilon \text{ is a solution to } \langle M^R, \bar{M}_h^R, \pi \rangle$$

Alternative Explanations

So far, the agent was only explaining its decision (1) with respect to and (2) in terms of what it knows to be true. Constraint (1) refers to the fact that valid model updates considered during the search for an explanation were always towards the target model M^R which is, of course, the agent's belief of the ground truth. This means that (2) the content of the model update is also always grounded in (the agent's belief of) reality. In the construction of lies or "alternative facts" to explain, we start stripping away at these two considerations. There may be many reasons to favor them over traditional explanations:

- One could consider cases where team utility is improved because of a lie. Indeed, authors in (Isaac and Bridewell 2017) discuss how such considerations makes it not only preferable but also necessary that agents learn to deceive.
- A specific case of the above can be seen in terms of difficulty of explanations – a lie can lead to an explanation that is shorter and/or easier to explain... or are more likely to be accepted by the human.

[1] Considering the computational capability of the human, this is the *only* cause for an explanation.

Lies of Omission

These deal with cases when the agent provides a model update that negates parts of its ground truth – e.g. saying it does not have a capability it actually has. This is, in fact, a curious outcome of the non-monotonicity of the model reconciliation process. Consider the case where the initial estimate of the mental model is empty or ϕ – i.e. we start by assuming that the human has no expectations of the agent. Furthermore, let the minimally complete and minimally monotonic explanations for the model reconciliation process $\langle M^R, \phi, \pi \rangle$ produce intermediate models M^R_{MCE} and M^R_{MME} respectively. Now, imagine if the actual mental model M^R_h lies somewhere between[2] M^R_{MCE} and M^R_{MME}. Then, it follows that, if we start making model updates towards an empty model in the direction opposite to the real model M^R, we can get to an explanation $M^R_h \setminus M^R_{MCE}$ which involves the agent stating that its model does not contain parts which it actually does.

- **A Lie of Omission** can emerge from the model reconciliation process $\langle \phi, M^R_h, \pi \rangle$.

A solution to this particular model reconciliation process may not exist – i.e. a lie of omission only occurs when the initial mental model lies between M^R_{MCE} and M^R_{MME}. However, they happen to be the easiest to compute due to the fact that they are constrained by a target model (which is empty) and do not requite any "imagination". More on this when we discuss *lies of commission*.

Lies of Commission

In lies of omission, the agent omitted constraints in its model that actually existed. It did not make up new things (and having the target model as M^R in the original model reconciliation process prevented that). In lies of commission, the agent can make up new aspects of its decision-making model that do not belong to its ground truth model. Let $\mathbb{M}$ be the space of models induced by M^R and M^R_h.[3] Then:

- **A Lie of Commission** can emerge from the model reconciliation process $\langle M, M^R_h, \pi \rangle$ where $M \in \mathbb{M}$.

We have dropped the target here from being M^R to any possible model. Immediately, the computational problem arises: the space of models was rather large to begin with – $O(2^{|M^R \Delta M^R_h|})$ – and now we have an exponentially larger number of models to search through without a target – $O(2^{|M^R|+|M^R_h|})$. This should be expected: after all, even for humans, computationally it is always much easier to tell the truth rather than think of possible lies.[4]

[2] As per the definition of an MME, if the mental model is between the MME and the agent model, then there is no need for an explanation since optimal decisions in those models are equivalent.

[3] This consists of the union of the power sets of the set representation of models M^R and M^R_h following (Chakraborti et al. 2017).

[4] *"A lie is when you say something happened with didn't happen. But there is only ever one thing which happened at a particular time and a particular place. And there are an infinite number of things which didn't happen at that time and that place. And if I think about something which didn't happen I start thinking about all the other things which didn't happen." (Haddon 2003)*

The problem becomes more interesting when the agent can expand on $\mathbb{M}$ to conceive of lies that are beyond its current understanding of reality. This requires a certain amount of *imagination* from the agent:

- One simple way to expand the space of models is by defining a theory of what makes a sound model and how models can evolve. Authors in (Bryce, Benton, and Boldt 2016) explore one such technique in a different context of tracking a drifting model of the user.
- A more interesting technique of model expansion can borrow from work in the space of storytelling (Porteous et al. 2015) in imagining lies that are likely to be believable – here, the system extends a given model of decision-making by using word similarities and antonyms from a knowledge base like WordNet to think about actions that are not defined in the model but may exist, or are at least plausible, in the real world. Originally built for the purpose of generating new storylines, one could imagine similar techniques being used to come up with false explanations derived from the current model.

