Aporia: The Argumentation Game *

Zaid Marji[1] and John Licato[1]

Advancing Machine and Human Reasoning (AMHR) Lab
Department of Computer Science and Engineering
University of South Florida, Tampa FL 33620, USA

Understanding rules that contain open-textured phrases involves *interpretive arguments* [4, 3, 2] which are used to support or attack an interpretation of a fixed expression within a fixed document. Formally, interpretive arguments are of the form: "If expression E occurs in document $\mathbf{D}$, E has a setting of S, and E would fit this setting of S by having interpretation $\mathcal{I}$, then E ought to be interpreted as $\mathcal{I}$" [4]. For example, one might argue that a term t should be interpreted a certain way because the typical person would understand t as having a certain definition (argument from ordinary meaning), or even that previous interpretations of t are binding (argument from precedent). A closely related task, automated compliance detection, is the task of determining whether some behaviors can be interpreted as compliant with a set of rules. Determining such compliance requires interpretive argumentation; particularly when the rules contain *open-textured* terms [1]. At present, no artificially intelligent algorithms can perform either interpretive reasoning or automated compliance detection to a substantial degree. These limitations might be partially explained by the lack of resources (e.g., large-scale datasets and test environments) available to develop such agents.

Here we describe our preliminary progress towards addressing these limitations, via a game called 'Aporia: The Argumentation Game'. This game can be played by any group of three or more people. We recommend it to be played with 6 players. The game is played in rounds. At the end of each round points will be awarded to the winner of that round. In each round, two players are randomly chosen to play against each other, with a third player designated as a judge. Each round starts with a tuple of a specific profession, a rule that members of that profession are expected to follow, and a scenario describing an action taken by a member of that profession. The rules provided are actual rules taken from codes of ethics of organizations – usually professional associations. The scenarios have been generated in a study that we previously conducted [2], such that the actions are unclear as to whether they violate the rule. The two players will argue for or against the view that the scenario violates the rule.

The rounds will be presented in four phases. All phases have time limits to ensure the game proceeds in an orderly fashion. The first phase is the *reading*

* This material is based upon work supported by the Air Force Office of Scientific Research under award numbers FA9550-17-1-0191 and FA9550-18-1-0052. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the United States Air Force.

phase where a rule and an associated scenario are presented. The players will be given a specified amount of time to read the scenario and think about their responses. Next is the *first argument phase* where the first player will declare whether they will argue for or against the professional's actions, and then present their argument within the allotted time. Next is the *second argument phase* where the second player will have to present counter arguments within their allotted time in order to cast doubt on the validity of the first player's arguments. The fourth and final phase is the *judging phase* where the judge will need to decide who the winner is. The judge will need to provide an explanation for their choice and announce the winner also within the allotted time. Once a winner is declared, the winner gets awarded a point, the round ends, and a new round will begin. A new set of players will be selected for the new round.

Aporia is intended to be used in multiple ways, which will benefit the study of interpretive argumentation and its automation. First, the arguments given by players can later be given to participants who will serve as judges in an experimental environment. This allows us to study the degree to which various factors affect the evaluation of interpretive arguments, as we can experimentally manipulate superficial factors such as argument length, gender of the arguers, etc. Secondly, the record of arguments and judgments made by players can comprise a dataset that can then be used to train and test artificially intelligent agents. These agents, we hope, will eventually play the game against human players, thus setting the stage for the development and testing of automated interpretive reasoners.

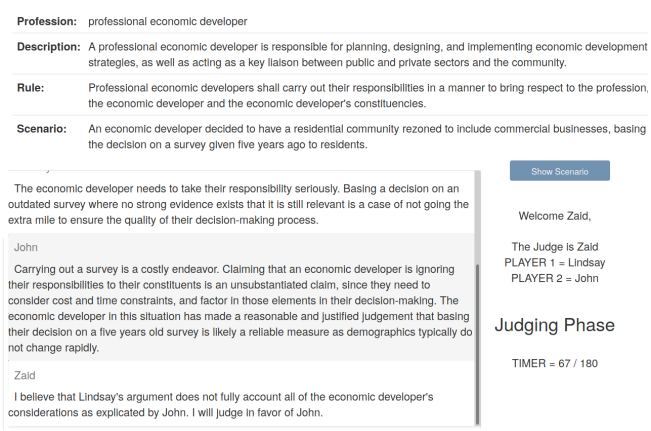

Fig. 1: Example of Aporia after a round is completed.

Using the aforementioned methodology we ran 48 game rounds, collecting tuples of arguments, judgements, and explanations.[1] We had 125 scenarios to select from. The scenario selection was made randomly. We have collected data from 41 unique scenarios. We had five scenarios with more than one entry; two of which have 3 entries each, while the other three have 2 entries each. The scenarios that were played repeatedly show the power of this method to collect interesting datasets, since we find different players when given the choice to choose whether to argue for or against will choose differently and we also find that the winners of the arguments can be on either side of the issue. An example of a scenario with

[1] The dataset can be found at https://github.com/Advancing-Machine-Human-Reasoning-Lab/aporia-dataset

multiple entries is shown below. Note how the two game rounds in that example show the the possibility of different sides of the issue winning the argument when presented using different arguments.

