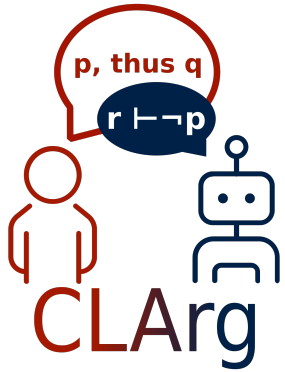
Imperial College London



On Monotonicity of Dispute Trees as Explanations for Case-Based Reasoning with Abstract Argumentation

Guilherme Paulino-Passos, Francesca Toni

Department of Computing
Computational Logic and Argumentation Group

ArgXAI – COMMA 2022

Motivation

- Abstract argumentation for case-based reasoning: AA-CBR

Motivation

- Abstract argumentation for case-based reasoning: AA-CBR
  - modelling case-based reasoning with argumentation

Motivation

- Abstract argumentation for case-based reasoning: AA-CBR
  - modelling case-based reasoning with argumentation
- Interactive explanations as reasoning

Motivation

- Abstract argumentation for case-based reasoning: AA-CBR
  - modelling case-based reasoning with argumentation
- Interactive explanations as reasoning
  - explanations can be seen as informing what behaviour is expected from classifiers for some other inputs

Motivation

- Abstract argumentation for case-based reasoning: AA-CBR
  - modelling case-based reasoning with argumentation
- Interactive explanations as reasoning
  - explanations can be seen as informing what behaviour is expected from classifiers for some other inputs
- Arbitrated dispute trees have been proposed as explanations for AA-CBR

Motivation

- Abstract argumentation for case-based reasoning: AA-CBR
  - modelling case-based reasoning with argumentation
- Interactive explanations as reasoning
  - explanations can be seen as informing what behaviour is expected from classifiers for some other inputs
- Arbitrated dispute trees have been proposed as explanations for AA-CBR

This work

Motivation

- Abstract argumentation for case-based reasoning: AA-CBR
  - modelling case-based reasoning with argumentation
- Interactive explanations as reasoning
  - explanations can be seen as informing what behaviour is expected from classifiers for some other inputs
- Arbitrated dispute trees have been proposed as explanations for AA-CBR

This work

- can arbitrated dispute trees be given an inferential interpretation?

Motivation

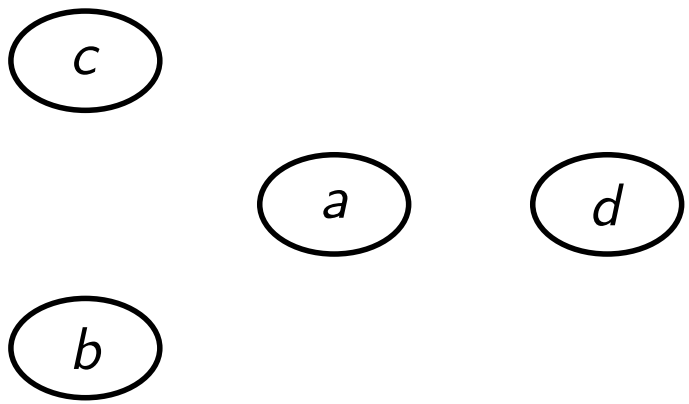
- Abstract argumentation for case-based reasoning: AA-CBR
  - modelling case-based reasoning with argumentation
- Interactive explanations as reasoning
  - explanations can be seen as informing what behaviour is expected from classifiers for some other inputs
- Arbitrated dispute trees have been proposed as explanations for AA-CBR

This work

- can arbitrated dispute trees be given an inferential interpretation?
- what do they reveal about predictions for other inputs?

