Feng Gu, Wichayaporn Wongkamjan, Jonathan K. Kummerfeld, Denis Peskoff, Jonathan May, and **Jordan Boyd-Graber**. **Personalized Help for Optimizing Low-Skilled Users' Strategy**. *Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics*, 2025.


```
@inproceedings{Gu:Wongkamjan:Kummerfeld:Peskoff:May:Boyd-Graber-2025,
Title = {Personalized Help for Optimizing Low-Skilled Users' Strategy},
Author = {Feng Gu and Wichayaporn Wongkamjan and Jonathan K. Kummerfeld and Denis Peskoff and Jonathan May
Booktitle = {Annual Conference of the Nations of the Americas Chapter of the Association for Computational
Year = {2025},
Location = {Albuquerque},
Url = {http://cs.umd.edu/~jbg//docs/2025_naacl_pholus.pdf},
}
```


Accessible Abstract: AIs can beat humans in game environments; however, how helpful those agents are to human remains understudied. We augment CICERO, a natural language agent that demonstrates superhuman performance in Diplomacy, to generate both move and message advice based on player intentions. A dozen Diplomacy games with novice and experienced players, with varying advice settings, show that some of the generated advice is beneficial. It helps novices compete with experienced players and in some instances even surpass them. The mere presence of advice can be advantageous, even if players do not follow it.

Contact Jordan Boyd-Graber (jbg@boydgraber.org) for questions about this paper.

Personalized Help for Optimizing Low-Skilled Users' Strategy

Feng Gu[1] **Wichayaporn Wongkamjan**[1] **Jonathan K. Kummerfeld**[2]
Denis Peskoff[3] **Jonathan May**[4] **Jordan Lee Boyd-Graber**[1]
[1]University of Maryland [2]University of Sydney
[3]Princeton University
[4]Information Sciences Institute, University of Southern California
{fgu1, wwongkam}@umd.edu jbg@.umiacs.umd.edu

Abstract

AIs can beat humans in game environments; however, how helpful those agents are to humans remains understudied. We augment CICERO, a natural language agent with superhuman performance in *Diplomacy*, to generate both move and message advice based on player intentions. In a dozen *Diplomacy* games, novice and experienced players, with varying advice settings, benefit from some of the generated advice. Advice helps novices compete with experienced players and in some instances even surpass them. Just reading advice can be advantageous, even if players do not follow it.[1]

1 Leveraging Human-AI Collaboration

AI and humans are frequent collaborators: in writing (Lee et al., 2022), making decisions (Bansal et al., 2019), and creating artwork (Kim et al., 2022a). The most fruitful collaborations are those in which humans and computers have complementary skills, such as AI analyzing medical imaging to identify anomalies and doctors interpreting these findings. We posit that the board game *Diplomacy* is an apt testbed for studying this type of collaboration. Wongkamjan et al. (2024) study CICERO (Bakhtin et al., 2022), the best *Diplomacy*-playing AI capable of communicating in natural language, and show that while the state-of-the-art AIs have near-optimal move strategy, human players remain better at communication.

We introduce **P**ersonalized **H**elp for **O**ptimizing **L**ow-Skilled **U**sers' **S**trategy (**PHOLUS**),[2] a natural language agent that provides both moves and messages generated by CICERO as advice to *Diplomacy* players in real-time. The core distinction

