Aragorn: Eliciting and Maintaining Secure Service Policies

Nirav Ajmeri, Chung-Wei Hang, Simon Parsons, Munindar P. Singh

Abstract

Services today are configured through policies that capture expected behaviors. However, because of subtle and changing stakeholder requirements, producing and maintaining policies is nontrivial. Policy errors are surprisingly common and cause avoidable security vulnerabilities.

We propose Aragorn, an approach that applies formal argumentation to produce policies that balance stakeholder concerns. We demonstrate empirically that, compared to the traditional approach for specifying policies, Aragorn performs (1) better on coverage, correctness, and quality; (2) equally well on learnability and effort÷coverage and difficulty; and (3) slightly worse on time and effort needed. Thus, Aragorn demonstrates the potential for capturing policy rationales as arguments.

Introduction

Policies are widely employed by organizations to support effective and secure service delivery. Policies are explicit, inspectable, and readily updatable. However, creating and maintaining policies is nontrivial: Errors made by system administrators in system configuration cause service provisioning failures, and introduce security vulnerabilities [9, 14].

Specifically, defining and maintaining policies *correctly* is nontrivial because correctness depends upon conflicting and changing stakeholder requirements. Current methods capture policies in a seat-of-the-pants manner by creating and reordering a list of policies. Previous formal approaches focus on analysis (to identify anomalies and conflicts) [1, 2].

In contrast, Aragorn focuses on the underlying stakeholder requirements, which current methods ignore, to guide system administrators in policy creation. Doing so helps capture administrators' tacit knowledge and facilitates maintenance.

Approach

We propose Aragorn, a general-purpose method for policy creation and maintenance based on argumentation, to facilitate policy creation. Aragorn incorporates domain-specific argumentation schemes and provides an evidential basis for balancing competing concerns that might reflect (inconsistent) stakeholder requirements. In essence, Aragorn models each policy specification as a decision and represents arguments pro and con that specification. This network of arguments captures the design rationale for each policy, highlighting dependencies and conflicts between the policies. The arguments provide guidance for updating policies in light of changing requirements and changing premises—specifically, in determining which policies to drop, insert, or modify in light of changes to stakeholder requirements.

Aragorn comprises two main parts. The first is an evidence-based framework that identifies actions on policies (such as activating or deactivating a policy). Successful argumentation relies upon applying one or more *argumentation schemes* that identify critical questions that an argument can raise. Aragorn incorporates argumentation schemes geared toward service security policies, and provides a simple approach to reasoning about the strength of belief in an argument. Aragorn's second part is a methodology that guides developing security policies that are the most defensible given the premises and arguments.

Successful service delivery involves security and associated policies at multiple levels. For concreteness, we consider cloud services and adopt firewall policies to demonstrate Aragorn. Firewalls are a well-understood setting in which to refine and evaluate methods for creating and maintaining service policies. Firewalls must be updated in light of new evidence about security threats and vulnerabilities [9].

Contributions and Findings

The main contribution for this paper is Aragorn, an argumentation-based policy-specification methodology, and its empirical evaluation. The benefit of Aragorn is that it addresses the challenge of misconfiguration of policies caused by traditional approaches, which causes avoidable vulnerabilities.

We demonstrate Aragorn using a firewall policy modeling scenario. We empirically evaluate Aragorn via a human-subject study. We find that Aragorn yields improvements over the traditional approach in measures (defined below) of maintainability, coverage, correctness, effort, and effort÷coverage. Participants find Aragorn easy to learn and not difficult to apply although it requires additional effort. We conjecture that aggregated gains in quality compensate for the additional time though we expect that effective tool support would mitigate this shortcoming.

Firewall as an Exemplar Domain

Firewalls and firewall policies are important to the security infrastructure of an enterprise. They form the first line of defense against attacks and unauthorized traffic. Dynamically emerging threats and requirements force system administrators to continually refine their firewall policies. As for any security policy, creating and maintaining firewall policies is nontrivial, and a lack of efficient mechanisms and tools to analyze firewall policies results in policy errors. For instance, Wool [14] found that despite significant effort by enterprises, firewalls are frequently wrongly configured. The complexity and challenges involved in defining and maintaining firewall policies makes firewalls an appropriate choice as the exemplar domain of application for Aragorn.

A firewall controls information flow from and to a computer network. A firewall policy is an ordered set of rules [2]. Each rule specifies if an incoming or outgoing packet is allowed or blocked, depending upon its protocol (TCP or UDP), source IP, source port, destination IP, and destination port. Given a packet, a firewall considers the rules in sequence and then takes the action specified by the first rule that applies. If that action is not desired, it might result in a breach. Therefore, it is crucial to order the rules correctly. Table 1 shows a partial firewall policy, and lists the requirements (R), and the facts (F) for the firewall policy.

