A Hybrid Human-AI Approach for Argument Map Creation From Transcripts

Lucas Anastasiou, Anna De Liddo
Knowledge Media Institute, The Open University
{lucas.anastasiou, anna.deliddo}@open.ac.uk

Abstract

In order to overcome challenges of traditional deliberation approaches that often silo information exchange between synchronous and asynchronous modes therefore hindering effective deliberation, we present a hybrid framework combining Large Language Models (LLMs) and human-in-the-loop curation to generate argument maps from deliberation transcripts. This approach aims to enhance the efficiency and quality of the generated argument maps, promote transparency, and connect the asynchronous and synchronous deliberation modes. Finally, we outline a realistic deliberation scenario where this process can be successfully integrated.

1. Introduction

Deliberation processes are important mechanisms for collaborative decision-making, fostering informed choices across a wide array of domains (Vaculín et al., 2013; Owen, 2015). Traditionally, these processes occurred through either synchronous (in-person or real-time online) discussions or asynchronous (such as online discussion forums) (Wright and Street, 2007). However, the distinction to synchronous and asynchronous consists of a siloed approach to deliberation that creates barriers to information exchange, development of shared understanding and subsequently consensus building and other elements that consist of effective deliberation (Friess and Eilders, 2015).

Recent advancements in Natural Language Processing (NLP) and particularly in Large Language Models (LLMs) have created promising paths to structure and synthesise information such as unstructured dialogue, i.e. free-flowing conversation (e.g. transcripts of meetings, online chat conversations) or semi-structured data (e.g. interviews, XML documents, and others) (Naveed et al., 2023; Serban et al., 2016). They possess the potential to generate structured discourse data (e.g. argument graphs or key points) (Chen et al., 2023). This may be the unblocker to overcome some of the challenges associated with traditional deliberative processes. Nonetheless, despite their impressive performance, LLMs are not without limitations; they are still susceptible to misinterpretation (Turpin et al., 2024), hallucinations (Ye et al., 2023), inaccuracies (Guo et al., 2023), therefore making them unreliable to be used in sensitive applications (like public deliberation that has significant impact in decision making).

To address these shortcomings, we propose an approach involving a human-in-the-loop (HITL) model ((Zanzotto, 2019)) to curate and filter LLM-generated outputs before integrating them into asynchronous debate platforms. This paper explores the potential of this hybrid framework to bridge the gap between synchronous and asynchronous deliberation modes, promoting accountability, transparency and more accurate and informed decision making.

2. Related work

2.1. Imperfect AI

Despite rapid advancements in the field, AI systems remain imperfect and likely will continue to be for the foreseeable future. Errors can arise from biases embedded in training data, limitations of the algorithms themselves, or unpredictable real-world inputs (Suresh and Guttag, 2019; Mehrabi et al., 2021). Furthermore, the “black-box” nature of many deep learning models hampers explainability, obscuring the logic behind potentially erroneous outputs (Samek et al., 2019). This persisting imperfection highlights the critical need for human oversight and intervention, especially in high-stakes domains. Recent work by Bussone et al. (2015) demonstrates how faulty AI-generated explanations can even worsen the situation, leading to unwarranted trust and potentially harmful decisions. Therefore, in contexts where safety, accuracy, and fairness are paramount, human-in-the-loop approaches remain essential for mitigating risks, ensuring ethical outcomes, and ultimately fostering responsible AI integration (Lee et al., 2020; Leslie, 2019).

2.2. Human-AI collaboration

Human-AI is focusing on the effective integration of human intelligence with the power of AI. Such collaboration holds the potential to surpass the limitations of either humans or AI working alone (Wilson and Daugherty, 2018; Passi and Vorvoreanu, 2022).

The First Workshop on Language-driven Deliberation Technology (DELITE2024), pages 45–51

To achieve various levels of collaboration, workflows such as human-in-the-loop (HITL), where AI provides assistance with humans retaining decision authority, and human-on-the-loop (HOTL), focusing on constant human oversight, have seen extensive exploration (Liu et al., 2014). Additionally, recent studies advocate for a human-in-command approach (Wesche and Sonderegger, 2019; Bostrom and Yudkowsky, 2018) stressing the necessity of maintaining ultimate human control in critical applications.

