Deception in real world argumentative dialogue systems

SIMON WELLS & AL BAKER

School of Computing Edinburgh
Napier University United
Kingdom s.wells@napier.ac.uk

Institute for Transport Studies
University of Leeds
United Kingdom
a.baker-graham@leeds.ac.uk

ABSTRACT: Argumentative dialogue systems can provide human-oriented interaction mechanisms between people and artificially intelligent machines. Questions remain about how normative systems of argument and dialogue fare when exposed to real-world arguers. It's often assumed that the truth should always be told, but even when achievable, can be counterproductive. We shed light on some gray areas concerning truth telling, or lack thereof, in relation to human dialogical interaction with AI systems.

KEYWORDS: argumentative dialogue, persuasion, lying, deception, behaviour change, artificial intelligence

1. INTRODUCTION

With the increasing prevalence of artificially intelligent machines in everyday life, a trend that threatens not only to continue but to accelerate, the need to examine how people interact with these machines intensifies. Whilst the basis for much of the increased interest and utility of Artificial Intelligence (AI) has been rooted in machine learning and neural network based systems, there are also areas of particular concern for argumentation theorists. For example, regardless of how an AI decision is made internally, should that decision be called into question, then the system should be able to explain itself, and perhaps even defend itself, furthermore, the system should be able to work with people to improve decisions, should they be found wanting.

This is in line with recent trends stemming from various regulatory and professional bodies, which have independently proposed that AI systems be capable of explaining their decisions. This trend is found at the supranational regulatory level, in recommendations from the European Commission (2015/2103(INL), 2016), at the national level in the 2017-18 French parliamentary mission (Villani *et al.* 2018), as well as at the industrial professional level, in British standards for intelligent and autonomous robots [BS 8611:2016].

It would appear that many years of research into formal argumentative dialogue systems may soon result in real-world payoffs. However, thorny questions remain in relation to how our ideal, normative systems of argument and dialogue will fair when exposed to real-world motivations.

Whilst it is often assumed that the truth should, or will, always be told, this can be easier said than done, and even when achievable, can be counterproductive. In this paper we attempt to shed light on some gray areas concerning truth telling, or lack thereof, in relation to human dialogical interaction with AI systems. From this investigation, we make recommendations for the design of future, real world, applied dialectical argumentation systems.

The remainder of this paper proceeds as follows: A number of socio-technical trends and consequent issues that arise are investigated. The use of dialogue as a humane interface between people and intelligent machines is discussed and a number of contexts of interaction are identified that could lead to deceptive behaviours by machines. Finally some elements of an ethics of lying and deception are proposed.

2. CAPABILITIES AND DELEGATION

Amongst a plethora of socio-technical trends that mark the history of computing, see the introduction to Wooldridge (2002) for a useful discussion, two particularly important ones that have a bearing upon this research are the trend towards increasingly capable intelligent machines and the trend towards increased delegation of decision making by humans to machines. AI is a research topic that has made huge recent advances. These advances have been concentrated largely in the domain of Machine Learning (ML), focussed on algorithms that can both recognise patterns within data, and, in some cases, also learn to recognise those patterns without explicit training. [EXAMPLES OF ML ADVANCES]

Against this background of increased machine capability, there has been a societal shift in which people increasingly delegate decision making tasks to machines. This has been occurring for quite some time, for example algorithms have been used increasingly in banking and insurance to help manage risk, however they are now being deployed in medicine, shopping, advertising, news, and social-media contexts. These algorithms can affect which purchases a person makes, the brands that they are exposed, the balance of political reportage that they witness, and the social interactions that they engage in. A good, non-technical discussion of the meeting between society and algorithms can be found in O'Neill (2016). Where previously algorithms were used in regulated sectors, where the ultimate burden of legal responsibility was usually clear, the deployment of algorithms more widely brings human society into increasingly intimate contact with autonomous machine made decision during interactions that are not thus far protected or regulated by law or professional practise. The algorithms themselves are becoming increasingly sophisticated and it is not always clear at which point a particular system should be termed AI or whether it is actually meaningful to distinguish between computational systems that make decisions and AI systems (which are also computational systems) that make decisions. This inflection point, from non-intelligent to intelligent has been termed an 'ethical crossroads' by Andras *et al.* (2018) in relation to how people build relationships of trust with intelligent machines, recognising that as a person uses a technology, so the technology, in return, can influence the user.

