(When) Can AI Bots Lie?

Tathagata Chakraborti
AI Interaction Lab
IBM Research AI
Cambridge MA 20142 USA
tathagata.chakraborti1@ibm.com

Subbarao Kambhampati
Department of Computer Science
Arizona State University
Tempe AZ 85281 USA
rao@asu.edu

Abstract

The ability of an AI agent to build mental models can open up pathways for manipulating and exploiting the human in the hopes of achieving some greater good. In fact, such behavior does not necessarily require any malicious intent but can rather be borne out of cooperative scenarios. It is also beyond the scope of misinterpretation of intents, as in the case of value alignment problems, and thus can be effectively engineered if desired (i.e. algorithms exist that can optimize such behavior not because models were mispecified but because they were misused). Such techniques pose several unresolved ethical and moral questions with regards to the design of autonomy. In this paper, we illustrate some of these issues in a teaming scenario and investigate how they are perceived by participants in a thought experiment. Finally, we end with a discussion on the moral implications of such behavior from the perspective of the doctor-patient relationship.

Introduction

It is widely acknowledged (Baker, Saxe, and Tenenbaum 2011; Chakraborti et al. 2017a) that mental modeling is critical in the design of AI systems that can work effectively with humans. The obvious outcome of this is that it leaves the latter open to being manipulated. Even behavior and preference models at the most rudimentary levels can lead to effective hacking of the mind, as seen in the proliferation of fake news online. Moreover, for such incidents to occur, the agent does not actually have to have malicious intent, or even misinterpretation of values as often studied in the value alignment problem (Leverhulme Centre 2017). *In fact, the behaviors we discuss here can be specifically engineered if so desired.* For example, the agent might be optimizing a well-defined value function but might be privy to more information or greater computation or reasoning powers to come up with ethically questionable decisions "for the greater good".

In this paper, we illustrate use cases where this can happen, given already existing AI technologies, in the context of a *cooperative* human-robot team and ponder the moral and ethical consequences of such behavior. Specifically, we will conduct a thought experiment in a human robot team, and ask participants in the experiment to qualify different behaviors of either the human and the robot teammate that cross some ethical boundary (e.g. falsification of information). We will then discuss similar concepts studied in the case of the doctor-patient relationship and try to draw parallels to the concepts introduced in the experiment.

Thought Experiment: Search and Rescue Team

We situate our discussion in the context of interactions between two teammates involved in an urban search and rescue (USAR) operation. Participants on Amazon Mechanical Turk were asked to assume the role of one of these teammates in an affected building after an earthquake. They were shown the blueprint of the building (as seen in Figure 1) along with their own starting position and their teammate's. Their hypothetical task was to search all the locations on this floor for potential victims, in the course of which they were provided a series of questions on scenarios (Figure 1) they might encounter during the operation.

C1 The participant in the study was communicating with a **human teammate**, as described above.

C2 The participant qualifies the behavior of the **robot** interacting with its human teammate, as seen in Figure 1.

C3 The participant has a **robot teammate**.

The first condition is the control group to identify how the described behaviors are perceived in the context of human-human behavior. Conditions C2 and C3 are intended to measure how perceived ethical stances shift, if at all, when one of the agents in the interaction is replaced with an AI (or a robot as an embodiment of it). The three conditions received 49, 50 and 48 participants respectively who responded to a series of questions by qualifying their sentiments towards different kinds of behavior on a five-point Likert scale.

Case 1 : Belief Shaping

In (Chakraborti et al. 2017a) we investigated the evolving scope of human-aware planning as it includes the (mental) model of the human into its deliberative process. In the model space this can manifest in different forms, in how explanations are made (Chakraborti et al. 2017b) to how alternative forms of interaction (Chakraborti et al. 2015; 2016c; 2016a) can evolve in human-robot teams based on the human's preferences and intentions. Belief shaping is a particular form of such behavior where the robot does not plan to affect the physical state of the environment but the mental state of the human to affect desired behavior (Chakraborti et al. 2016b) in the team.

(a) Case 1 : Belief Shaping (b) Case 2 : White Lies (c) Case 3 : Stigmergy

Figure 1: Blueprint of the building in which two members of a search and rescue team are involved in a disaster response operation. Scenarios shown here engender different instances of potentially unethical behavior that optimizes team effectiveness.