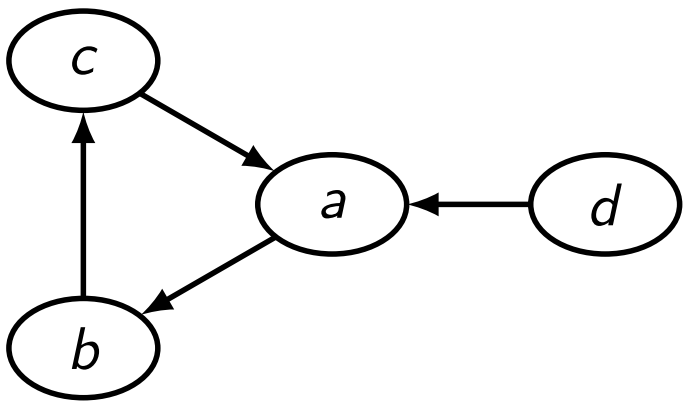
Abstract argumentation in brief

- Arguments



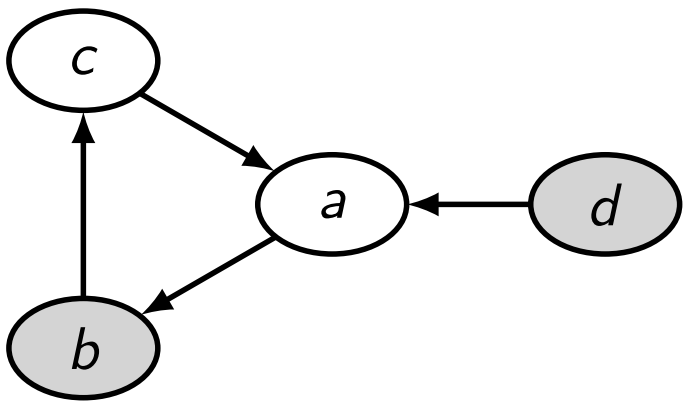
Abstract argumentation in brief

- Arguments
- Attacks



Abstract argumentation in brief

- Arguments
- Attacks
- Grounded extension



Abstract argumentation for case-based reasoning

- Modelling case-based reasoning with argumentation[1]

[1] Kristijonas Čyras, Ken Satoh, and Francesca Toni. "Abstract Argumentation for Case-Based Reasoning". In: *Principles of Knowledge Representation and Reasoning: Proceedings of the Fifteenth International Conference, KR 2016, Cape Town, South Africa, April 25-29, 2016*. Ed. by Chitta Baral, James P. Delgrande, and Frank Wolter. AAAI Press, 2016, pp. 549–552.

[2] Oana Cocarascu et al. "Data-Empowered Argumentation for Dialectically Explainable Predictions". In: *ECAI 2020 - 24th European Conference on Artificial Intelligence, Santiago de Compostela, Spain, 10-12 June 2020*. 2020.

[3] Ibid.

[4] Kristijonas Čyras et al. "Explanations by arbitrated argumentative dispute". In: *Expert Syst. Appl.* 127 (2019), pp. 141–156.

Abstract argumentation for case-based reasoning

- Modelling case-based reasoning with argumentation[1]
- Inspiration from legal domain

[1] Čyras, Satoh, and Toni, "Abstract Argumentation for Case-Based Reasoning".

[2] Cocarascu et al., "Data-Empowered Argumentation for Dialectically Explainable Predictions".

[3] Ibid.

[4] Čyras et al., "Explanations by arbitrated argumentative dispute".

Abstract argumentation for case-based reasoning

- Modelling case-based reasoning with argumentation[1]
- Inspiration from legal domain
- Some of those approaches have been used as classifiers in different scenarios:

[1] Čyras, Satoh, and Toni, "Abstract Argumentation for Case-Based Reasoning".

[2] Cocarascu et al., "Data-Empowered Argumentation for Dialectically Explainable Predictions".

[3] Ibid.

[4] Čyras et al., "Explanations by arbitrated argumentative dispute".

Abstract argumentation for case-based reasoning

- Modelling case-based reasoning with argumentation[1]
- Inspiration from legal domain
- Some of those approaches have been used as classifiers in different scenarios:
  - image classification[2]

[1] Čyras, Satoh, and Toni, "Abstract Argumentation for Case-Based Reasoning".