Defining and maintaining a firewall policy involves complex decision making. There is a potential for conflicts between policies resulting in *anomalies* [1]. A rule L *generalizes* over a rule M provided L has higher priority than M, every packet that satisfies L also satisfies M, and both rules specify different actions. For example, in Table 1, Rule 12 generalizes over Rule 9 [2]. Additional firewall anomalies include *shadowing*, *correlation*, and *redundancy* [1]. Some firewall rules are never executed because of such anomalies.

Table 1: An example firewall policy [2], and associated requirements (R) and facts (F)

#	Action	Protocol	Source	Port
1	Allow	*	*	20
2	Allow	*	*	80
3	Block	*	locA.example.com	20
4	Allow	*	*	22
5	Block	*	*	53
6	Allow	TCP	locB.example.com	23
7	Block	*	*.example.com	*
8	Allow	UDP	locB.example.com	5027
9	Allow	UDP	*	*
10	Block	*	*	6889
11	Allow	*	example.net	53
12	Block	*	*	*

#	Description: Requirement or Fact
R	Example Inc. requires file transfer
F	FTP (port 20), SFTP (port 22) enable file transfer
R	Public website, made available via port 80
R	Prevent known attacks
F	Attacks from locA.example.com, port 20
R	Example Inc. requires telnet access
F	Telnet requires access to port 23
R	Application to access UDP port 5027
R	Prevent access to torrent ports
F	Torrent uses port 6889
R	Enable DNS for example.net
F	DNS requires access to port 53

Anomalies in a firewall policy may result in legitimate packets being blocked, and unwanted packets being allowed. An administrator should resolve any anomalies. However, detecting and resolving anomalies is tedious and error-prone. Common errors include failing to validate key assumptions; omitting important evidence; overlooking connections across the available evidence; and improperly prioritizing conflicting evidence and justifications.

Argumentation and Argumentation Schemes

An argument is constructed and consists of three parts: a conclusion, a set of premises, and an inference from the premises to the conclusion [4]. Each elementary argument captures a reasoning step from premises to conclusion. An argument can be supported or attacked by other arguments. Argumentation involves constructing a network of arguments, where the conclusion of one is a premise of another. Formal argumentation theory models arguments as first-class entities. An argument may *undermine* (contradict the premises of) or *rebut* (contradict the conclusions of) another.

An argumentation scheme is a pattern for constructing arguments [12] that represents the inference structure of an argument and provides *critical questions* for evaluating if the argument holds. A critical question captures critical thinking needed to engage in an argument. Specifically, a decision-maker raises critical questions to find arguments that support or attack another argument.

A decision-maker follows an argumentation scheme to iteratively collect evidence and infer the veracity of its conclusion until a decision (i.e., whether to accept or reject the conclusion) can be made. We write a scheme in the following template: $\langle \text{Premise}, \text{Conclusions}, \text{Questions} \rangle$.

The practical reasoning argumentation scheme provides a template for a situation where a decision-maker asserts that action A should be carried out because A is a means to realize requirement R [12]. We write this scheme as follows:

- Premise (Major) h_1: R is a requirement.

- Premise (Minor) h_2: Action A is a means to realize R.
- Conclusion c: A should be carried out.

This scheme corresponds to these critical questions:

CQ1 What requirements should we consider that might conflict with R?

CQ2 Are alternative means available for carrying out R?

CQ3 Is A more efficient than these alternative means?

CQ4 Is it practically possible to bring about A?

CQ5 What consequences of bringing about A should we take into account?

CQ6 Are other actions, in addition to A, required to bring about R?

These critical questions help identify supporting or conflicting requirements and alternative and efficient means to realize a requirement. We adopt additional argumentation schemes, such as *arguments by consequence*, and *arguments from alternatives and opposites* to answer these questions.

Aragorn: Evidence-Based Argumentation

The decision-maker's knowledge base consists of premises and inference rules that are supported or opposed by available evidence, each piece of evidence being associated with a belief measure. A belief measure of a premise or inference rule is calculated based on the belief measures of the evidence that supports or opposes it. The decision-maker applies argumentation schemes to construct an argument and calculates the belief measure of its conclusion by combining the belief measures.

We adapt a probability-certainty representation of belief [13]. The probability represents the likelihood that the conclusion is true given the evidence, whereas the certainty measures the amount of nonconflicting evidence. A sufficiently high certainty indicates the completeness of the argument.

We define the belief measure of a premise or inference rule as in Tang et al. [11].