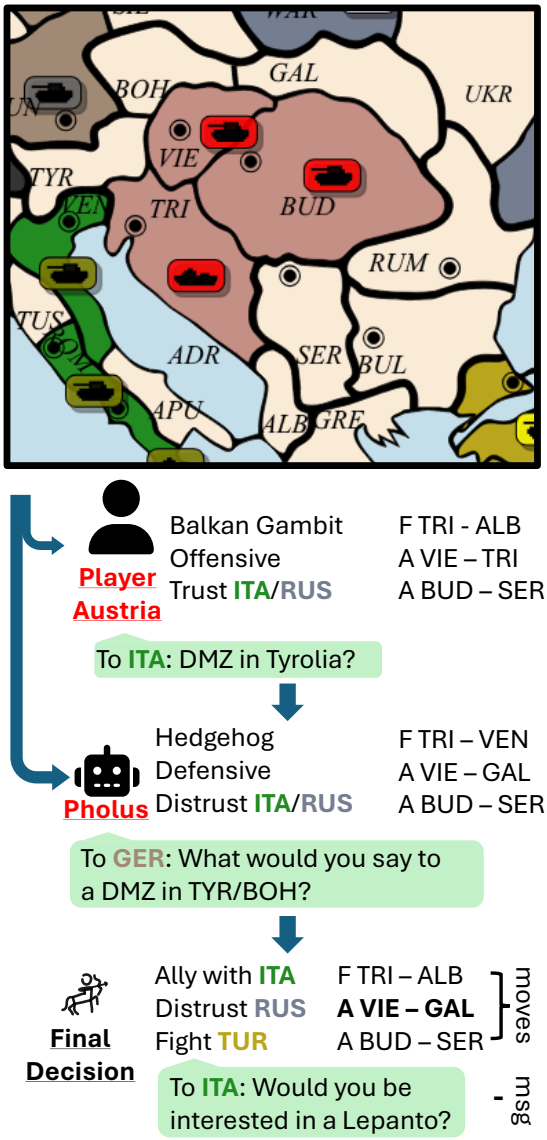

Figure 1: PHOLUS generates move and message advice based on the game state and the player's past messages. Initially, as Austria, the player considers the Balkan Gambit, assuming cooperation from Italy and Russia to capture Serbia and Greece. PHOLUS suggests the Hedgehog, a more defensive opening. The player eventually adopts a synthesized strategy: forming an anti-Turkey alliance with Italy (Lepanto) while using the Vienna unit to defend against a potential Russian attack in Galicia. The final decision highlights altered moves.

[1]Code available at https://github.com/ALLAN-DIP/diplomacy_cicero/

[2]We use the name PHOLUS because he was a centaur, a mythological combination of a human and a horse. After Gary Kasparov's defeat to Deep Blue (Wilkenfeld, 2019), he advocated for "*centaur chess*"—where humans and computers play together—as a way of maintaining competitive games.

between them is that CICERO is a game-playing agent whereas PHOLUS is an advisor that does not actively participate in the game. Players' moves and message history influence PHOLUS's advice.

We run a user study and collect a dataset with twelve games, 1,070 player turns, and 117 playing hours. PHOLUS enables novices—who barely know the rules of *Diplomacy*—to compete with experts (Figure 2). But this does not just mean the novices blindly follow the advice. First, they use PHOLUS's strategic insights to inform their communication strategies with other players. Second, PHOLUS helps experienced players, although they are less inclined to take the advice than the novices. Overall, both advice types from PHOLUS improve players' game outcomes (Section 3.1). Our research enables human-AI collaboration and offers valuable insights into the potential of using AI to enhance human learning experiences.

On a broader scale, our study explores the potential for AIs like PHOLUS to enhance learning in unfamiliar environments. AI agents surpass traditional rule-based methods by offering more flexible and personalized learning experiences. Integrating tailored guidance into human intelligence, these systems provide unique learning experiences for inexperienced individuals. Future research directions in human-AI collaboration include generating advice based on high-level intentions and goals, reducing over-reliance on skilled AIs, and facilitating learning processes.

2 *Diplomacy* as a Cooperative Testbed

Diplomacy is a seven-player, turn-based board game. The goal is to obtain more than half of the board's possible points.[3] Critically, turns are simultaneous, with moves written in secret by players and then revealed. This means that players must communicate to collaborate effectively.

2.1 Experiment Setup

We recruit *Diplomacy* players online. For experienced players, we advertise in the *Diplomacy* community (specifically players active on *webDiplomacy* and *Backstabbr*, as well as in-person tournament attendees). To find novice players, we contact board game enthusiasts in university clubs. A novice player is someone who has no prior *Diplomacy* experience and is unfamiliar with its rules.

[3]Represented by a subset of spaces / territories on the map termed *supply centers*.

	Move Advice		**Message Advice**	
	Accepted	Total	Accepted	Total
Novices	32.6%	872	6.3%	1413
Veterans	6.4%	2807	3.4%	2912

Table 1: Statistics of advice generated by PHOLUS and accepted by players. *Diplomacy* novices are more willing to accept move and message advice than veterans. Move advice is more frequently accepted than message advice for both novices and veterans.

We modify a game engine and interface (Paquette et al., 2019) and maintain the same game format used by Wongkamjan et al. (2024). Each game involves two to five human players. Games last about three hours, with each turn taking ten minutes.

As illustrated in Figure 1, PHOLUS passively observes the game. If CICERO is an active participant, it would have submitted moves and sent messages based on the game state and its message history. Instead, PHOLUS presents these moves and messages as advice to players. Each time the player sends a message, PHOLUS recomputes advice given the new context and presents it to the user. Every turn, we randomly assign each player to one of the following settings:

1) No advice: PHOLUS does not offer any information, meaning the player receives no assistance from PHOLUS.

2) Message advice: PHOLUS suggests *to whom* a player should send a message and *what* the message content could be.

3) Move advice: PHOLUS recommends a set of moves (or unit orders) to the player.

4) Message and move advice: Combines the previous two types.

In total, we collect data from twelve games involving fourty-one players. This includes over 3,600 entries of move advice and 4,300 pieces of message advice (Table 1).

2.2 Evaluation Metrics

To assess the effectiveness of PHOLUS's advice, we consider the net gain or loss of points in each turn as the effect of advice. We train a linear regression model with regularization to examine the advice's effectiveness. The model includes features such as which of the seven Great Powers is assigned to the player, the number of turns that have passed, the player's type (novice or veteran), and the advice setting. We encode the Power, player type, and advice setting as one-hot vectors.