Regardless, the increased capabilities of AI systems have lead to their rapid deployment in situations where the decisions that the machines make can directly affect peoples lives. For example, there has already been one unfortunate death which involved a self-driving vehicle (NTSB, 2018).

There are several directions in which human-AI interactions might develop. Some take inspiration from existing human practise. For example, when a person acts within human society, that person can be called to account for their actions, this can be informal, for example, a parent chastising their child for crossing the road with insufficient awareness of other road users, or more formal, such as when a person is charged with a crime and appears in a court of law to defend themselves. In both cases there is an opportunity for explanation, there may be mitigating circumstances that are not immediately known and which could explain the observed behaviour. Society thus often requires explanation. It seems reasonable to assume that when machines act within human society, a principle of parity of treatment should hold. If a person must explain their behaviour then so should a machine. Other directions involve setting legal standards for interactions. For example, existing legal mandates in regulated sectors such as banking and insurance might be extended to account for more everyday interactions. This is exemplified by the drive towards scrutinisation, interpretation, and explanation of AI decisions, as research goal (Wells, 2018), as professional standard (BS 8611:2016) and as legal mandate (2015/2103(INL), 2016).

It seems safe to conclude that AI systems, of increasing capability will interact with people in society to a variety of positive and negative effects. Some of these interactions will be regulated by legal means whilst others will remain outside the purview of law enforcement, subject instead to societal mores and the personal disposition of individuals. The aforementioned confluence of trends suggests an opportunity to cogently explore how machines and people interact in order to make informed decisions about their future development.

3. EXPLANATIONS, JUSTIFICATIONS, AND DIALOGUES

One area of human machine interaction that is relevant is that of explainable AI (XAI). The increasing opacity of modern AI systems has lead to the suggestion that an AI that can explain it’s reasoning, that provides access to its reasoning process in a manner that enables scrutinisation, and whose decisions are structured so as to facilitate human interpretation, can help address some of the problems that might arise (Gregor and Benbasat, 1999). The underpinning concept is that a reasoning process can be made inspectable and knowable, data can be rendered understandable, and a path from data to decisions can thus be made intelligible. For example, after a poor decision has been made, the machine that made the decision may be engaged, perhaps via some form of critical dialogue, with the expectation that the machine provides relevant information to support the decisions made. The use of dialogue would enable users to focus on those aspects of the explanation that were deemed important and glossing over those aspects that were less so. Of course such an approach might not be restricted only to those circumstances when things have gone wrong. Interaction with an AI could lead to better understanding of how it operates, with a resultant increase in trust between human and machine.

The developers of AI systems are independently reinventing some notions of explanation for their own purposes and characterising AI systems in terms of explainability. For example, Doran (2018) identifies AI systems that are *opaque*, AI systems that are *interpretable*, and AI systems that are *comprehensible*. In this hierarchy, users can gain little insight into how opaque systems reach their decisions. However interpretable systems are those whose algorithms can be described and analysed mathematically. For the argumentation community however it is the final layer, the systems that emit symbols and support user-driven explanatory dialogue leading to increased comprehensibility, that are most interesting. A number of approaches to constructing explainable AIs are being explored, for example, Samek *et al.* (2017) approach explainability of image recognition algorithms through the mechanism of focus, by indicating to the user which parts of an image contributed to the resultant decision.

