Jun 4th, 3:00 PM - 4:00 PM

Assessing Evidence Relevance by Disallowing Assessment

John Licato
University of South Florida

Michael Cooper
University of South Florida

Assessing Evidence Relevance by Disallowing Assessment

JOHN LICATO
Department of Computer Science and Engineering
Advancing Machine and Human Reasoning (AMHR) Lab
University of South Florida
Tampa, FL
United States of America
licato@usf.edu

MICHAEL COOPER
Department of Philosophy
Advancing Machine and Human Reasoning (AMHR) Lab
University of South Florida
Tampa, FL
United States of America
michaelcoop@usf.edu

Abstract: Guidelines for assessing whether potential evidence is relevant to some arguments tend to rely on criteria that are subject to well-known biasing effects. We describe a framework for argumentation that does not allow participants to directly decide whether evidence is potentially relevant to an argument---instead, evidence must prove its relevance through demonstration. This framework, called WG-A, is designed to translate into a dialogical game playable by minimally trained participants.

Keywords: Analogy, evidence, reasoning, relevance, testimony, WG-A

1. Introduction

In evidential reasoning, establishing that a piece of evidence is relevant to proving some fact may require the elaboration of an *evidential hypothesis*, typically a general premise showing how a candidate piece of evidence supports, or is relevant to proving, a target inference (Ball, 1980; Mueller & Kirkpatrick, 2011). The elaboration of an evidential hypothesis may be unnecessary in cases where the relevance of a proposed fact is obvious, but in many scenarios, establishing relevance might require additional inferential steps. Furthermore, even in cases where the relevance of a fact appears obvious at first glance, the practice of making an evidential hypothesis explicit may help a reasoner to catch and minimize the influence of cognitive, argumentative, or stereotyping biases.

For example, consider the use of witness testimony in determining whether someone is responsible for a crime. We may have previous knowledge that in a previous case, witness testimony s_f was allowed and determined to be relevant, but the determination of why s_f was considered relevant was not fully made explicit. If we nevertheless have a suspicion that the reasoning used to determine the relevance of s_f might apply to the witness testimony in the present case, a decision procedure to test this suspicion by simultaneously generating and evaluating an evidential hypothesis is highly desirable.

In this paper, we will present such a decision procedure. More precisely, assume we are given:

- a factum probans / facta probantia $\mathbf{s_f}$ (evidence given in support of $\mathbf{s_h}$ whose relevance is in question);
- a factum probandum $\mathbf{s_h}$ (fact to be proved);
- a factum probans / facta probantia $\mathbf{t_f}$ from a previous case **C**; and
- a factum probandum $\mathbf{t_h}$ also from **C**, such that (1) $\mathbf{t_f}$ was previously established to be relevant to $\mathbf{t_h}$, but a minimal or no evidential hypothesis was made explicit, and (2) the pair ($\mathbf{s_f}$, $\mathbf{s_h}$) are similar to ($\mathbf{t_f}$, $\mathbf{t_h}$).

Such scenarios will be familiar to proponents of case-based or analogical reasoning. The task we then propose to solve with WG-A is one of analogical generalization: How might one systematically and algorithmically use $\mathbf{s_f}$, $\mathbf{s_h}$, $\mathbf{t_f}$, $\mathbf{t_h}$, and **C** to generate an evidential hypothesis which simultaneously establishes (1) how $\mathbf{s_f}$ is relevant to proving $\mathbf{s_h}$, and (2) how $\mathbf{t_f}$ is relevant to proving $\mathbf{t_h}$?

The framework we will apply to this task is called WG-A (for Warrant Game -- Analogy), which is based on the Articulation Model (AM) of analogical reasoning (Bartha. 2010). WG-A allows for two participants, an advocate and a critic, to participate in a highly structured dialogical exchange in which a warrant (in this case, an evidential hypothesis) is made explicit and the relevance of facts to that warrant are tested iteratively. In other words, the relevance of facts is tested continually as the participants take their turns, thus offering a solution to the concern that any fact can be shown relevant to any other by simply creating an appropriately manufactured evidential hypothesis, e.g., "if (fact 1) then (fact 2)" (Michael & Adler, 1931; Tillers, 2005).

