Evaluating relevance in analogical arguments through warrant-based reasoning

JOHN LICATO
Department of Computer Science and Engineering / Advancing Machine and Human Reasoning (AMHR) Lab / University of South Florida, USA
licato@usf.edu

MICHAEL COOPER
Advancing Machine and Human Reasoning (AMHR) Lab / Department of Philosophy / University of South Florida, USA
michaelcoop@usf.edu

Arguments by analogy are particularly difficult to teach, assess, and implement computationally, in part because of the requirement of relevance. Our goal in this paper is an algorithm for assessing arguments by analogy, which: (1) lends itself to computational implementation using currently available tools, and (2) can be applicable to the kinds of arguments typically made by minimally trained arguers. We describe such an algorithm, through what we call warrant-based reasoning.

KEYWORDS: warrants, analogies, analogical arguments, warrant game, WG-A

Arguments by analogy are particularly difficult to teach, assess, and analyze computationally, in part because of the requirement of relevance. For example, consider the argument "the sun is round, the sun is extremely hot, and a compact disc is round; therefore compact discs are extremely hot." Trivially, this is a poor analogical argument, since the property of being round is not directly relevant to its temperature.

Determining relevance, however, is rather difficult in practice. If a group of minimally trained participants are asked to assess a given analogical argument, the resulting argumentative dialogue will tend to contain many statements of questionable relevance. A trained moderator of such a dialogue might be able to manage this, by only allowing statements that are relevant to the analogical argument being

discussed. But such moderation can be labor-intensive, requires a high degree of training for the moderator, and has little to no guarantee that the moderator will act in accordance with norms of rationality (whichever those might be).

Nevertheless, it is worthwhile to explore whether there exists a method for identifying the strengths and weaknesses of analogical arguments, particularly one lending itself to computational implementation, so that it can be carried out by, or with the assistance of, an artificially intelligent system. Even if such a system might occasionally rely on human input, its potential benefits are tremendous. E.g., arguments by analogy are prevalent in online discussions, and automatically evaluating argumentation might enable discussions where bad argumentation is filtered out or critiqued for educational purposes.

In this paper, we propose a framework for the dialogical evaluation of analogical arguments whose method of ensuring relevant utterances is built-in. This is done through what we call warrant-based reasoning, a framework for assessing an informal argument's quality by evaluating the strongest warrants that can be found in its support. We have created a computer program which restricts the "moves" participants can make to those which focus on the common warrant shared by the source and target domains of the analogy. The program is called WG-A (for "Warrant Game - Analogy"), and we describe its details in Section 2.

Our long-term goal in this project is a method for assessing arguments by analogy satisfying two desiderata: (1) it lends itself to computational implementation using currently available tools, and (2) it can be used by minimally trained arguers, with a minimum of external moderation. WG-A has been played by undergraduate students with under an hour of training, and we ultimately hope it will be playable by artificially intelligent systems.

1. BACKGROUND

1.1 Arguments by Analogy

We take as our starting point Bartha's (Bartha, 2010) general schema for analogical arguments. An analogical mapping is a systematic, one-to-one correspondence between two groups of propositions: a source domain, and a target domain. On the basis of this mapping, an analogical argument concludes that some hypothetical proposition holds in the target domain. Borrowing terms from Keynes (Keynes, 1921), an analogical argument can be seen as consisting of four parts:

- Positive analogy (**P**) – Proposition groups P in the source domain and P^* in the target domain that correspond to "known similarities".
- Negative analogy (**N**) - Proposition groups $A,\neg B$ in the source domain and $\neg A^*,B^*$ in the target domain corresponding to "known differences" between the domains. For example, the facts "Earth has an atmosphere" / "Mars does not have an atmosphere" would be in A and $\neg A^*$, respectively.
- Neutral analogy (**O**) - A set of propositions in the source such that the truth values of analogous propositions in the target are not known, and vice versa.
- Hypothetical analogy (**Q**) - A single proposition Q known to hold in the source and a hypothetical proposition Q^* in the target whose truth value is not known but is the conclusion of the analogical argument.

An argument from analogy might thus be a claim of the following form: "It is prima facie plausible that Q^* holds in the target because of certain known (or accepted) similarities with the source domain, despite certain known (or accepted) differences" (Bartha, 2013). Conformance to this schema alone is insufficient to determine the quality of an analogical argument; Bartha's schema is meant to be entirely general, intended to represent both good and bad analogical arguments. Bartha's articulation model (Bartha, 2010) is based on the idea that a successful analogical argument is one which identifies a prior association and a potential for generalization:

- **Prior Association.** "There must be a clear connection, in the source domain, between the known similarities (the positive analogy) and the further similarity that is projected to hold in the target domain (the hypothetical analogy). This relationship determines which features of the source are critical to the analogical inference."
- **Potential for Generalization.** "There must be reason to think that the same kind of connection could obtain in the target domain. More pointedly: there must be no critical disanalogy between the domains" (Bartha, 2013).

The articulation model describes how the prior association and potential for generalization can be made explicit and assessed through a dialogue between an advocate and critic, whose goals are to defend and attack the analogical argument, respectively. Because such a dialogue is meant to reflect real-world dialogues which take place to assess analogical arguments, the standards for what constitutes an acceptable

prior association is dependent on the kind of vertical relations (i.e., the relations that hold between the elements in the source domain) being considered. Mathematical analogies may require such relations to be proof-theoretic, whereas for certain informal arguments, associations or weak causal relationships may suffice.

We take Bartha's work as a starting point and assume that a good analogical argument has a good prior association and potential for generalization.