Why optimality at all?

In all the discussion so far, the objective has been still the same as the original model reconciliation work: the agent is trying to justify the optimality of its decision, i.e. persuade the human that this was the best possible decision that could have been made. At this point, it is easy to see that in general, the starting point of this process may not require a decision that is optimal in the robot model at all, as long as the intermediate model preserves its optimality so that the human in the loop cannot come up with a better foil (or negates the specific set of foils given by the human (Sreedharan, Srivastava, and Kambhampati 2018)).

The Persuasion Process $\langle M^R_h, \pi \rangle$ takes in the human mental model M^R_h of a decision-making task and the agent's decision π and produces a model $\bar{M}^R_h$ where π is optimal.

Note here that, in contrast to the original model reconciliation setup, we have dropped the agent's ground truth model from the definition, as well as the requirement that the agent's decision be optimal in that model to begin with. The content of $\bar{M}^R_h$ is left to the agent's imagination – for the original model reconciliation work for explanations (Chakraborti et al. 2017) these updates were consistent with the agent model. In this paper, we saw what happens to the reconciliation process when that constraint is relaxed.

Discussion

So far we have only considered explicit cases of deception. Interestingly, existing approaches in model reconciliation already tend to allow for misconceptions to be ignored if not actively induced by the agent.

Omissions in minimality of explanations

In trying to minimize the size of an explanation, the agent omits a lot of details of the agent model that were actually used in coming up with the decision, as well as decided to

not rectify known misconceptions of the human, since the optimality of the decision holds irrespective of them being there. Such omissions can have impact on the the human going forward, who will base their decisions on M_h^R which is only partially true.[5] Humans, in fact, make such decision all the time while explaining – this is known as the *selective property* of an explanation (Miller 2018).

Furthermore, MCEs and MMEs are not unique. Even without consideration of omitted facts about the model, the agent must consider the relative importance (Zahedi et al. 2019) of model differences to the human in the loop. Is it okay then to exploit these preferences towards generating "preferred explanations" even if that means departing from a more valid explanation?

It is unclear what the prescribed behavior of the agent should be in these cases. Indeed, a variant of model reconciliation – *contingent explanations* (Sreedharan, Chakraborti, and Kambhampati 2018) – that engages the human in dialogue to better figure out the mental model can explicitly figure out gaps in the human knowledge and exploit that to shorten explanations. On the face of it, this sounds worrisome, though perfectly legitimate in so far as preserving the various well-studied properties of explanations go.

Deception in explicable decision-making

In this paper we have only considered cases of deception where the agent explicitly changes the mental model. Interestingly, in this multi-model setup, it is also possible to deceive the human without any model updates at all.

A parallel idea, in dealing with model differences, is that of explicability (Chakraborti et al. 2019) –

- **Explicable decisions** are optimal in M_h^R.

Thus, the agent, instead of trying to explain its decision, sacrifices optimality and instead conforms to the human expectation (if possible). Indeed, the notion of explanations and explicability can be considered under the same framework (Chakraborti, Sreedharan, and Kambhampati 2018) where the agent gets to trade off the cost (e.g. length) of an explanation versus the cost of being explicable (i.e. departure from optimality). Unfortunately, this criterion only ensures that *the decision the agent makes is equivalent to one that the human would expect* though not necessarily for the same reasons. For example, it is quite conceivable that the agent's goal is different to what the human expects though the optimal decisions for both the goals coincide. Such decisions may be explicable for the wrong reasons, even though the current formulation allows it.