Our goals for WG-A are twofold: First, we seek to create a well-defined model of good analogical reasoning that can be used by, and used to teach, human reasoners. Second, WG-A is part of a larger research project whose goals are to create highly structured modes of argumentative interactions which lend themselves to computational *implementation* (they can be implemented as computer programs) and *automation* (artificially intelligent reasoners can be developed which are able to participate in these interactions). By studying the kinds of reasoning tasks that can be carried out by WG-A, such as the target task of this paper, we hope to develop algorithms for automated reasoners and datasets to train them. Our present work thus seeks to complement existing related work (e.g., Bex et al., 2003; Bex, 2015; Verheij et al., 2016) by focusing on a model of analogical generalization which can lead to the automated generation of evidential hypotheses used to establish relevance.

We will proceed by introducing Bartha's Articulation Model, upon which WG-A is based. We will then explain WG-A itself, and provide summarized arguments for how WG-A establishes relevance. Finally, we will close with a detailed example of WG-A solving a problem from the target task introduced above.

2. Background

To provide the necessary context, Sections 2 and 3 will closely follow (Licato and Cooper, 2019).

2.1 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, p. 202). 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 that 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, 2019).

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.

2.2 Warrants

A warrant, in Stephen Toulmin's model of argumentation, is a statement connecting the premises and conclusion of an argument, showing how the former permits the inference of the latter (Toulmin et al., 1984; Toulmin, 1958). 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.

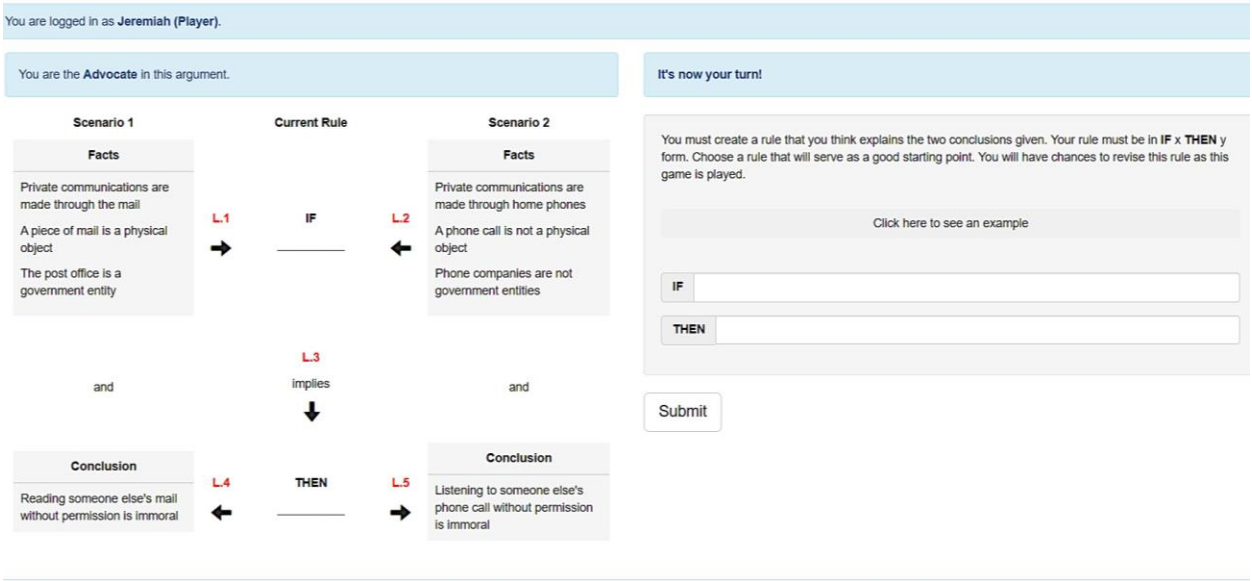
The warrant, when made explicit, makes it easier to determine key features typically associated with argument strength, not limited to: (1) what kind of attacks can be used against the argument, (2) whether the premises are relevant or necessary to the argument, and (3) whether, and with what strength, 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 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 (Harrell & Wetzel, 2015; Lunsford et al., 2002; Rex et al., 2010). It has been observed that this omission is to the detriment of automated reasoning in the former case, and to students in the latter (Beach et al., 2016; Warren, 2010;).