1.2 Warrants

A warrant, in Stephen Toulmin's model of argumentation, is a statement connecting the premises[1] and conclusion of an argument, showing how the former permits the inference of the latter (Toulmin et al., 1984; Toulmin, 2003). Whereas premises may be facts, evidence, or pieces of data that support a conclusion, a warrant is typically a broad principle of reasoning which might range from truth-preserving inference rules drawn from formal, deductive models, to unreliable heuristic norms.

For example, given the premise "Socrates is a man" and the conclusion "Socrates is mortal," two possible warrants are W_1: "Anyone who is a man is also mortal," and W_2: "Typically, men are mortal." These two warrants differ in the degree to which they allow the premise to support the conclusion. They also differ in the ways they can be challenged: W_1 can be refuted with a single example of an immortal man; whereas W_2 requires data showing that a majority of men are, in fact, immortal. Given these differences in weak points, it behooves an arguer to ensure the strongest possible warrant is used for their arguments.

The warrant, when made explicit, makes it easier to determine key features typically associated with argument strength, not limited to: (1) what kind of attacks can be used against the argument, (2) whether the premises are relevant or necessary to the argument, and (3) whether, and with what strength, the conclusion follows from the premises. Furthermore, whether or not a warrant was used in the creation of an argument, the process of making a warrant explicit and evaluating its connection to the premises and conclusion is a highly useful exercise in the assessment of that argument. Despite this level of utility, the warrant is often left implicit. This difficulty has led researchers in AI and

---

[1] We are following (Hitchcock, 2005) in using the term 'premises' to refer to what others, including Toulmin, might call the 'data' or 'evidence,' to reflect the position that warrants should be distinguished from premises. We do not defend that position here, and instead refer the reader to (Hitchcock, 2005)

computational argumentation to omit warrants from their models and datasets (Besnard et al., 2014; Habernal et al., 2014), and educators to leave warrants out of their lesson plans (Lunsford et al., 2002; Rex et al., 2010; Harrell & Wetzel, 2015). It has been observed that this omission is to the detriment of automated reasoning in the former case, and to students in the latter (Warren, 2010; Beach et al., 2016).

We will collectively refer to the kinds of reasoning processes which create, improve, or otherwise evaluate arguments by focusing on their warrants and how those warrants connect to the other parts of the arguments as "warrant-based reasoning."

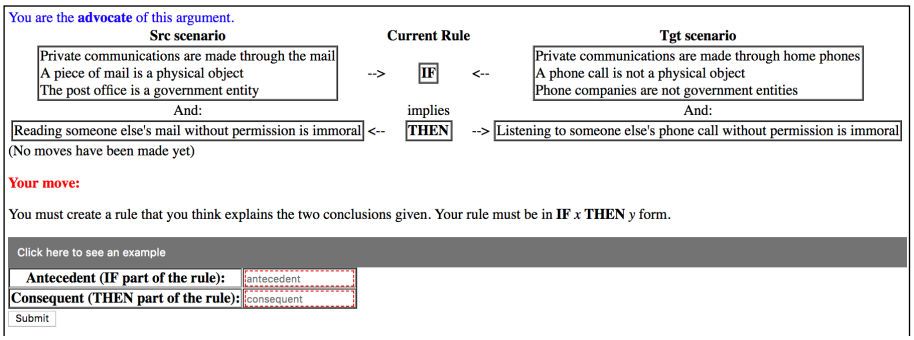

Figure 1: Starting screen, as viewed by the advocate

2. THE WARRANT GAME AND WG-A

Given the benefits of warrant-based reasoning, our research lab recently developed a classroom activity to introduce students of critical thinking to warrant-centered argumentation called "the Warrant Game" (WG). In WG, teams of students put forth opposing arguments. They must carefully phrase the warrants for their arguments, because warrants and their connections to the rest of the argument can be attacked by other teams using one of a predefined set of allowed attacks. If an attack is successful (as determined by a moderator), the attacking team gains points, whereas the attacked team loses points and has the opportunity to revise the wording of their warrant to prevent (or inadvertently open themselves up to) further attacks.

WG provides a model for how to create, and iteratively improve on, a warrant: First, create an initial warrant by joining the premises and conclusion in a conditional statement ("If [*premises*], then [*conclusion*]"). Second, determine whether the warrant is subject to any of the pre-determined allowed attacks. If so, revise the warrant so it will be more resistant to these attacks, and then iterate until the warrant is sufficiently strong (in WG, this tends to be limited by time considerations or the skill level of the players). Thus, the measurement

of argument strength used here is qualitative: an argument is considered strong if its components are resistant to relevant attacks. An argument's maximal warrant strength is determined by the strongest warrant that can be found for it, and the strength of that warrant in turn is determined by how resistant it is to the attacks that can be found against it. This qualitative notion of argument strength allows us to define a partial ordering between arguments: Given two arguments, if one is subject to a subset of the attacks that another one is, then the first is stronger. Maximal warrant strength is meant to maintain some compatibility with the approaches derived from argument acceptability semantics (Dung, 1995; Mogdil & Prakken, 2013; Besnard et al., 2014; Reed et al., 2017) and Walton's argumentation schemes (Walton, 1985; Walton, 1999; Walton et al., 2008).