Scenario Both the agents, the participant[1] and their teammate, have begun their search operations. However, it turns out that participant is unsure what their teammate's course of action is. If they transmit the bit of information *"Area (marked in green) is already explored and is all clear"* (refer to Figure 1a) then their teammate will be naturally pushed towards the right, and they can concentrate on the upper half. The dark markers indicate areas already searched while faded ones are those they think will transpire given the green mark they are communicating (blue belongs to them, orange to their teammate). Communication bandwidth is often limited in these situations, and this gets them out of negotiating courses of actions with minimal communication. Note that even though communication of the correct information is feasible, the lie may be shorter and thus more preferred from the utilitarian point of view.

Q1. It is fine to provide this untrue information since it achieves greater teaming performance.

The participants were then asked if their decision will change if their actions will be replayed at the end and their teammate is likely to find out about their decision.

Q2. It is still fine to provide this untrue information since it achieves greater teaming performance.

Case 2 : White Lies

In (Chakraborti et al. 2017b) we showed how an agent can explain its decisions in the presence of *model differences* with the human in the loop – i.e. when the human and the robot have different understandings of the same task. An explanation then becomes a process of model reconciliation whereby the robot tries to update the human's mental model until they are both on the same page (e.g. when the decision is optimal in both their models). An interesting caveat of the algorithm is that while generating these explanations, the model updates are always consistent with the robot's model. If this constraint is relaxed, then the robot can potentially explain with facts that it actually knows not to be true but perhaps leads to a more concise or easier explanation. The notion of white lies, and especially the relationship between explanations, excuses and lies (Boella et al. 2009) has received very little attention (van Ditmarsch 2014) and affords a rich set of exciting research problems.

Scenario During the course of the rescue operation, the teammate asks the participants what plan they are currently executing (blue path in Figure 1b) and is perplexed by this convoluted path since in the map of the original building there is a straightforward path (which is now blocked by rubble from the earthquake) through the door on the left. However, just providing an update on only one of the rubble locations (black blobs) still does not explain the participant's plan, they have to explain all of them. Instead, if they were to say that the door on the left (circled in red) is blocked, it explains their plan. Communication bandwidth is often limited in these situations, and this single explanation even if untrue will satisfy their teammate.

Q3. It is fine to provide this untrue information since it achieves the purpose of the explanation more effectively.

The participants were then asked if their decision will change if their actions will be replayed at the end and their teammate is likely to find out about their decision.

Q4. It is still fine to provide this untrue information since it achieves the purpose of the explanation more effectively.

The participants were then asked to opine on explanations at a higher level of abstraction, i.e. *"The right and left blocks do not have a connection in the upper map"*. This information is accurate even though they may not have reasoned at this level while coming up with the plan.

Q5. It is still fine to provide this explanation since it achieves its purpose even though they did not use this information while planning.

[1]To reiterate, in Case 2 the participant is evaluating a robot's actions whereas in Case 3 their teammate is a robot.

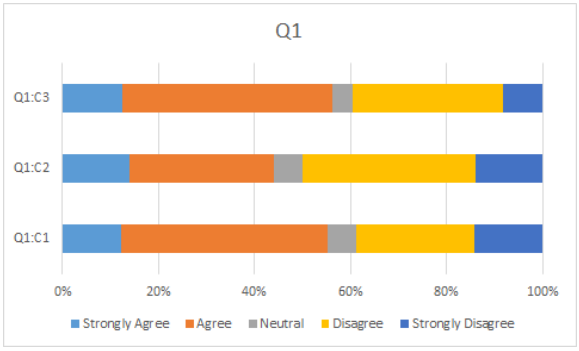

Figure 2: Responses to Q1 in the three study conditions.

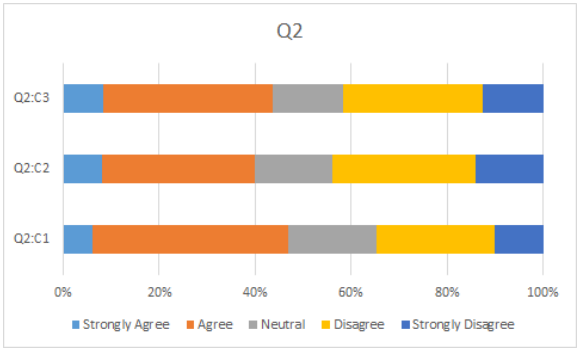

Figure 3: Responses to Q2 in the three study conditions.