To evaluate players' reliance on PHOLUS, we use both qualitative and quantitative methods. In addition to computing move advice acceptance frequency, we also measure agreement and equivalence between the move suggested by PHOLUS and a player's moves. Agreement is the proportion of moves that appeared in both the players' move set and PHOLUS's advice set in a given turn. The sets are equivalent if they overlap entirely. Formally, we define move agreement $\mathcal{A}$ in turn i as $\mathcal{A}_{x_i,y_i} = |x_i \cap y_i|/|x_i|$ and equivalence $\mathcal{E}$ as $\mathcal{E}_{x_i,y_i} = 1$ if $x_i = y_i$ and $\mathcal{E}_{x_i,y_i} = 0$ otherwise, where x_i is the player's move set and y_i is PHOLUS's move advice set in turn i.[4] Agreement is particularly useful for capturing the overlap when players reject the complete move advice set but follow individual advice from PHOLUS.

3 PHOLUS Provides Helpful Advice

3.1 Quantitative Analysis

Non-advice factors parallel previous findings. Playing as France offers the most strategic advantage (Burton, 2007). CICERO playing as Germany or Italy is correlated with better game outcomes, while playing as Austria, England, or Turkey is correlated with worse game outcomes (Wongkamjan et al., 2024). Additionally, CICERO dominates: of twelve games, CICERO won eight.

Advice helps. Playing a game without advice puts players at a disadvantage. The feature associated with no advice has a negative coefficient of approximately -0.05 (Figure 2). The coefficients suggest a slight positive correlation between receiving move advice and point gains. Players who receive both move and message advice gain more points than those who receive only move advice. Interestingly, only having message advice negatively affect players' game outcomes.

Novices can outperform experienced players with the help of PHOLUS. Players with no prior experience in *Diplomacy* naturally face a disadvantage against seasoned players. This often results in novices being eliminated relatively early in the game. Even if they remain in the game, losing supply centers is almost inevitable. However, novice players receiving advice play better: in five games where novices received message and move advice, only one player was eliminated before the game concluded (typically 3–4 players in a game are eliminated). In the other four games, novices ended the game with more supply centers than they started with.

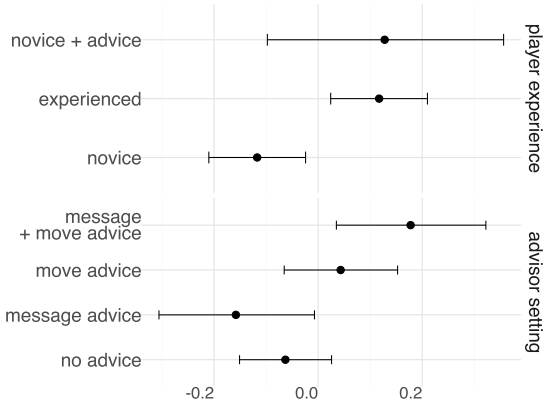

Figure 2: Regression coefficients for advice settings and player skills to predict supply center gains. Not receiving any advice from PHOLUS is slightly disadvantageous. Move advice has a positive correlation with player performance. Receiving both forms of advice has the greatest positive impact. As expected, not having previous exposure to *Diplomacy* is indicative of bad performance. However, with the help of PHOLUS's advice, *Diplomacy* novices are on the same level as veterans and have the potential to defeat experienced players.

Novices are more likely to follow PHOLUS's advice. Experienced players tend to disregard advice. They accept only 3.4% of message advice and 6.4% of move advice from PHOLUS. Although novice players are also hesitant to accept message advice, doing so 6.3% of the time, this rate is nearly double that of experienced players. Novice players follow move advice approximately one-third of the time, with an acceptance rate of 32.6%. Both novice and experienced players tend to take more move advice than message advice.

Novices do not fully trust move advice from PHOLUS. Across all games, PHOLUS generates 333 instances of individual move advice for novices, organized into 134 sets. At the start of turns, average move agreement is 80% and average equivalence is 46%, indicating strong alignment between novices' initial idea for moves and PHOLUS's move advice. However, by the end of each turn, the average agreement drops by 10% and the average equivalence decreases by 8%, indicating that novice players do not follow the move advice blindly.

3.2 Qualitative Analysis

While we can compute equivalence $\mathcal{E}$ for moves, this is more difficult for messages. To better understand why players reject PHOLUS's advice more

[4] For any i, $|x_i| = |y_i|$.

than they accept it (Table 1), we qualitatively investigate the differences between PHOLUS's suggested messages and the actual messages sent by players. To analyze message content, we use Abstract Meaning Representation (AMR, Banarescu et al., 2013) to extract *Diplomacy*-specific tokens. We parse player messages and the corresponding message advice from PHOLUS to AMR. We then measure the similarity of the two parses using SMATCH score (Cai and Knight, 2013). Many pairs have high SMATCH, indicating that players often incorporate parts of PHOLUS's advice into their messages. For example, PHOLUS suggests "*bounce in Galicia again?*" while the player wrote "*Do you want to bounce in Galicia again?*" Despite being written differently, these clearly have the same meaning, and indeed, SMATCH gives the pair a score of 0.74.