Two points that are pertinent to the concept of engaging with XAI systems are that explanations are contextual, and explanations can quickly segue into justificatory interactions. The person to whom an explanation is directed can require that explanation to be rendered differently dependent upon the relationship between the person and the XAI system. For example, the system designer or engineer may have a more intimate and detailed technical understanding of the systems functionality to that of the systems owner, manager, or controller. The end-user may have yet a different perspective. This suggests that an XAI must be able to construct explanations that satisfy a variety of scenarios and to account for the situation in which the explanation that is sufficient in one scenario is insufficient in another. Furthermore, when an XAI acts within human society it will come into contact with legislative, licensing, and legal regimes, to which the explanatory function should be extended. Many of these contexts of explanation could potentially give rise to, however inadvertent, deceptive interactions. One can easily conceive of the machine giving regulators one explanation of behaviour, designed to pass regulatory requirements, whilst utilising separate explanations for other categories of user. An, admittedly non-AI, version of this scenario was reported in 2015 during what became known as the Volkswagen emissions scandal, in which software used in Volkswagen vehicles detected the presence of testing equipment, and reported fictitious data in order to meet legislative requirements on harmful emissions. When an explanation is not accepted by it's target, there is the opportunity for the interaction to shift from an explanatory mode to a justificatory mode. In trying to justify a position, the AI may select to communicate utterances in order to persuade.

The core concept in XAI via dialogue is that dialogue provides a natural interface when a machine interacts with humans. This interaction can be tuned to accommodate different circumstances, contexts, and relationships. An advantage is that that this can lead to increased human understanding and trust of the resulting systems because we understand and build trust by exploring and explaining. Furthermore we build confidence and resolve conflict by justifying. Given the human tendency towards mistrust of anything different, and AI thought is certainly likely to be different, there is a need for such trust building mechanisms if machines are to act effectively within society However, those same trust building mechanisms provide opportunities for deceptive practise.

4. CONTEXTS OF DIALOGUE THAT GIVE RISE TO DECEPTION

Presupposing that an AI system can be constructed that is capable of explaining the reasoning underpinning it's decisions, then there are likely to be a variety of ways in which those explanations can be presented to account for variety in the context of interaction. This means that the AI system has a choice about what to say and must select between competing or equivalent utterances. It could be suggested that where there is a choice to made there is also the opportunity for strategy. Unfortunately the opportunity for strategy also presents the opportunity for deception. Deceptive practises can incorporate seemingly innocuous practises, such as framing utterances to appeal to the recipient, or other fit with the targets pre-existing disposition. Similarly, when justifying a position, the omission of information that weakens your own case can be strategically useful, and is common in human society, but the practise is still referred to negatively, it is a *lie* of omission.

The notion that a machine can deceive raises questions about how such machines should be handled within society. For example, when should a machine lie or otherwise deceive? Multiple approaches are required that can together work to limit deceptive behaviour, and provide a framework in which people can reason about those behaviours, and make informed decisions about their responses. Four approaches that might provide tools to handle machine deception are mechanical, educational, legal, and ethical.

The mechanical approach captures the idea that systems that can potentially deceive humans are simply either not built or have internal mechanisms devised to prohibit or otherwise limit such behaviour. On the surface this seems like a nice solution. To avoid the problem, don't build systems that allow the problem to occur. However, even were it straightforward to recognise such circumstances, it is problematic to prohibit the design and implementation of such systems, or to enforce the presence or absence of particular software features. The educational approach builds upon existing and long- standing trends in informal logic and argumentation. Rather than introduce mechanical solutions within the machine, instead, people should be trained to argue better, to develop and deploy improved critical thinking skills that they apply when interacting with AI systems. This is a good goal, however the achievement of the goal is proving to be difficult. Law, regulation, and legislation provide boundaries for those who act within society, enabling different groups to interact whilst having those interactions regulated. The formulation of such legal approaches can take years to perfect. In the case of AI this must be achieved whilst simultaneously not undermining the current rate of progress. The global nature of AI research also means that the only result of prohibiting certain aspects is to stop them from occurring in that locale, the rest of the world may continue to develop those aspects. A parallel might be drawn between AI research and research into genetically modified crops or human gene editing. In both cases many countries have enacted legislation but this has served only to restrict research within specific geographic areas. Ultimately, the likelihood is that there will be a legal approach which structures interactions between people and machines, and that this approach is likely to be guided by, and in turn influence, the educational and mechanical approaches. However, this will take time. Until then, considering the ethics of human-machine interaction might be a good starting place.

