Improving Explanations by Integrating Preferences*

AnneMarie Borg[1][0000−0002−7204−6046] and Floris Bex[1,2][0000−0002−5699−9656]

[1] Department of Information and Computing Sciences, Utrecht University
[2] Tilburg Institute for Law, Technology, and Society, Tilburg University
{A.Borg,F.J.Bex}@UU.nl

Abstract. The use of preferences is an important aspect of human reasoning and the modeling of preferences is one of the main research branches of the formal study of non-monotonic reasoning. However, from research in the social sciences it is known that explanations based on preferences are not as effective as explanations based on causes. As a result, explanations for argumentation-based conclusions usually do not take preferences into account. We will discuss the importance of accounting for preferences in argumentation-based explanations and then integrate preferences into an existing explanation method.

In the formal study of non-monotonic reasoning the use of preferences and uncertainty has extensively been studied, including in the context of computational argumentation [1]. The result is a wealth of research results, investigating and solving a wide variety of problems that can and should be applied in other research areas of modeling (non-monotonic) reasoning as well.

Interestingly, one of the main conclusions in [8, p. 4] is that "probabilities probably don't matter" when explaining an AI-based decision. Rather, humans prefer explanations based on causes. Consequently, in many explainable artificial intelligence (XAI) approaches, including those based on computational argumentation [4, 12], the use of preferences in the decision making process is mostly ignored. However, preferences play an important role in the derivation and evaluation of argumentation-based conclusions.

Abstract argumentation frameworks [5] are pairs $\mathcal{AF} = \langle \text{Args}, \mathcal{A} \rangle$ where Args is a set of arguments and $\mathcal{A} \subseteq \text{Args} \times \text{Args}$ is an attack relation on these arguments. Given two arguments $A, B \in \text{Args}$, A *attacks* B if $(A, B) \in \mathcal{A}$ and A *defends* B if A attacks an attacker of B or defends an argument that defends B. A set of arguments $\mathsf{S} \subseteq \text{Args}$ is a *complete extension* if there are no $A, B \in \mathsf{S}$ such that $(A, B) \in \mathcal{A}$, it defends all its arguments and contains all the arguments it defends. A *preferred extension* of $\mathcal{AF}$ is a maximal (w.r.t. $\subseteq$) complete extension.

Example 1. Consider an abstract argumentation framework [5] with two arguments A and B that attack each other ($\mathcal{AF}_1$ in Fig. 1). In this case a skeptical

* This research was partially funded by the Dutch Ministry of Justice and the Netherlands Police.

reasoner will not accept any argument (i.e., the minimal (w.r.t. $\subseteq$) complete extension is empty), while a credulous reasoner might choose one of the arguments to accept (i.e., there are two preferred extensions: $\{A\}$ and $\{B\}$). Now, if A is preferred over B (denoted by $B < A$) and, as a consequence, the attack from B to A is no longer successful ($\mathcal{AF}_2$ in Fig. 1), both skeptical and credulous reasoners might accept A as the only conclusion.

Fig. 1. Graphical representation of the argumentation frameworks from Example 1.

The preference for A over B in the above scenario is not taken into account in most of the literature on explanations for argumentation-based conclusions (see, e.g., [3, 6, 7, 11]), yet it is the preference relation that changes the acceptability of the arguments in a crucial way. We therefore propose an explanation method in which the preference relation is part of the explanation.

In this work we extended the basic framework for explanations from [3] with a preference component. We choose this particular explanations framework since both acceptance and non-acceptance for both abstract argumentation [5] and ASPIC$^+$ [10] can be provided by it. Given the space restrictions we discuss here only explanations for a credulously accepted argument in abstract argumentation under preferred semantics. We denote by $\mathsf{Defending}(A, \mathcal{E}) = \{B \in \text{Args} \mid B \text{ defends } A\}$ the set of all arguments that defend A in the preferred extension $\mathcal{E}$. An acceptance explanation for an argument A contains all the arguments that defend A in a preferred extension.