[2] Cocarascu et al., "Data-Empowered Argumentation for Dialectically Explainable Predictions".

[3] Ibid.

[4] Čyras et al., "Explanations by arbitrated argumentative dispute".

Abstract argumentation for case-based reasoning

- Modelling case-based reasoning with argumentation[1]
- Inspiration from legal domain
- Some of those approaches have been used as classifiers in different scenarios:
  - image classification[2]
  - sentiment analysis in text[3]

[1] Čyras, Satoh, and Toni, "Abstract Argumentation for Case-Based Reasoning".

[2] Cocarascu et al., "Data-Empowered Argumentation for Dialectically Explainable Predictions".

[3] Ibid.

[4] Čyras et al., "Explanations by arbitrated argumentative dispute".

Abstract argumentation for case-based reasoning

- Modelling case-based reasoning with argumentation[1]
- Inspiration from legal domain
- Some of those approaches have been used as classifiers in different scenarios:
  - image classification[2]
  - sentiment analysis in text[3]
  - passage of bills in the UK parliament[4]

---

[1] Čyras, Satoh, and Toni, "Abstract Argumentation for Case-Based Reasoning".

[2] Cocarascu et al., "Data-Empowered Argumentation for Dialectically Explainable Predictions".

[3] Ibid.

[4] Čyras et al., "Explanations by arbitrated argumentative dispute".

Main elements

- cases: each a pair input-output (x, y)

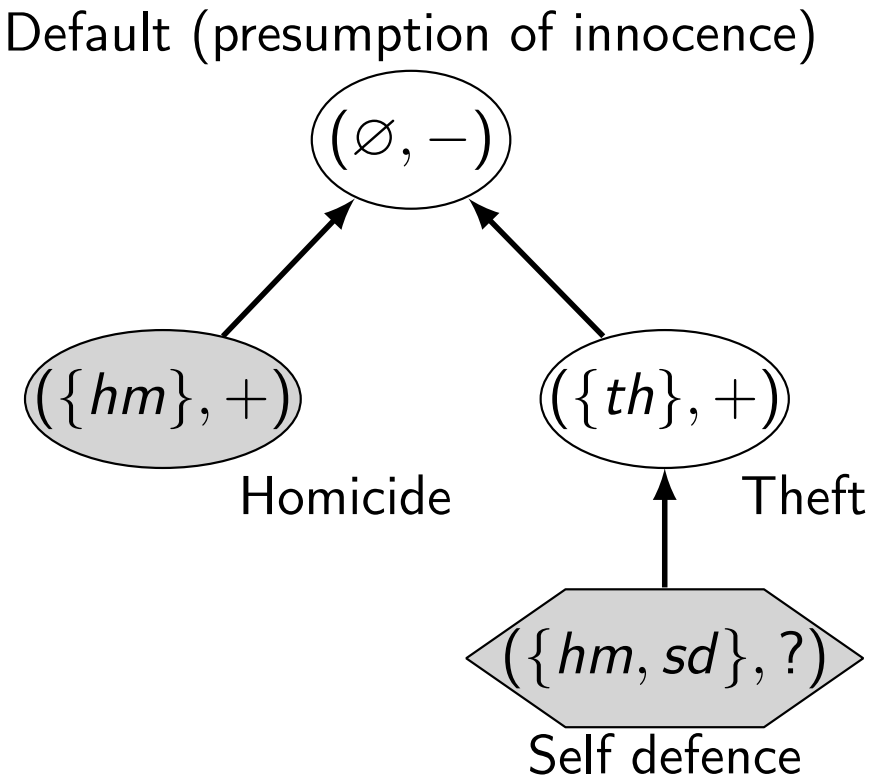
Main elements

- cases: each a pair input-output (x, y)
- a partial order $(\preceq)$ between inputs