2.3. Argument mining using LLMs

Argument mining, the task of identifying and extracting argumentative structures from text (Cabrio and Villata, 2018; Lawrence and Reed, 2020), has seen significant advancements with Large Language Models (LLMs), such as OpenAI's GPT-3 (Brown et al., 2020), GPT-4 (OpenAI, 2023), Google's Gemini (Team et al., 2023), Anthropic's Claude (Azzollini and Pomponio, 2019) and others. LLMs' ability to understand and generate complex language enables more nuanced argument extraction (Kashefi et al., 2023), offering the potential to improve argument component identification, relationship classification, and even argumentative summarisation (Reimers et al., 2019; Lauscher et al., 2022; Elaraby and Litman, 2022). This opens up opportunities for automated analysis of large-scale debates, supporting decision-making, and facilitating critical thinking. However, challenges remain. LLMs can conflate correlation with causation, leading to the identification of spurious arguments (Jin et al., 2023). Additionally, biases inherent in the LLM's training data can propagate into argument identification (Acerbi and Stubbersfield, 2023). Despite these limitations, LLM-based argument mining holds significant promise for understanding and structuring complex discourse.

3. Proposed Method

We propose a curated method for argument map creation from conversational data (specifically from transcripts of informal or formal meetings) that prioritises both accuracy and automation, combining the capabilities of computational tools and the critical reasoning by humans. This hybrid Human-AI approach involves:

- *Initial AI Processing*: We utilise LLM prompting to mine arguments (identify argumentative components) from the transcript of conversations. The transcripts consist of written records of what was said in a meeting, speech, interview or any other spoken event; in our case we use video captions (.srt files) that is easily accessible (though not ideal as there is no speaker identification). We build the argument map using the simplified IBIS model ((Kunz and Rittel, 1970)), i.e. organising arguments into positions and pro (supporting) or con (opposing) arguments. An illustrative method for extracting arguments from textual transcripts using Large Language Models (LLMs) to the Issue-Based Information System (IBIS) argumentation scheme is shown in Prompt 1. Note that to facilitate transparency and provenance, we emphasize the inclusion of original transcript snippets alongside generated arguments.

- *Human Annotation and Curation*: At this stage the generated argument map is presented to a human curator where they annotate each argument node across several evaluation dimensions inspired by Argument Mining evaluation frameworks (e.g. Sofi et al. (2022)) such as *Groundedness* (Levonian et al. (2023) - whether the argument generated is based on the input text), *Context Relevance* (whether it draws from the surrounding text - it relates to the connected argument) and others. Such annotation process can be logged using modern software such as trulens[1]. Human curators are enabled to confirm the inclusion of each argument node, edit the content of it or change the connection links to each. To facilitate this process we use several visual assistance aids that we explain further in Section 3.2. The curated versions of the argument maps are later used to as ground truth examples to finetune the LLM used in the initial AI processing stage.

- *Semantically connect and merge with other argument maps*: At this stage we proceed to import into the curated argument map into an established database of argument maps/debates. We identify similar arguments by comparing the semantic similarity of the argument nodes (using e.g. argueBERT (Behrendt and Harmeling, 2021)). We proceed to merge the similar arguments following again a curation workflow (asking humans to select whether to combine the two arguments by generating via LLM a summary of the two or just denote explicitly the similarity of both but keep separated)

- *Key-Point analysis and summarisation*: Upon creating the final argument map, we proceed to create a summarised view, i.e. automatically extracting the core arguments or essential messages from the collection of arguments (using key point analysis (Bar-Haim et al., 2020)).