Definition 1. *Let $\mathcal{AF} = \langle Args, \mathcal{A} \rangle$ be an abstract argumentation framework and suppose that $A \in Args$ is part of a preferred extension. Then:* $\mathsf{Acc}(A) = \{\mathsf{Defending}(A, \mathcal{E}) \mid \mathcal{E}$ *is a preferred extension of* $\mathcal{AF}$ *and* $A \in \mathcal{E}\}$.

At the Netherlands Police several argumentation-based applications have been implemented [2]. These applications are aimed at assisting the police at working through high volume tasks, leaving more time for tasks that require human attention. We will illustrate the notions in this abstract with an example based on a real-life ASPIC$^+$-based application for online trade fraud (see [9] for more details on the application). This application can receive a variety of input information, for the sake of simplicity we provide here arguments rather than the underlying knowledge base and set of rules. The goal of the system is to determine whether it is a case of fraud or not.

Example 2. Consider the following arguments: the complainant delivered (A_1) or not (A_6); the counterparty delivered (A_3) or not (A_2); and the received product is fake (A_4) or not (A_5). From this input, based on Dutch Criminal Law (i.e.,

Article 326) additional arguments can be derived: if the complainant delivered and the counterparty did not it is a case of fraud (B_1); if the counterparty did deliver but the product is fake, the received product is fake (B_2) and if then the complainant did deliver as well it is a case of fraud (B_5); if the complainant did not deliver it is not a case of fraud (B_3); and if the complainant delivered and the received product is not fake it is not a case of fraud (B_4). The attacks between the arguments are based on the underlying structure: see Fig. 2.

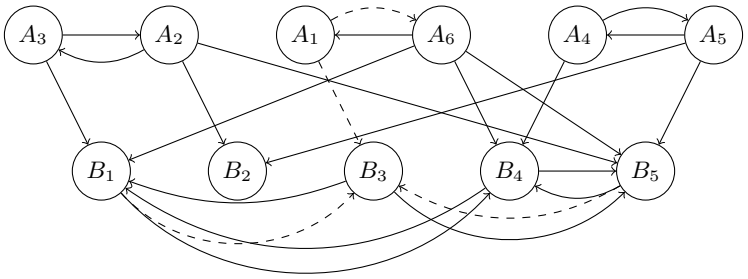

Fig. 2. Graphical representation of the argumentation framework from Example 2.

A credulous reasoner under preferred semantics can accept arguments for a case of fraud as well as arguments for not a case of fraud. This follows since we have the following preferred extensions: $\{A_1, A_2, A_4, B_1\}$, $\{A_1, A_2, A_5, B_1\}$, $\{A_1, A_2, A_5, B_4\}$, $\{A_1, A_3, A_4, B_2, B_5\}$, $\{A_1, A_3, A_5, B_4\}$, $\{A_2, A_4, A_6, B_3\}$, $\{A_2, A_5, A_6, B_3\}$, $\{A_3, A_4, A_6, B_2, B_4\}$ and $\{A_3, A_5, A_6, B_3\}$. Some example explanations are then: $\mathsf{Acc}(B_1) = \{\{A_1, A_2, B_1\}, \{A_1, A_2, A_4, B_1\}\}$ and $\mathsf{Acc}(B_3) \in \{\{A_3, A_5, A_6, B_3\}, \{A_5, A_6, B_3\}, \{A_6, B_3\}\}$.

In order to integrate the relevant preference relations between arguments into the explanations, we extend the basic explanations from [3] such that explanations become pairs: the first element contains the set of arguments as in Definition 1 and the second element contains the relevant preference relations.

