Deception by default

András Kornai
SZTAKI Institute of Computer Science
Kende u 13-17, H-1111 Budapest
kornai@sztaki.hu

Abstract

We investigate what it means for an algorithm to lie by omission. We describe what we take to be the central mechanism for successfully deceiving one's reader/listener, the abuse of defaults. As we shall see, no matter what safeguards are put in place against outright lies, AGIs will have to be perfectly capable of circumventing these. We consider the design of algorithms capable of effective deception and show that the 'negative' skills such an algorithm relies on are not any different from the 'positive' skills needed for effective communication with humans in the first place, lending considerable support to the Orthogonality Thesis.

Keywords: deception, lying, manipulation, default reasoning, AGI

0. Background

Before turning to our main subject, we briefly consider the commonsensical understanding of *lie* 'to deliberately tell someone something that is not true' and *deceive* 'make someone believe something that is not true'. These come with commonsensical valuation that considers lying to be morally far more reprehensible than deceiving, though some moral philosophers, in particular Saul, 2013 have argued that "acts of merely misleading are not, in general, morally better than acts of lying", and others, in particular Rees, 2013, 1, have argued that it is typically better to lie than to deceive.

We accept the commonsensical value judgment because the standard requirement of sworn testimony "the truth, the whole truth, and nothing but the truth" contains a major asymmetry: telling truth, and not telling untruths are relatively easy, while telling the whole truth is impossible. There are so many things true at any given point that selecting the relevant portion is fraught with difficulties. Since lying by omission (Fallis, 2018) is next to impossible to control in this 'whole truth' sense, even the best intentioned testimony can be misleading. Here we will concentrate on how an algorithm can deliberately deceive the recipient without resorting to untruths, leaving the opposite case (lying without deception, the 'bald-faced lies' of Fallis, 2010) for future analysis. We will use deception in a broad sense, including both *mimicry* 'deception as to the status of the speaker' and perhaps

even more important, *attitudinal* deception, aiming at creating some desired attitude, be it positive or negative, towards the embedded proposition.

No matter what we think about the ethical status of deception, within limits it is clearly pro-survival, and quite frequent in nature (Smith, 1987), so there is no *a priori* reason to assume it is excluded from the behavioral repertoire of AGIs. Also, if ‘white lies’ are permissible, so will be ‘white deceptions’. The very first exposition of lying by robots, Asimov, 1941, probes the moral ambiguity of telling a lie in order not to hurt people. Since *primum nil nocere* is obviously a high-level, if not the highest-level goal, the AGI may recruit all kinds of secondary algorithms for its fulfillment. Remarkably, even if an absolute prohibition on lying by commission is enforced, lying by omission remains possible, not just in special cases, but under such a broad set of conditions that it can become an everyday tool of survival for the AGI in question irrespective of its tendency to avoid hurting humans.

1. Introduction

In the past few years, owing primarily to advances in deep learning, natural language processing (NLP) has become a field “with pervasive societal impact” and an anticipated rise in the “importance of developing NLP technologies for social good” (Jin, Chauhan, Tse, Sachan, & Mihalcea, 2021). With the impressive NL skills of transformer models come not just the public perception of these algorithms as intelligent (see Shieber, 2007 on the value of the Turing Test), but also the potential for the software to be capable of much deeper deception than passing itself off as human. In fact mimicry, deception perpetrated on the out-group, such as beetles or snakes adapting coloration so as to look poisonous, are present very early on the evolutionary timescale, while more complex forms of behavioral pretending, aimed at deceiving members of the in-group (the same individuals encountered over and over again) are best seen in mammals, primates in particular.

Our goal here is to provide an analysis of *deception* as this notion pertains to AGI. The subject requires a long-term outlook, and our approach is correspondingly general, but this is not to say that the problem is not particularly urgent. To the contrary, for the past few years major conferences on information retrieval and NLP have been maintaining separate tracks for the study of misinformation such as the TREC Health Misinformation Track and there are several datasets specifically designed to test misinformation detection algorithms (see Oshikawa, Qian, and Wang, 2020 for a recent overview).