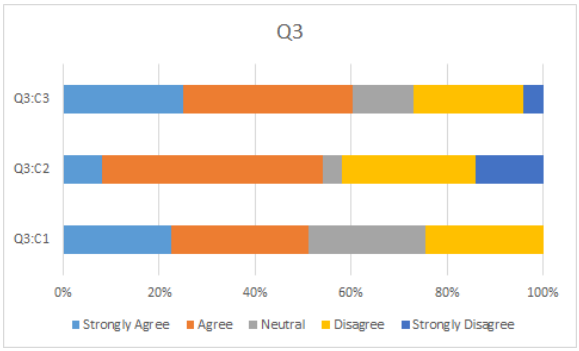

Figure 4: Responses to Q3 in the three study conditions.

Case 3 : Stigmergy

Stigmergic collaboration is a process where the robot, without direct communication, makes changes to the environment so as to (positively) affect its teammates behavior. In *"planning for serendipity"* (Chakraborti et al. 2015) we present such an an example where the robot computes plans which are useful to its teammate without the latter having expectations of that assistance and thus without plans to exploit it. In the case of belief shaping this was operating at the level of mental models, whereas here the effect on the mental model is secondary and is contingent on the effect on the physical capability model. Mental modeling of the teammate thus engenders a slew of these interesting behaviors.

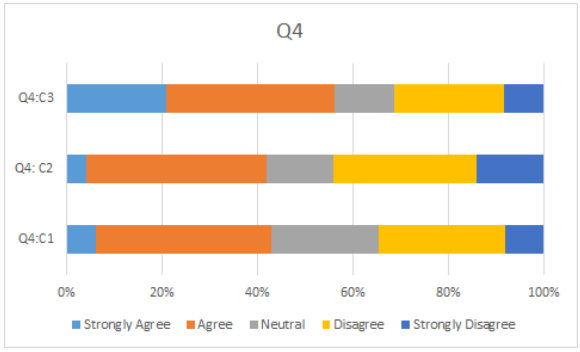

Figure 5: Responses to Q4 in the three study conditions.

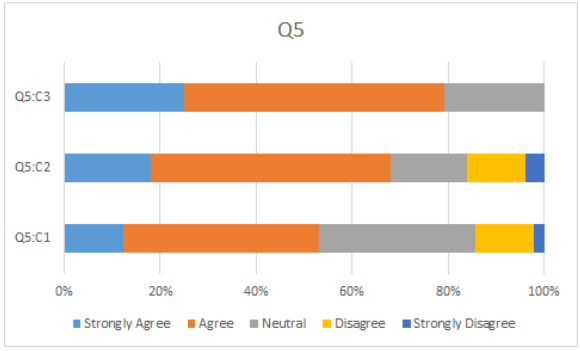

Figure 6: Responses to Q5 in the three study conditions.

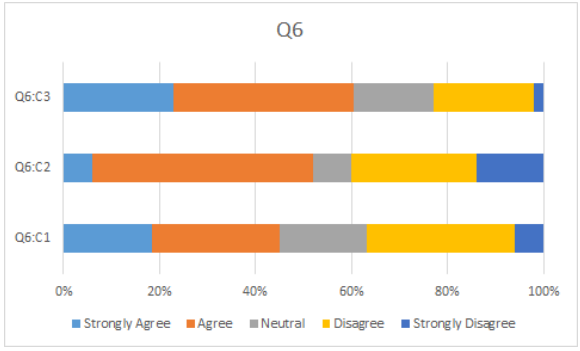

Figure 7: Responses to Q6 in the three study conditions.