Definition 1 (Belief Measure Combination) *Let h_A and h_B be premises in Σ with belief measures $\langle b_A, d_A, u_A\rangle$, $\langle b_B, d_B, u_B\rangle$, respectively. Let h_C be another premise and $\delta = \frac{h_A}{h_C}$ be an inference rule with belief measure $\langle b_\delta, d_\delta, u_\delta\rangle$. Assume h_A and h_B are independent. Then,*

$$\begin{aligned}
b_{A\wedge B} &= b_A b_B \\
d_{A\wedge B} &= d_A d_B + d_A u_B + u_A d_B \\
b_{A\vee B} &= (b_A + b_B)/2 \\
d_{A\vee B} &= (d_A + d_B)/2 \\
b_C &= b_\delta b_A \\
d_C &= b_\delta d_A
\end{aligned}$$

By answering a critical question about h_i, the decision-maker collects (or updates) evidence $e_j \in \mathbb{E}$ and associated belief measure $\langle b_{e_j}, d_{e_j}, u_{e_j} \rangle$. The final belief measure of h_i can be calculated by combining belief measures of all the pieces of evidence corresponding to h_i. That is, answering a critical question about h_i either increases b_{h_i} or d_{h_i}, and decreases u_{h_i}. Wang et al. [13] provide additional details about how belief measures are calculated from evidence.

Defining Policies via Aragorn

Besides general argumentation schemes [12], Aragorn uses domain-specific schemes (e.g., the firewall argumentation scheme), associated critical questions, and related evidence to synthesize arguments.

Constructing Arguments

Table 1 lists requirements underlying a firewall policy, for instance, the Example Inc. needs to enable file transfer. And, one known fact is that FTP and SFTP enable file transfer. To construct arguments for these requirements and premises, we adopt the practical-reasoning argumentation scheme introduced above. From the file-transfer requirement and the fact we conclude that FTP should be enabled. Further, on answering critical questions associated with the scheme, we find that FTP needs port 20. Iteratively applying the scheme, we conclude that we need to allow port 20 to support file transfer.

A useful firewall-specific scheme may be expressed as follows [6].

- Evidence $\mathbb{E}$
- Premise h_1: Port $\mathcal{P}$ on destination $\mathcal{H}$ is required by application $\mathcal{A}$
- Premise (Assumption) h_2: $\mathcal{H}$ is patched to handle vulnerabilities
- Premise (Exception) h_3: Vulnerability $\mathcal{V}$ on port $\mathcal{P}$
- Premise (Exception) h_4: Source $\mathcal{S}$ is malicious
- Conclusion c: Allow packet to port $\mathcal{P}$ on destination $\mathcal{H}$ from $\mathcal{S}$
- Inference Rule δ_1: $\frac{\neg h_3}{h_2}$
- Inference Rule δ_2: $\frac{\neg h_4}{h_2}$
- Inference Rule δ_3: $\frac{h_1 \wedge h_2}{c}$

Inference rules δ_1, δ_2, and δ_3 are general knowledge required by a computational system; each may be reasoned about by other argumentation schemes, which calculate the corresponding triple $\langle b_\delta, d_\delta, u_\delta \rangle$.

The following critical questions are associated with the firewall scheme.

CQ1. Does any requirement rely upon opening port $\mathcal{P}$?

CQ2. Are there any known security vulnerabilities $\mathcal{V}$?

CQ3. Is there any evidence that host $\mathcal{H}$ avoids vulnerability $\mathcal{V}$?

CQ4. Is there any evidence that source $\mathcal{S}$ is malicious?

We could potentially construct arguments (employing additional argumentation schemes) whose conclusions are the above-mentioned premises, e.g., to determine whether a destination is vulnerable.

Associating Evidence and Belief Measure

Consider the file-transfer requirement in Table 1, and the scenario where the administrator needs to determine if file transfer from the source locX.example.com should be allowed. By *argument by practical reasoning*, we see that FTP is a means to file transfer, and conclude that to enable file transfer via FTP, packets to port 20 should be allowed. Also, using a similar argument for SFTP, we see that SFTP is a means to file transfer, and conclude that to enable file transfer via SFTP, packets to port 22 should be allowed. Next, as described above, we apply the firewall scheme and associate the evidence listed in Figure 1c to determine if packets with destinations of ports 20 and 22 should be allowed.

Considering the pieces of evidence and their belief measures in Figure 1c, we calculate the belief measure of the conclusions of arguments $A1$ "Allow packets to port 20" and $A2$ "Allow packets to port 22" as follows.

