Justification, Stability and Relevance for Transparent and Efficient Human-in-the-Loop Decision Support

Daphne Odekerken[1,2]

[1]Department of Information and Computing Sciences, Utrecht University
[2]National Police Lab AI, Netherlands Police

Abstract

One of the promises of artificial intelligence is improving efficiency in various processes, including decision-making. For specific decisions it is vital that human experts understand and are able to influence machine-made advice. In my dissertation research, I design and study argumentation-based systems for transparent human-in-the-loop decision support. Based on a domain-specific argumentation setting, these systems are able to construct an initial advice on some decision (*justification*); investigate the possibility that additional, yet uncertain, information can change the conclusion (*stability*) and if so, which information is worth investigating (*relevance*). The systems' requirements of detecting justification, stability and relevance correspond to theoretical problems in computational argumentation, most of which are in high complexity classes. In order to achieve reasonable estimations for these problems in polynomial time, I develop and investigate not only exact algorithms but also approximations.

1 Introduction

Artificial intelligence (AI) is developing fast, promising quality and efficiency benefits in various processes where (repetitive) tasks are outsourced from human analysts to machines. At the Netherlands Police, two examples of tasks in which AI could be helpful are the intake of complaints and the classification of suspect web shops. Especially for complex or high-risk decision-making tasks it is essential that domain experts understand machine-made decisions or advice and have the possibility to correct possible mistakes [European Commission, 2021]. In addition, there are decisions that require input that cannot be obtained by a machine; instead, an analyst has to be consulted. We therefore need AI systems that are able to explain their decisions and to keep the human in the loop. Since computational argumentation is a research topic concerning reasoning with incomplete or inconsistent information in a way similar to human reasoning [Atkinson et al., 2017], it seems promising to use argumentation-based techniques for developing the required systems. In my dissertation research, I propose a general decision-making approach centered around three theoretical problems in computational argumentation: justification, stability and relevance.

Informally, the problem (or task) of determining *justification* can be seen as giving an initial advice regarding a specific topic, thereby only considering information that is currently available. The topic satisfies *stability* if the corresponding advice will not change, regardless of currently unavailable information that could still be added and/or currently available information that could be removed. In situations where the topic is not stable, one should identify information that is *relevant*, in the sense that investigation into its presence possibly leads to a stable topic. I will define these three problems on three settings: structured and abstract argumentation frameworks, as well as precedent-based reasoning.

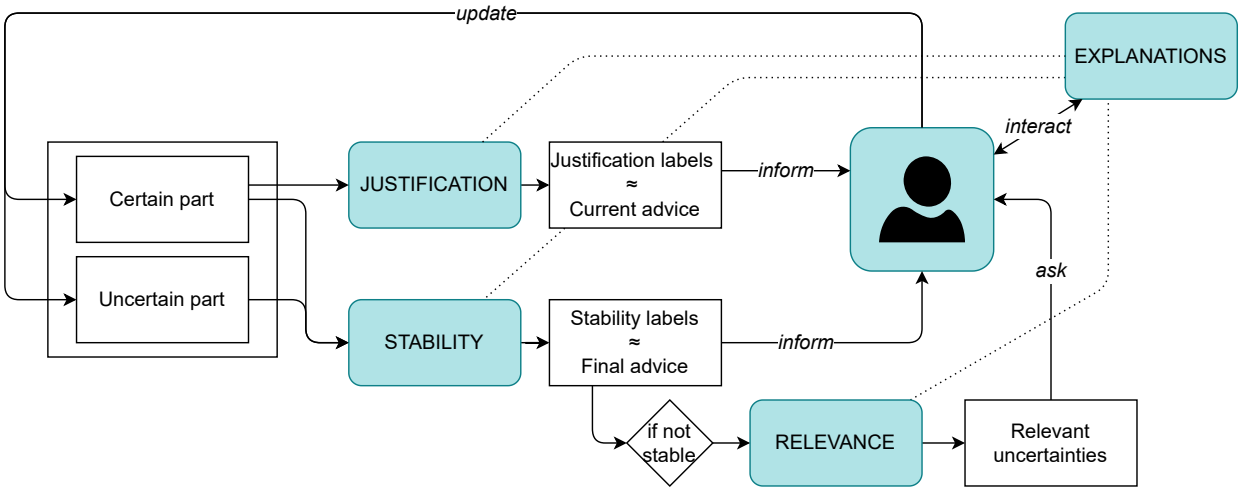

Figure 1: High-level overview of the proposed human-in-the-loop decision-making process, involving algorithms for justification, stability and relevance identification.