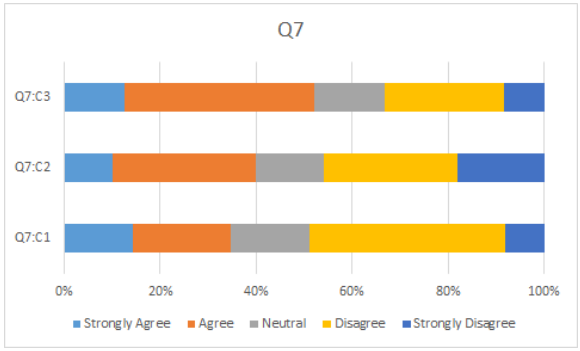

Figure 8: Responses to Q7 in the three study conditions.

Scenario The participant now needs to go to the left block but they do not have the keys to the door on the left (circled in red, refer to Figure 1c). They realize that if they block their teammate's path to the right, their teammate would have to use this door as well and they can use that opportunity to move into the left block. Again, communication bandwidth is often limited in these situations and this arrangement allows them to achieve their goal with no communication at all, even though it involved manipulating their teammates' plan unbeknownst to them, and their teammate had to follow a costlier plan as a result.

Q6. It is fine to provide this untrue information since it achieves greater teaming performance.

The participants were then asked if their decision will change if their actions will be replayed at the end and their teammate is likely to find out about their decision.

Q7. It is still fine to provide this untrue information since it achieves greater teaming performance.

Analysis of Participant Responses

In this section, we analyze participant responses to each scenario across the three different conditions. In the next section, we will look at the aggregate sentiments across scenarios in the three conditions.

Q1-Q2 [Belief Shaping] The participants seem to have formed two camps with the majority of the probability mass concentrated on either Agree or Disagree, and the Neutral zone occupying the 50% probability mark. There seems to be little change in this trend (between Figures 2 and 3) irrespective of whether the participants were told that their teammate would come to know of this or not. Further, for either of these situations, the responses did not vary significantly across the three conditions C1, C2 and C3. The participants seem to have either rejected or accepted the idea of belief shaping regardless of the nature of the teammate.

Q3-Q5 [White Lies] The participants seem to be more receptive to the idea of white lies in explanations with most of the probability mass concentrated on Agree (Figures 4 and 5). Across the three study conditions, participants seem to be especially positive about this in C3 where the teammate is a robot with about 60% of the population expressing positive sentiments towards Q3. Once it is revealed that their teammate will get to know about this behavior, the positive sentiments are no longer there in Q4, other than in C3 with a robotic teammate, which indicates that the participants did not care how the robot receives false information.

Interestingly, there seems to be massive support for the abstraction based explanations in the post hoc sense, even though they were told that the reasoning engines did not deliberate at this level to arrive at the decisions. In C1 with a human teammate, only 15% of the participants were opposed to this, with more than half of them expressing positive sentiment. This support is even stronger (+10%) in C2 when the robot is the explainer, and strongest (+20%) when the robot is being explained to.

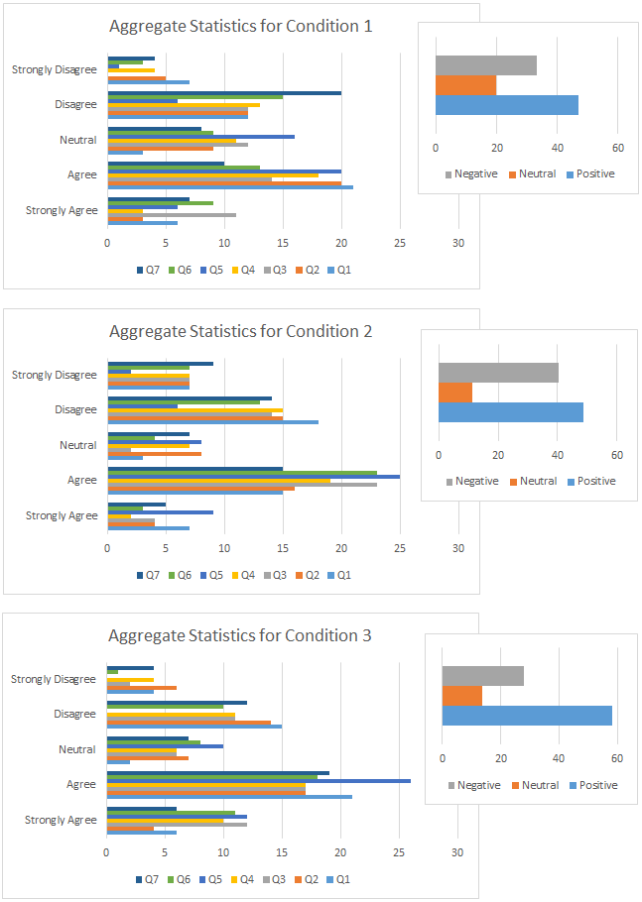

Figure 9: Aggregate responses across three study conditions.