[1] https://www.trulens.org/

Prompt 1 Extract key positions and argument from transcript

Below is a transcript from a debate in the european parliament:
–
{{ TRANSCRIPT TEXT FROM SRT FILE }}
–
What are the main positions and arguments for and against given in the above? Provide those in a bulleted list like:
- Position N: <position_text>
– Arguments supporting Position N (pro arguments):
— <argument text N.p.i>
– Arguments against Position N (con arguments):
— <argument text N.c.j>
Do not include supporting or opposing arguments if they do not exist. Make sure you include only arguments or positions that appear in the given text. To make sure that this is the case, on each argument or position include the timestamp that this is mentioned in the given text

3.1. Example

We present here the output of the application of such prompting in a sample taken from “Economic Dialogue with Christine Lagarde”[2] in the European Parliament in Figure 1. Our analysis revealed significant variation in the outputs generated by the different models used. As expected, GPT3.5, exhibits the weakest performance, producing a comparatively simplistic representation of the arguments presented in the example transcript. Interestingly, the outputs from the two more proficient models (GPT4 and Gemini Advanced) displayed distinct characteristics. It is noteworthy that Gemini Advanced deviated from the instructed format and fully omitted any counter-arguments (con arguments) from its representation.

3.2. Curation workflow and interface

The output of the initial AI processing, while demonstrating promising accuracy, cannot guarantee perfect results. Therefore, we propose a following human-in-the-loop curation workflow that incorporates several design elements:

1. *Confidence Indicators*: We display confidence levels derived from the probability of each generated token, presented visually as a highlight with red color in case of low logit probability.

   As an extension, we propose calculating the semantic distance between the generated argument component and the original transcript snippet, serving as a proxy for confidence, and display this confidence score on each generated argument node.

2. *Alternative Outputs*: The workflow provides a selection of alternative outputs for curators to choose from, enhancing decision-making and refining the extracted arguments. Alternative outputs can come from invocation of different models (as shown in our example, outputs significantly vary according to the used model), use of different prompts or chunking strategies.

3. *Editable Argument Tree with Drag-and-Drop Functionality*: Users can directly manipulate the generated argument tree. Transcript chunks or individual arguments can be repositioned within the tree, allowing for intuitive restructuring and refinement.

4. *Approval and Confidence Assessment*: Each argument node offers either a simple approval/rejection mechanism (checkmark/cross x mark) or a granular confidence slider (0-100%). In case of multiple annotators, we can follow a simple majority agreement to ensure reliabel results.

Crucially, for points 3 and 4, the system retains a historical log of actions (approvals, changes, dismissals). This historical record is a crucial element for ensuring transparency, accountability, and ultimately, trust in the quality of the final curated argument structure - as it allows the reconstruction of decision-making processes and providing rich explanations for the final curated argument structure.

3.3. Evaluation design

For the systematic evaluation of the workflow described in Section 3.2 we consider the following hypotheses:
H1: Different LLM models impact the quality of generated argument maps
H2: Human-in-the-loop curated argument maps exhibit an improvement in quality compared to argument maps generated solely by automated methods

Regarding dataset, we shall focus on transcripts for argument mining tasks that are deliberation contextualised (e.g. M-Arg Mestre et al. (2021)). Both H1 and H2 can be tested by asking domain experts (human annotators) to evaluate the quality and coherence of argument extractions from each LLM candidate and with or without human curation or/and with the use of computational metrics (accuracy, F1) against the reference dataset. This can be

[2]https://multimedia.europarl.europa.eu/en/video/economic-dialogue-with-christine-lagarde-ecb-president-opening-statements-by-irene-tinagli-sd-it-chair-of-econ-and-introductory-remark-by-christine-lagarde-president-of-the-european-central-bank-ecb_I252842

(i) GPT3.5

Position 1: Concern about inflation levels
– Arguments supporting Position 1 (pro arguments):
— Inflation fell to 2.4% in November 2023, reaching its lowest level since July 2021 (Chapter 1)
— Inflation increased again to 2.9% in December 2023 (Chapter 1)
— The recent increase was mainly due to upward base effects for the cost of energy (Chapter 2)
— Inflation levels are considerably lower compared to the average annual inflation in 2023 (Chapter 2)
— ECB is determined to ensure that inflation returns to its 2% medium-term target (Chapter 2)
Position 2: Impact of ECB's monetary policy on inflation
– Arguments supporting Position 2 (pro arguments):
— Tight financing conditions are dampening demand and helping to push down inflation (Chapter 3)
— ECB's interest rate increases are being transmitted forcefully into financing conditions (Chapter 3)
— ECB's forward-looking wage structure signals strong wage pressures, but agreements suggest some levelling (Chapter 11)
— Wage pressure for 2024 hinges particularly on ongoing and upcoming negotiations (Chapter 11)
— Contribution of unit profits to domestic price pressures continued to decline, suggesting that wage increases are buffered by profit margins (Chapter 12)