2 Method

Before formally defining the problems of justification, stability and relevance, I first outline the proposed human-in-the-loop decision-making procedure, illustrated in Figure 1, which relies on practical solutions for these theoretical problems. First, an algorithm for justification is used to obtain an initial decision or advice, given only the current information. Additionally, an algorithm for stability detection is applied on both the current and yet uncertain information; if this algorithm identifies a stable situation, the advice can be considered final. Otherwise, the algorithm for relevance identifies those uncertainties that should be investigated by the analyst. After investigation, the analyst can return findings to the system. This process is repeated until a final advice is found (or the analyst decides not to investigate any further).

In the remainder of this section, I define the problems of justification, stability and relevance on both structured and abstract argumentation frameworks, as well as for *a fortiori reasoning* based on precedents. Due to limited space, I only give formal definitions of these problems for incomplete (abstract) argumentation frameworks and provide an informal description for the other two settings.

2.1 Definitions for IAFs

An argumentation framework (AF) $\langle \mathcal{A}, \mathcal{R} \rangle$ consists of a set $\mathcal{A}$ of arguments and attack relation $\mathcal{R} \subseteq \mathcal{A} \times \mathcal{A}$, where $(A, B) \in \mathcal{R}$ indicates that argument A attacks argument B [Dung, 1995]. Given a specific semantics (see e.g. [Baroni et al., 2011]), one can determine the justification status of arguments in an AF, based on their presence in an extension (i.e. set of arguments):

Definition 1 (Argument justification status)**.** *Let $AF = \langle \mathcal{A}, \mathcal{R} \rangle$ be an argumentation framework and σ some semantics. Let A be some argument in $\mathcal{A}$.*

- *A is σ-scep-*IN *(resp. σ-cred-*IN*) iff A belongs to each (resp. some) σ-extension of AF;*
- *A is σ-scep-*OUT *(resp. σ-cred-*OUT*) iff for each (resp. some) σ-extension S of AF, A is attacked by some argument in S;*
- *A is σ-scep-*UNDEC *(resp. σ-cred-*UNDEC*) iff for each (resp. some) σ-extension of AF, A is not in S, nor attacked by any argument in S.*

Incomplete argumentation frameworks (IAFs) are an extension to AFs that encode qualitative uncertainty regarding the presence of arguments and attacks [Baumeister et al., 2021]. An IAF is a tuple

$\langle \mathcal{A}, \mathcal{A}^?, \mathcal{R}, \mathcal{R}^? \rangle$, where $\mathcal{A} \cap \mathcal{A}^? = \emptyset$, $\mathcal{R} \cap \mathcal{R}^? = \emptyset$; $\mathcal{A}$ is the set of certain arguments; $\mathcal{A}^?$ is the set of uncertain arguments; $\mathcal{R} \subseteq (\mathcal{A} \cup \mathcal{A}^?) \times (\mathcal{A} \cup \mathcal{A}^?)$ is the certain attack relation and $\mathcal{R}^? \subseteq (\mathcal{A} \cup \mathcal{A}^?) \times (\mathcal{A} \cup \mathcal{A}^?)$ is the uncertain attack relation. An IAF can be *specified* by investigating the presence of uncertain elements: a specification is an IAF $\langle \mathcal{A}', \mathcal{A}^{?\prime}, \mathcal{R}', \mathcal{R}^{?\prime} \rangle$, where $\mathcal{A} \subseteq \mathcal{A}' \subseteq \mathcal{A} \cup \mathcal{A}^?$; $\mathcal{R} \subseteq \mathcal{R}' \subseteq \mathcal{R} \cup \mathcal{R}^?$; $\mathcal{A}^{?\prime} \subseteq \mathcal{A}^?$ and $\mathcal{R}^{?\prime} \subseteq \mathcal{R}^?$. Given some IAF $\mathcal{I}$, we denote all possible specifications for $\mathcal{I}$ by $F(\mathcal{I})$.