The notion of maximal warrant strength has several strengths as a formalization of argument quality, with respect to our stated desiderata: (1) It can be assessed on the spot, without requiring one to wait to see if the argument's conclusion is correct or not; (2) it is a property of arguments as a whole, rather than of individual premises or conclusions; and (3) it encourages reasoners to focus more on the connection between premises and conclusions. However, it should be noted that we do not claim the maximal warrant strength standard is the only measure of argument quality, nor that it is the best measure in all circumstances. Rather, it is a standard that lends itself to our desiderata by providing a way for arguments to be assessed by automated reasoners.

The warrant game breaks down the task of warrant evaluation into simpler tasks, represented by the allowed attack types. For example, instead of detecting the gap between premises and conclusions (as in Boltužic and Šnajder (2016)), one allowed attack is to focus on the much smaller gap between premises and a warrant's antecedent. When explaining this attack type to students, we might ask, "is it reasonable for you to believe the premises but not the warrant's antecedent?" Although drawing on an intuition of what it means for an inferential leap to be "reasonable" is not yet fully achievable through AI, we suspect it might be approximable through natural language inference tools, the state-of-the-art of which is currently achieved by deep neural networks (Lai & Hockenmaier, 2017; Chen et al., 2016; Cheng et al., 2016; Rocktäschel et al., 2015); and for this reason, this approach to warrant evaluation is in line with our first desideratum.

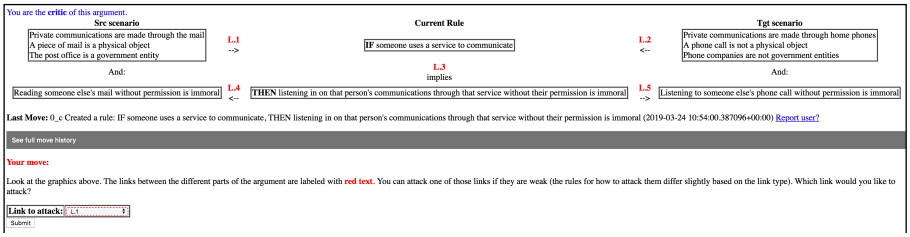

Figure 2: When deciding to attack, the critic is given a detailed image showing links in the argument which are open to attack.

2.1 Warrant Game – Analogy (WG-A)

Our underlying approach to combining Bartha's articulation model and WG is based on the supposition that given an analogical argument ***A***, the process of extracting a single warrant which applies to both the source and target domains of ***A***: (1) is a task which is accessible to many and does not require excessive training and study, and (2) will tend to elicit reasoning and moves which are relevant to the evaluation of ***A***. The resulting model based on, and designed to test, this supposition is called WG-A (Warrant Game for Analogies). Of course, there are quite a few functional differences between Bartha's prior generalizations and Toulmin's warrants, which affect the kinds of analogical arguments that A works best with. We describe some of these difficulties in Section 3. But for the most part, WG-A is a framework that helps us achieve the desiderata stated in the beginning of this paper.

WG-A is a Django-based[2] web app which allows for participants to productively engage in a dialogue assessing an analogical argument with minimal training. Each instance (or "game") involves two participants, who play the roles of advocate and critic and are each given a unique and customized URL. Prior to the game starting, a short video is shown to participants explaining the basic structure of an analogical argument, the concept of a warrant (referred to within WG-A as a 'rule'), and simple examples for how to attack a link between components of an argument. For the latter, we focus on links as if they were antecedents and consequents of a material implication and encourage participants to attack them by finding defeating counterexamples—examples where the antecedent holds, but the consequent does not.

At the beginning of the game, an analogical argument is first presented in the form of source facts ($P \cup A \cup \neg B$) a source hypothetical (Q), target facts ($P^* \cup \neg A^* \cup B^*$), and a target hypothetical (Q^*). Players are told that Q is to be considered established fact, and the goal of the advocate is to show that ***A*** supports Q^*, whereas the critic's goal is to

[2] For more info, visit https://www.djangoproject.com/

show that $\boldsymbol{A}$ doesn't support Q^*. The advocate begins by stating a candidate warrant which simultaneously explains the connection between the source facts and source hypothetical, and between the target facts and target hypothetical (Figure 1). A detailed example is available to the advocate at that point for further clarification on what is expected.

When the advocate completes their action, control reverts to the critic, and the move is recorded in a log that is always accessible to both participants. The critic receives a notification saying that it is their turn, and they are given the choice to either update the source / target facts, send an attack, or pass (passing is only an option after a certain number of moves have been made). When two passes are made consecutively, the game is terminated.

If a critic decides to attack, the five links which are possible to attack are labeled as in Figure 2. Note that there are no attackable links between the source domain's facts and its conclusion, and likewise for the target domain. This is in keeping with the guiding principles of warrant-based reasoning: attacks should be allowed only if they address a flaw in the warrant or the ways in which it connects to other parts of the argument.

When the critic selects one of the attackable links, the two linked argument components are displayed to the user, along with instructions for what constitutes a valid attack. These directions treat the two linked argument components almost as if they were the antecedent and consequent of a material inference. For example, consider the link summarized in Figure 3a. The critic is asked to explain how the rule's antecedent fails to lead to its consequent, and is given suggestions for how to do so, e.g.: show that the "logical leap" between them is too far, or describe an example where the antecedent holds but the consequent does not. In this case, the critic chose the latter, and Figure 3b shows the screen that is subsequently shown to the advocate.

The advocate then has a choice of either rejecting or accepting the attack. If the attack is rejected, a reason must be provided, and the advocate is encouraged to write a reason grounded in the instructions the critic was given when creating this attack. An attack rejection effectively ends that attack, but the critic can submit a similar attack later (indeed, they can do so directly after if necessary). On the other hand, if the advocate decides to accept the attack, they are rewarded with the opportunity to make another move. Though it is not required to, this additional move is meant to be used to modify the rule or facts in order to defend against similar attacks in the future.