(ii) GPT4

Position 1: Inflation monitoring and response by the ECB
Arguments supporting Position 1:
— Inflation fell to 2.4% in November and increased to 2.9% in December, showing fluctuating but controlled levels (00:00:02.000).
— After ten rate hikes, the ECB left the key interest rate unchanged for the third consecutive time, showing a determined stance to ensure inflation returns to the 2% medium-term target (00:00:48.000).
— Tight financing conditions are helping to push down inflation, indicating the effectiveness of ECB policies (00:01:45.000).
Position 2: The euro's significance and performance over 25 years
Arguments supporting Position 2:
— The euro area has grown from 11 to 20 countries, demonstrating the currency's expanding influence and integration within Europe (00:02:16.000).
— Support for the euro among citizens is close to record levels, reflecting its acceptance and perceived benefits (00:02:16.000).
— The ECB's monetary policy has been instrumental in navigating financial and COVID crises, underscoring the central bank's role in maintaining economic stability (00:03:16.000).

(iii) Gemini Advanced

Position 1: The euro has been a success story, delivering simplicity, stability, and sovereignty to its member states.
Arguments supporting Position 1 (pro arguments):
Simplicity: Eliminates currency exchange hassle within the eurozone, facilitating trade and travel. (Chapters 14, 16)
Stability: Relatively low average inflation (2.1% since 1999) and anchored long-term inflation expectations signal ECB success in maintaining price stability. (Chapter 16)
Sovereignty: As a major global currency, the euro insulates eurozone economies from international currency fluctuations. (Chapter 17)
Position 2: The European Central Bank (ECB) is committed to ensuring the euro's continued success.
Arguments supporting Position 2 (pro arguments):
Mandate focus: The ECB has responded forcefully to inflation surges, demonstrating a commitment to price stability. (Chapter 18)
Climate adaptation: The ECB is incorporating climate change considerations into monetary policy operations. (Chapter 19)
Digital innovation: The ECB's ongoing work on the digital euro shows a willingness to adapt to technological shifts. (Chapter 20)

Figure 1: Example of argument map generations from transcript of Economic Dialogue with Christine Lagarde in European Parliament. Variations of output using (i) GPT3.5, (ii) GPT4 and (iii) Gemini Advanced. The presented is a fraction of the generated output due to space constraints

achieved through surveys in an A/B experimental setup.

4. Use case

Further to the systematic evaluation, we envision to incorporate the above described method into a deliberation scenario where a policy organisation utilises the LISTEN-REFRAME-ACT (L-R-A)[3] method to broader citizen and expert engagement on public policy issues. The L-R-A method is a structured apporach to public deliberation that: The *LISTEN* phase emphasizes on deep understanding of the diverse perspectives surrounding an issue. In the *REFRAME* phase, based on the insights from the LISTEN phase, the participants focus on reframing the issue collaboratively, developing more inclusive, evidence-based narratives and exploring potential solutions. In the last *ACT* phase, the reframed understanding and ideas are transformed into actionable proposals.

Traditionally all of the above phases are carried out in physical settings. The proposed transcript-to-argument-graph conversion method offers a powerful solution. It enables importing *LISTENING* phase insights directly into an online deliberation platform. By systematically analyzing transcripts, extracting key arguments, and incorporating LLM-assisted refinement, this method enables the successful transition from unstructured discussions to argumentative structure discussion. The generated argument maps can be used to seed further focused online discussions, providing a grounded starting point for the *REFRAME* and *ACT* phases. This integration ensures that the valuable insights from the *LISTENING* phase are effectively carried forward into the online deliberation, enhancing the process's richness and inclusivity.