We consider analogical reasoning to be a form of *substantive argumentation*, analysis of which renders of judgment of the plausibility of the comparison between domains. To elucidate this distinction, Toulmin offers his analysis of an eighteenth-century story of the Count and the Abbé. In the story, the Count tells his audience that he was the Abbé's first penitent, and the Abbé later claims that his first penitent was a murderer. Toulmin points out that treating this story as a formal argument leads to misleading conclusions:

> We have only to hear this story to jump to the conclusion: 'The Count was a murderer'; and truly, if we take the two statements at face value—'The Count was the Abbé's first penitent' and "The Abbé's first penitent was a murderer"—they lead as they stand, by a formal argument, to the conclusion: 'The Count was a murderer.' Yet the same story can be parsed, instead, as a piece of *substantive argumentation*. What guarantee have we that either the Count or the Abbé is telling the truth? (2001, p. 16)

Though formal argumentation suggests that the result that the Count is a murderer, in real argumentation this result would need to be held in doubt due to the two different sources and worries about their veracity. Despite its limitations, substantive argumentation is useful in justifying adherence to a warrant. 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



3. The Warrant Game and WG-A

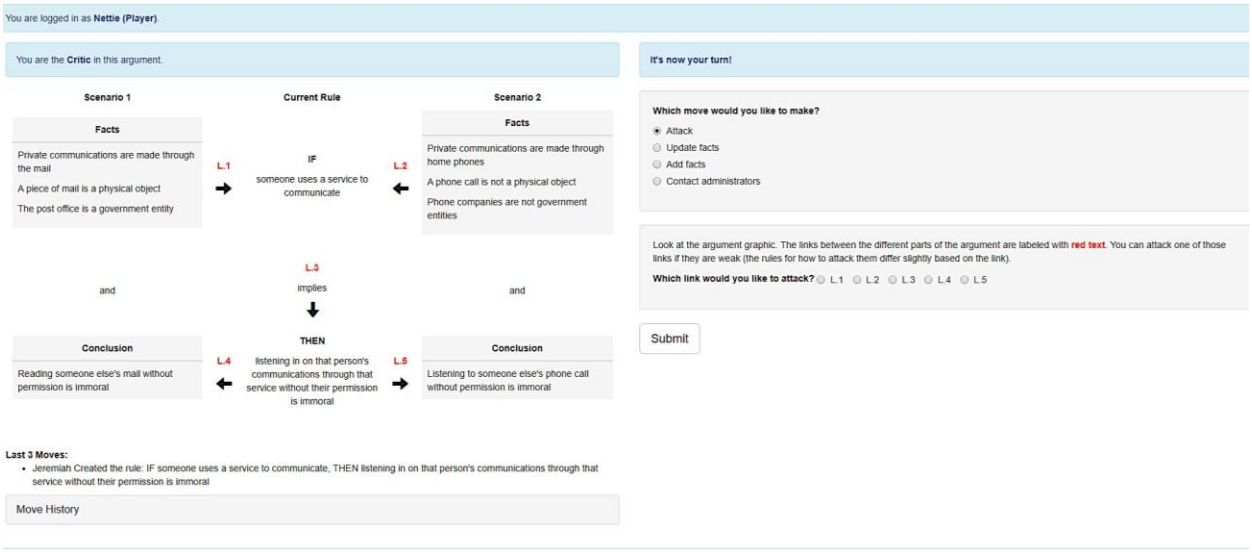
Given the benefits of warrant-based reasoning, our previous work developed a classroom activity to introduce students of critical thinking to warrant-centered argumentation called "the Warrant Game" (WG). In WG, teams of students put forth opposing arguments. They must carefully phrase the warrants for 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 (Besnard et al., 2014; Dung, 1995; Modgil & Prakken, 2013; Reed et al., 2017) and Walton's argumentation schemes (Walton, 1985; Walton et al., 2008).

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; and for this reason, this approach to warrant evaluation is in line with the long-term goals of our research program.

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



3.1 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).

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 show that **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[1] who is given the choice to either update the source / target facts, send an attack, or pass.[2] 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.

See figures (3a, 3b) on following page (p. 7).

[1] The move is recorded in a log that is always accessible to both participants.

[2] Passing is only an option after a certain number of moves have been made.

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.