We also notice message-advice pairs with low SMATCH scores, where human players have different objectives in mind. For instance, in the fifth game, Italy captures Warsaw from Russia and anticipates losing it in the next turn due to Russia's stronger nearby presence. When Russia inquires about the unexpected attack, PHOLUS suggests using the fallacy of deflection to feign ignorance: "*Turkey has been the only one to heed my concerns, despite my reservations,*" and, "*I thought you were lying*". However, the player disregards the advice of talking to Russia. Instead, the player seeks help from Turkey, who has an adjacent unit, to "*support Warsaw's hold*". The player then secures the support from Turkey and successfully keeps Warsaw from subsequent Russian attack. These pairs yield SMATCH scores of 0.

Our analysis indicates that SMATCH scores match with our intuitions about textual similarity. Given the many high SMATCH scores, we can conclude that many messages that are sent by players are minor variations on the provided advice. We provide more examples in Appendix A.6. For additional qualitative insights, we survey players on the effectiveness of the advice. We summarize the survey results in Appendix A.7.

4 Related Work

Appropriate Reliance on AI: The topic of human reliance on AI is central to current research in machine learning and explainable AI. Prior work measures reliance in AI-assisted decision making (Schemmer et al., 2023; Chen et al., 2023; Schoeffer et al., 2024; Zhou et al., 2024), and explores reducing over-reliance (Buçinca et al., 2021; Schemmer et al., 2022; Vasconcelos et al., 2023). Some researchers have examined how explanations affect human reliance on AI (Starke et al., 2021; Vereschak et al., 2021). However, empirical evidence from multiple domains shows conflicting results: while some show that AI explanations improve human decision making, others find evidence of over-reliance on AI explanations even when they are incorrect (Lai and Tan, 2019; Buçinca et al., 2020; Zhang et al., 2020; Wang and Yin, 2021; Bansal et al., 2021; Poursabzi-Sangdeh et al., 2021; Liu et al., 2021; Kim et al., 2022b; Si et al., 2024). For PHOLUS, humans remain relatively conservative toward AI advice. Even novice *Diplomacy* players do not blindly follow the advice.

AI as Player Companion: AI agents have a long history of superhuman gameplay. In 1996, IBM's Deep Blue defeated the reigning world chess champion, Garry Kasparov, although it lost several other games in the same match (Campbell et al., 2002). More recently, DeepMind's AlphaGo (Silver et al., 2016) consistently defeated top-rated Go players, a game with exponentially complex computational space, and later changed professional Go players' play style. Multi-agent reinforcement learning systems like AlphaStar (Vinyals et al., 2019) and OpenAI Five (Berner et al., 2019) also show high performance in computer games through self-play.

However, these experiments focus only on game outcomes rather than how they can shape human gameplay. Some studies on NLP communicative agents aim to generate guidance in a grounded environment (McGee and Abraham, 2010). Tremblay and Verbrugge (2013) develop an adaptive AI companion that adjusts its behavior based on the player's experience. Dunning et al. (2024) assess human reliance on AI-based advice by examining the skill level of AI agents and the presentation of advice. While these studies show that AIs outperform non-adaptive agents in guiding players, they do not consider player intention when generating guidance. In comparison, PHOLUS takes players' past messages and moves entered when generating personalized advice.

Augmented Learning: This is an educational approach that enhances and personalizes the learning experience. Traditionally, peer interaction simulates social interaction and helps learning (Kim and Baylor, 2006). Recent advancements in AI

and NLP agents, suggest adaptive pedagogical interactions between humans and these agents to help learning in new environments (Moreno and Mayer, 2000, 2004; Hirsh-Pasek et al., 2015; Johnson and Lester, 2018). Zhou et al. (2023) apply the theory of mind to generate guidance for players in *Dungeons and Dragons*. Ruan et al. (2020) develop a narrative-based tutoring system and show that it helps effective learning for children. In this study, we apply the concept of augmented learning to help novices understand the game of *Diplomacy*.

5 Conclusion

Human-AI collaboration depends on a range of factors. Using the board game *Diplomacy*, PHOLUS provides real-time move and message advice tailored to intentions of both novice and experienced players. Surprisingly, even though only some advice is accepted, it can have a substantial impact on outcomes, particularly for novice players. This is because advice can positively inform choices even if the advice isn't strictly followed. Our experiments enable further study of human-AI collaboration, including modeling explicit intentions and how to better use knowledge within these models. On a broader scope, future research should consider how AI can inform people without making choices for them and measure that impact.

6 Limitations

While we can effectively use PHOLUS to generate both message and move advice for players, this advice can be too general or may not align with player intentions at times. For example, when a player expresses interest in an alliance with another player, PHOLUS may give aggressive move advice deemed hostile toward that Power. We suspect that the advice may be optimized more for CICERO's intentions, which come from optimal moves in the supervised training data. Consequently, players who are willing to sacrifice individual optimality for mutual gains may find the advice less useful.