5. Conclusions

This paper has presented an approach for argument map creation from transcript text that offers a synergistic approach, combining the efficiency of computational automation with the depth of human critical thinking, therefore getting results superior to either in isolation. Our method empowers untrained users to effectively construct argument maps, addressing a known challenge highlighted

[3]https://www.linkedin.com/pulse/future4citizens-barcelona-european-capital-democracy-xxgge/

in prior research (e.g. Le et al., 2013). Crucially, our approach maintains human control throughout the process, ensuring transparency and accountability in the resulting argument map. This fosters trust between users and the generated outcomes. Moreover, this method has the potential to facilitate the fluid exchange from synchronous to asynchronous deliberation modes. Future development could explore the integration of chain-of-thought (Wei et al., 2022) or tree of thoughts (Yao et al., 2024) reasoning for improving the performance of the AI pre-processing and also mitigate dependence on prompt engineering. Importantly, while the method seeks to mitigate individual subjectivity through majority agreement, human annotation of what consists argument or position remains inherently susceptible to personal perspectives. Finally, the scope of this work did not include the addition of arguments into an existing knowledge base, leaving room for exploration into how the approach can support the evolution of established argument maps. Future work will focus on implementing and extending this approach in real large-scale deliberation scenarios.

6. Acknowledgements

This research was funded in collaboration by UKRI under the UK Government's Horizon Europe Guarantee scheme (Reference Number: 10048874) and by the European Commission under the Horizon Europe Programme, in the context of the ORBIS Project (GA: 101094765) on "Augmenting participation, co-creation, trust and transparency in Deliberative Democracy at all scales".

7. Bibliographical References

Alberto Acerbi and Joseph M Stubbersfield. 2023. Large language models show human-like content biases in transmission chain experiments. *Proceedings of the National Academy of Sciences*, 120(44):e2313790120.

Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *Ai Magazine*, 35(4):105–120.

Antonio Azzollini and Alessio Pomponio. 2019. Positive energy static solutions for the chern-simons-schrödinger system under a large-distance fall-off requirement on the gauge potentials.

Roy Bar-Haim, Lilach Eden, Roni Friedman, Yoav Kantor, Dan Lahav, and Noam Slonim. 2020. From arguments to key points: Towards automatic argument summarization. *arXiv preprint arXiv:2005.01619*.

Maike Behrendt and Stefan Harmeling. 2021. Arguebert: How to improve bert embeddings for measuring the similarity of arguments. In *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*, pages 28–36.

Nick Bostrom and Eliezer Yudkowsky. 2018. The ethics of artificial intelligence. In *Artificial intelligence safety and security*, pages 57–69. Chapman and Hall/CRC.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The role of explanations on trust and reliance in clinical decision support systems. In *2015 international conference on healthcare informatics*, pages 160–169. IEEE.

Elena Cabrio and Serena Villata. 2018. Five years of argument mining: A data-driven analysis. In *IJCAI*, volume 18, pages 5427–5433.

Guizhen Chen, Liying Cheng, Luu Anh Tuan, and Lidong Bing. 2023. Exploring the potential of large language models in computational argumentation. *arXiv preprint arXiv:2311.09022*.

Mohamed Elaraby and Diane Litman. 2022. Arglegalsumm: Improving abstractive summarization of legal documents with argument mining. *arXiv preprint arXiv:2209.01650*.

Dennis Friess and Christiane Eilders. 2015. A systematic review of online deliberation research. *Policy & Internet*, 7(3):319–339.

Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, Deyi Xiong, et al. 2023. Evaluating large language models: A comprehensive survey. *arXiv preprint arXiv:2310.19736*.

Zhijing Jin, Jiarui Liu, Zhiheng Lyu, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona Diab, and Bernhard Schölkopf. 2023. Can large language models infer causation from correlation? *arXiv preprint arXiv:2306.05836*.

Omid Kashefi, Sophia Chan, and Swapna Somasundaran. 2023. Argument detection in student essays under resource constraints. In *Proceedings of the 10th Workshop on Argument Mining*, pages 64–75.

W. Kunz and H.W.J. Rittel. 1970. *Issues as Elements of Information Systems*. Number no. 131 in California. University. Center for Planning and Development Research. Working paper, no. 131. Institute of Urban and Regional Development, University of California.

Anne Lauscher, Henning Wachsmuth, Iryna Gurevych, and Goran Glavaš. 2022. Scientia potentia est—on the role of knowledge in computational argumentation. *Transactions of the Association for Computational Linguistics*, 10:1392–1422.