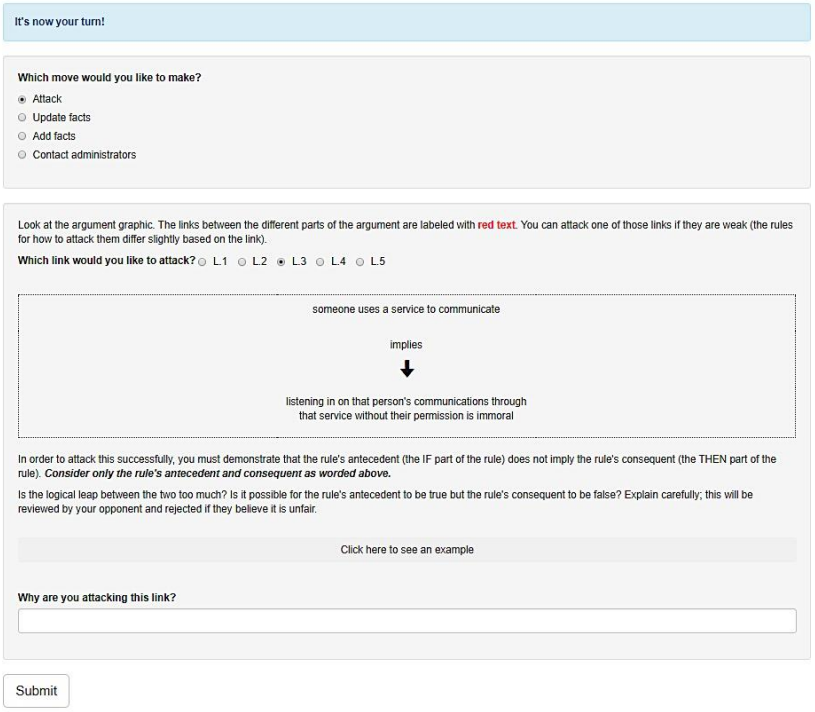


Figure 3b: When attacked, the advocate is given a summary and asked whether or not the attack is in accordance with the guidelines for that attack type.

It's now your turn!

You opponent was given the instructions inside the dashed box and has pointed out a weakness in the argument: **"the communicator might be a known terrorist and might be giving information that could save thousands of lives"**

someone uses a service to communicate

implies

listening in on that person's communications through that service without their permission is immoral

In order to attack this successfully, you must demonstrate that the rule's antecedent (the IF part of the rule) does not imply the rule's consequent (the THEN part of the rule). ***Consider only the rule's antecedent and consequent as worded above.***

Is the logical leap between the two too much? Is it possible for the rule's antecedent to be true but the rule's consequent to be false? Explain carefully; this will be reviewed by your opponent and rejected if they believe it is unfair.

Click here to see an example

Decide whether to accept or reject this critique. If you decide to reject, explain your decision carefully as it will be reviewed by your opponent. If you decide to accept, you will be given another turn to fix the problem in the future.

Is their critique reasonable?

○ Accept ○ Reject

Submit

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 consistent 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.

3.2 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: (1) Predictive analogies ($P \rightarrow Q$). The hypothetical *Q* is a consequence of *P*.[3] (2) Explanatory analogies ($P \leftarrow Q$). *Q* explains *P*.[4] (3) Functional analogies ($P \leftrightarrow Q$). There is an association in each direction (but not necessarily the same type).[5] (4) Correlative analogies (*P* and *Q* have no known direction of priority).[6]

The above list suggests that WG-A is best suited to non-functional, and perhaps non-explanatory analogical arguments, since those better fit our formalization of the warrant. 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.

---

[3] 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."

[4] We can approximately capture this with the warrant "If P_G, then it can be explained by Q_G."

[5] 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.

[6] For example, we might have no more than knowledge of a statistical correlation between *P* and *Q* and express this as "If P_G, then it's likely that Q_G."

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 simultaneously to source and target domains.

3.3 Ensuring Relevance Without Assessing It Directly

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 **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).[7]

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). 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.

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 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).

[7] 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.)

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. 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.[8]

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 (**L.1** and **L.4** in Figure 2) identify flaws in the rule's applicability to the source domain, whereas attacking links **L.2** and **L.5** 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.

3.3.1 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 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. 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 for the advocate to modify **Q** may introduce too much of a temptation to make them easier to defend, and it is therefore disallowed in the competitive version of WG-A (but in Section 4, we will introduce a case where enabling the advocate to edit Q^* is allowed, and perhaps even encouraged).

4. Developing an evidential hypothesis: Examples

[8] 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.