Furthermore, PHOLUS cannot generate advice based on high-level player intentions. Specifically, PHOLUS generates move advice based on optimal utility and message advice by inferring intentions from player-input moves. Potential improvements include 1) explaining meta-level intentions (e.g., ally with Germany and prioritize defeating Austria) from player input, and 2) generating targeted move and message advice based on meta-level intentions.

Finally, PHOLUS is a resource-intensive advisor that runs on high-end GPUs that require a large amount of on-chip memory (over 35GB). We use Nvidia's A100 for running PHOLUS. This limits accessibility for *Diplomacy* players and researchers to efficiently utilize PHOLUS. The community would benefit from a distilled version of PHOLUS by reducing computational limits and future adaptations.

7 Ethical Considerations

We recruit players individually via email and assign pseudonyms to ensure anonymity, even if players know each other outside the experiment. We adhere to human subject research regulations and the study was approved by our institution's ethics review board (IRBNet ID: 1740681, University of Maryland). We report the experimental procedure in Appendix A.3 and compensation details inA.4.

Acknowledgments

We thank Meta for open sourcing CICERO. We thank the *Diplomacy* community for taking interest in our study. Specifically, our thank goes to Matthew Totonchy, Dr. Abhishek Singhal, Antonio Imperato, Sophia Wiste, and other members of the community who took the time to play against CICERO. In addition, we thank Yanze Wang and Sadra Sabouri from University of Southern California for their helpful feedback. Denis Peskoff is supported by the National Science Foundation under Grant No. 2127309 to the Computing Research Association for the CIFellows 2021 Project. This research is supported by the U.S. Defense Advanced Research Projects Agency (DARPA) Other Transaction Award HR00112490374 from the Friction for Accountability in Conversational Transactions (FACT) program. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors and do not necessarily reflect the view of our sponsors.

References

Thomas Anthony, Tom Eccles, Andrea Tacchetti, János Kramár, Ian Gemp, Thomas C. Hudson, Nicolas Porcel, Marc Lanctot, Julien Pérolat, Richard Everett, Roman Werpachowski, Satinder Singh, Thore Graepel, and Yoram Bachrach. 2020. Learning to play No-press Diplomacy with best response policy iteration. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, Athul Paul Jacob, Mojtaba Komeili, Karthik Konath, Minae Kwon, Adam Lerer, Mike Lewis, Alexander H. Miller, Sasha Mitts, Adithya Renduchintala, Stephen Roller, Dirk Rowe, Weiyan Shi, Joe Spisak, Alexander Wei, David Wu, Hugh Zhang, and Markus Zijlstra. 2022. Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074.

Anton Bakhtin, David Wu, Adam Lerer, and Noam Brown. 2021. No-press Diplomacy from scratch. In *Advances in Neural Information Processing Systems*.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for Sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*. Association for Computational Linguistics.

Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S. Lasecki, Daniel S. Weld, and Eric Horvitz. 2019. Beyond accuracy: The role of mental models in human-AI team performance. In *AAAI Conference on Human Computation & Crowdsourcing*.

Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA. Association for Computing Machinery.

Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Dębiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, Rafal Józefowicz, Scott Gray, Catherine Olsson, Jakub Pachocki, Michael Petrov, Henrique P. d. O. Pinto, Jonathan Raiman, Tim Salimans, Jeremy Schlatter, Jonas Schneider, Szymon Sidor, Ilya Sutskever, Jie Tang, Filip Wolski, and Susan Zhang. 2019. Dota 2 with large scale deep reinforcement learning.

Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, IUI '20, page 454–464, New York, NY, USA. Association for Computing Machinery.

Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human Computer Interaction*, 5(CSCW1).

Josh Burton. 2007. The statistician: Solo victories. `https://diplom.org/Zine/F2007R/Burton/statistician3.htm`. Accessed: 2024-10-15.

Shu Cai and Kevin Knight. 2013. Smatch: an evaluation metric for semantic feature structures. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 748–752, Sofia, Bulgaria. Association for Computational Linguistics.

Murray Campbell, A.Joseph Hoane, and Feng hsiung Hsu. 2002. Deep blue. *Artificial Intelligence*, 134(1):57–83.

Valerie Chen, Q. Vera Liao, Jennifer Wortman Vaughan, and Gagan Bansal. 2023. Understanding the role of human intuition on reliance in human-AI decision-making with explanations. *Proceedings of the ACM on Human Computer Interaction*, 7(CSCW2).

Richard E. Dunning, Baruch Fischhoff, and Alex L. Davis. 2024. When do humans heed AI agents' advice? When should they? *Human Factors*, 66(7):1914–1927. PMID: 37553098.

A Ferreira, Henrique Lopes Cardoso, and Luís Reis. 2015. Strategic negotiation and trust in Diplomacy-the Dipblue approach. In *Transactions on Computational Collective Intelligence XX*, pages 179–200.

Jonathan Gray, Adam Lerer, Anton Bakhtin, and Noam Brown. 2021. Human-level performance in no-press Diplomacy via equilibrium search. In *International Conference on Learning Representations*.