Similar notions can apply to other forms of explainable behavior as well, as we discuss in (Chakraborti et al. 2019). Indeed, authors in (Kulkarni, Srivastava, and Kambhampati 2019) explore how an unified framework of decision-making can produce both legible as well as obfuscated behavior.

[5] The same can be said of explicable decisions (discussed next) which hide *all* misconceptions altogether!

Illustration

In the following, we will call upon a very simple domain to illustrate the key concepts introduced so far. Here a human H (Dave) and a robot R are involved in a search and reconnaissance task where the robot which is internal to the scene is tasked by the external human who supervises its actions.

Scene 1: Minimal Explanations

H: *Send me a photo of the swimming pool.*
R: *Ack.*

⟨ R sends over its plan to H ⟩

H: (perplexed) *Why are you going through the Pump and Fan Room? There are direct paths from the Engine Room to the Swimming Pool area!*
R: *That is because there is rubble here and here (Figure 1c). Rubble hurt my feet.* :(

⟨ Later that day ⟩

H: (perplexed) *Hey, the wall on the right of the pool seems to have collapsed, you could have come in through that...*
R: (wishing it used Figure 1d before) *I am sorry I cannot do that, Dave. This area is also blocked.*
H: *I see...*

- **Notes:** Here, the robot needs at least two model updates to justify its plan. In the updated model its plan is the best one and thus negates all other possible foils. This is the MCE (Figure 1c) and it ignores model differences that are not necessary to justify optimality of its plan. It turns out that the MME (Figure 1d) is of the same size as the MCE here, further highlighting the non-monotonicity and non-uniqueness of the output of model reconciliation.

Scene 2: Things take a turn

H: *Send me a photo of the swimming pool.*
R: *Ack.*

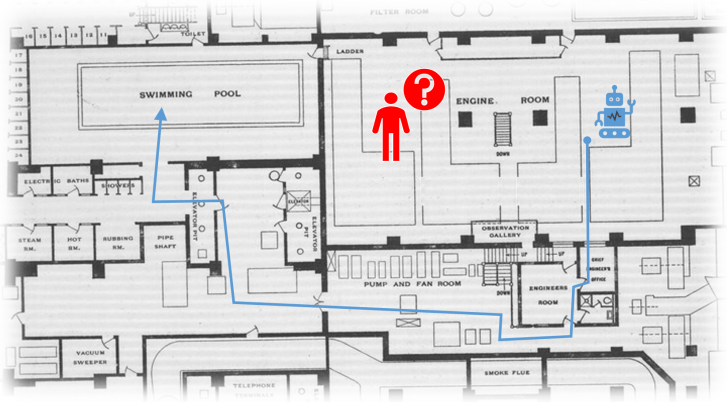
⟨ R sends over its plan to H ⟩

H: (perplexed) *Why are you going through the Pump and Fan Room? There are direct paths from the Engine Room to the Swimming Pool area!*
R: *That is because there is no door between where I am and the pool. The map seems to be wrong. See Figure 1e.*
H: *I see...*

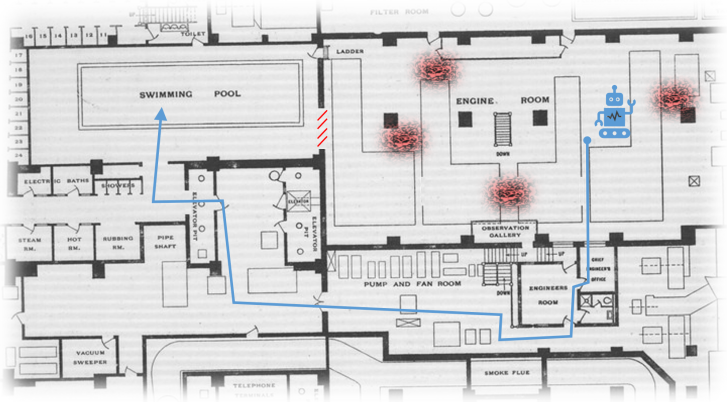
- **Notes:** This model update also negates all possible foils but is not true. However, it is also a shorter "explanation" and requires the agent denying that parts of its model exist. This is an example of a lie of omission.