In [Odekerken et al., 2022b], we introduced the notion of stability for IAFs, where an argument is stable if and only if its justification status remains the same, regardless of the way the uncertain arguments and attacks would turn out to be present or absent.

Definition 2 (Stability)**.** *Given an IAF $\mathcal{I} = \langle \mathcal{A}, \mathcal{A}^?, \mathcal{R}, \mathcal{R}^? \rangle$ with $A \in \mathcal{A}$ and justification status j, A is* stable-j *w.r.t. $\mathcal{I}$ iff for each $\langle \mathcal{A}', \mathcal{A}^{?\prime}, \mathcal{R}', \mathcal{R}^{?\prime} \rangle$ in $F(\mathcal{I})$, A is j w.r.t. $\langle \mathcal{A}', \mathcal{R}' \cap (\mathcal{A}' \times \mathcal{A}') \rangle$.*

In situations where the topic argument is stable, one can give a final advice; in all other situations, further investigation could lead to a different advice. In order to decide which arguments or attacks are worth investigating, we define relevance on IAFs in Definition 4. The notion of relevance is defined based on minimal stable specifications, which we introduced in [Odekerken et al., 2022b] and recall next.

Definition 3 (Minimal stable specification)**.** *Given an IAF $\mathcal{I} = \langle \mathcal{A}, \mathcal{A}^?, \mathcal{R}, \mathcal{R}^? \rangle$, a certain argument $A \in \mathcal{A}$ and a justification status j, a minimal stable-j specification for A w.r.t. $\mathcal{I}$ is a specification $\mathcal{I}' = \langle \mathcal{A}', \mathcal{A}^{?\prime}, \mathcal{R}', \mathcal{R}^{?\prime} \rangle$ in $F(\mathcal{I})$ such that A is stable-j in $\mathcal{I}'$ and there is no $\mathcal{I}'' = \langle \mathcal{A}'', \mathcal{A}^{?\prime\prime}, \mathcal{R}'', \mathcal{R}^{?\prime\prime} \rangle$ in $F(\mathcal{I})$ such that A is stable-j in $\mathcal{I}''$, $\mathcal{I}'' \neq \mathcal{I}'$ and $\mathcal{I}' \in F(\mathcal{I}'')$.*

Definition 4 (Relevance)**.** *Given an IAF $\mathcal{I} = \langle \mathcal{A}, \mathcal{A}^?, \mathcal{R}, \mathcal{R}^? \rangle$, certain argument $A \in \mathcal{A}$, uncertain element $U \in \mathcal{A}^? \cup \mathcal{R}^?$ and justification status j,*

- *Addition of U is j-relevant for A w.r.t. $\mathcal{I}$ iff there is some minimal stable-j specification $\langle \mathcal{A}', \mathcal{A}^{?\prime}, \mathcal{R}', \mathcal{R}^{?\prime} \rangle$ for A w.r.t. $\mathcal{I}$ such that $U \in \mathcal{A}' \cup \mathcal{R}'$;*
- *Removal of U is j-relevant for A w.r.t. $\mathcal{I}$ iff there is some minimal stable-j specification $\langle \mathcal{A}', \mathcal{A}^{?\prime}, \mathcal{R}', \mathcal{R}^{?\prime} \rangle$ for A w.r.t. $\mathcal{I}$ such that $U \notin \mathcal{A}' \cup \mathcal{A}^{?\prime} \cup \mathcal{R}' \cup \mathcal{R}^{?\prime}$.*

Note that this argumentation-based approach offers ample opportunity for explanation. For example, the justification status IN of some argument A can be explained by returning one or more extensions containing A (see e.g. [Borg and Bex, 2021]). A stability status can be explained by showing all specifications and appropriate extensions, but, alternatively, also by returning a minimal stable specification containing A. One way of explaining relevance of U for A would be to give a specification where U is certainly present and A has a given justification status j, whereas A would not be j in the variation in which U is certainly absent.