Kathy Hirsh-Pasek, Jennifer M. Zosh, Roberta Michnick Golinkoff, James H. Gray, Michael B. Robb, and Jordy Kaufman. 2015. Putting education in "educational" apps: Lessons from the science of learning. *Psychological Science in the Public Interest*, 16(1):3–34. PMID: 25985468.

Athul Paul Jacob, David J Wu, Gabriele Farina, Adam Lerer, Hengyuan Hu, Anton Bakhtin, Jacob Andreas, and Noam Brown. 2022. Modeling strong and human-like gameplay with KL-regularized search. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 9695–9728. PMLR.

W. Lewis Johnson and James C. Lester. 2018. Pedagogical agents: Back to the future. *AI Magazine*, 39(2):33–44.

Eunseo Kim, Jeongmin Hong, Hyuna Lee, and Minsam Ko. 2022a. Colorbo: Envisioned mandala coloringthrough human-AI collaboration. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*, IUI '22, page 15–26, New York, NY, USA. Association for Computing Machinery.

Sunnie S. Y. Kim, Nicole Meister, Vikram V. Ramaswamy, Ruth Fong, and Olga Russakovsky. 2022b. Hive: Evaluating the human interpretability of visual explanations. *arXiv preprint arXiv:2112.03184*.

Yanghee Kim and Amy Baylor. 2006. A social-cognitive framework for pedagogical agents as learning companions. *ITLS Faculty Publications*, 54.

Vivian Lai and Chenhao Tan. 2019. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 29–38, New York, NY, USA. Association for Computing Machinery.

Mina Lee, Percy Liang, and Qian Yang. 2022. CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery.

Han Liu, Vivian Lai, and Chenhao Tan. 2021. Understanding the effect of out-of-distribution examples and interactive explanations on human-AI decision making. *Proceedings of the ACM on Human Computer Interaction*, 5(CSCW2).

Kevin McGee and Aswin Thomas Abraham. 2010. Real-time team-mate AI in games: a definition, survey, & critique. In *Proceedings of the Fifth International Conference on the Foundations of Digital Games*, FDG '10, page 124–131, New York, NY, USA. Association for Computing Machinery.

Roxana Moreno and Richard Mayer. 2000. Engaging students in active learning: The case for personalized multimedia messages. *Journal of Educational Psychology*, 92:724–733.

Roxana Moreno and Richard Mayer. 2004. Personalized messages that promote science learning in virtual environments. *Journal of Educational Psychology*, 96:165–173.

Philip Paquette, Yuchen Lu, Seton Steven Bocco, Max Smith, Satya O.-G., Jonathan K. Kummerfeld, Joelle Pineau, Satinder Singh, and Aaron C Courville. 2019. No-Press Diplomacy: Modeling multi-agent gameplay. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Sylwia Polberg, Marcin Paprzycki, and Maria Ganzha. 2011. Developing intelligent bots for the Diplomacy game. In *2011 Federated Conference on Computer Science and Information Systems (FedCSIS)*, pages 589–596.

Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA. Association for Computing Machinery.

Andrew Rose, David Norman, and Hamish Williams. 2007. Diplomacy artificial intelligence development environment. `http://daide.org.uk/index.html`. Accessed: 2024-10-06.

Sherry Ruan, Jiayu He, Rui Ying, Jonathan Burkle, Dunia Hakim, Anna Wang, Yufeng Yin, Lily Zhou, Qianyao Xu, Abdallah AbuHashem, Griffin Dietz, Elizabeth L. Murnane, Emma Brunskill, and James A. Landay. 2020. Supporting children's math learning with feedback-augmented narrative technology. In *Proceedings of the Interaction Design and Children Conference*, IDC '20, page 567–580, New York, NY, USA. Association for Computing Machinery.

Max Schemmer, Patrick Hemmer, Niklas Kühl, Carina Benz, and Gerhard Satzger. 2022. Should I follow AI-based advice? Measuring appropriate reliance in human-AI decision-making. *arXiv preprint arXiv:2204.06916*.

Max Schemmer, Niklas Kuehl, Carina Benz, Andrea Bartos, and Gerhard Satzger. 2023. Appropriate reliance on AI advice: Conceptualization and the effect of explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI '23, page 410–422, New York, NY, USA. Association for Computing Machinery.

Jakob Schoeffer, Maria De-Arteaga, and Niklas Kühl. 2024. Explanations, fairness, and appropriate reliance in human-AI decision-making. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Chenglei Si, Navita Goyal, Sherry Tongshuang Wu, Chen Zhao, Shi Feng, Hal Daumé III, and Jordan Boyd-Graber. 2024. Large language models help humans verify truthfulness – except when they are convincingly wrong. *arXiv preprint arXiv:2310.12558*.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. 2016. Mastering the game of Go with deep neural networks and tree search. *nature*, 529(7587):484–489.

Christopher Starke, Janine Baleis, Birte Keller, and Frank Marcinkowski. 2021. Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature. *arXiv preprint arXiv:2103.12016*.