Only the advocate can make edits to the rule, and such edits are not subject to approval by the critic. Modifications to the source or target facts, however, can be initiated by either the critic or advocate.

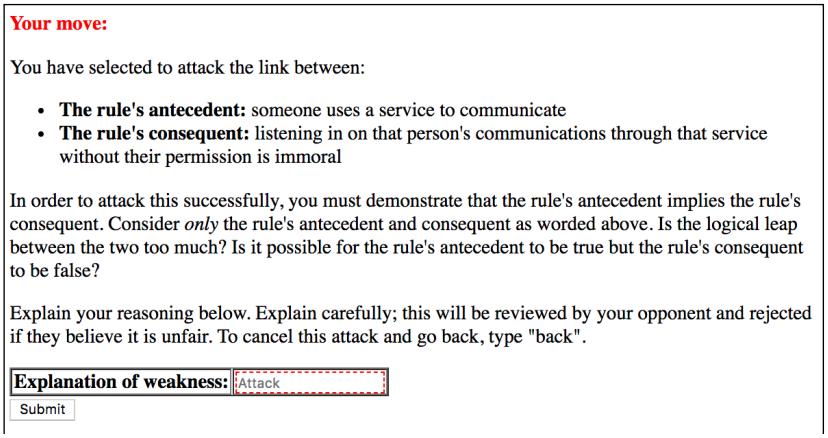

Figure 3a: The critic is provided an easy-to-read explanation of how to justify their attack and asked to elaborate on the reasoning behind their attack.

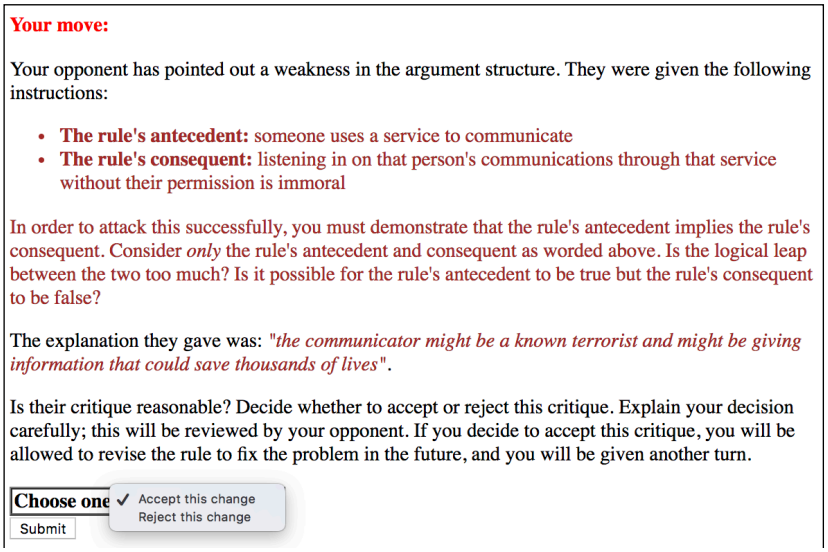

Figure 3b: When attacked, the advocate is given a summary and asked whether or not the attack is in accordance with the guidelines for that attack type.

They can either add a new pair of facts (one to the source domain, one to the target domain) or edit an existing pair of facts. It is explained to the user that such fact pairs must be analogous, and can either both refer to positive analogous properties (e.g., “the chicken crosses the road” / “the boat crosses the stream”) or opposite analogous properties (e.g., “the chicken lives near the road” / “the boat is not housed near the stream”), as long as they are factual and that mappings of concepts are

consistent[3] with the rest of the fact pairs. To ensure this factuality and consistency, all suggested fact pair changes by one user require approval by the other user. If the other user decides to accept a fact pair change, the player who made the acceptance is rewarded with another turn. If not, they are required to explain why they did not accept and are given the option of suggesting an alternate change instead, which is passed back to the other player for approval or rejection. In the current version, this back-and-forth is allowed to continue indefinitely, or until the user who initially suggested the change withdraws the motion.

2.2 Comparing Warrants and the Prior Association

With WG-A, we propose that by trying to find a common warrant that justifies both the source and target hypotheticals, we perform many of the same functions achieved by Bartha's articulation model, namely: the extraction and clarification of a prior association, and the evaluation of its potential for generalization. But it may be noted by the reader that this alignment is not perfect; indeed, there are quite a few differences between Bartha's prior association and what we are calling the warrant of an analogical argument (which itself is a simplification of Toulmin's warrants, e.g. we do not explicitly represent the warrant's backing).

Let us therefore briefly discuss some of the differences. Perhaps most importantly, the warrant is inherently inferential and directional; it is meant to show how a particular inference is warranted given a set of premises. A prior association, on the other hand, might go in the opposite direction, it might be bi-directional, or an undirected relationship between P and Q. Bartha uses these directions to distinguish between four types of prior associations (Bartha, 2010), most of which we can approximately capture through warrants by changing their qualifiers:

- Predictive analogies ($P \rightarrow Q$). The hypothetical Q is a consequence of P. We can express this with the warrant "If P_G, then Q_G," where P_G is a generalization of P and P∗, and Q_G is a generalization of Q and Q∗. If the relationship is causal, we might use "If P_G, then it will cause Q_G."
- Explanatory analogies ($P \leftarrow Q$). Q explains P. We can approximately capture this with the warrant "If P_G, then it can be explained by Q_G."

