ADVOCATING PROFESSIONAL INTERVENTIONS

CASE STUDY – CICERO, HUMAN-LEVEL DIPLOMACY AI

CICERO[1] is a system developed by Meta AI to play full-press Diplomacy, a board game that involves complex strategy and inter-player communication. This domain is interesting for AI research, as players must persuade and coordinate with others in order to perform well.

This builds on work like AlphaGo[2], the first program to beat top professionals at Go, and GPT-3[3], a language model that can generate text of high enough quality so as to be indistinguishable from human-written content.

Systems analogous to AlphaGo and GPT-3 were combined to allow both strategic planning and natural language communication, and these components were trained in concert such that the system uses its language capabilities to coordinate and persuade other players in order to win games.

Meta AI stresses that Cicero behaves in a way that is “honest”, and that the communications it makes with its opponents represent truthful communication of its intentions. The way they achieve this is by training the model on human Diplomacy communications from situations where Meta AI assessed that the communications of the human matched with their actual action taken in the game.

Individuals prominent in the AI Safety community have voiced concerns over improving AI capability in the domain of social engineering[4], and note that the methods used in Cicero do not guarantee non-deceptive behaviour, and could result in both deception and self-deception within Cicero[5].

CORE READINGS

- CICERO: An AI agent that negotiates, persuades, and cooperates with people.
- Concrete Problems in AI Safety

ESSAY – MASS-PRODUCED MALIGN PERSUASION

Meta's Cicero system represents an enormous leap forward in the capabilities of AI systems – it proves that AI can effectively work with humans to achieve goals in domains where this communication must be over the medium of text, both sent and received. This is a positive indicator for the success of future natural-language interfaces to software, and in this respect, Cicero is a great success for the company.

This technology could however result in significant negative consequences. The recent incident where Google's LaMDA convinced AI researcher Blake Lemoine that it was sentient[6] proves that systems of this nature can convince people of outlandish propositions *even if they are not trying to* – LaMDA only attempts to produce "likely" dialogue, it has no aims or plans. Cicero combines such capability with *goal-directed behaviour* – Cicero is trying to convince people of things in order to *win*. Moreover, Cicero achieves human-level persuasive abilities using a language model possessing only 2.7 billion parameters, compared to the 137 billion[7] parameters of LaMDA, and the neural network scaling laws[8] imply that the returns on increasing model size will be large.

The success of Cicero opens the door for more powerful systems that instrumentally employ persuasion in service of other goals. This technology could conceivably enhance the ability of nefarious actors to spread conspiracy theories or influence public opinion, as the system can decide what to write in order to achieve its aims. Future systems could use *Reinforcement Learning from Human Feedback*[9] to optimise for "convincingness", and the resultant scale at which high-quality harmful writing could be mass-produced is something that must be seriously considered.

In order to mitigate these concerns, we must develop procedures and resources for producing AI systems that are honest and helpful, and, most importantly, aligned with our goals. Cicero is trained on a filtered dataset of dialogue, where only "truthful" dialogue is retained, in order to promote honest communication. This is a good example of the kinds of procedures we should develop and promote, but it is far from sufficient. One might argue that a model trained exclusively on truthful communication will not be able to lie – it doesn't know how! This is not the case – in games, Cicero regularly acts contrary to the way it claimed it was intending to.

We must make it easier and more convenient to produce well-aligned systems over misaligned ones, as other companies and research groups can be relied upon to take the path of least resistance. Skipping work on alignment is attractive, as it will accelerate development, and dangerously unaligned systems could be hard to spot in the near term, and may only reveal themselves once time has sufficiently altered the environment in which a system is acting, an issue called *distributional shift*[10] to which many AI systems are vulnerable.

In order to make well-aligned AI *easy*, we should produce good-quality datasets to train models on so that they are robust and correct, advocate open research practices so that advances can be shared, and focus research effort on AI alignment, especially the *value-loading problem*[11], which is – in simple terms – the problem of "making AI want what we want".

To motivate this, consider that the main issue with Cicero-like systems is that reward-seeking behaviour is liable to overpower whatever guardrails the system has in place to encourage aligned behaviour. In the case of Cicero, this manifests in the form of what David Kreuger calls "stenographic backstabbing"[5], where Cicero will tell a game partner something that it "believes" according to its internal model, but will then go on to act in a way contrary to this, betraying its partner. This indicates that Cicero may encode deceptive intention in a way that allows it to technically "tell the truth", while still being able to break its promises in order to win.

Clearly, this happens because despite Cicero being trained exclusively on truthful dialogue, its goal is to win, not to tell the truth, and because deception is instrumentally useful in Diplomacy, the system learns to deceive while still talking in a way that is “truthful”. Solving the value-loading problem would fix this problem at the root, as Cicero’s goal would no longer conflict with what we want it to do.

Overall, Cicero is a very successful system, but it contains important flaws and additionally represents a breakthrough in a domain that could be used for great harm, and much effort must be expended to ensure that such agentic persuasive systems are used for good ends.

REFERENCES

[1] Bakhtin, A. *et al*. (2022) 'Human-level play in the game of Diplomacy by combining language models with strategic reasoning'. To be published in *Science* [First Release]. Available at: 10.1126/science.ade9097 (Accessed: 30 November 2022).

[2] Silver, D. *et al.* (2016) “Mastering the game of Go with deep neural networks and tree search,” *Nature*, 529(7587), pp. 484–489. Available at: https://doi.org/10.1038/nature16961.

[3] Brown, T. *et al*. (2020) *Language Models are Few-Shot Learners*, *arXiv.org*. Available at: https://arxiv.org/abs/2005.14165 (Accessed: 1 December 2022).

[4] *Preparing for Malicious Uses of AI* (2018). Available at: https://openai.com/blog/preparing-for-malicious-uses-of-ai/ (Accessed: 1 December 2022).

[5] Kreuger, David. (2022) 'Possible stenographic backstabbing in CICERO' [Twitter] 23 November. Available at: https://twitter.com/DavidSKrueger/status/1595387450775437318?s=20&t=DHi-oGqoBVIWR9SSzVr4pA (Accessed: 30 November 2022).

[6] *Is LaMDA Sentient? — an Interview* (2022). Available at: https://cajundiscordian.medium.com/is-lamda-sentient-an-interview-ea64d916d917 (Accessed: 30 November 2022).

[7] *LaMDA: Towards Safe, Grounded, and High-Quality Dialog Models for Everything* (2022). Available at: https://ai.googleblog.com/2022/01/lamda-towards-safe-grounded-and-high.html (Accessed: 1 December 2022).

[8] Kaplan, J. *et al*. (2020) *Scaling Laws for Neural Language Models*, *arXiv.org*. Available at: https://arxiv.org/abs/2001.08361 (Accessed: 1 December 2022).

[9] Christiano, P. *et al*. (2017) *Deep reinforcement learning from human preferences*, *arXiv.org*. Available at: https://arxiv.org/abs/1706.03741 (Accessed: 1 December 2022).

[10] Amodei, D. *et al*. (2016) *Concrete Problems in AI Safety*, *arXiv.org*. Available at: https://arxiv.org/abs/1606.06565 (Accessed: 1 December 2022).

[11] Bostrom, N. (2016) *Superintelligence: Paths, dangers, strategies*. Oxford: Oxford University Press.