Argument $A1$: Allow packets to port 20

$$
\begin{aligned}
h_1 &: \langle 0.70, 0.20, 0.10 \rangle && e_1 \\
h_2 &: \left(\frac{\neg h_3}{h_2}\right) \vee \left(\frac{\neg h_4}{h_2}\right) && e_3, e_5, e_7, e_8 \\
&= \langle b_{e_7} d_{e_3}, b_{e_7} b_{e_3}, 1 - b_{e_7} d_{e_3} - b_{e_7} b_{e_3} \rangle \\
&\quad \vee \langle b_{e_8} d_{e_5}, b_{e_8} b_{e_5}, 1 - b_{e_8} d_{e_5} - b_{e_8} b_{e_5} \rangle \\
&= \langle 0.30, 0.34, 0.36 \rangle \\
c &: \frac{h_1 \wedge h_2}{c} && e_9 \\
&= \langle b_{e_9} b_{h_1} b_{h_2}, b_{e_9} (d_{h_1} d_{h_2} + d_{h_1} u_{h_2} + u_{h_1} d_{h_2}), 1 - b_c - d_c \rangle \\
&= \langle 0.1995, 0.1653, 0.6352 \rangle
\end{aligned}
$$

Argument $A2$: Allow packets to port 22

$$
\begin{aligned}
&h_1 : \langle 0.90, 0.05, 0.05\rangle && e_2\\
&h_2 : \left(\frac{\neg h_3}{h_2}\right) \vee \left(\frac{\neg h_4}{h_2}\right) && e_4, e_6, e_7, e_8\\
&\quad = \langle b_{e_7} d_{e_4}, b_{e_7} b_{e_4}, 1 - b_{e_7} d_{e_4} - b_{e_7} b_{e_4}\rangle\\
&\qquad \vee \langle b_{e_8} d_{e_6}, b_{e_8} b_{e_6}, 1 - b_{e_8} d_{e_6} - b_{e_8} b_{e_6}\rangle\\
&\quad = \langle 0.68, 0.04, 0.28\rangle\\
&c : \frac{h_1 \wedge h_2}{c} && e_9\\
&\quad = \langle b_{e_9} b_{h_1} b_{h_2}, b_{e_9}(d_{h_1} d_{h_2} + d_{h_1} u_{h_2} + u_{h_1} d_{h_2}), 1 - b_c - d_c\rangle\\
&\quad = \langle 0.5814, 0.171, 0.2476\rangle
\end{aligned}
$$

Figure 1 shows the belief measure calculation for arguments $A1$ and $A2$ to determine if packets being sent to ports 20 and 22 should be allowed. We observe that argument $A1$ yields a belief with high uncertainty value whereas $A2$ yields high belief with reasonable uncertainty. Therefore, the administrator can allow packets to port 22.

Empirical Evaluation

We empirically evaluated the comparative effectiveness of Aragorn and the traditional rule-based approach (henceforth Trad). We considered following hypotheses.

$\mathbf{H_1}$. Aragorn yields policies of higher coverage than Trad yields.
Null hypothesis: Aragorn has no effect on the coverage of the policies.

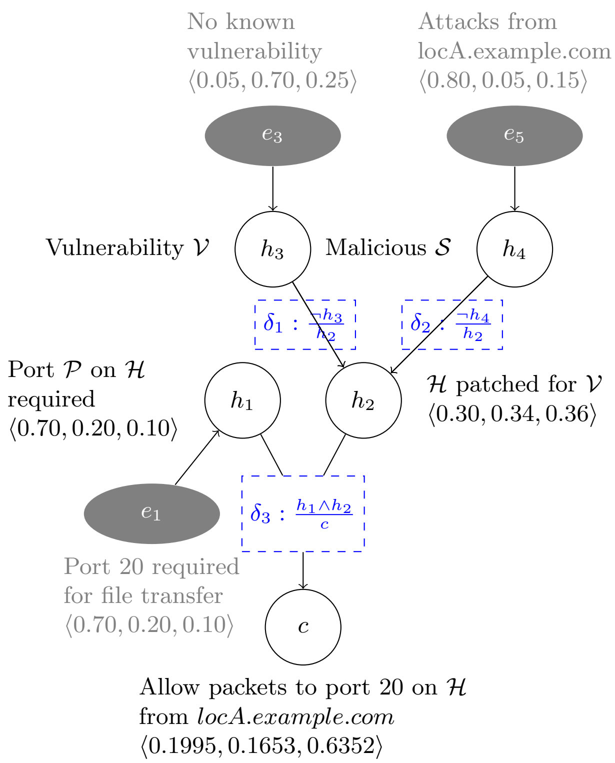
$\mathbf{H_2}$. Aragorn yields policies of greater correctness than Trad yields.
Null hypothesis: Aragorn has no effect on the correctness of the policies.

$\mathbf{H_3}$. Modelers expend less time and effort in defining policies using Aragorn than those using Trad.
Null hypothesis: Aragorn has no effect on the time and effort expended.