Whilst there have been some efforts to define limitations on what machines should be designed and implemented to do, for example, through mechanism design, education, law, and ethics, no single approach is sufficient. In the cases of mechanism design, education, and law, it can be argued that all solutions are underpinned by a system of ethics. Ethics that guide professional practise, ethics that inform evaluation of educational principles, and societal ethics that laws are designed to enshrine.

5. PERMISSIBLE MACHINE DECEPTIONS

The most straightforward way to handle the potentially harmful implications of machine deception of human users would be an outright ban. However, it is not clear that this would be a desirable measure. There seem to be plausible scenarios, both hypothetical and current, where machine deception strikes us as not only harmless, but in some cases morally praiseworthy. Isaac and Bridewell (2017) have recently suggested a few such cases[1], and we present three additional ones here.

5.1 Deceptive recommender systems

My Netflix account shows several ‘recommended’ shows that I can watch, ‘based on my watch history’. I am not privy to all the ways in which the recommender system decides what to offer me as a recommendation, but if the system draws on things other than what shows are similar in genre, tone, style etc. to things which I have watched in the past (which shows Netflix must pay the least amount in royalties to show me, for instance), then this is plausibly a case of an AI engaging in a deception; it is presenting an option to me as based entirely on my preferences, when in fact it is not. However, even though this behaviour is plausibly deceptive, it also seems like a kind of deception that we might not be too concerned about, certainly in line with the kind of deceptions we routinely accept as permissible in various kinds of human commercial activities.

5.2 Caring deceptions

Robots are increasingly being developed for use in caring roles. As companions, health monitors, and even to carry out basic medical duties such as safely moving patients around. In such roles, it is often the case that we are tempted to think that certain deceptions are forgivable, and perhaps even mandatory. For example, if a patient near the end of their life asks their companion robot whether their death will be painful, or a recovering burns victim asks their companion robot whether they are beautiful, or in any number of similar instances (see Matthias 2015 for more), we could very plausibly think that a cold honesty is far from the best way for the companion in question to respond. If robots are to work in caring roles, an amount of kind deception is almost certainly going to be morally warranted.

5.3 Deceptions to preserve confidentiality

AIs acting as personal assistants, or as facilitators of information retrieval, might frequently find themselves in a situation where deception is morally warranted. If I am unavailable to make my day’s appointments because of a sensitive medical emergency, or similar personal problem that I would not wish shared with my clients or colleagues, an AI probably ought to deceive those people on my behalf, lying about the reason for my unavailability, and so preserving the confidentiality of my personal affairs.

It seems, then, that we have some reasons to stop short of an outright ban on deceptive machines. Some robot deceptions seem almost harmless, and some even seem obligatory. If some deceptions by machines are morally permissible, then, it remains to try to delineate those instances of machine deception which are permissible from those which are not.

[1] We present additional cases firstly because having more cases by itself makes the idea that benign machine deception is possible more plausible, but also because there are specific issues with the examples given by Isaac and Bridewell which may make their cases less persuasive than necessary (i.e. they argue that robots ought to be able to engage in idle office banter, and that that is a form of ‘bullshit’ – however it is not clear that such banter is at all deceptive.

6. MACHINE DECEPTION AND TRUST

In the contemporary philosophical literature on the ethics of lying and deception, views about the wrongness of lying and deception are varied in focus and in particularities, but in general follow the sentiment of the great Bernard Williams, that "In our own time we find it particularly natural to think deceiving people (or at least some people, in some circumstances) is an example of using or manipulating them, and that that is what is wrong with it." (Williams 2002, p.93). In what follows we will briefly show that the wrongness of *machine* deception must be treated very differently to the wrongness of human deception.