Definition 2. *Let $\mathcal{AF} = \langle Args, \mathcal{A} \rangle$ be an abstract argumentation framework, $<$ a preference relation over Args and suppose that $A \in Args$ is part of a preferred extension. Then* $\mathsf{Acc}(A) = \{(\mathsf{Defending}(A, \mathcal{E}), \Theta) \mid \mathcal{E}$ *is a preferred extension of* $\mathcal{AF}$*,* $A \in \mathcal{E}$ *and* $(B, C) \in \Theta$ *iff* $C \in \mathsf{Defending}(A, \mathcal{E})$ *or* $C = A$*,* $(B, C) \in \mathcal{A}$ *and* $B < C\}$.

In words, the second element of an explanation is a set of pairs of arguments, such that the second argument is part of the basic explanation and the attack from the first argument is not successful due to the preference relation.

Example 3. According to Dutch Criminal Law, there is only a case for online trade fraud if the complainant delivered (e.g., paid or sent the agreed upon

goods). We therefore assign a higher priority to A_6 than to $\{A_1, \ldots, A_5\}$. In the resulting framework $\mathcal{AF}'$ we have that B_3 is no longer attacked (i.e., the dashed attacks in Fig. 2 are no longer successful). Moreover, there are only four preferred extensions left: $\{A_2, A_4, A_6, B_3\}$, $\{A_2, A_5, A_6, B_3\}$, $\{A_3, A_4, A_6, B_2, B_3\}$ and $\{A_3, A_5, A_6, B_3\}$. For the acceptance of not a case for fraud, i.e., the acceptance of B_3, we obtain the following explanation: $\mathsf{Acc}(B_3) = (\emptyset, \{B_1 < B_3, B_5 < B_3\})$. Intuitively, this explanation shows that B_3 is not attacked because of the preferences $B_1 < B_3$ and $B_5 < B_3$.

The resulting explanation in the above example is still rather difficult to parse. In our talk we will present this example with the underlying structure and show how the introduced integration of preferences will help to explain that it is not a case of fraud because the complainant did not deliver, which is a requirement by law.

References

1. Beirlaen, M., Heyninck, J., Pardo, P., Straßer, C.: Argument strength in formal argumentation. Journal of Applied Logics – IfCoLog **5**(3), 629–676 (2018)
2. Bex, F., Testerink, B., Peters, J.: AI for online criminal complaints: From natural dialogues to structured scenarios. In: Workshop proceedings of Artificial Intelligence for Justice at ECAI 2016. pp. 22–29 (2016)
3. Borg, A., Bex, F.: A basic framework for explanations in argumentation. IEEE Intelligent Systems **36**(2), 25–35 (2021)
4. Čyras, K., Rago, A., Albini, E., Baroni, P., Toni, F.: Argumentative XAI: A survey pp. 4392–4399 (2021)
5. Dung, P.M.: On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. Artificial Intelligence **77**(2), 321–357 (1995)
6. Fan, X., Toni, F.: On computing explanations in argumentation. In: Bonet, B., Koenig, S. (eds.) Proceedings of AAAI'15. pp. 1496–1502. AAAI Press (2015)
7. Liao, B., van der Torre, L.: Explanation semantics for abstract argumentation. In: Prakken, H., Bistarelli, S., Santini, F., Taticchi, C. (eds.) Proceedings of COMMA'20. pp. 271–282. IOS Press (2020)
8. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence **267**, 1–38 (2019)
9. Odekerken, D., Borg, A., Bex, F.: Estimating stability for efficient argument-based inquiry. In: Prakken, H., Bistarelli, S., Santini, F., Taticchi, C. (eds.) Proceedings of COMMA'20. pp. 307–318. IOS Press (2020)
10. Prakken, H.: An abstract framework for argumentation with structured arguments. Argument & Computation **1**(2), 93–124 (2010)
11. Saribatur, Z., Wallner, J., Woltran, S.: Explaining non-acceptability in abstract argumentation. In: Proceedings of ECAI'20. pp. 881–888. IOS Press (2020)
12. Vassiliades, A., Bassiliades, N., Patkos, T.: Argumentation and explainable artificial intelligence: A survey. The Knowledge Engineering Review **36**, e5 (2021). https://doi.org/10.1017/S0269888921000011