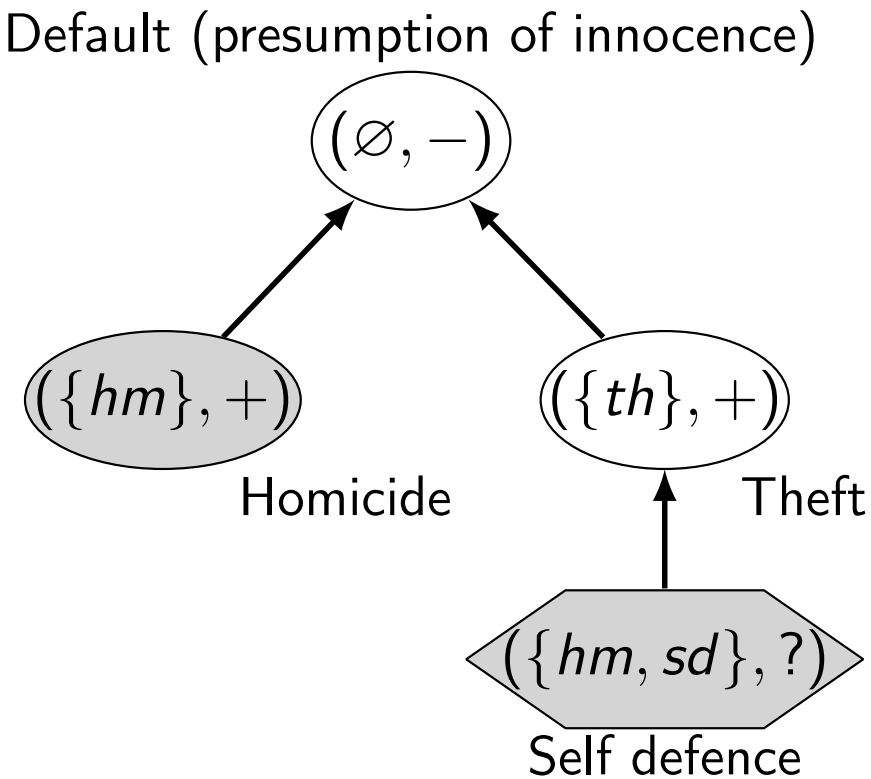
Main elements

- cases: each a pair input-output (x, y)
- a partial order $(\preceq)$ between inputs
- A new case to be predicted