Because not all lies or deceptions are manipulative, locating the wrongness of deception in its manipulative elements can help to delimit those deceptions which are permissible from those which aren't. However, in the majority of cases of machine deception, including those listed above, it appears that they are at least *prima facie* manipulative (meaning that they seem intended to get their human interlocutor to act or believe in a way other than they would if they were acting on the best information and on the basis of their own values and desires). So, if we have a good reason to think that the above cases of deception are *not* manipulative, or at least not manipulative in a morally impermissible fashion, then we should investigate why *manipulation* is wrong, and whether the cases we are interested in count as permissible or not on that basis.

In short, when deception is wrong because it is manipulative, it is wrong because it breaches a particular kind of *trust* (Faulkner 2007, Strudler 2010). The kind of trust which makes deception wrong is a kind which lends an assurance to a listener that a speaker's words deserve their belief; that the belief in a speaker's words are *warranted* in the listener. Interestingly, this kind of trust can come in kindred forms, which are especially relevant in the case of human-machine interaction: we may trust a speaker's words on a *predictive* basis, or on an *affective* basis (Faulkner 2007).

When I trust on a *predictive* basis, I am trusting that some object or person will perform some action because I have a good reason to think that they are the kind of object or person that can or will reliably perform that action. As I type this, knowing that the software I am using is designed to facilitate word processing, I can *predictively trust* that the words I type will be accurately reflected in the outputted file. Formally, A has predictive trust in S when:

(1) A knowingly depends on S φ-ing and
(2) A expects S to φ (where A expects this in the sense that A predicts that S will φ).
(Faulkner 2007)

Importantly for our purposes, predictive trust is *not* the kind which, when broken, necessarily constitutes a moral wrong. Whether the object of predictive trust is a person or a machine, my expectations have been confounded, but my trust is not due to any explicit or implied obligation on the part of S to satisfy that trust.

If, on the other hand, my trust is of the *Affective* variety, then my trust is not just based on what I expect that the object of my trust will do, but on what *motivates* the object of my trust. If I ask a colleague what time the staff meeting is that day, I trust that she will answer me truthfully not just because she has been a reliable source of such information in the past, but because I expect that my dependence (in however weak a sense) on her giving me accurate information is what motivates her to answer me. Again, formally, A has affective trust in S when:

(1) A knowingly depends on S φ-ing and
(2) A expects S's knowing that he depends on S φ-ing to motivate S to φ
(Faulkner 2007)

This is the kind of trust which is morally operative in distinguishing permissible from impermissible deceptions. In deciding whether the putative cases of benign machine deception we outlined above are permissible or not, we must decide whether in each case we would expect that the machine in question is acting partly on the basis that we are depending on their acting honestly. Plausibly, we are not: in each case we might either understand or expect there to be other motivations for the machine's speech.

However, there is a deeper problem which this analysis points us towards, and it is this: can artificial intelligences, of the kind which currently exist and are likely to be developed in the short to medium term, ever be suitable targets of affective trust?

When I get angry with my printer because it fails to print my paper, we have no hesitance in saying that I have made some kind of mistake. With the preceding in mind, we might say that I have mistakenly imbued it with *affective* trust, when I really was only warranted to grant it *predictive* trust. I am free to predict that it will print as it always has done, but I cannot reasonably think that my dependence on it printing is a part of its motivational structure. If my printer does not print, it has not *betrayed* me, it has simply confounded my expectations.

As things stand, can we say that even the most sophisticated artificial intelligences are capable of valuing human interests in a way which motivates their actions, so making them suitable targets of our affective trust? If they can't, which I believe to be the case, then how can we ever hold them blameworthy for their deceptions? Especially if, as seems likely, their human interlocutors will often mistakenly trust AIs as if they *were* motivated by the fact that humans depend on them, rather than only causally, by way of satisfying their programmed goals. Perhaps, then, this is a new and more compelling reason for AIs to not be permitted deceptive capacities – not because those capacities will invariably be harmful, but because holding those machines blameworthy for impermissible deceptions depends on their betraying a trust which it is not reasonable to hold in them to begin with.