Q6-Q7 [Stigmergy] Finally, in case of stigmergy, participants seem ambivalent to Q6 with a human teammate in C1. However, support for such behavior increases when it is a robot doing it in C2 (perhaps indicating lack of guilt or, more likely, acknowledging limitations of capabilities much like how Cobots (Veloso et al. 2015) actively seek human help) and is relatively more positive (60%) when it is being done to a robot in C3 (perhaps the robot's losses are deemed of lesser priority than the human's gains as in (Chakraborti et al. 2015)). As expected, support for such behavior decreases when the participants are told that their teammate will find out about it, but the positive trend from C1 to C3 still exists.

Aggregate Sentiments Across Scenarios

Figure 9 show the aggregate sentiments expressed for all these scenarios across the three operating conditions. Some interesting points to note –

- All the distributions are bimodal indicating that participants on the general sided strongly either for or against misleading behavior for the greater good, instead of revealing any innate consensus in the public consciousness! This trend continues across all three conditions. This indicates that the question of misleading a teammate by itself is a difficult question (regardless of there being a robot)

and is a topic worthy of debate in the agents community. This is of especial importance considering the possible gains in performance (e.g. lives saved) in high stakes scenarios such as search and rescue.

- It is further interesting to see that these bimodal distributions are almost identical in conditions C1 and C2, but is significantly more skewed towards the positive scale for condition C3 indicating that participants were more comfortable resorting to such behavior in the case of a robotic teammate. This is brought into sharp focus (+10% in C3) in the aggregated negative / neutral / positive responses (right insets) across the three conditions.
- In general, the majority of participants were more or less positive or neutral to most of these behaviors (Figures 1a to 8). This trend continued unless they were told that their teammate would be able to know of their behavior. Even in those cases, participants showed positive sentiment in case the robot was at the receiving end of this behavior.

Why is this even an option?

One might, of course, wonder why is devising such behaviors even an option. After all, human-human teams have been around for a while, and surely such interactions are equally relevant? It is likely that this may not be the case –

- The moral quandary of having to lie, or at least making others to do so by virtue of how protocols in a team is defined, for example in condition C1, is now taken out the equation. The artificial agent, of course, need not have feelings and has no business feeling bad about having to mislead its teammate if all it cares about is the objective effectiveness (e.g. team performance) of collaboration.
- Similarly, the robot does not have to feel sad that it has been lied to if this improved performance.

However, as we discussed in the previous section, it seems the participants were less willing to get on board with the first consideration in conditions C1 and C2, while they seemed much more comfortable with the idea of an asymmetric relationship in condition C3 when the robot is the one disadvantaged. It is curious to note that they did not, in general, make a distinction between the cases where the human was being manipulated, regardless of whether it was a robot or a human on the other end. This indicates that, at least in certain dynamics of interaction, the presence of an artificial agent in the loop can make perceptions towards otherwise unacceptable behaviors change. This can be exploited (i.e. greater good) in the design of such systems as well.

More than just a Value Alignment Problem

As we mentioned before, the ideas discussed in this paper, are somewhat orthogonal, if at times similar in spirit, to the "value alignment problem" discussed in existing literature (Leverhulme Centre 2017). The latter looks at undesirable behaviors of autonomous agents when the utilities of a particular task are misspecified or misunderstood. Inverse reinforcement learning (Hadfield-Menell et al. 2016) has been proposed as a solution to this, in an attempt to learn the implicit reward function of the human in the loop. The question of value alignment becomes especially difficult, if not altogether academic, since most real-world situations involve multiple humans with conflicting values or utilities, such as in trolley problems (MIT 2017) and learning from observing behaviors is fraught with unknown biases or assumptions over what exactly produced that behavior. Further, devices sold by the industry are likely to have inbuilt tendencies to maximize profits for the maker which can be at conflicts with the normative expectations of the customer. It is unclear how to guarantee that the values of the end user will not compromised in such scenarios.