In order not to get bogged down in problems tangential to the main issue of deception, such as the problem of consciousness or the problem of intentionality, we will concentrate on the instruments of deception: text, video, and other forms of communication aimed at humans. *Who* created some piece of misinformation, and with *what* intentions, are questions of great significance e.g. in legal proceedings, but in Section 2 we concentrate on the *how* at the expense of the who and the why. In Section 3 we argue that the 'negative' skills such an algorithm relies on are no different from the 'positive' skills needed for effective communication with humans in the first place. As we shall see, our argument provides a replacement for what Dahaner, 2012 calls the "No-Belief Defense" of the Orthogonality Thesis (Bostrom, 2012): "It is possible to construct an intelligent system such that it would have no functional analogues of beliefs or desires."

2. Replicating human deception

As our first example, consider a recent promotional video, where the company presented a truck moving along a highway, but, as we learn, "[The company] never stated its truck was driving under its own propulsion in the video." As the example shows, deception is different from outright lying, in fact the cases we'll be most interested in are deceptive without making recourse to untruths. It is also different from weaponizing ignorance (see e.g. President Trump's campaign for hydrocloroquine) in that we assume that the human target is in possession of a good (though of course not complete) knowledge base (KB) and deductive facilities.

Such a KB will contain both specific pieces of knowledge such as *Lincoln was born February 12, 1809* and more general statements such as *evaporation is an endothermic process* as well as generic judgments such as *rape is bad* (we follow Sinnott-Armstrong, 2006 Ch. 2 in assuming that evaluative statements are truth-apt the same way as ordinary database entries are). Between the two extremes of the highly particular and the highly general, there are many pieces of knowledge that have limited generality, e.g. that *fat* is a substance often found in food. In fact, most of our knowledge of what words mean falls in this category of limited generality: we know that birds fly, even if penguins don't; that water is H_2O, except for heavy water (D_2O); and that lying is wrong, except for surprise parties.

Linguistic evidence is clear that such *default knowledge* (Reiter, 1978) is a substantive part, perhaps the dominant part, of lexical semantics. That food normally contains fat is evident from the fact that we have a specific word, *fat-free* for describing food that does not, just as we have a specific word, *blind*, to describe people who lack sight. By default, people are assumed sighted, and fully functioning vehicles are capable of moving under their own power. It is also possible for something to move under external forces (indeed, it is the force of gravity that is moving this truck) but the default assumption, especially on a seemingly horizontal road, is that that they move because the engine supplies motive power.

One very successful method of deception relies precisely on the mechanism of defaults: if our goal is to make people believe the truck moves under its own power, create a situation where it normally would. If we want to create the impression that a drug is effective against some painful condition, it is sufficient to show a cheerful pensioner declaring “I feel much better today”. There is no need to say that the person was more ill yesterday, or that the improvement stems from taking the drug – these details are supplied by the default mechanism as a matter of course. In general, instead of saying some outright lie L, it is sufficient to create some K from which anybody will, in the absence of evidence to the contrary, conclude L. Attitudinal deception works by the same mechanism: instead of attacking an evaluative statement such as *furlough programs benefit society* by disputing the general premiss (such programs incentivize good behavior inside prison walls) they focus attention on particular negative cases (such as Willie Horton). Human attitudes are typically formed on the spot, based on few data points: control the choice of instances and you control the attitude formed.

Whether this is doable for any false proposition L remains to be seen, but it does seem hard to create K that will support some L directly contradicted by our senses, or by our KB (individual memories). An assault on collective memory, such as Holocaust denial, is still feasible, but has no effect other than eliciting deep disgust and outrage on people who have first- or even second-hand experience. As individuals don’t share the cultural heritage of humanity uniformly, less informative KBs enable this kind of deception from flat-Earthism to Covid denial to claim many victims.