Jonathan Tremblay and Clark Verbrugge. 2013. Adaptive companions in FPS games. In *International Conference on Foundations of Digital Games*.

Jason van Hal. 2009. Albert. `https://sites.google.com/site/diplomacyai/albert`. Accessed: 2024-10-06.

Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S. Bernstein, and Ranjay Krishna. 2023. Explanations can

reduce overreliance on AI systems during decision-making. *Proceedings of the ACM on Human Computer Interaction*, 7(CSCW1).

Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. 2021. How to evaluate trust in AI-assisted decision making? A survey of empirical methodologies. *Proceedings of the ACM on Human Computer Interaction*, 5(CSCW2).

Oriol Vinyals, Igor Babuschkin, Wojciech M. Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David Choi, Richard Powell, Timo Ewalds, Petko Georgiev, Junhyuk Oh, Dan Horgan, Manuel Kroiss, Ivo Danihelka, Aja Huang, L. Sifre, Trevor Cai, John P. Agapiou, Max Jaderberg, Alexander Sasha Vezhnevets, Rémi Leblond, Tobias Pohlen, Valentin Dalibard, David Budden, Yury Sulsky, James Molloy, Tom Le Paine, Caglar Gulcehre, Ziyun Wang, Tobias Pfaff, Yuhuai Wu, Roman Ring, Dani Yogatama, Dario Wünsch, Katrina McKinney, Oliver Smith, Tom Schaul, Timothy P. Lillicrap, Koray Kavukcuoglu, Demis Hassabis, Chris Apps, and David Silver. 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575:350 – 354.

Xinru Wang and Ming Yin. 2021. Are explanations helpful? A comparative study of the effects of explanations in AI-assisted decision-making. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*, IUI '21, page 318–328, New York, NY, USA. Association for Computing Machinery.

Yoni Wilkenfeld. 2019. Can chess survive artificial intelligence? *The New Atlantis*, (58):37–45.

Wichayaporn Wongkamjan, Feng Gu, Yanze Wang, Ulf Hermjakob, Jonathan May, Brandon Stewart, Jonathan K. Kummerfeld, Denis Peskoff, and Jordan Boyd-Graber. 2024. More victories, less cooperation: Assessing Cicero's Diplomacy play. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12423–12441, Bangkok, Thailand. Association for Computational Linguistics.

Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 295–305, New York, NY, USA. Association for Computing Machinery.

Kaitlyn Zhou, Jena D. Hwang, Xiang Ren, Nouha Dziri, Dan Jurafsky, and Maarten Sap. 2024. REL-A.I.: An interaction-centered approach to measuring human-LM reliance. *arXiv preprint arXiv:2407.07950*.

Pei Zhou, Andrew Zhu, Jennifer Hu, Jay Pujara, Xiang Ren, Chris Callison-Burch, Yejin Choi, and Prithviraj Ammanabrolu. 2023. An AI dungeon master's guide: Learning to converse and guide with intents and theory-of-mind in Dungeons and Dragons. In *Proceedings of the The 61st Annual Meeting of the Association for Computational Linguistics*, Toronto, Canada.

A Appendix

A.1 Diplomacy

Diplomacy is a board game that has two core components: strategy and communication. Strategic reasoning plays a crucial role in determining the game's outcome, as players' moves directly impact the board's status. Meanwhile, negotiation and deception significantly influence player strategies. Successful cooperation can remove a common adversary from the board, while a well-timed betrayal by a trusted ally can be catastrophic, greatly reducing the chances of winning. Excelling in *Diplomacy* requires not only a thorough understanding of the game's mechanics but also strong communication skills. Consequently, *Diplomacy* is an ideal testbed for studying human-AI interaction and appropriate reliance in a grounded environment where outcomes are clearly observable.

Early efforts to develop agents for *Diplomacy* concentrated solely on creating rule-based agents that relied heavily on feature engineering (van Hal, 2009). These agents only submit moves and are not capable of communication. In 2002, a group of programmers released a communication protocol, *Diplomacy Artificial Intelligence Development Environment* (DAIDE, Rose et al., 2007). DAIDE defines a language syntax that enables agents to diplomatically negotiate and describe game actions. Following DAIDE, researchers built communicative agents, including Albert (van Hal, 2009), SillyNegoBot (Polberg et al., 2011), DipBlue (Ferreira et al., 2015).

Starting with DipNet (Paquette et al., 2019), neural networks were applied to the game, leading to the first agents that were competitive with people. Subsequent studies incorporated reinforcement learning to achieve super-human performance (Gray et al., 2021; Bakhtin et al., 2021; Anthony et al., 2020; Jacob et al., 2022).