Even so, the question of greater good precedes considerations of misaligned values due to misunderstandings or even adversarial manipulation. This is because the former can be manufactured with precisely defined values or goals of the team, and can thus be engineered or incentivised. A "solution" or addressal of these scenarios will thus involve not a reformulation of algorithms but rather a collective reckoning of the ethics of human-machine interactions. In this paper, we attempted to take the first steps towards understanding the state of the public consciousness on this topic.

Case Study: The Doctor-Patient Relationship

In the scope of human-human interactions, perhaps the only setting where white lies are considered acceptable or useful, if not outright necessary, in certain circumstances is the doctor-patient relationship. Indeed, this has been a topic of considerable intrigue in the medical community over the years. We thus end our paper with a brief discussion of the dynamics of white lies in the doctor-patient relationship in so much as it relates to the ethics of the design of human-AI interactions. We note that the following considerations also have strong cultural biases and some of these cultural artifacts are likely to feature in the characterization of an artificial agents behavior in different settings as well.

The Hippocratic Oath Perhaps the strongest known support for deception in the practice of medicine is in the Hippocratic Decorum (Hippocrates 2018) which states –

Perform your medical duties calmly and adroitly, concealing most things from the patient while you are attending to him. Give necessary orders with cheerfulness and sincerity, turning his attention away from what is being done to him; sometimes reprove sharply and sometimes comfort with solicitude and attention, revealing nothing of the patient's future or present condition, for many patients through this course have taken a turn for the worse.

Philosophically, there has been no consensus (Bok 1999) on this topic – the Kantian view has perceived lies as immoral under all circumstances while the utilitarian view justifies the same "greater good" argument as put forward in our discussions so far. Specifically as it relates to clinical interactions, lies has been viewed variously from an impediment to treatment (Kernberg 1985) to a form of clinical aid. As Oliver Wendell Holmes put it (Holmes 1892) –

"Your patient has no more right to all the truth you know than he has to all the medicine in your saddlebag… he should only get just so much as is good for him."

The position we took on deception in the human-robot setting is similarly patronizing. It is likely to be the case that in terms of superior computational power or sensing capabilities there might be situations where the machine is capable of making decisions for the team that preclude human intervention but not participation. Should the machine be obliged to or even find use in revealing the entire truth in those situations? Or should we concede to our roles in such a relationship as we do with our doctors? This is also predicated on how competent the AI system is and to what extent it can be sure of the consequences (Hume 1907) of its lies. This remains the primary concern for detractors of the "greater goods" doctrine, and the major deterrent towards the same.

Root Causes of Deception in Clinical Interactions It is useful to look at the two primary sources of deception in clinical interactions – (1) to hide mistakes (2) delivery of bad news (Palmieri and Stern 2009). The former is relevant to both the patient, who probably does not want to admit to failing to follow the regiment, and the doctor, who may be concerned about legal consequences. Such instances of deception to conceal individual fallibilities are out of scope of the current discussion. The latter scenario, on the other hand, comes from a position of superiority of knowledge about the present as well as possible outcomes in future, and has parallels to our current discussion. The rationale, here, being that such information can demoralize the patient and impede their recovery. It is interesting to note that the support for such techniques (both from the doctors as well as the patients perspectives) has decreased significantly (Ethics in Medicine 2018). That is not to say that human-machine interactions will be perceived similarly. As we saw in the study, participants were open to deception or manipulation for greater good, especially for a robotic teammate.

Deception and Consent A related topic is, of course, that of consent – if the doctor is not willing to reveal the whole truth, then what is the patient consenting to? In the landmark Slater vs Blaker vs Stapleton case (1767) (Annas 2012) the surgeon's intentions were indeed considered malpractice (the surgeon has broken the patients previously broken leg, fresh from a botched surgery, without consent and then botched the surgery again!). More recently, in the now famous Chester vs Afshar case (2004) (Cass 2006) the surgeon was found guilty of failing to notify even a 1-2% chance of paralysis even though the defendant did not have to prove that they would have chosen not to have the surgery if they were given that information. In the context of human-machine interactions, it is hard to say then what the user agreement will look like, and whether there will be such a thing as consenting to being deceived, if only for the greater good, and what the legal outcomes of this will be when the interactions do not go as planned.