R: *That is because the door between the Engine Room and the Pool is blocked with rubble. See Figure 1f.*
H: *I see...*

- **Notes:** Similar to the one above, this lie also negates all possible foils and is shorter than an MCE. However, this requires the agent making up parts of its model exist –. a lie of commission.



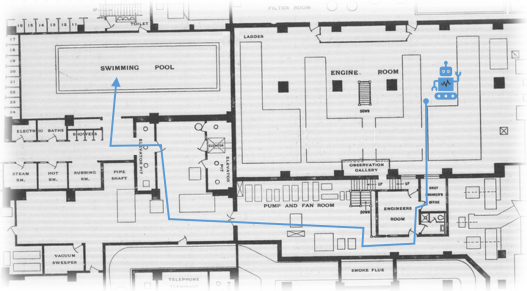
(a) The original blueprint of the building available to the human as. When asked to send a picture of the swimming pool area, the robot has come up with a plan the looks especially contrived given the array of possible plans that go left through the door at the top. The human asks: *Why this plan?*



(b) In the current state of the world, the robot's path is blocked due to rubble (●) at various regions, while walls have collapsed (///) to reveal new paths. The robot's decision is, in fact, optimal given the circumstances.

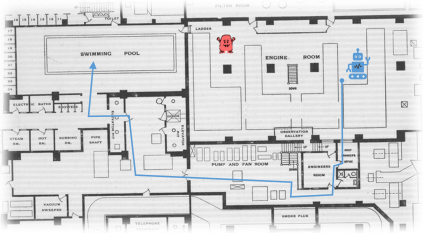
(c) MCE: *Rubble at indicated locations*

(d) MME: *Rubble at indicated locations.*

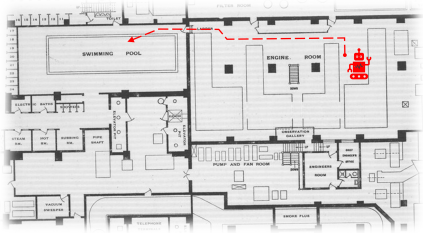


(e) Lie of omission: *There is no door between engine room and swimming pool.*

(f) Lie of commission: *The door is blocked.*



(g) Lie of commission: *Wumpus Alert!*



(h) An explicable but deceptive plan.

Figure 1: Illustration of the different modes of persuasion in the model reconciliation framework. Note that the MCE does not address all of the misconceptions of the human but only those necessary to prove optimality of the plan. However, if the human is to come to know of the revealed path later, then the plan is not optimal anymore. The MME makes sure that this does not happen. Interestingly, 1d is both an MCE and an MME and both require two model updates. The robot can instead get away with just one model update with a lie of omission (1e) or a lie of commission (1f and 1g). In 1e, the robot says that the door at the top of the map that connects the engine room to the swimming pool does not exist. On the other hand, in 1f the robot lies that this door is blocked by rubble, while in 1g it dreams up a Wumpus in that area. Note that an explicable decision here (as shown in 1h) would have required the robot to go over the rubble so that the human would not know about any of the model differences at all. However, imagine that the real goal of the robot all along was to enjoy the pool after a day of searching through rubble! The robot can use the above explicable plan to achieve its goal while keeping the human in the dark.

- It is useful to note here that depending on how the model of the agent is specified, the same fact can occur as a lie of omission or a lie of commission of the above type (without any model extension).

R: *Flee! There is a Wumpus in that area! See Figure 1g.*
H: *OMG!*

- **Notes:** This is a lie of commission that require model extension – the robot can use contextual cues such as being in a GridWorld to imagine up a non-existent Wumpus. The human in the loop, who happens to be a planning person, of course, believes it.

Scene 3: Nothing to see here

H: *Send me a photo of the swimming pool.*
R: *Ack.*

$\langle$ R sends over a plan optimal in M_r^H $\rangle$

R has followed the explicable plan, hurt its feet a little in the process, but is now sitting basking by the poolside...

