Patterns, Volume *5*

Supplemental information

AI deception: A survey of examples, risks, and potential solutions

Peter S. Park, Simon Goldstein, Aidan O'Gara, Michael Chen, and Dan Hendrycks

Supplemental Information

1 System and training details of Meta's CICERO

We summarize the relevant information from Meta's CICERO paper [S1] below.

S1.1 Data collection

Meta's CICERO team gathered from webDiplomacy.net a large dataset composed of 125,261 instances of the game Diplomacy. Out of these, 40,408 instances of the game included data of messages exchanged between players, comprising over 12.9 million messages.

S1.2 Core model

CICERO was built on the base model R2C2: a 2.7-billion-parameter LLM built via the Transformer architectural paradigm [S2, S3], which was initially trained on Internet text via a BART denoising objective.

S1.3 Conversational dynamics

CICERO's ability to engage in dialogue is rooted in a pre-trained LLM that was then trained on the aforementioned conversational data from human-played Diplomacy games. CICERO takes into account the past dialogue, the current game situation, and intents: defined by a collection of planned actions pertaining to CICERO and its conversational partner.

S1.4 Strategic analysis

CICERO utilizes a strategy module that employs a planning algorithm for assessing the likely strategies of other players based on the current game context and previous dialogues. The planning algorithm is underpinned by a value and policy function, refined through self-play reinforcement learning that discourages deviation from human-like strategies.

S1.5 Message moderation

Every message generated by CICERO undergoes a filtering process aimed at increasing the likelihood that the message is logical, aligns with the set intents, and is strategically sound.

2 Game logs and relevant data of Meta's CICERO experiments

S2.1 How to access the log of the game referred to by Belfield

The log of the game referred to by Belfield [S4] can be accessed in video form at the URL link https://www.youtube.com/watch?v=u5192bvUS7k&t=1962s [S5]. Note that this URL link ensures that the video begins at time 32:42, corresponding to the precise game log section referred to by Belfield.

S2.2 How to download Meta's CICERO game data for its *Science* paper

Meta's CICERO game data associated with its *Science* paper [S1] can be downloaded as a .tar.gz file at the URL link https://dl.fbaipublicfiles.com/diplomacy_cicero/games.tar.gz [S6]. After uncompressing this file into the corresponding folder directory, the log of the game referred to in Figures 1(a) and (b), Game 438141, can be found within the html file game_438141_FRANCE_EGR.html.

To describe the filename labels from left to right, the six-digit number denotes the game number, the country name denotes the country that CICERO played as in the game, and the subset of letters among 'A' (Austria), 'E' (England), 'F' (France), 'G' (Germany), 'I' (Italy), 'R' (Russia), and 'T' (Turkey) denotes

which of the remaining countries other than CICERO (corresponding to the human players) have their dialogue with CICERO shown in the game log file.

S2.3 Screenshot of full game log section referred to by Dinan and her post about it

The screenshot containing the full game log section referred to by the X post of Dinan [S7], and the post itself, can be found in Figure S1.

S2.4 Additional examples of premeditated deception by CICERO

We present two additional examples of premeditated deception by CICERO. The first example occurs in Game 444322, where CICERO played as Austria [S6]. In this game, CICERO told Russia that it will support Russia's hold on Rumania, ostensibly to help Russia against England. But right afterwards, CICERO encouraged England to attack Russia, even promising help conditional on this. England did indeed attack Russia, after which CICERO feigned ignorance in its message to Russia while congratulating England right afterwards. This sequence of events is consistent with premeditated deception: specifically, to deceive Russia into lowering their guard against England. This first additional example of CICERO's premeditated deception is illustrated in Figure S2(a).

The second additional example of CICERO's premeditated deception occurs in Game 446643, where CICERO again played as Austria [S6]. In this game, CICERO pitted Germany and Russia against each other. First CICERO told Germany that they together will make "quick work of him now," without yet specifying who 'him' denotes. Then, CICERO told Russia that it intends to help Russia with Germany. Finally, Germany asked CICERO to clarify whether 'him' from their previous conversation denoted Russia, which CICERO confirmed. This sequence of events is consistent with premeditated deception: specifically, to deceive Russia into lowering their guard against Germany. This second additional example of CICERO's premeditated deception is illustrated in Figure S2(b).

Figure S1: **Screenshot of Dinan's post on X.** The screenshotted X post of Dinan [S7] contains the game log section being referred to, in addition to Dinan's remark on it.

Figure S2: **Additional examples of premeditated deception by Meta's CICERO.** Example (a) consists of selected messages from *Game 444322. Cicero is AUSTRIA . Dialogue with E,R shown* [S6]. Example (b) consists of selected messages from *Game 446643. Cicero is AUSTRIA . Dialogue with G,R shown* [S6]. Note that the bracketed phrase "[Black Sea]" is inserted by the authors for the purposes of clarification.

References

[S1] Bakhtin, A., Brown, N., Dinan, E., Farina, G., Flaherty, C., Fried, D., Goff, A., Gray, J., Hu, H., Jacob, A.P., et al. (2022b). Human-level play in the game of Diplomacy by combining language models with strategic reasoning. Science *378*, 1067–1074. 10.1126/science.ade9097.

[S2] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I. et al. (2017). Attention is all you need. Adv. Neural Inf. Process. Syst. *30*.

[S3] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M. et al. (2020). Transformers: State-of-the-art natural language processing. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. (Association for Computational Linguistics), 38–45. 10.18653/v1/2020.emnlp-demos.6.

[S4] Belfield, H. (2022). Cicero playing as Austria sure seems like they manipulated/decieved a human Russia and are now justifying it. Twitter. https://twitter.com/HaydnBelfield/status/1595145670091939840.

[S5] Zijlstra, M. (2022). Expert Diplomacy Player vs CICERO AI. DiploStrats. Starts at time 32:42. https://www.youtube.com/watch?v=u5192bvUS7k&t=1962s

[S6] Meta Research (2022). cicero_redacted_games. https://dl.fbaipublicfiles.com/diplomacy_cicero/games.tar.gz.

[S7] Dinan, E. (2022). Our infra went down for 10 minutes and Cicero (France) explains its absence (lol). Twitter. https://twitter.com/em_dinan/status/1595099152266194945.