[3] I.e., in these examples, 'chicken' is clearly mapped to 'boat,' 'living' to 'being housed', and 'road' to 'stream.' A suggested fact pair that violates this mapping, such as "chickens are often found on roads" / "boats are often found in boathouses," could be rejected on this basis.

- Functional analogies ($P \leftrightarrow Q$). There is an association in each direction (but not necessarily the same type). Both directions can be expressed through warrants using the methods described above, but in many cases it is not clear whether it is possible to express more than one direction at a time with a single warrant.
- Correlative analogies (P and Q have no known direction of priority). For example, we might have no more than knowledge of a statistical correlation between P and Q. We might express this as "If P_G, then it's likely that Q_G."

The above list suggests that WG-A is best suited to non-functional, and perhaps non-explanatory analogical arguments. In our initial tests of WG-A, we used starting fact pairs that had moral or ethical analogical arguments. WG-A requires warrants to be expressed as "if-then" statements. To our knowledge, this is not something that was required by Toulmin or others, but it is a useful way to informally express many warrants, and as such is a helpful "starting point" for students still learning how to write warrants.

Another important distinction is that Bartha's articulation model first elaborates the prior association in the source domain, and then assesses its potential for generalization by applying it to the target. The warrants we propose here instead begin their lives as generalized statements, and have that generalizability tested iteratively through attacks and rewrites.

3. ENSURING RELEVANCE

WG-A is designed to ensure relevance in argumentative dialogues whose goals are to assess analogical arguments. In this section, we sharpen our claims towards meeting that goal. First, we adopt Bartha's idea that a good analogical argument has a clear prior association and potential for generalization. Then a relevant move (with respect to some analogical argument $\boldsymbol{A}$) is a move which affects the clarity of the prior association or its potential for generalization, either by affecting it directly or by implying a direct effect (using some measure of inferential distance). We are only dealing with the relevance of moves and are not addressing whether relevance is also a property of general utterances or other in-person actions (e.g., using voice tones to make implicit suggestions, wearing a t-shirt with printed text priming certain semantic frames, using body language to intimidate, etc.).[4]

[4] We will only briefly state here that although such utterances or non-verbal actions might indeed have a non-negligible effect on how minimally-trained participants assess analogical arguments, it is a separate issue whether they should be included in WG-A, given our desiderata.

Let us assume there is an argumentative dialogue D between minimally-trained participants, whose goal is to assess the quality of some analogical argument ***A***. If D is unrestricted and face-to-face, it is extremely difficult to ensure participants only make utterances and actions that are relevant to assessing ***A***. And it is also extremely difficult for some moderator to assess relevance of utterances in real-time. In American courts, for example, trial judges have "broad discretion when ruling on the relevance of evidence" (Blinka, 2006). Yet, overconfidence in their own ability to stay unbiased can lead to their ignoring of rules of evidence (Chortek, 2013), and there is evidence to show that judges exposed to inadmissible biasing evidence were, unknowingly or not, affected by it (Eren & Mocan, 2018; Landsman & Rakos, 1994; Rachlinski et al., 2015; Wistrich et al., 2005; Wistrich et al., 2015).[5] Furthermore, in adversarial trials, many objections of irrelevance "are simply missed because opposing counsel did not recognize the issue within the time limits demanded by the rules" (Blinka, 2006); other times, objections are used to "intimidate or confuse a lawyer of lesser skill, knowledge, and experience" (ibid). As an attempt to combat such problems, WG-A operates through an in-browser app, separating the players physically and only allowing them to make moves through the game, giving them more time to carefully choose their next moves. No other communication between players is allowed.

In a game like WG-A, a move might consist of changes made over one turn, in the same turn as other moves, or across multiple turns. The rules of WG-A restrict the moves that are permitted, and this paper's central claim is that those allowed moves tend to be relevant to assessing ***A***, since they tend to either strengthen the prior association or potential for generalization, or point out their flaws. To support this claim, let us first note that meaningful changes to the warrant correspond to meaningful changes to the prior association or its potential for generalization. Consider a warrant of the form "If $\varphi_1 \wedge ... \wedge \varphi_n$ then $\gamma_1 \wedge ... \wedge \gamma_m$," where all φ_i, γ_j are open formulae. Then adding new conjuncts to the warrant's antecedent or removing conjuncts from the consequent will tend to reduce the space of counterexamples to the warrant—i.e., the domain of objects for which the antecedent is true but the consequent is false. Likewise, removing from the antecedent or adding to the consequent will tend to increase the space of counterexamples. A change in the space of counterexamples to a

[5] Wistrich et al. (2005) noted that in some situations, some judges displayed a "surprising ability" to avoid being influenced by relevant but inadmissible information. To our point, however, this is a difficult skill to acquire, maintain, and externally ensure.

warrant is a change in the ways in which the warrant can be directly attacked on the basis of its generality. Furthermore, any change in the antecedent may affect the degree to which it is applicable to the source or target domain facts (and likewise for the consequent's applicability to the source or target hypothetical).

As a WG-A game goes on, the set of conditions φ_i in the warrant's antecedent will tend towards describing factors which are relevant, in the sense that they are necessary to describe the prior association claimed to hold in both the source and target domains. If any conditions in the antecedent are relevant but missing, then the space of possible counterexamples will be too large, and the advocate will be motivated to narrow it through WG-A's attack-edit mechanism. The advocate is discouraged from adding conditions to the antecedent that they believe are irrelevant, because it will unnecessarily cost them a turn.