AA-CBR: example

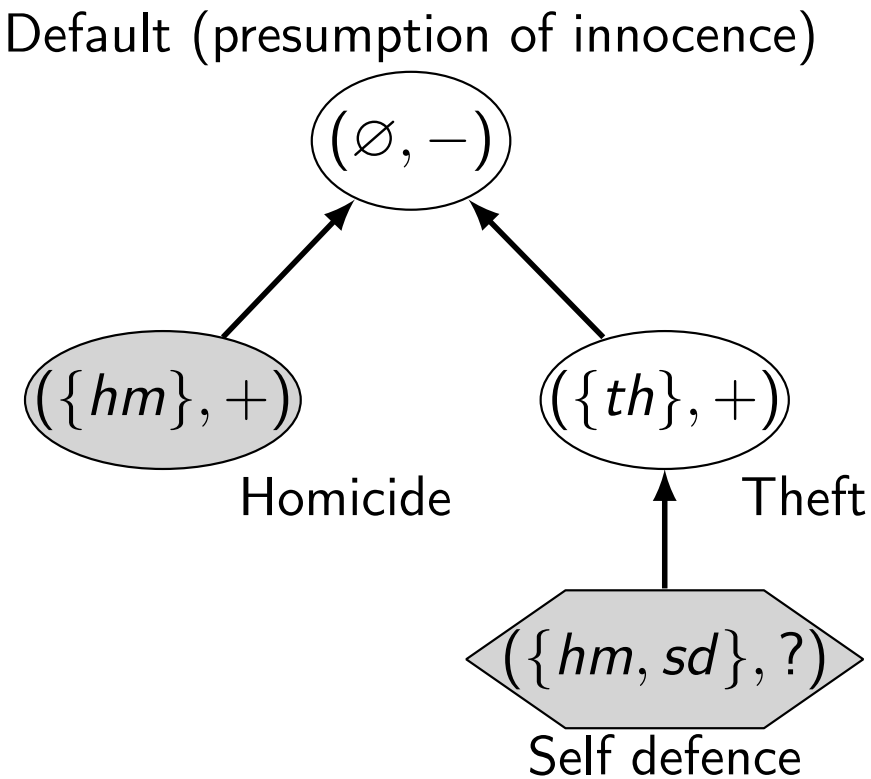


AA-CBR: example



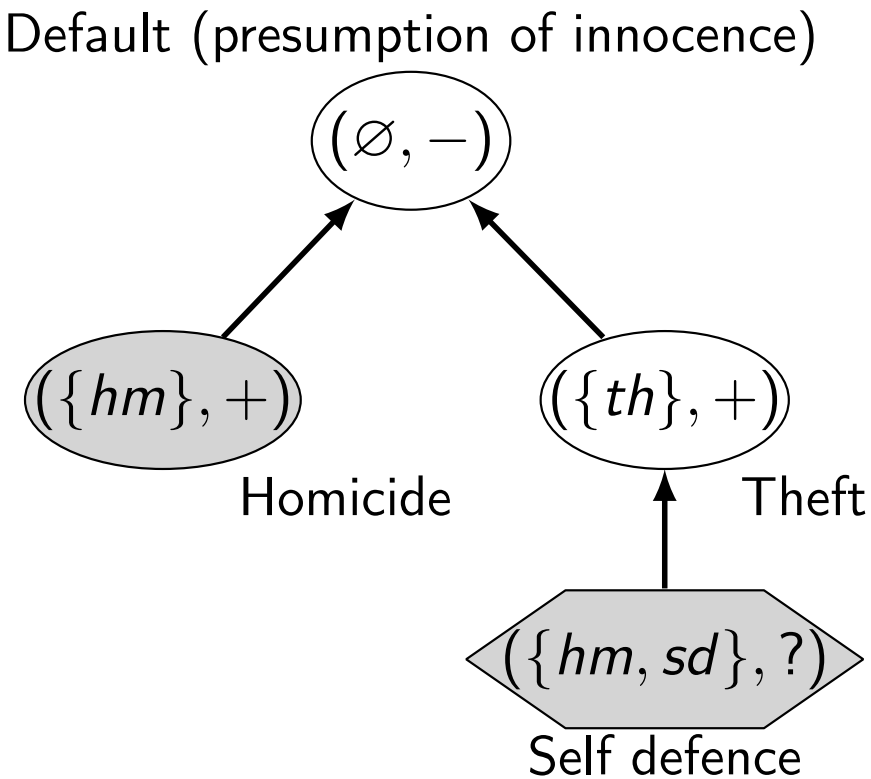
- case descriptions are partially ordered ($\preceq$)