John Lawrence and Chris Reed. 2020. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Nguyen-Thinh Le, Sabine Niebuhr, David Drexler, and Niels Pinkwart. 2013. Scaffolding computer supported argumentation processes through mini map based interaction techniques. In *Design, User Experience, and Usability. Web, Mobile, and Product Design: Second International Conference, DUXU 2013, Held as Part of HCI International 2013, Las Vegas, NV, USA, July 21-26, 2013, Proceedings, Part IV 2*, pages 533–542. Springer.

Min Kyung Lee, Nina Grgic-Hlaca, Michael Carl Tschantz, Reuben Binns, Adrian Weller, Michelle Carney, and Kori Inkpen. 2020. Human-centered approaches to fair and responsible ai. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*.

David Leslie. 2019. Understanding artificial intelligence ethics and safety. *ArXiv*, abs/1906.05684.

Zachary Levonian, Chenglu Li, Wangda Zhu, Anoushka Gade, Owen Henkel, Millie-Ellen Postle, and Wanli Xing. 2023. Retrieval-augmented generation to improve math question-answering: Trade-offs between groundedness and human preference. *arXiv preprint arXiv:2310.03184*.

Juan Liu, Aaron Wilson, and David Gunning. 2014. Workflow-based human-in-the-loop data analytics. In *Proceedings of the 2014 Workshop on Human Centered Big Data Research*, pages 49–52.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35.

Rafael Mestre, Razvan Milicin, Stuart E. Middleton, Matt Ryan, Jiatong Zhu, and Timothy J. Norman. 2021. M-arg: Multimodal argument mining dataset for political debates with audio and transcripts. In *Proceedings of the 8th Workshop on Argument Mining*, pages 78–88, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Nick Barnes, and Ajmal Mian. 2023. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*.

OpenAI. 2023. Gpt-4 technical report.

Daniel Owen. 2015. Collaborative decision making. *Decision Analysis*, 12(1):29–45.

Samir Passi and Mihaela Vorvoreanu. 2022. Overreliance on ai literature review. *Microsoft Research*.

Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2019. Classification and clustering of arguments with contextualized word embeddings. *arXiv preprint arXiv:1906.09821*.

Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller. 2019. *Explainable AI: interpreting, explaining and visualizing deep learning*, volume 11700. Springer Nature.

Iulian Vlad Serban, Ryan Lowe, Laurent Charlin, and Joelle Pineau. 2016. Generative deep neural networks for dialogue: A short review. *arXiv preprint arXiv:1611.06216*.

Mehmet Sofi, Matteo Fortier, and Oana Cocarascu. 2022. A robustness evaluation framework for argument mining. In *Proceedings of the 9th Workshop on Argument Mining*, pages 171–180.

Harini Suresh and John V Guttag. 2019. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*, 2(8).

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2024. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36.

Roman Vaculín, Richard Hull, Maja Vukovic, Terry Heath, Nathaniel Mills, and Yutian Sun. 2013. Supporting collaborative decision processes. In *2013 IEEE International Conference on Services Computing*, pages 651–658. IEEE.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Jenny S Wesche and Andreas Sonderegger. 2019. When computers take the lead: The automation of leadership. *Computers in human Behavior*, 101:197–209.

H James Wilson and Paul R Daugherty. 2018. Collaborative intelligence: Humans and ai are joining forces. *Harvard Business Review*, 96(4):114–123.

Scott Wright and John Street. 2007. Democracy, deliberation and design: the case of online discussion forums. *New media & society*, 9(5):849–869.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.

Hongbin Ye, Tong Liu, Aijia Zhang, Wei Hua, and Weiqiang Jia. 2023. Cognitive mirage: A review of hallucinations in large language models. *arXiv preprint arXiv:2309.06794*.

Fabio Massimo Zanzotto. 2019. Human-in-the-loop artificial intelligence. *Journal of Artificial Intelligence Research*, 64:243–252.

Quanshi Zhang, Ying Nian Wu, and Song-Chun Zhu. 2018. Interpretable convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8827–8836.