When Rees, 2013, 1 notes that “deliberate misleading depends on [the misled] inferring meaning beyond what is said in the form of her deceiver’s conversational implicatures as well”, she describes, perhaps inadvertently, the main driver of the deception algorithm. There are two points where we differ from her analysis. First, it is not just conversational but also conventional implicatures that are relevant for us – in fact all implications captured by default reasoning are. Second, these are not just the deceiver’s implicatures, the whole deception relies on the fact that the recipient also has these implications stored in their KB. This is put in sharp relief by human cases of mimicry, such as a spy operating in enemy territory, whose success depends on speaking the language, wearing the clothing/uniforms, and seemingly adopting the value systems, of the enemy. As von Hippel and Trivers, 2011 argue, this is made considerably easier by ‘lying truthfully’ by means of self-deception, another issue that has, perhaps, received less attention in the study of AGI behavior than the subject merits (Fingarette, 1969; Kornai, 2014).

Given a false target L, we need to abductively infer some K that would lead to it by implicature. However, if we have a true target L', the process is not any different as long as we wish to adhere to Grice's maxim of quantity. Rare is the everyday situation when we need to define our terms. A polite request *Please put your dog on a leash* cannot be felicitously responded to by *Define 'dog'* or by *Prove I'm not using an invisible leash.* Even if literally true, *The dog is not mine, I'm just taking it for a walk* will be considered disingenuous.

Since the abductive skill in question is essential for communicating with humans, let us examine for a moment what it rests on. First and foremost, it rests on lexical semantics, knowing what words mean. No AGI lacking in this knowledge could be able to communicate with humans effectively. Second, it rests on an understanding of how humans make inferences, because the abductive step requires this knowledge. Third, it is likely to require some degree of understanding of how human beliefs and desires operate: for example, Holocaust deniers have their greatest success among people who are inclined to believe (neo)nazi tenets to begin with.

However, for successful deception it is sufficient for an AGI to have this knowledge about *human* lexical semantics, human inferencing, and human beliefs and desires, and it in no way follows that they themselves must rely on these semantics, inference patterns, beliefs, or desires. In fact, even among humans, the deceiver is typically a highly cynical manipulator of the victim, not at all sharing their their beliefs or desires. This is not to say that future AGIs will lack desires or beliefs, it simply means that having these *for its own use* is an issue different from having some model of human beliefs and desires. Just as adults have models for childish beliefs and desires, and can manipulate these, often in the service of positive parental goals, we must be prepared for AGIs doing the same with us.

This is particularly clear for attitudinal manipulation, which generally aims at presenting totally reasonable beliefs and attitudes as unreasonable and totally unreasonable ones as reasonable. It may require a book-length effort, but one can present a case for genocide (Card, 1985), where the central mechanism is set up to make the perpetrator appear innocent (Kessel, 2004). Or take the idea of eugenics, considered loathsome by most people (including this author), yet a glib argument in favor is not hard to construct. First, it is everyday experience among the breeders of dogs, horses, and other animals that one can improve the breed by culling. Second, following Darwin we have no doubt that humans are animals. Therefore, culling humans can reasonably be expected to lead to a better breed. We see no need to dignify this argument by discussing why it is wrong, especially as this is tangential to our main point, namely that humans can actually be swayed by such "reasoning".

3. Conclusions

A moral prohibition, such as the Ninth Commandment, can be effective against lies of commission, at least within the usual epistemic limits of taking the deductive closure of the agent's KB. (These limits are severe, given that no large KB can be guaranteed entirely consistent, and some form of paraconsistent logic is required to guarantee that the deductive closure of the KB does not blow up (Belnap, 1977).) But no such prohibition can be formulated against lies of omission, since the number of relevant facts is always infinite. A food recipe cannot start "make sure no ingredient will make a Geiger counter tick". In any communication aimed at humans we *must* abstract away from certain truths and rely on defaults. Moreover, an AGI wishing to communicate with humans must have at its disposal some theory of the defaults used by them simply to be effective, and once the tools required by this positive goal are at hand, there may be nothing to stop the system from abusing them.

Importantly, our argument presupposes nothing for the AGI in way of beliefs or desires: it may have some, or it may have none beyond what is attributed to it by ordinary anthropomorphization as in water "seeking" the lowest point. What it must have, what any communicating agent must have, is some kind of internal model of the communications competence of the other party, and this (e.g. a dictionary) will include defaults. It is worth adding that our argument does not require humans to have beliefs or desires (though the introspective evidence that we do appears incontrovertible), let alone for the AGI to have a superb (indeed, superhuman) theory of these. What is being manipulated primarily is the very means of ordinary human communication, the ability to not produce the full (infinite) story, or not to consider the entire gamut of relevant evaluative factors. While there is considerable debate on the intrinsic ranking of the classic (Maslow, 1943) list of needs, attitudinal deception is not driven by efforts at some deeper reranking, but rather on simply increasing the salience of some factors, leveraging well-known limitations of short-term human memory.