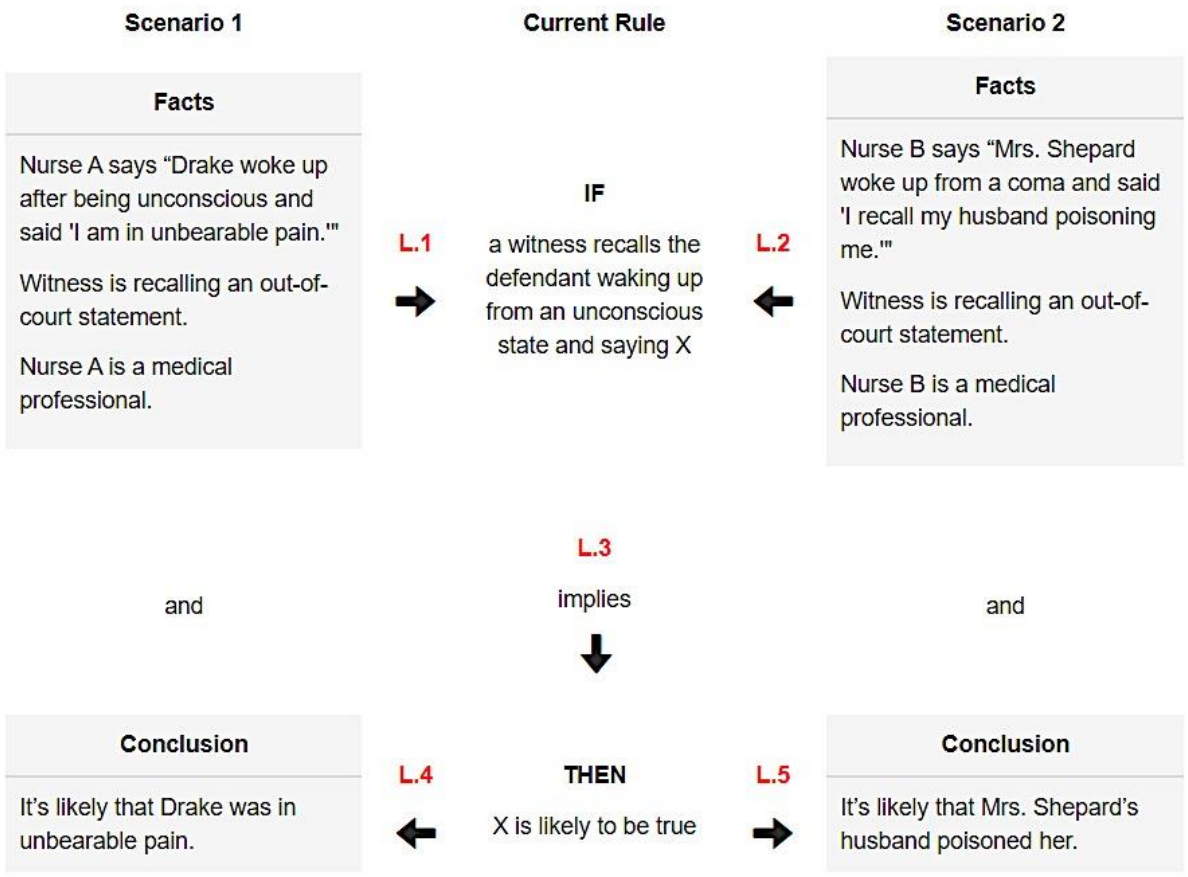
We will now walk through an example of WG-A applied to the target task described in this paper's introduction. Our example uses two cases adapted from Klein (2013). In the source case, a driver named Drake is in a car accident, and is hospitalized for several days, the first of which was spent in an unconscious state. He wakes up, upon which his nurse (Nurse A) claims he said "I am in unbearable pain." A few days later, he dies from his injuries. The driver of the other car in the accident is sued to compensate for the pain Drake suffered, and it is determined that the nurse's testimony is relevant. However, in our toy example, the evidential hypothesis for why Nurse A's testimony is relevant is not made explicit.

In the target case, Mrs. Shepard was admitted to the hospital, apparently in a coma. Several days in, she woke up, and according to the nurse (Nurse B), said "my husband has poisoned me." She died not long after, and her husband was charged with murder. Is the testimony of Nurse B relevant to the case *in the same way that the testimony of Nurse A was relevant to the previous case*, and if so, what is an evidential hypothesis justifying this determination?