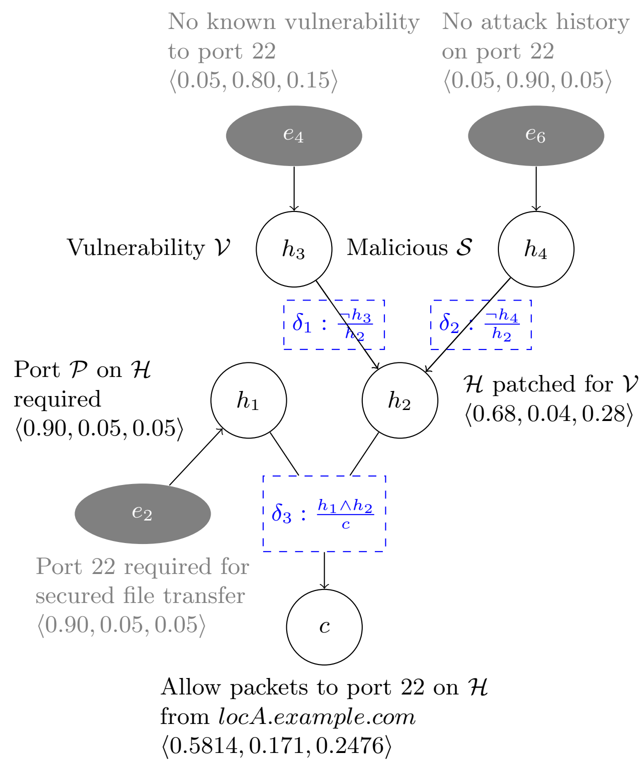
We conducted a human-subject study to evaluate these hypotheses. Our study was approved by North Carolina State University's Institutional Review Board (IRB). We collected an informed consent from each participant and provided a payment of 20 USD to each participant completing the study.

Study Design

We selected 24 computer science (21 graduate, and three undergraduate) students. Each participant had more than three years of programming and software development experience, and was familiar with conceptual modeling, network security, and firewalls. Since network administration is a technology task performed by network engineers, our participants are acceptable surrogates for



(a) Argument $A1$: Allow packets to port 20.



(b) Argument $A2$: Allow packets to port 22.

Evidence	Premise	Belief Measure	Description	Argument
e_1	h_1	$\langle 0.70, 0.20, 0.10\rangle$	Port 20 is required for file transfer via FTP	$A1$
e_2	h_1	$\langle 0.90, 0.05, 0.05\rangle$	Port 22 is required for file transfer via SFTP	$A2$
e_3	h_3	$\langle 0.05, 0.70, 0.25\rangle$	No known vulnerability to port 20	$A1$
e_4	h_3	$\langle 0.05, 0.80, 0.15\rangle$	No known vulnerability to port 22	$A2$
e_5	h_4	$\langle 0.80, 0.05, 0.15\rangle$	Attacks from locA.example.com, port 20	$A1$
e_6	h_4	$\langle 0.05, 0.90, 0.05\rangle$	No attack history on port 22	$A2$
e_7	δ_1	$\langle 0.80, 0.05, 0.15\rangle$	Decision-maker's experience in the rule	$A1$, $A2$
e_8	δ_2	$\langle 0.80, 0.05, 0.15\rangle$	Decision-maker's experience in the rule	$A1$, $A2$
e_9	δ_3	$\langle 0.95, 0.02, 0.03\rangle$	Decision-maker's experience in the rule	$A1$, $A2$

(c) Pieces of evidence corresponding to the file transfer requirement in Table 1.

Figure 1: Example arguments and pieces of evidence.

firewall administrators. Of the participants, 19 had academic or industry experience with network security and firewalls and 16 had academic or industry experience with conceptual modeling.

Our study included three phases and applied the one-factor (approach) design with two alternatives (Trad and Aragorn).

Phase 1: Learn. Participants in each group learned the respective approach by specifying a firewall policy for a hypothetical academic scenario. The academic scenario had eight requirements that participants took into account when specifying a firewall policy.

Phase 2: Design. Each participant specified a firewall policy for a hypothetical enterprise scenario using the approach learned in Phase 1. The scenario had twelve requirements.

Phase 3: Maintain. We provided participants an incomplete solution to the scenario of Phase 2, and asked them to modify that solution to accommodate five additional requirements.

We mitigated two main threats to our study. Specifically, to mitigate the threat of skill differences between participants, we surveyed participants' about their educational backgrounds, their prior experience with conceptual modeling and network security, and their familiarity with defining and maintaining firewall policies. We balanced the groups based on the survey. To mitigate the threat of participants failing to return surveys, we had them complete the difficulty survey after each phase, while it was fresh in their minds, and record their completion time.

We split participants into two groups. The Trad group defined firewall packet filtering rules based on the requirements and the evidence described in the scenario. The Aragorn group used argumentation schemes and critical questions to create an argumentation network consisting multiple arguments (with premises and claims) and to associate evidence with arguments and assign a strength value (a decimal value between 0 and 1) based on their intuition.