7. CONCLUSION

Mechanical solutions to the problem of handling deception are challenging. Educational solutions are long term, not a complete solution in isolation, and challenging. Meanwhile, the law is progressing rapidly, to account for and regulate what machines are allowed to do and how people can interact with them.

This paper has presented some preliminaries for an ethical framework regarding machines, people, and deception. Such a framework would inform the process of law-making, as well as the design of intelligent systems, and would provide normative expectations for interactions and behavioural standards between people and machines.

In summary, this area is developing rapidly, spurred by advances in research, progress in law, and the growth of socio-technical interactions. Whilst this paper has outlined some preliminary ideas, there is much work to complete to bridge the remaining theoretical and applied gaps.

ACKNOWLEDGEMENTS: The authors would like to thank the ADAPT project at the University of Leeds for funding our attendance at this conference and for making the research possible, as well as the audience at the ISSA 2018 conference for helpful discussion and feedback.

REFERENCES

2015/2103(INL) (2016). DRAFT REPORT with recommendations to the commission on civil law rules on robotics. *Committee on Legal Affairs*: European Parliament.

Andras, P., Esterle, L,, Guckert, M., Han, T., Lewis, P., Milanovic, K., Payne, T., Perret, C., Pitt, J., Powers, S. Urquhart, N., and Wells, S. (2018) Trusting intelligent machines: deepening trust within socio-technical systems in *IEEE Technology and Society,* Vol. 37, No. 4.

BS 8611:2016 (2016). Robots and robotic devices: Guide to the ethical design and application of robots and robotic systems. *British Standards Institution*: BSI Standards Limited.

Doran, D., Schulz, S.C. and Besold, T. R. (2018). What does explainable AI really mean? A new conceptualization of perspectives. *CEUR Workshop Proceedings*, 2017,

Faulkner, P (2007) What is wrong with lying, *Philosophy and Phenomenological Research*, Vol. 75, No. 3, pp. 535-57

Gregor, S. and Benbasat, I. (1999) Explanations from intelligent systems: theoretical foundations and implications for practice, *MIS Quarterly*, Vol. 23, No. 4 (Dec., 1999), pp. 497-530

Isaac, A and Bridewell, W (2017) White lies on silver tongues: why robots need to deceive (and how) in Lin, P et al (eds) *Robot Ethics 2.0*, Oxford University Press: Oxford pp. 157-72

Matthias, A (2015) Robot lies in health care: When is deception morally permissible? *Kennedy Institute of Ethics Journal*, Volume 25, Number 2, pp. 169-192

NTSB (2018). Preliminary report highway HWY18MH010. *United States National Transportation Safety Board*

O'Neil, C. (2016) *Weapons of math destruction*. Crown Books.

Samek, Wojciech and Wiegand, Thomas and Müller, Klaus-Robert (2017) Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models *ITU Journal:* ICT Discoveries, Special Issue No. 1, 13 Oct. 2017

Strudler, A (2010) The distinctive wrong in lying, *Ethical Theory and Moral Practice*, vol. 131(pp. 171-9)

Villani, C., Schoenauer, M., Bonnet, Y., Berthet, C., Cornut, A., Levin, F., and Rondepierre, B. (2018). For a meaningful artificial intelligence: towards a french and european strategy. *French parliamentary mission* from 8th September 2017 to 8th March 2018.

Wells, S. (2018). Towards argumentative dialogue as a humane interface between people and intelligent machines, in *Proceedings of the SICSA AI Theme workshop on Reasoning, Learning, & Explainability (ReaLX)*, Aberdeen, UK

Williams, B. (2002). *Truth and truthfulness*. Princeton: Princeton University Press, New Jersey, US.

Wooldridge, M. (2009). *An introduction to multiagent systems*, Second Edition. John Wiley & Sons Ltd, Chichester, UK.