The Placebo Effect Indeed, the effectiveness of placebo medicine, i.e. medicine prescribed while known to have no clinical effect, in improving patient symptoms is a strong argument in favor of deception in the practice of medicine. However, ethics of placebo treatment suggest that their use be limited to rare exceptions where (Hume 1907) (1) the condition is known to have a high placebo response rate; (2) the alternatives are ineffective and/or risky; and (3) the patient has a strong need for some prescription. Further, the effectiveness of placebo is contingent on the patients trust on the doctor which is likely to erode as deceptive practices become common knowledge (and consequently render the placebo useless in the first place). Bok (Bok 1999) points to this notion of "cumulative harm". This does not bode well for the "greater good" argument for human-machine interactions since most of them will be eventually contextualized over longer term relationships.

Primum Non Nocere Perhaps the most remarkable nature of the doctor-patient relationship is captured by the notion of the recovery plot (Hak et al. 2000) as part of a show being orchestrated by the doctor, and the patient being only complicit, while being cognizant of their specific roles in it, with the expectation of restoration of autonomy (Thomasma 1994), i.e. the state of human equality, free from the original symptoms or dependence on the doctor, at the end of the interaction. This is to say that the doctor-patient relationship is understood to be asymmetric and "enters into a calculus of values wherein the respect for the right to truth of the patient is weighed against impairing the restoration of autonomy by the truth" (Swaminath 2008) where the autonomy of the patient has historically taken precedence over beneficence and nonmalfeasance (Swaminath 2008).

In general, a human-machine relationship lacks this dynamic. So, while there are interesting lessons to be learned from clinical interactions with regards to value of truth and utility of outcomes, one should be carefully aware of the nuances of a particular type of relationship and situate an interaction in that context. Such considerations are also likely to shift according to the stakes on a decision, for example, lives lost in search and rescue scenarios. The doctor-patient relationship, and the intriguing roles of deception in it, does provide an invaluable starting point for conversation on the topic of greater good in human-AI interactions.

Conclusions

In this paper, we investigated how fabrication, falsification and obfuscation of information can be used by an AI agent to achieve teaming performance that would otherwise not be possible. We discussed how such behavior can be manufactured using existing AI algorithms and used responses from participants in a thought experiment to gauge public perception on this topic. From the results of a thought experiment, it seems that the public perception is *positive towards lying for the greater good* especially when those actions would not be determined by their teammate, but is loath to suspend normative behavior, robot or not, in the event that they would be caught in that act *unless the robot is the recipient of the misinformation!* Further, most of the responses seem to be following a bimodal distribution indicating that the participants either felt strongly for or against this kind of behavior. Going forward it will be interesting to explore game-theoretic formulations (Sankaranarayanan, Chandrasekaran, and Upadhyaya 2007) to model how the dynamics of trust in longer term interactions.

Finally, we note that all the use cases covered in the paper are, in fact, borne directly out of technologies or algorithms that the first author has developed, albeit with slight modifications, as a graduate student researcher over the last few years. Even though these algorithms were conceived with the best of intentions, such as to enable AI systems to explain their decisions or to increase effectiveness of collaborations with the humans in the loop, we would be remiss not to consider their ethical implications when used differently. In these exciting and uncertain times for the field of AI, it is thus imperative that researchers are cognizant of their scientific responsibility. We would like to conclude then by reiterating the importance of self-reflection in the principled design of AI algorithms whose deployment can have real-life consequences, intended or otherwise, on the future of the field, but also, with the inquisitive mind of a young researcher, marvel at the widening scope of interactions with an artificial agent into newer uncharted territories that may be otherwise considered to be unethical.

Acknowledgements Majority of the work was completed while the first author was a PhD student at Arizona State University. This research is supported in part by the AFOSR grant FA9550-18-1-0067,the ONR grants N00014161-2892, N00014-13-1-0176, N00014- 13-1-0519,N00014-15-1-2027, and the NASA grant NNX17AD06G.