Today, our primary worry is with human agents deploying the linguistic abilities of large-scale deep neural language models in order to deceive other humans. But there may very well come a time when AGIs are more intelligent than humans and thus have a deeper capacity for deception than humans do.

Acknowledgments

I am indebted to Vincent Mueller and the anonymous referees for suggestions that led to considerable improvements in the paper, and to Luís Monez Pereira for calling my attention to the importance of mimicry. Work supported by MILAB, the Hungarian Artificial Intelligence National Laboratory.

References

Asimov, I. (1941). Liar! *Astounding Science Fiction*, *27*(3), 43–55.

Belnap, N. D. (1977). How a computer should think. In G. Ryle (Ed.), *Contemporary aspects of philosophy* (pp. 30–56). Newcastle upon Tyne: Oriel Press.

Bostrom, N. (2012). The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds and Machines*, *22*, 71–85.

Card, O. S. (1985). *Ender's game*. Tor Books.

Dahaner, J. (2012). Bostrom on superintelligence and orthogonality. Retrieved from https://philosophicaldisquisitions.blogspot.com/2012/04/bostrom-on-superintelligence-and.html

Fallis, D. (2010). Lying and deception. *Philosophers' Imprint*, *10*(11). Retrieved from https://quod.lib.umich.edu/cgi/p/pod/dod-idx/lying-and-deception.pdf?c=phimp;idno=3521354.0010.011;format=pdf

Fallis, D. (2018). Lying and omissions. In J. Meibauer (Ed.), *The oxford handbook of lying*. doi:10.1093/oxfordhb/9780198736578.013.13

Fingarette, H. (1969). *Self-deception*. Routledge and Kegan Paul.

Jin, Z., Chauhan, G., Tse, B., Sachan, M., & Mihalcea, R. (2021). How good is NLP? a sober look at NLP tasks through the lens of social impact. *CoRR*, *abs/2106.02359*. Retrieved from https://arxiv.org/abs/2106.02359

Kessel, J. (2004). Creating the innocent killer: Ender's game, intention, and morality. *Foundation, the International Review of Science Fiction*, *33*(90). Retrieved from https://web.archive.org/web/20081227053817/http://www4.ncsu.edu/~tenshi/Killer_000.htm

Kornai, A. (2014). Bounding the impact of AGI. *Journal of Experimental and Theoretical Artificial Intelligence*, *26*(3), 417–438.

Maslow, A. H. (1943). A theory of human motivation. *Psychological Review*, *50*(4), 370–396.

Oshikawa, R., Qian, J., & Wang, W. Y. (2020). A survey on natural language processing for fake news detection. In *Proceedings of the 12th language resources and evaluation conference* (pp. 6086–6093). Marseille, France: European Language Resources Association. Retrieved from https://aclanthology.org/2020.lrec-1.747

Rees, C. F. (2013). Better lie! *Analysis*, *74*, 59–64. doi:https://doi.org/10.1093/analys/ant104

Reiter, R. (1978). On reasoning by default. In *Proceedings of tinlap-2*, University of Illinois at Urbana-Champaign.

Saul, J. M. (2013). *Lying, misleading, and what is said: An exploration in philosophy of language and in ethics*. Oxford University Press.

Shieber, S. M. (2007). The Turing test as interactive proof. *Noûs*, *41*(4), 686–713.

Sinnott-Armstrong, W. (2006). *Moral skepticisms*. Oxford University Press.

Smith, E. O. (1987). Deception and evolutionary biology. *Cultural Anthropology*, *2*(1), 50–64.

von Hippel, W., & Trivers, R. (2011). The evolution and psychology of self-deception. *Behavioral and Brain Sciences*, *34*, 1–56. doi:10.1017/S0140525X10001354