We structure the initial state of WG-A using the facts and hypotheticals listed in Figure 4. An initial, simplistic warrant is provided: "If a witness recalls the defendant waking up from an unconscious state and saying X, then X is likely to be true." An immediate attack on the warrant (attack point **L.3**) seems an obvious next step. A counterexample to the warrant would suffice: "A witness might simply be lying about what the defendant said."

Such an attack can easily be responded to by the advocate, by updating the warrant to "If a **highly credible** witness recalls the defendant waking up from an unconscious state and saying X, then X is likely to be true." However, this leaves attack points **L.1** and **L.2** open, as the critic might claim that the nurses haven't been established as "highly credible." This again is easy to fix, by proposing that facts be updated on both sides to include "The nurse is a medical professional**, and thus highly credible.**"

Figure 4
Starting State and Initial Warrant for Example 1

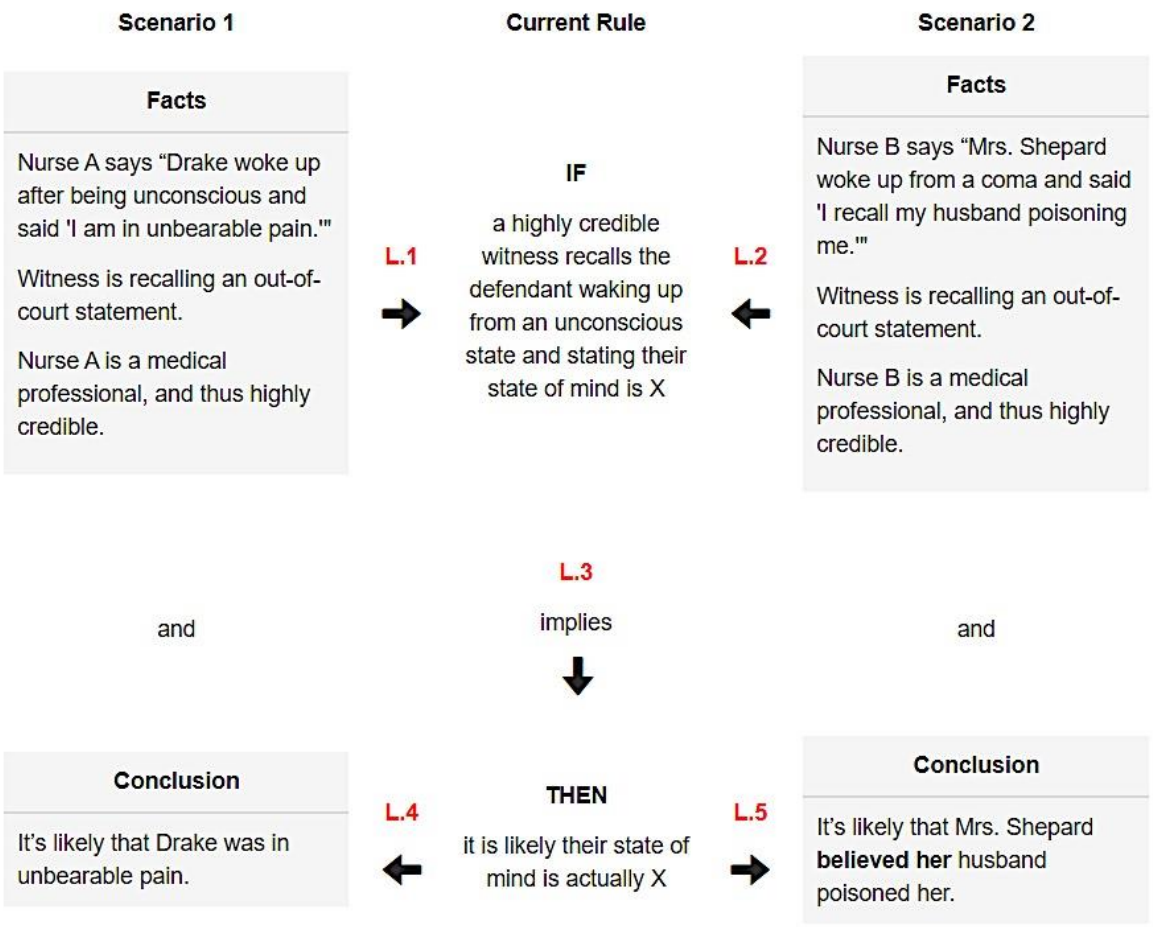


Such a claim is disputable by the critic, but for simplicity let us assume they accept these proposed fact changes. Another attack can now be launched on the warrant: "Defendants waking up from unconscious states are often confused, and they have been known to have distorted perceptions." Examples of patients waking from comas and exhibiting paranoid behavior about their surroundings might be cited. In any case, the advocate is forced to respond with a strengthening of the warrant, and the result may be: "If a highly credible witness recalls the defendant waking up from an unconscious state and stating **their state of mind is X**, then **it is likely their state of mind is actually X.**" This warrant shift specifies that the testimony of the witness is relevant only insofar as it helps to establish the state of mind of the defendant, to which presumably, the defendant has privileged access.

However, this shift in meaning of the warrant opens the warrant's connections to the source and target to attack. Specifically, the warrant's consequent now speaks of the defendant's "state of mind," and although the source conclusion (Drake's feeling of unbearable pain) can be classified as such, Mrs. Shepard's memory cannot. Instead, the target warrant's consequent must be edited, e.g. to "It's likely that Mrs. Shepard **believed her** husband poisoned her." Now, the argument on the target side is not that the witness's testimony is relevant to (directly) establishing whether her husband poisoned her, but rather to whether *she recalls* her husband doing so. As noted earlier, the direct editing of conclusions was not allowed in earlier versions of WG-A, out of concern that

advocates would make the conclusion trivial and thus easier to defend. That concern, however, does not apply in the present example, since a trivial conclusion would be of little use to the advocate.

Figure 5
Final State for Example 1