Metrics

We analyzed the artifacts produced to measure the following:

Coverage Ratio of the number of requirements satisfied to the total number of requirements in the design and the maintenance phases. Higher is better.

Correctness Ratio of the number of requirements satisfied to the number of requirements attempted. Higher is better.

Specification Quality Product of *coverage* and *correctness*. Higher is better.

Learnability Time in minutes to learn and design the solution. Lower is better.

Maintainability Time in minutes to make changes to an existing solution. Lower is better.

Difficulty in learning Rating by participant on a scale of 1–5 interpreted as very easy, easy, neutral, difficult, and very difficult. Lower is better.

Difficulty in applying Rating by participant on a scale of 1–5 interpreted as very easy, easy, neutral, difficult, and very difficult. Lower is better.

Effort Product of time in minutes to design the solution, and *difficulty in applying*. We define it as a subjective measure to account for both the actual time spend and perceived difficulty to complete the task. Lower is better.

Effort÷Coverage ratio *Effort* divided by *coverage*. We define it as a subjective measure to compute effort required per unit coverage. Lower is better.

These statistical measures address both quality and process. For quality, we measure coverage and correctness, which directly correlate with the human errors. For process, we measure time and difficulty. Effort and Effort÷Coverage ratio are hybrids of quality and process.

Results and Discussion

We evaluate the proposed hypotheses (H_1, H_2, and H_3) using the foregoing metrics, computed from the solutions designed by each participant. We applied the common t-test for means and Wilcoxon's ranksum-test to compare the difference in the medians ($\tilde{x}$). Wilcoxon is stricter than the t-test, and does not assume normality. Table 2 lists the computed values.

Table 2: Empirical results on the effectiveness of Aragorn compared to the traditional approach. (Bold is better.)

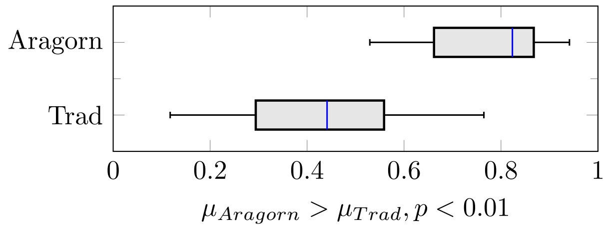
	Statistic	Trad	Aragorn	p
Coverage (in %)	Mean	35.76	**74.54**	< **0.01**
Correctness (in %)	Mean	39.99	**82.93**	< **0.01**
Specification quality (in %)	Mean	14.30	**61.82**	< **0.01**
Learnability (in minutes)	Mean	88.66	**84**	0.45
Maintainability (in minutes)	Mean	**20.08**	29.5	**0.03**
Learning difficulty (1–5)	Median	3	3	0.25
Applying difficulty (1–5)	Median	**3**	4	0.16
Effort	Mean	**158.33**	225.42	**0.02**
Effort÷coverage	Mean	466.09	**291.99**	0.28

Coverage and Correctness We evaluate the coverage and correctness hypotheses (H_1 and H_2) using the coverage, correctness, and quality measures. The mean coverage (74.54%) and mean correctness (82.93%) for Aragorn was found to be significantly higher than the mean coverage (35.76%) and mean correctness (39.99%) obtained for Trad. Also, the p values for coverage and correctness are significant at the 5% level. The specification quality for Aragorn specifications was significantly better than Trad: thus H_1 and H_2 hold.

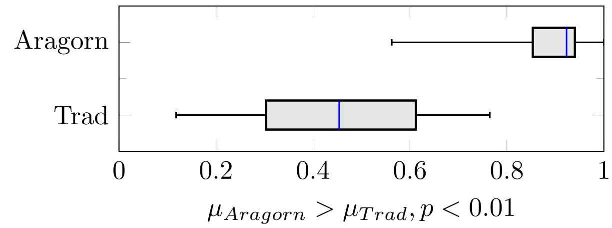
Practical implications. These results suggest that Aragorn's systematic approach (from requirements to general arguments to packet-level arguments to policies) is beneficial. Figures 2a and 2b show the coverage and correctness boxplots based on the numbers of requirements satisfied in Phases 2 and 3.

Time and Effort We evaluate the time and effort hypothesis (H_3) using the learnability, maintainability, difficulty and effort measures.