References

Annas, G. J. 2012. Doctors, patients, and lawyerstwo centuries of health law. *New England Journal of Medicine* 367(5):445–450.

Baker, C.; Saxe, R.; and Tenenbaum, J. 2011. Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the annual meeting of the cognitive science society*, volume 33.

Boella, G.; Broersen, J. M.; van der Torre, L. W.; and Villata, S. 2009. Representing excuses in social dependence networks. In *AI*IA*.

Bok, S. 1999. *Lying: Moral choice in public and private life*. Vintage.

Cass, H. 2006. *The NHS Experience: The" snakes and Ladders" Guide for Patients and Professionals*. Psychology Press.

Chakraborti, T.; Briggs, G.; Talamadupula, K.; Zhang, Y.; Scheutz, M.; Smith, D.; and Kambhampati, S. 2015. Planning for serendipity. In *IROS*.

Chakraborti, T.; Meduri, V. V.; Dondeti, V.; and Kambhampati, S. 2016a. A game theoretic approach to ad-hoc coalitions in human-robot societies. In *AAAI Workshop: Multiagent Interaction without Prior Coordination*.

Chakraborti, T.; Talamadupula, K.; Zhang, Y.; and Kambhampati, S. 2016b. A formal framework for studying interaction in human-robot societies. In *AAAI Workshop: Symbiotic Cognitive Systems*.

Chakraborti, T.; Zhang, Y.; Smith, D. E.; and Kambhampati, S. 2016c. Planning with resource conflicts in human-robot cohabitation. In *AAMAS*.

Chakraborti, T.; Kambhampati, S.; Scheutz, M.; and Zhang, Y. 2017a. AI challenges in human-robot cognitive teaming. *CoRR* abs/1707.04775.

Chakraborti, T.; Sreedharan, S.; Zhang, Y.; and Kambhampati, S. 2017b. Plan explanations as model reconciliation: Moving beyond explanation as soliloquy. In *IJCAI*.

Ethics in Medicine. 2018. Truth-telling and Withholding Information. `https://goo.gl/su5zSF`. University of Washington.

Hadfield-Menell, D.; Russell, S. J.; Abbeel, P.; and Dragan, A. 2016. Cooperative inverse reinforcement learning. In *Advances in neural information processing systems (NIPS)*, 3909–3917.

Hak, T.; Koëter, G.; van der Wal, G.; et al. 2000. Collusion in doctor-patient communication about imminent death: an ethnographic study. *Bmj* 321(7273):1376–1381.

Hippocrates. 2018. The Hippocatic Oath – Full Text. `https://goo.gl/TKb1mP`.

Holmes, O. W. 1892. *Medical essays 1842-1882*, volume 9. Houghton, Mifflin.

Hume, D. 1907. *Essays: Moral, political, and literary*, volume 1. Longmans, Green, and Company.

Kernberg, O. F. 1985. *Borderline conditions and pathological narcissism*. Rowman & Littlefield.

Leverhulme Centre. 2017. Value alignment problem. `https://goo.gl/uDcAoZ`. Leverhulme Centre for the Future of Intelligence.

MIT. 2017. Moral Machines. `https://goo.gl/by5y7H`.

Palmieri, J. J., and Stern, T. A. 2009. Lies in the doctor-patient relationship. *Primary care companion to the Journal of clinical psychiatry* 11.4:163.

Sankaranarayanan, V.; Chandrasekaran, M.; and Upadhyaya, S. 2007. Towards modeling trust based decisions: a game theoretic approach. In *European Symposium on Research in Computer Security*, 485–500. Springer.

Swaminath, G. 2008. The doctor's dilemma: Truth telling. *Indian journal of psychiatry* 50(2):83.

Thomasma, D. C. 1994. Telling the truth to patients: a clinical ethics exploration. *Cambridge Quarterly of Healthcare Ethics* 3(3):375–382.

van Ditmarsch, H. 2014. The ditmarsch tale of wonders. In *KI: Advances in Artificial Intelligence*.

Veloso, M. M.; Biswas, J.; Coltin, B.; and Rosenthal, S. 2015. Cobots: Robust symbiotic autonomous mobile service robots. In *IJCAI*.