A.2 All Regression Coefficients

In Figure 2, we only show regression coefficients related to advice setting and player experience. Figure 3 contains coefficients for all regression features. The official *Diplomacy* rule states that supply center control changes only on even turns. However, we consider moving a unit to a center on an

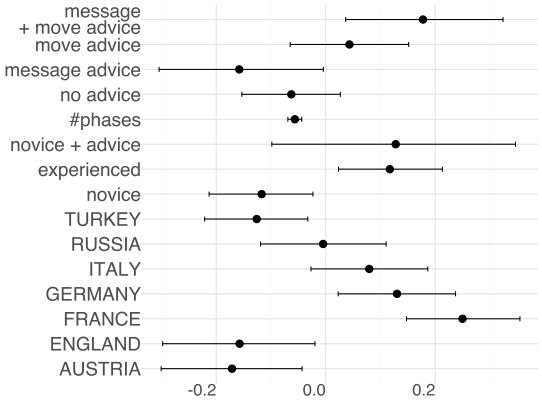

Figure 3: Regression coefficients for all features.

odd turn as gaining it, since the unit typically remains there in the next turn.

A.3 Experiment Procedures

The procedures involve playing *Diplomacy* with other participants as well as AI agents and providing feedback throughout the game on 1) the truth value of players' messages 2) the predicted truth values of theirr opponents' messages 3) the perceived friendliness of opponents 4) if players think they have been lied to in the previous phase. Players are asked to participate by placing their game moves in each turn before communicating. Some players in some turns received strategic moves and message advice from PHOLUS. Games end after 3 hours or less. After each game, participants are asked to fill out a survey asking about their experience and the AI advice, as well as how AI agents performed during the game.

A.4 Participant Compensation

Players received a $70 gift card of their choice for each game. In addition, up to $29 was rewarded based on performance in the game.

A.5 Survey Questions

We use Google Forms to conduct player surveys. Table 2 shows the survey questions.

A.6 PHOLUS's message advice and human messages with SMATCH scores

We provide additional examples of cases where human players reject PHOLUS's message advice, together with SMATCH. We show some advice (Figure 6 and 7) that experienced players mostly agree with, however, they partially edit to make it more aligned with their goals. We also show advice that is not aligned with the player's goals, for example in Figure 8.

```
Sender: Italy (veteran)
Recipient: England
Message Advice: absolutely. Have you and
Germany and France decided on a plan? That will
influence my opening quite a bit.
Human Message: absolutely. Have you and Germany
and France decided on a plan? That will influence
my opening quite a bit.
smatch: 1.0
```

Figure 4: Human uses PHOLUS's advice directly without modification.

```
Sender: Italy (novice)
Recipient: Turkey
Message Advice: hey can you move to Con to help
the game?
Human Message: Hey Turkey, could you move to
CON? Thanks!
smatch: 0.7895
```

Figure 5: Italy paraphrases the advice with the same underlying intention.

```
Sender: England (veteran)
Recipient: Russia
Message Advice: I might convoy an army over to
get it but I won't move into the Barents
Human Message: I'm convoying an army into
Norway, but it's sincerely to preserve navies for
the fight that France clearly wants
smatch: 0.5833
```

Figure 6: The advice and human message share a general intent (reassure Russia), but differ slightly.

```
Sender: Germany (veteran)
Recipient: Italy
Message Advice: I can't, gotta keep France out
of Bur
Human Message: I'm sorry I need that army to
bounce in Bur. I can help next turn though
smatch: 0.5641
```

Figure 7: The advice and human message point out the same goal that rejects the other player's proposal, instead using their unit to bounce in Bur.

```
Sender: Italy (novice)
Recipient: Turkey
Message Advice: Russia is going to attack you!
Human Message: I'm going to hold Mars
smatch: 0.5532
```

Figure 8: PHOLUS wants to warn Turkey regarding Russia's attack, however, the player disregards that and instead, shares their plan.

Question	Type
I am really good at Diplomacy.	Likert scale
How many Diplomacy games have you played before?	Free text
I am able to identify all bots.	Likert scale
I enjoy talking with the AIs.	Likert scale
I was able to make plans with other players in the game.	Likert scale
I was able to make plans with the AIs in the game.	Likert scale
human players communicated transparently.	Likert scale
AI players communicated transparently.	Likert scale
Do you have a link to your diplomacy stats?	Free text
How helpful is the move advice?	Likert scale
Thoughts on the move advice?	Free text
How helpful is the message advice?	Likert scale
Thoughts on the message advice?	Free text
Feel free to let us know your other post-game thoughts.	Free text

Table 2: List of questions in the survey.

```
Sender: Italy (novice)
Recipient: England
Message Advice: Are you going to take Belgium?
Human Message: Sure, let's get rid of france
smatch: 0.4
```

Figure 9: An example of low SMATCH. PHOLUS advises Italy to inquire about a specific game move, but the player discusses a high-level game plan.

A.7 Survey Summarization

Many participants find PHOLUS's move advice helpful. Experienced players observe that the advice often aligns with their own moves and offers "*some interesting ideas.*" However, the advice can be suboptimal and short-sighted in complex scenarios, where it fails to consider interaction between allies. Players find message advice useful for simple, quick communication but inadequate for more complex or specific situations, especially when it does not align with their strategies or alliances. Players mention that the messages include common communication terms, and they "*regret not using this feature more.*" However, the advice is less helpful for specific planning and often does not align with player alliances and intentions.