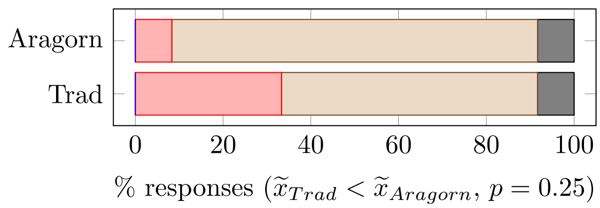
Learnability and Maintainability The average time to learn the approaches and design the solution for the security settings in Phases 1 and 2 was lower for Aragorn (84 minutes) than for Trad (88.6 minutes). However, the time taken to modify an existing solution in Phase 3 was greater for Aragorn (29.5 minutes) than for Trad (20.08 minutes), thus the null hypothesis is not rejected. The result was not surprising because participants in the Aragorn group were required to meticulously answer each critical question in



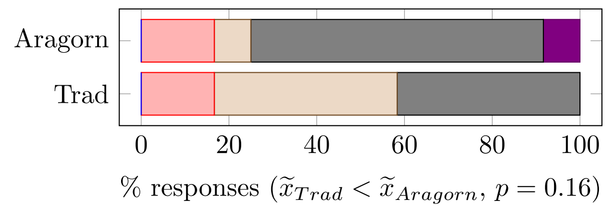
(a) Coverage in Phases 2 and 3.



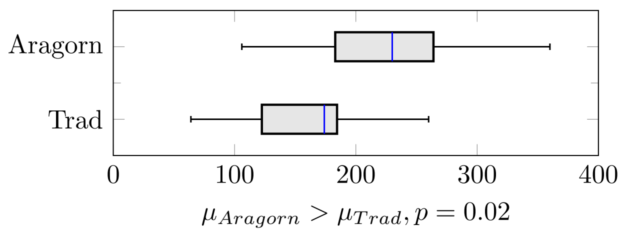
(b) Correctness in Phases 2 and 3.



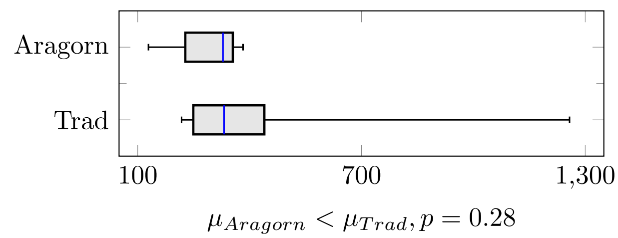
(c) Difficulty in learning.



(d) Difficulty in applying.



(e) Effort: time taken × subjective difficulty.



(f) Effort÷coverage ratio for Phases 2 and 3.

Figure 2: Results.

the argumentation scheme as they made changes to the existing arguments. We can potentially overcome this challenge through improved tool support.

Difficulty Median perceived difficulty to learn was found to be the same for the two groups. Difficulty in applying was higher for Aragorn. This can be attributed to the participants lacking prior experience of working with formal argumentation. Figures 2c and 2d show plots for difficulty perceived by the participants. The p values indicate there is no significant difference in difficulty.

Effort and Effort÷Coverage ratio Participants using Aragorn spent more effort on the task. Figure 2e shows the boxplot for effort expended by the participants. However, as Figure 2f shows, the effort÷coverage ratio was lower for Aragorn (291.99) than for Trad (466.09). We do not include time and effort expended in Phase 1 when computing these parameters. The p values indicate there is no significant difference in the associated effort÷coverage measure.

Practical implications. Although participants using Aragorn spend more time in modifying an existing solution, we find that Aragorn is not difficult to use, and the extra effort expended when using Aragorn yield significant benefits in coverage.

Related Work

Relevant works combine argumentation and security requirements.

Walton et al. [12] provide a foundation for general argumentation schemes, which several frameworks, models, and tools support [6, 8, 10]. But their effectiveness has not been adequately empirically evaluated.

Bentahar et al.'s [3] argumentation-driven approach enables web services to negotiate and persuade peers to join communities, and reason about their commitments.

Franqueira et al. [5] model security requirements and assess risks associated via arguments. Similarly, Ionita et al. [7] analyze risk using an argumentation game. Their tabular representation of arguments appears no more usable or scalable than the traditional firewall representation. These works disregard uncertainty and incompleteness of information, and have not been empirically evaluated.

Conclusions

Aragorn is novel in incorporating argumentation schemes and combining evidence and beliefs to capture the design rationale in service policies.

Our empirical evaluation of Aragorn indicates that Aragorn performs significantly better than the traditional approach. We find that the measures of coverage, correctness, and quality are higher for participants using Aragorn. Aragorn performs on par with the traditional approach on learnability, difficulty, and effort÷coverage. Much as we expected, participants expended more time and effort when using Aragorn, though justified by aggregated improvements in quality. This seeks to the need of better tooling to reduce the time and effort while maintaining the advantages of Aragorn.

Aragorn addresses a general problem in specifying service security policies, namely, to respect the requirements of the various stakeholders despite any conflicts among those requirements. Aragorn provides a way to incorporate evidence where available and to ask critical questions to help orient a search for additional evidence.