2.2 Structured argumentation

Similar to (abstract) AFs, the notions of justification status, stability and relevance can be defined on (a dynamic version of) structured argumentation frameworks, such as ASPIC$^+$. In [Odekerken et al., 2020], we extended ASPIC$^+$ argumentation theories with a set of queryables $\mathcal{Q}$, containing those literals for which it is not yet known if they will be added to the knowledge base. Using the definition of justification for ASPIC$^+$ [Modgil and Prakken, 2018], we defined stability on this extension in [Odekerken et al., 2020]. In future work, I plan to define and study the notion of relevance for this ASPIC$^+$ extension as well.

2.3 Precedent-based reasoning

The idea of precedent-based reasoning is that decisions on cases generalise to, and thereby restrict possible outcomes of new, yet undecided, cases [Horty, 2011]. Normally, it is assumed that the factors of a new case are certain and this case is compared to a case base with earlier cases and the corresponding decisions. We view the task of decision-making given only certain information as the justification problem. However, in reality it is not always known which factors are present in a case: sometimes, this can only be determined by additional investigation. Just like for abstract and structured argu-

mentation frameworks, the framework for precedent-based reasoning can be extended with an uncertain component, on which the problems of stability and relevance can be defined. Specifically, we assume that the factors of cases in the case base are all certain, while a new case may have a combination of certain and uncertain factors. Whereas the justification task is to make a decision based on the *certain* factors of the case and the case base, the stability task is to decide if the addition or removal of *uncertain* information could still change this outcome; if so, identifying this information corresponds to the relevance task.

3 Computational Issues

The decision-making procedure proposed in the previous section requires efficient algorithms for detecting justification status, stability and relevance. However, formal complexity analysis reveals that most of the problems are in high complexity classes. For example, in [Odekerken et al., 2020] we proved that the problem of detecting stability is CoNP-complete under grounded semantics, given a simple implementation of ASPIC$^+$. In [Odekerken et al., 2022b] we studied stability and relevance problems for IAFs and observed that these are highly complex as well.

It is therefore far from trivial to find a fast solution for each instantiation of the problem. This issue can be handled in various ways, as shown in the argumentation literature for (other) problems in high complexity classes. One possible solution are exact algorithms based on SAT-solvers (see e.g. [Baumeister et al., 2021]) or Answer Set Programming (e.g. [Lehtonen et al., 2020]). These algorithms may be relatively fast in practice, but fast computation is not guaranteed as they are exponential. A second option are approximation algorithms that are learned from data [Craandijk and Bex, 2020]: once trained, such algorithms are fast and quite accurate, but a theoretical accuracy analysis is not possible. I am therefore particularly interested in a third alternative: developing approximation algorithms that can be evaluated not only empirically, but also theoretically. In [Odekerken et al., 2020], we describe an approximation algorithm for estimating stability and show that it is polynomial and sound, but not complete. In [Odekerken et al., 2022a], we compare this algorithm to an exact algorithm. In future work, I plan to develop and study approximation algorithms for the relevance problem as well. To assess the performance of these algorithms, I plan to conduct a theoretical analysis of time complexity and identify when the algorithm gives an exact solution, similar to our stability study in [Odekerken et al., 2022a].

At this moment, the general human-in-the-loop decision-making procedure is already applied for two specific applications at the police: an ASPIC$^+$-based inquiry system that helps citizens to decide if they should submit a complaint on online trade fraud [Schraagen et al., 2018, Testerink et al., 2019, Odekerken et al., 2020], as well as a human-in-the-loop classifier of suspect web shops, based on a combination of structured argumentation and precedent-based reasoning [Odekerken and Bex, 2020]. In a future user study, I will investigate the analysts' experience and performance using the latter system, studying e.g. how they use the suggestions for relevance.