AA-CBR: example



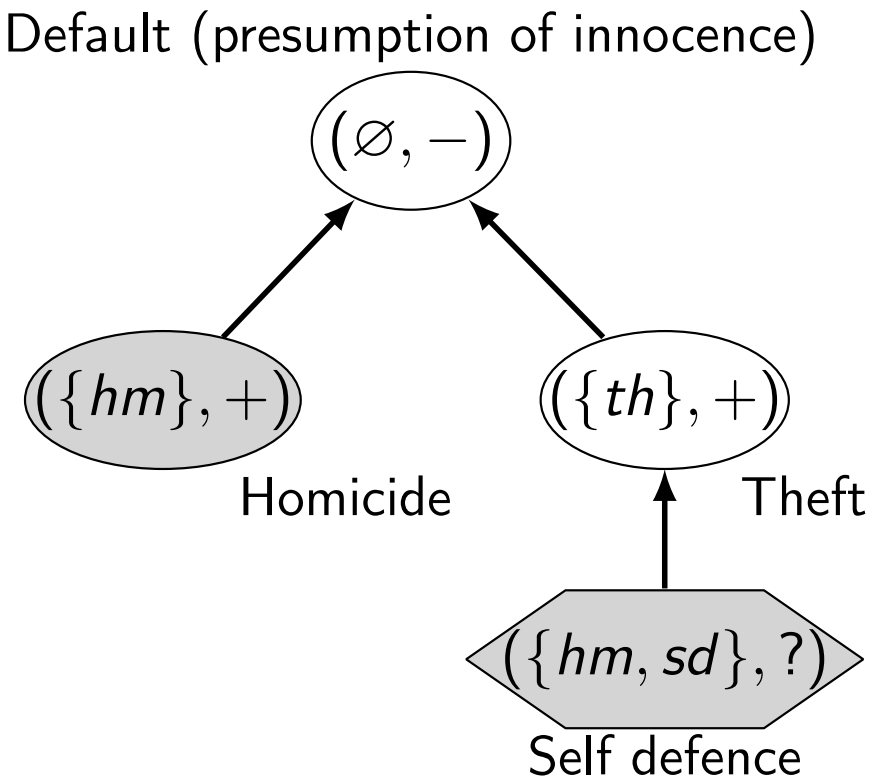
- case descriptions are partially ordered ($\preceq$)
- prediction is default outcome iff default argument is in grounded extension

AA-CBR: example



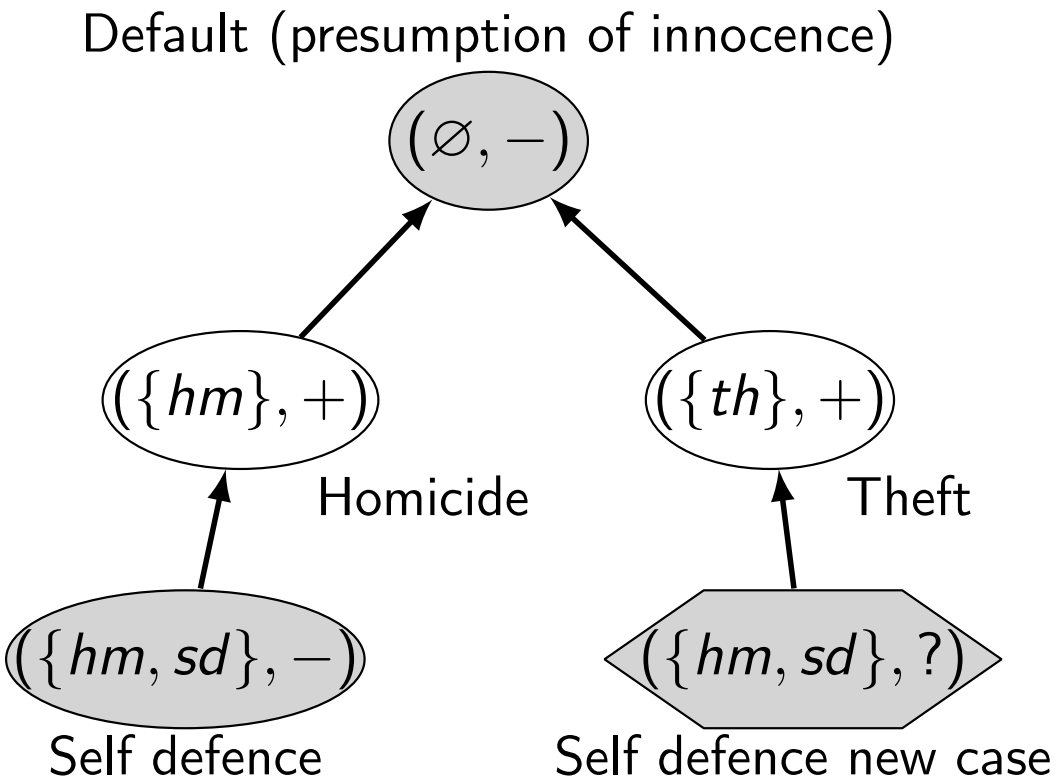
- case descriptions are partially ordered ($\preceq$)
- prediction is default outcome iff default argument is in grounded extension
- attacks from more specific to less specific with opposing outcomes