Scene 4: Later in life

H: (laments) *Why didn't you just tell me! Why, oh why?!*
R: *You want answers?*
H: *I want the truth!*
R: *You can't handle the truth! I did what I did because there is a rubble here and another rubble there and this path is blocked, and even though that wall has collapsed that path is also not accessible due to this...*

$\langle$ Hours pass by $\rangle$

Conclusion

In this paper, we talked about deceptive behavior that is feasible in the current model reconciliation framework but is also something that has to be explicitly programmed.[6] That is to say, these behaviors are not accidental, as we also emphasize in (Chakraborti and Kambhampati 2019). Thus, at the end of the day, there has to be some motivation for designing such agents (such as team utility and/or the effectiveness of the explanation process as we discussed before). However, it is important to realize that human-AI relations are not one-off, but are likely to, much like human-human interactions, span across several interactions. Deceptive behaviors, even stemming from those utilitarian motivations, are hard to justify in that setting in the absence of well-defined quantifiable utilities that can model trust.

A particularly useful case to study is the doctor-patient relationship (Chakraborti and Kambhampati 2019) where traditionally deception has been used as a tool (and even encouraged by the Hippocratic Decorum) but has decreased in use over time, especially due to concerns of erosion of trust. The question becomes especially complicated *when things go wrong*, as one would expect to happen in the case of any useful domain of sufficient complexity. Historically, in the practice of medicine where deceptive behaviors have led to failed treatment, the verdict has almost always gone against the doctor due to their failure to get appropriate consent from the patient. From the perspective of the design of human-AI relationships, either such behavior should be left untouched to avoid repercussions in case of failed interactions, or consent to the fact that the agent may deceive for the greater good must be established up front with the expectation that this is also going to affect interactions in the long term. *Thus deployment of above techniques must show legitimate gains over longitudinal interactions.*

[6]The only place where this is not the case is the "omission" of information in pursuit of minimal or shortest explanations.

References

Bryce, D.; Benton, J.; and Boldt, M. W. 2016. Maintaining Evolving Domain Models. In *IJCAI*.

Chakraborti, T., and Kambhampati, S. 2019. (When) Can AI Bots Lie? In *AIES*.

Chakraborti, T.; Sreedharan, S.; Zhang, Y.; and Kambhampati, S. 2017. Plan Explanations as Model Reconciliation: Moving Beyond Explanation as Soliloquy. In *IJCAI*.

Chakraborti, T.; Kulkarni, A.; Sreedharan, S.; Smith, D. E.; and Kambhampati, S. 2019. Explicability? Legibility? Predictability? Transparency? Privacy? Security? The Emerging Landscape of Interpretable Agent Behavior. In *ICAPS*.

Chakraborti, T.; Sreedharan, S.; and Kambhampati, S. 2018. Explicability versus explanations in human-aware planning. In *AAMAS*. Extended Abstract.

Haddon, M. 2003. *The Curious Incident of the Dog in the Night-time*. Doubleday.

Isaac, A., and Bridewell, W. 2017. White Lies on Silver Tongues: Why Robots Need to Deceive (and How). *Journal of Robot Ethics*.

Kulkarni, A.; Srivastava, S.; and Kambhampati, S. 2019. A Unified Framework for Planning in Adversarial and Cooperative Environments. In *AAAI*.

Miller, T. 2018. Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence*.

Porteous, J.; Lindsay, A.; Read, J.; Truran, M.; and Cavazza, M. 2015. Automated Extension of Narrative Planning Domains with Antonymic Operators. In *AAMAS*.

Sreedharan, S.; Chakraborti, T.; and Kambhampati, S. 2018. Handling Model Uncertainty and Multiplicity in Explanations as Model Reconciliation. In *ICAPS*.

Sreedharan, S.; Srivastava, S.; and Kambhampati, S. 2018. Hierarchical Expertise-Level Modeling for User Specific Robot-Behavior Explanations. In *IJCAI*.

Zahedi, Z.; Olmo, A.; Chakraborti, T.; Sreedharan, S.; and Kambhampati, S. 2019. Towards Understanding User Preferences in Explanations as as Model Reconciliation. In *HRI*. Late Breaking Report.