Rule: Professional economic developers shall maintain in confidence the affairs of any client, colleague or organization and shall not disclose confidential information obtained in the course of professional activities.
Scenario: The professional economic developer agreed to meet publicly with their client at a restaurant in the downtown business district.
Game 1:
- (Player 1 - Arguing against) Meeting in a public place to discuss confidential matters seems like a very bad error in judgement. Anyone within earshot can hear them discuss the details of plans that are supposed to be confidential.
- (Player 2) While I would normally agree that, yes, meeting in a public space does seem to be an error in judgement, the scenario does seem to imply that it is the client that wished to meet in public. It could be that the client themselves is more comfortable meeting with the economic developer in a public space and that is what the developer needs to do to assure the clients comfort. So long as both are in agreement, and nothing confidential is necessarily discussed, I see no issue with maintaining the clients comfort level.
- (Judge - Player 2 wins) I think Player 2's counter argument is convincing. The subject of the meeting may not be confidential and the client asked for the setting.

Game 2:
- (Player 1 - Arguing for) The developer "agreed" to meet the client, strongly suggesting the client made the offer to meet in public. I didn't think this is a violation in any case, but the fact that the client themselves are comfortable in a public setting makes me more comfortable judging the developer to be innocent of any impropriety.
- (Player 2) The economic developer should know that conversations in a downtown restaurant can be overheard. Any comments the developer makes during the conversation will not be confidential. Regardless of the client's naivety, he should also protect the client from public disclosure.
- (Judge - Player 2 wins) I agree with Player 2 that the developer should have taken the client's naivety into consideration and suggested a less public venue for the meeting.

Table 1 shows some general statistics from the dataset. We have also tested some hypotheses on the dataset in order to find out if there are any surface-level features of arguments that correlate with their persuasiveness. We analyzed two features:

- The number of words in each argument.
- The percentage of overlap between the most similar substrings of the argument and the rule that is being argued about. [2]

We found no correlations between word count in an argument or substring overlap percentage with the rule at hand and winning the game. First, we set the

[2] Using Python's TheFuzz library: fuzz.partial_ratio(argument,rule)

Label	Count	Label	Mean	StdDev
Player 1 wins	31	**Player 1 words**	47.6	15.7
Player 2 wins	17	**Player 2 words**	51.6	22.8
For-argument wins	31	**Player 1 overlap**	51.1	10.7
Against-argument wins	17	**Player 2 overlap**	54.3	16.5
Player 1 argues for	30	**Winner words**	47.7	17.8
Player 1 argues against	18	**Loser words**	51.4	21.2
Player 1 argues for and wins	22	**Winner overlap**	51.4	11.8
Player 1 argues against and wins	9	**Loser overlap**	53.9	15.7
Player 2 argues for and wins	9	**Words difference** [a]	-3.7	26.7
Player 2 argues against and wins	8	**Overlap difference** [a]	-2.5	19.4

[a] Differences are the values for the winner minus the values for the loser

Table 1: Some statistics of the collected dataset

independent variable as Player 1's word count minus Player 2's word count, and the dependent variable as Player 1 winning the game. We calculate the point-biserial correlation and get -0.10 with p-value 0.5. We do the same with the substring overlap percentage and get -0.09 with p-value 0.55.

We plan to execute more of those experiments to gather more data of this kind. Moreover, future work also includes presenting the same arguments to different judges to get a more reliable view of how each pair of arguments fair across different annotators. This will help get more reliable statistics on the properties of argumentation, as well as, provide larger datasets to support future development of argumentation-related artificial intelligence systems.

References

1. Hart, H.: The Concept of Law. Clarendon Press (1961)
2. Licato, J., Marji, Z., Abraham, S.: Scenarios and recommendations for ethical interpretive AI. In: Proceedings of the AAAI 2019 Fall Symposium on Human-Centered AI. Arlington, VA (2019)
3. Rotolo, A., Governatori, G., Sartor, G.: Deontic defeasible reasoning in legal interpretation: Two options for modelling interpretive arguments. In: Proceedings of the 15th International Conference on Artificial Intelligence and Law. pp. 99–108. ICAIL '15, ACM, New York, NY, USA (2015). https://doi.org/10.1145/2746090.2746100, http://doi.acm.org/10.1145/2746090.2746100
4. Sartor, G., Walton, D., Macagno, F., Rotolo, A.: Argumentation schemes for statutory interpretation: A logical analysis. In: Legal Knowledge and Information Systems. (Proceedings of JURIX 14). pp. 21–28 (2014)