AA-CBR: example

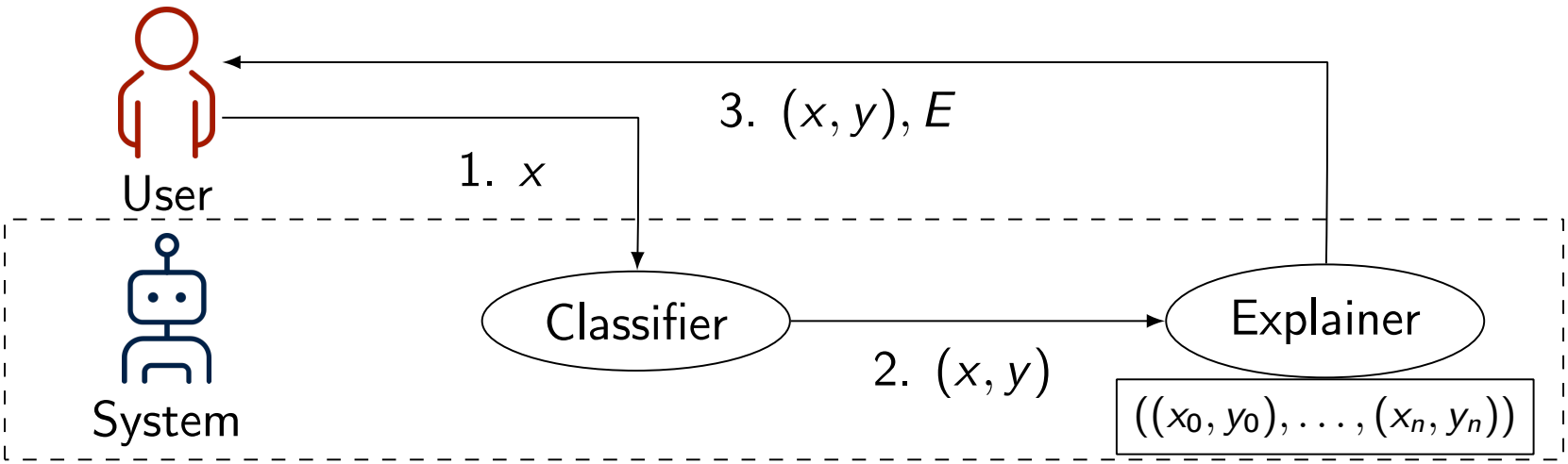


- case descriptions are partially ordered ($\preceq$)
- prediction is default outcome iff default argument is in grounded extension
- attacks from more specific to less specific with opposing outcomes
- new case attacks irrelevant past cases (not less specific)