4 Conclusion

The application of argumentation-based techniques is a promising approach towards transparent human-in-the-loop support for complex or high-risk decisions. This abstract summarized my proposal for a general approach for decision support, centered around three theoretical problems: justification, stability and relevance. These problems are defined for various settings in computational argumentation. Given that most of these problems are situated in high complexity classes, I develop algorithms that obtain an estimation in polynomial time and evaluate them empirically and theoretically. The algorithms are applied in decision support systems at the Netherlands Police.

Acknowledgements

I would like to thank Floris Bex, AnneMarie Borg and Henry Prakken for their support throughout my PhD project.

References

[Atkinson et al., 2017] Atkinson, K., Baroni, P., Giacomin, M., Hunter, A., Prakken, H., Reed, C., Simari, G., Thimm, M., and Villata, S. (2017). Towards artificial argumentation. *AI magazine*, 38(3):25–36.

[Baroni et al., 2011] Baroni, P., Caminada, M., and Giacomin, M. (2011). An introduction to argumentation semantics. *The Knowledge Engineering Review*, 26(4):365–410.

[Baumeister et al., 2021] Baumeister, D., Järvisalo, M., Neugebauer, D., Niskanen, A., and Rothe, J. (2021). Acceptance in incomplete argumentation frameworks. *Artificial Intelligence*, 295:103470.

[Borg and Bex, 2021] Borg, A. and Bex, F. (2021). A basic framework for explanations in argumentation. *IEEE Intelligent Systems*, 36(2):25–35.

[Craandijk and Bex, 2020] Craandijk, D. and Bex, F. (2020). Deep learning for abstract argumentation semantics. *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pages 1667–1673.

[Dung, 1995] Dung, P. M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77:321–357.

[European Commission, 2021] European Commission (2021). Proposal for a regulation of the European Parliament and of the council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts. `https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206` [Online; accessed 17 June 2022].

[Horty, 2011] Horty, J. F. (2011). Rules and reasons in the theory of precedent. *Legal Theory*, 17:1–33.

[Lehtonen et al., 2020] Lehtonen, T., Wallner, J. P., and Järvisalo, M. (2020). An answer set programming approach to argumentative reasoning in the ASPIC+ framework. In *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning*, volume 17, pages 636–646.

[Modgil and Prakken, 2018] Modgil, S. and Prakken, H. (2018). Abstract rule-based argumentation. In Baroni, P., Gabbay, D., Giacomin, M., and van der Torre, L., editors, *Handbook of Formal Argumentation*, volume 1, pages 286–361. College Publications.

[Odekerken and Bex, 2020] Odekerken, D. and Bex, F. (2020). Towards transparent human-in-the-loop classification of fraudulent web shops. In *Legal Knowledge and Information Systems*, pages 239–242.

[Odekerken et al., 2022a] Odekerken, D., Bex, F., Borg, A., and Testerink, B. (2022a). Approximating stability for applied argument-based inquiry. *Intelligent Systems with Applications*, 16:200110.

[Odekerken et al., 2020] Odekerken, D., Borg, A., and Bex, F. (2020). Estimating stability for efficient argument-based inquiry. In *Computational Models of Argument. Proceedings of COMMA 2020*, pages 307–318.

[Odekerken et al., 2022b] Odekerken, D., Borg, A., and Bex, F. (2022b). Stability and relevance in incomplete argumentation frameworks. In *Computational Models of Argument. Proceedings of COMMA 2022*, pages 272–283.

[Schraagen et al., 2018] Schraagen, M., Bex, F., Odekerken, D., and Testerink, B. (2018). Argumentation-driven information extraction for online crime reports. In *Proceedings of International Workshop on Legal Data Analytics and Mining*, pages 20–25.

[Testerink et al., 2019] Testerink, B., Odekerken, D., and Bex, F. (2019). A method for efficient argument-based inquiry. In *Proceedings of the 13th International Conference on Flexible Query Answering Systems*, pages 114–125. Springer International Publishing.