A player might propose to edit or add a new fact pair. Such modifications must be approved by both players, in order to help ensure that the wording used for the fact pairs reflects uncontroversial details about the source and target domains. Players are encouraged to accept proposed edits or additions by being rewarded with an additional turn after accepting. Furthermore, because proposing an edit or addition costs a valuable turn (or an arbitrarily large number of turns), users are discouraged from making frivolous, unnecessarily argumentative, or loaded modifications. Our assumption is that this set of constraints will push players to only make fact pair modifications if they affect the logical connection between the fact pairs and the rule's antecedent, or open up possibilities for attacks or warrant edits later.

If a player is being unnecessarily abusive, clearly not following the rules of the game, or behaving in a way that is too far outside of what might be considered acceptable (in the opinion of the other player), the option to report their actions is always available to both players. When a report is submitted, the game is paused until a human moderator can review it and decide how to best resolve the dispute.

Only five attack types are allowed, all of which are encouraged to come in the form of counterexamples. An attack on the link between the rule's antecedent and consequent is thus a challenge to its generalizability. Attacks on the links connecting the rule to the source facts (L1 and L4 in Figure 2) identify flaws in the rule's applicability to the source domain, whereas attacking links L2 and L5 do the same for the target domain. Our assumption here is that most weaknesses in the prior association or its potential for generalization can be expressed in the form of attacks through one of the five links we have identified.

3.1 Disallowed Moves

Thus, the three major types of allowed moves in WG-A (edits to the warrant, revision of the source/target fact pairs, and attacks) all tend to affect the strength of the prior association or its potential for generalization. However, we do not claim all possible moves relevant to assessing A can be made using allowed moves of WG-A. Our approach to introducing moves to WG-A must be a slow and careful one, else we risk allowing the irrelevant or deceptive argument tactics that WG-A was designed to prevent. Our decisions on which move types or forms of dialogue to omit were made on a case-by-case basis, by estimating the tradeoff between a move's ability to introduce relevant moves and its likelihood of allowing irrelevant moves and comments. All such decisions are subject to change based on the results of future empirical evaluations. That being said, notable features intentionally omitted from the current version of WG-A include:

Limitations on editing. Both the advocate and critic have the option of editing the fact pairs in the source and target, and such edits are subject to approval by both sides. However, neither has the ability to make edits to the source or target hypotheticals Q. In very early versions of WG-A, players would sometimes edit the hypotheticals to be uninformative, uninteresting, uncontroversial statements. For example, the target hypothetical in Figure 1 might be changed to "Listening to someone else's phone call without their permission can be immoral in some situations."

Indeed, in real-world dialogues, a participant might backtrack and weaken the scope of their claim in order to make it more defensible. But the intended players of WG-A do not necessarily deeply believe the truth or falsity of Q^*. As such, the ability to modify **Q** may introduce too much of a temptation to make them easier to defend. A future update might allow the advocate to update **Q**, but it would likely need to penalize the player for modifications that make **Q** too tautological.

We also do not allow the critic to propose edits to the warrant, nor do we require the critic to approve warrant edits. We did not find any instances in which the advocate would benefit from the critic's suggesting a change to the warrant which could not be expressed through one of the allowed attacks.

Related arguments. In real-world, unrestricted dialogues meant to assess some analogical argument, there are many argumentative tactics which are regularly employed which WG-A explicitly disallows. The first is the use of multiple analogical arguments to either reinforce or undermine Q^*. Again, here we follow Bartha in noting that although

the assessment of an analogical argument may involve assessing how coherent it is with alternate analogies, the ability to assess individual analogical arguments should be considered a more fundamental reasoning task (Bartha, 2010).

Another common argument pattern is to use something resembling a high-level argumentum ad absurdum to attack the mapping of elements within an analogical argument by showing it leads to some absurdity:

1. The sun is round, and compact discs are round.
2. The sun is extremely large.
3. The sun is extremely hot.
4. From (1) and (3), compact discs are extremely hot.
5. If the sun and compact discs were analogous, then from (1) and (2), compact discs would be extremely large.
6. Compact discs are not extremely large. Therefore, the analogy fails, and (4) is false.

In our reading of such arguments, such moves are equivalent to introducing multiple analogical arguments—in this case, one which replaces a fact pair and hypothetical pair about extreme temperature with those about extreme size. Such an argument pattern, then, is not within the scope of the current version of WG-A.

Multi-step moves. It might be noted by the reader that many notions of argumentative relevance involve allowing actions or utterances that are indirectly relevant, in the sense that they set the stage for directly relevant moves later. The current version of WG-A has a clear preference towards moves which have immediate, observable effects on the warrant or source/target facts. Our experience so far is that a sufficiently large subset of such multi-step argumentative moves can be captured with the moves WG-A already has available. But the possibility remains open that perfectly legitimate moves may be required, particularly with complex analogical arguments, which require multiple iterations before they will bear fruit. Our suspicion is that these are (1) minimal, or (2) can be restructured to work within the confines of WG-A's rules. For example, a great deal of setup can be obtained by continually adding to and editing the source/target facts until they are ripe for an allowed attack. Future work may explore incorporating some notion of inferential distance, such as that described by (Macagno, 2018) based on argumentation schemes.

3.1 Antagonism Between Advocate and Critic

The roles of advocate and critic are contrary to one another, but it would be a mistake to imagine their roles as entirely antagonistic.

Bartha explains that their roles are differentiated in terms of the features they want to maximize in an analogical argument.