The resulting state of our example WG-A game is pictured in Figure 5. Players WG-A can theoretically continue after this, making attacks and edits *ad infinitum*, but this is a natural stopping point, particularly because the very distinction just discussed was the one Klein (2013) sought to make in contrasting these two example cases. If the conclusion of the target case is shifted to the one in Figure 5, Klein argues, "[t]his is precisely the kind of evidence the state of mind exception [to the hearsay rule] is designed to allow in."

Also note that Figure 5 still contains a fact pair (relating to the witness's statements being out-of-court) that is not made use of in the warrant. Implicitly, then, that fact pair has been determined to be irrelevant to the present case, at least according to the warrant-centric definition of relevance we adopt here. Many other fact pairs may be present but un-used in an instance of WG-A (e.g., the gender of the nurse, or the amount of time Drake and Mrs. Shepard were unconscious). It is thus that WG-A approaches determination of which factors are relevant to a

warrant: *not by requiring direct assessment of their relevance, but by putting their relevance to the test, and determining relevance based on the final product of the WG-A game*.

5. Conclusion

WG-A is a recent contribution to a long line of highly structured argumentation games, which break down complex and sometimes opaque reasoning processes into easier-to-understand, algorithmic steps. This may lead to better artificially intelligent reasoners, if performing well at those simpler steps become within reach of AI. They may also lead to AI that is better at justifying and explaining its reasoning, particularly if we require them to break down their reasoning in similar ways.

The educational implications of the above are clear as well. Experienced reasoners may look at some of the visualizations and moves in WG-A as obvious and trivial, but those who are new to legal or argumentative reasoning can benefit from seeing WG-A's possible points of attack clearly presented. Such breakdowns can also make it easier to design automated tutoring systems to help students by giving suggestions of applicable attacks and responses to those attacks.

In recent work [NOTE: this is currently under publication consideration, and will be updated in this paper's final version], our lab explored whether minimally trained participants using WG-A performed better on a task of argumentation and critical reasoning than participants who were simply asked to discuss an analogy in an open-ended dialogue format. Our initial results suggest that a benefit does in fact exist, and interestingly, this effect may be delayed: the participants who used WG-A did not perform much better at the critical reasoning tests immediately after using WG-A, but did perform better a week later than the control group. Future work will further explore this and related ideas.

Acknowledgements: 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. We also wish to thank Marc Badilla, an undergraduate researcher who contributed significantly to the WG-A project.

References

Ball, V. C. (1980). *The myth of conditional relevancy*. Georgia Law Review, 435.

Bartha, P. F. A. (2010). *By parallel reasoning: The construction and evaluation of analogical arguments*. Oxford University Press.