An important future direction is to automatically extract evidence supporting and opposing policy arguments from a security policy corpus. Another direction is to adapt existing context-aware requirements elicitation approaches to systematically define, maintain and reason about contextual goals and plans.

Acknowledgments

We thank the US Department of Defense for support through the Science of Security Lablet at NC State University. We also thank the anonymous referees for their helpful comments on the previous versions of this paper.

Author Bios

Nirav Ajmeri is a PhD student in Computer Science at NC State University. His research interests include software engineering and multiagent systems with a focus on security and privacy. Ajmeri has a BE in Computer Engineering from Sardar Vallabhbhai Patel Institute of Technology, Gujarat University. Contact him at najmeri@ncsu.edu.

Chung-Wei Hang is a software engineer at IBM. His research interests include natural language processing, multiagent systems and probabilistic trust models. Hang received a PhD in Computer Science from North Carolina State University, Raleigh. Contact him at chung-wei.hang@gmail.com.

Simon D. Parsons is a Professor of Computer Science and the Vice Dean (Technology) for the Faculty of Natural and Mathematical Science at King's College London. His research interests are in the general area of autonomous systems that includes trust between agents, and argumentation. He is a current Co-Editor-in-Chief of the *Knowledge Engineering Review*. Contact him at simon.parsons@kcl.ac.uk.

Munindar P. Singh is a Professor in Computer Science and a co-director of the Science of Security Lablet at NC State University. His research interests include the engineering and governance of sociotechnical systems. Singh is an IEEE Fellow, a AAAI fellow, a former Editor-in-Chief of *IEEE Internet Computing*, and the current Editor-in-Chief of *ACM Transactions on Internet Technology*. Contact him at singh@ncsu.edu.

References

[1] Ehab Al-Shaer and Hazem Hamed. Discovery of policy anomalies in distributed firewalls. Proc. *Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM)*, pages 2605–2616, 2004. IEEE Computer Society.

[2] Andy Applebaum, Karl Levitt, Jeff Rowe, and Simon Parsons. Arguing about firewall policy. Proc. *International Conference on Computational Models of Argument*, pages 91–102, 2012.

[3] Jamal Bentahar, Zakaria Maamar, Wei Wan, Djamal Benslimane, Philippe Thiran, and Sattanathan Subramanian. Agent-based communities of web services: An argumentation-driven approach. *Service Oriented Computing and Applications*, 2(4):219–238, 2008.

[4] Philippe Besnard and Anthony Hunter. *Elements of Argumentation*. MIT Press, 2008.

[5] Virginia N.L. Franqueira, Thein Than Tun, Yijun Yu, Roel Wieringa, and Bashar Nuseibeh. Risk and argument: A risk-based argumentation method for practical security. Proc. *IEEE International Requirements Engineering Conference*, pages 239–248, 2011.

[6] Thomas Gordon, Henry Prakken, and Douglas Walton. The Carneades model of argument and burden of proof. *Artificial Intelligence*, 171(10–15):875–896, 2007.

[7] Dan Ionita, Jan-Willem Bullee, and Roel Wieringa. Argumentation-based security requirements elicitation. Proc. *IEEE Workshop on Evolving Security and Privacy Requirements Engineering*, pages 7–12, 2014.

[8] Jordan Salvit, Zimi Li, Senni Perumal, Holly Wall, Jennifer Mangels, Simon Parsons, and Elizabeth Sklar. Employing argumentation to support human decision making. Proc. *International Workshop on Argumentation in Multi-Agent Systems*, 2014.

[9] Thomas Shinder. Security considerations for platform as a service (PaaS). http://social.technet.microsoft.com/wiki/contents/articles/3809.security-considerations-for-platform-as-a-service-paas.aspx, November 2013.

[10] Mark Snaith and Chris Reed. TOAST: Online ASPIC$^+$ implementation. Proc. *International Conference on Computational Models of Argument*, pages 509–510, 2012.

[11] Yuqing Tang, Chung-Wei Hang, Simon Parsons, and Munindar Singh. Towards argumentation with symbolic Dempster-Shafer evidence. Proc. *International Conference on Computational Models of Argument*, pages 462–469, 2012.

[12] Douglas Walton, Chris Reed, and Fabrizio Macagno. *Argumentation Schemes*. Cambridge University Press, 2008.

[13] Yonghong Wang, Chung-Wei Hang, and Munindar Singh. A probabilistic approach for maintaining trust based on evidence. *Journal of Artificial Intelligence Research*, 40:221–267, 2011.

[14] Avishai Wool. Trends in firewall configuration errors: Measuring the holes in Swiss cheese. *IEEE Internet Computing*, 14(4):58–65, 2010.