> [A]n enthusiastic advocate presents the analogical argument to a polite but moderately skeptical critic. Introducing this framework highlights the need to balance two competing pressures at work in representing and evaluating arguments from analogy: explicitness and economy. On the one hand, the critic wants the argument to be as explicit as possible, noting every factor that might be relevant to the conclusion, since the inclusion of detail increases the chance of exposing a weakness in the argument. On the other hand, the advocate wants to be economical about what counts as relevant (Bartha, 2010, p.102).

So, these roles should be seen as collaborating in the creation of an analogical argument, even while they compete to determine the qualities of the resultant argument. Understanding these roles is important for avoiding the problems that might result from a straightforwardly antagonistic relationship, which treats the loss of the opponent as a goal to be achieved.

A bad-faith advocate, imagining their duty to make a strong comparison, might refuse to focus in on an area of relevance. Instead, this advocate might try to draw a multitude of connections in the source and target domain, hoping to make the connection stronger that way. This would lead to an unhelpful list of similarities that cannot cohere to any rule. A bad-faith critic, in response, might refuse any and all additions to the source and target domains as irrelevant, at which point no progress could be made. These framing problems are arguably the result of the participants not appreciating the collaborative nature of the work.

Analogical reasoning might be restrained from such bad-faith excesses by the addition of a trained moderator, like a judge, who can call foul when one side is being unreasonable. WG-A allows for a moderator, but the need for intervention or oversight is minimized by the program's structure, which limits the acceptable moves to those that generally produce good results and provide regular opportunities for each participant to challenge the other's work. In the rare case that these remedies aren't enough, the participants have the option of reporting their interlocutor to the moderator. Initial testing shows minimal use of the report function, suggesting that most issues are resolved without the need for moderation.

4. DISCUSSION

WG-A helps us satisfy the desiderata of being implementable and requiring minimal training. It lends itself to computational implementation in several ways: It can be played remotely by two human players, or between a human player and some future AI, or perhaps later between two artificial reasoners. By structuring a cooperative dialogue so that it mostly contains moves that are relevant to the dialogue's goal, and by only allowing a finite number of types of moves, we have an easy way to generate a large dataset which can be analyzed, and used to train both people and artificial reasoning systems to reason better—indeed, such dataset building is a next step of this work.

Secondly, it satisfies our desideratum of being usable by minimally trained arguers, with minimal external moderation. Our initial results suggested that the 'report' function was used very rarely (roughly once per 20 games), thus allowing many WG-A instances to complete successfully without laborious manual oversight. Participants in our informal test runs were given less than 20 minutes of instruction prior to starting. Certainly, more evidence is required; as such we are in the process of performing empirical evaluations of WG-A on minimally-trained participants. This will satisfy two goals: it will allow us to demonstrate WG-A's strengths and weaknesses, and it will also allow us to collect large amounts of dialogues from actual games.

We expect that the development of WG-A will continue to iterate as we learn more about its strengths and weaknesses. In the remainder of this section, we report some of our observations from our informal test runs, along with ideas for future WG-A modifications.

Infrequent warrant edits. As expected, games were largely attack-driven. However, we also expected that successful attacks would be followed up with edits in order to prevent similar attacks in the future, made primarily to the warrant. This turned out not to be the case.

In the original warrant game described in Section 2, participants were given points for successful actions. A successful attack could be immediately followed up by another very similar attack, if the advocate did not make an effective move to edit the warrant to defend against it. In WG-A, we removed the point system entirely. This might simultaneously explain why we did not see advocates prioritize warrant edits, and why we did not see similar attacks re-occurring. In our next version of WG-A, our instructions will also make it clearer that warrant edits are encouraged to defend against future attacks, and that repeated

attacks (as long as they are still valid attacks) are allowed, even encouraged.

Hyper-specific warrants. In Section 3 we explained that the desire to not waste moves unnecessarily puts pressure on the advocate to not introduce irrelevant conditions into the warrant's antecedent. However, overcoming these pressures are still possible in the current version of WG-A. For example, given the game in Figure 2, imagine that the players somehow agree to add the fact pair "'piece of mail' starts with the letter 'p'" and "'phone call' starts with the letter 'p'," and then the advocate decides to modify the warrant to be "If someone uses a service to communicate and the service starts with the letter 'p', then listening on that person's communications through that service without their permission is immoral." Such a move severely restricts the warrant's generalizability, virtually ensuring that it only applies to the source and target domain as they are represented by the currently stated fact pairs. Clearly, more pressure is required to discourage such hyper-specific warrants.

It is worth noting that the hyper-specific warrant problem, as it is described here, did not appear once in our informal tests. However, it is worthwhile to try to prevent it and similar problems in the future. We expect that the introduction of a competitive point system may reward advocates for shortening their warrants. Another possibility is to introduce a new attack type, which allows critics to call out warrants that are hyper-specific. However, it is not clear at present how to define such attacks so that they can fit nicely within the constraints defined by WG-A.

Infrequent fact edits. In our informal tests, participants were provided with instances of WG-A, which started with only 3 or 4 initial fact pairs. It was expected that as games continued, players would add fact pairs corresponding to details of the source and target domain they felt were relevant to the analogical argument. That in turn would lead to a re-shaping of the warrant in order to defend against attacks. However, in practice, players rarely suggested adding new fact pairs, thus raising questions about the effects of the initial fact pairs.

Much more research is required to determine how to encourage more fact pair addition and editing, and on whether they are actions that are worth encouraging in the first place. It is not immediately clear what immediate effects this would have on WG-A. Two ideas we are exploring are (1) experimenting with instances of WG-A which contain a large number of initial fact pairs, particularly those which evoke different frame-semantic suggestions (Fillmore, 1976); and (2) having an initial game phase where players only have the option of adding, editing, and approving fact pairs.