Bartha, P. F. A. (2013). *Analogical arguments in mathematics*. In A. Aberdein, & I. Dove (Eds.), *Logic, epistemology, and the unity of science*. Springer.

Bartha, Paul, "Analogy and analogical reasoning", *The Stanford Encyclopedia of Philosophy* (Spring, 2019 ed.), E. N. Zalta (Ed.), https://plato.stanford.edu/archives/spr2019/entries/reasoning-analogy/

Beach, R., Thein, A. H., & Webb, A. (2016). *Teaching to exceed the English Language Arts Common Core Standards: A critical inquiry approach for 6-12 classrooms* (2). Routledge.

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

Bex, F. J., Prakken, H., Reed, C., & Walton, D. N. (2003). Towards a formal account of reasoning about evidence: Argumentation schemes and generalisations. *Artificial Intelligence and Law*, *11*(2/3), 125-165.

Bex, F. (2015). *An integrated theory of causal stories and evidential arguments* [Proceedings].15th International Conference on Artificial Intelligence and Law (ICAIL).

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

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

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

Dung, P. M. (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. & Mocan, N. (2018). Emotional judges and unlucky juveniles. *American Economic Journal: Applied Economics*, *10*(3), 171-205. http://www.aeaweb.org/articles?id=10.1257/app.20160390

Habernal, I., Eckle-Kohler, J., & Gurevych, I. (2014). *Argumentation mining on the Web from information seeking perspective* [Proceedings]. The Workshop on Frontiers and Connections Between Argumentation Theory and Natural Language Processing.

Harrell, Marallee & Wetzel, Danielle. (2015). Using argument diagramming to teach critical thinking in a first-year writing course. In M. Davies, & R. Barnett (Eds.), *The palgrave handbook of critical thinking in higher education*. Macmillan.

Keynes, J. M. (1921). *A treatise on probability*. London: Macmillan.

Klein, K. S. (2013). The enduring quality of an alluring mistake: Why one person's intentions cannot --- and never could --- be evidence of another person's conduct. *American Journal of Trial Advocacy*, *37*(2).

Landsman, S. & 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. https://onlinelibrary.wiley.com/doi/abs/10.1002/bsl.2370120203

Licato, J. & Cooper, M. (2019). *Evaluating relevance in analogical arguments through warrant-based reasoning* [Proceedings]. The European Conference on Argumentation (ECA 201).

Lunsford, K. J. (2002). Contextualizing Toulmin's model in the writing classroom: A case study. *Written Communication*, *19*(1), 109-174. https://doi.org/10.1177/074108830201900105

Michael, J. & Adler, M. J. (1931). *The nature of judicial proof: An inquiry into the logical, legal, and empirical aspects of the law of evidence*. (Ad Press).

Modgil, S. & Prakken, H. (2013). A general account of argumentation with preferences. *Artificial Intelligence*, *195*, 361-397.

Mueller, C. B. & Kirkpatrick, L. C. (2011). *Evidence Under the Rules: Text, Cases, and Problems*, (7th ed.). Wolters Kluwer Law and Business.

Rachlinski, J. J., Wistrich, A. J., & 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., & 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., & Engel, S. (2010). Applying Toulmin: Teaching logical reasoning and argumentative writing. *English Journal*, *99*(6), 56-62.
Tillers, Peter. (2005). *Rethinking relevancy*. Available at SSRN: http://dx.doi.org/10.2139/ssrn.692627
Toulmin, S. E. (1958). *The uses of argument*. Cambridge University Press.
Toulmin, S., Rieke, R., & Janik, A. (1984). *An introduction to reasoning*. New York, New York: Macmillan Publishing Company.
Toulmin, S.E. (2001). *Return to reason*. Cambridge, Mass: Harvard University Press.
Verheij, B., Bex, F., Timmer, S. T., Vlek, C. S., Meyer, J.-J. C., Renooj, S., & Prakken, H. (2016). Arguments, scenarios and probabilities: Connections between three normative frameworks for evidential reasoning. *Law, Probability and Risk*, *15*, 35-70.
Walton, D. (1985). *Arguer's position: A pragmatic study of Ad Hominem attack, criticism, refutation, and fallacy*. Greenwood Press.
Walton, D., Reed, C., & 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., Rachlinski, J. J., & Guthrie, C. (2015). Heart versus head: Do judges follow the law or follow their feelings? *Texas Law Review*, *93*(4), 855 - 923.