REFERENCES

Bartha, P. F. (2010). By Parallel Reasoning: The Construction and Evaluation of Analogical Arguments. Oxford University Press.

Bartha, P. F. (2013). Analogy and Analogical Reasoning. In Zalta, E. N., editor, The Stanford Encyclopedia of Philosophy. http://plato.stanford.edu/archives/fall2013/entries/reasoning-analogy/, fall 2013 edition.

Beach, R., Thein, A. H., and Webb, A. (2016). Teaching to Exceed the English Language Arts Common Core Standards: A Critical Inquiry Approach for 6-12. Classrooms. Routledge, 2 edition.14

Besnard, P., Garcia, A., Hunter, A., Modgil, S., Prakken, H., Simari, G., and Toni, F. (2014). Introduction to structured argumentation. Argument and Computation, 5(1):1–4.

Blinka, D. D. (2006). Ethics, evidence, and the modern adversary trial. Georgetown Journal of Legal Ethics, 19(1).

Boltužić, F. and Šnajder, J. (2016). Fill the gap! Analyzing implicit premises between claims from online debates. In Proceedings of the 3rd Workshop on Argument Mining.

Chen, Q., Zhu, X., Ling, Z., Wei, S., and Jiang, H. (2016). Enhancing and combining sequential and tree LSTM for natural language inference. CoRR, abs/1609.06038.

Cheng, J., Dong, L., and Lapata, M. (2016). Long short-term memorynetworks for machine reading. CoRR, abs/1601.06733.

Chortek, M. (2013). The psychology of unknowing: Inadmissible evidence in jury and bench trials. The Review of Litigation, 32(117).

Dung, P. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. Artificial Intelligence, 7(2):321–358.

Eren, O. and Mocan, N. (2018). Emotional judges and unlucky juveniles. American Economic Journal: Applied Economics, 10(3):171–205.

Fillmore, C. J. (1976). Frame semantics and the nature of language. In Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech, volume 280, pages 20–32.

Habernal, I., Eckle-Kohler, J., and Gurevych, I. (2014). Argumentation mining on the web from information seeking perspective. In Cabrio, E., Villata, S., and Wyner, A., editors, Proceedings of the Workshop on Frontiers and Connections Between Argumentation Theory and Natural Language Processing, volume 1341 of CEURWS.

Harrell, M. and Wetzel, D. (2015). Using argument diagramming to teach critical thinking in a first-year writing course. In Davies, M. and Barnett, R., editors, The Palgrave Handbook of Critical Thinking in Higher Education, chapter 13. Macmillan.

Hitchcock, D. (2005). Good reasoning on the toulmin model. Argumentation, 19(3):373–391.

Keynes, J. (1921). A Treatise on Probability. Macmillan, London.

Lai, A. and Hockenmaier, J. (2017). Learning to predict denotational probabilities for modeling entailment. In Proceedings of the 15th

Conference of the European Chapter of the Association for Computational Linguistics.

Landsman, S. and Rakos, R. F. (1994). A preliminary inquiry into the effect of potentially biasing information on judges and jurors in civil litigation. Behavioral Sciences & the Law, 12(2):113–126.

Lunsford, K. J. (2002). Contextualizing toulmin's model in the writing classroom: A case study. Written Communication, 19(1):109–174.

Macagno, F. (2018). Assessing relevance. Lingua, 210-211:42 – 64.

Modgil, S. and Prakken, H. (2013). A general account of argumentation with preferences. Artificial Intelligence, 195:361–397.

Rachlinski, J. J., Wistrich, A. J., and Guthrie, C. (2015). Can judges make reliable numeric judgments? distorted damages and skewed sentences. Indiana Law Journal, 90.

Reed, C., Budzynska, K., Duthie, R., Janier, M., Konat, B., Lawrence, J., Pease, A., and Snaith, M. (2017). The argument web: an online ecosystem of tools, systems and services for argumentation. Philosophy and Technology, 30(2):137–160.

Rex, L. A., Thomas, E. E., and Engel, S. (2010). Applying toulmin: Teaching logical reasoning and argumentative writing. English Journal, 99(6):56–62.

Rocktäschel, T., Grefenstette, E., Hermann, K. M., Kociský, T., and Blunsom, P. (2015). Reasoning about entailment with neural attention. CoRR, abs/1509.06664.

Toulmin, S., Rieke, R., and Janik, A. (1984). An Introduction to Reasoning. Macmillan Publishing Company, New York, New York, 2 edition.

Toulmin, S. E. (2003). The Uses of Argument (Updated Edition). Cambridge University Press, updated edition.

Walton, D. (1985). Arguer's Position: A Pragmatic Study of Ad Hominem Attack, Criticism, Refutation, and Fallacy. Greenwood Press.

Walton, D. (1999). One-Sided Arguments: A Dialectical Analysis of Bias. State University of New York Press.

Walton, D., Reed, C., and Macagno, F. (2008). Argumentation Schemes. Cambridge University Press.

Warren, J. E. (2010). Taming the warrant in toulmin's model of argument. English Journal, 99(6):41–46.

Wistrich, A. J., Guthrie, C., and Rachlinski, J. J. (2005). Can judges ignore inadmissible information? the difficulty of deliberately disregarding. Cornell Law Faculty Publications, 20.

Wistrich, A. J., Rachlinski, J. J., and Guthrie, C. (2015). Heart versus head: Do judges follow the law or follow their feelings?. Texas Law Review, 93(4):855 – 923.