AA-CBR: example

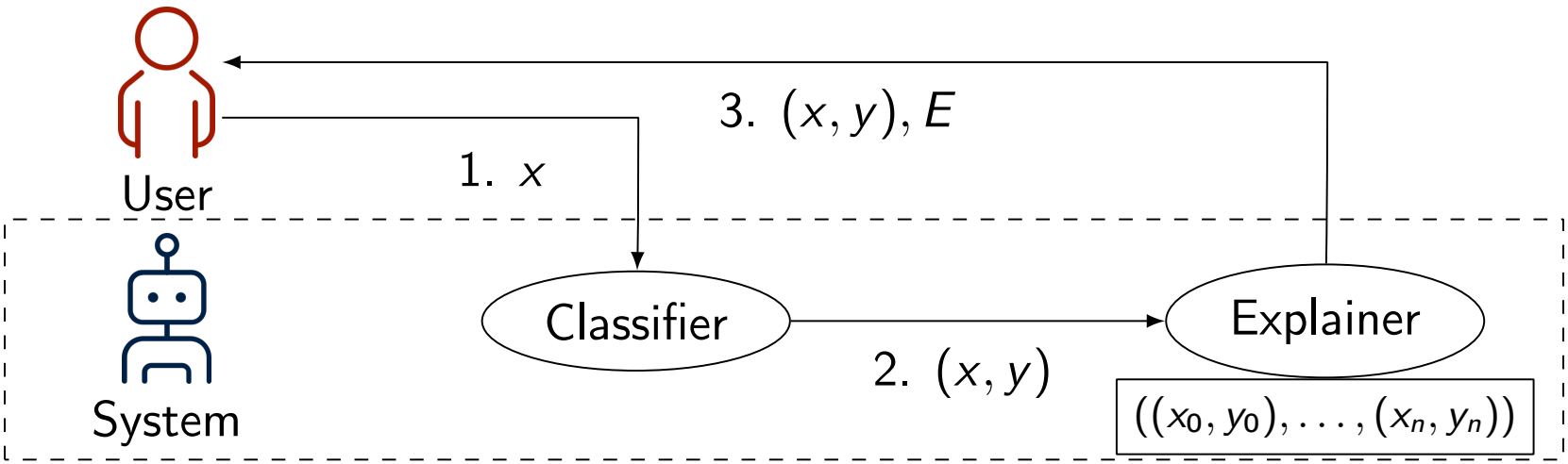


The interactive process



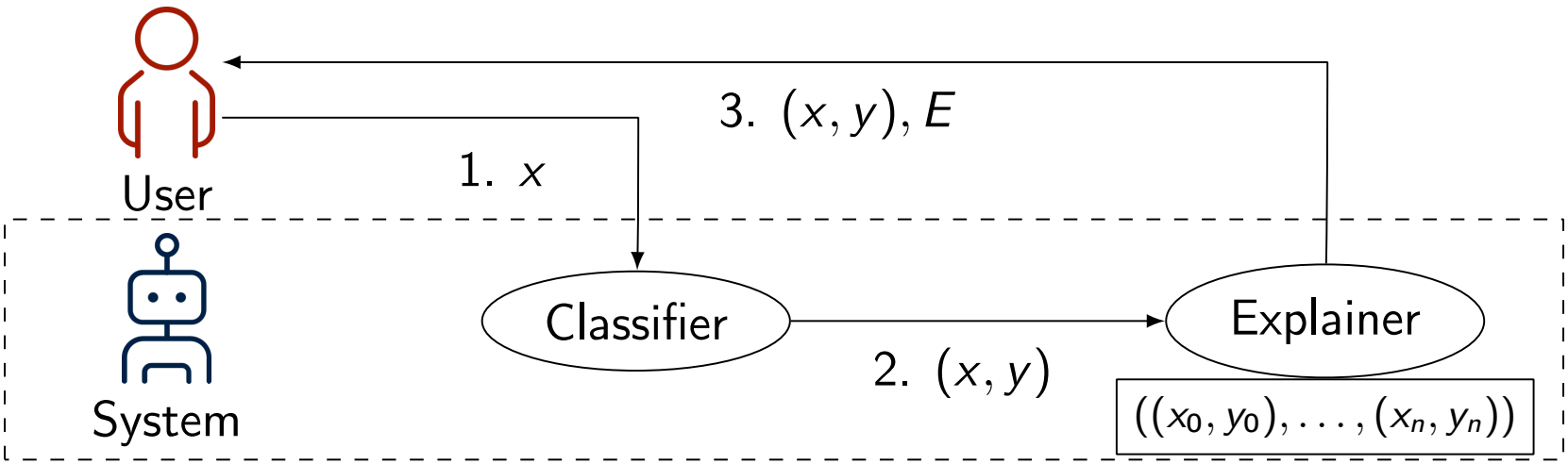
1. user queries the system with an input x, which goes to classifier

The interactive process



1. user queries the system with an input x, which goes to classifier
2. classifier produces output y, and sends pair (x, y) to explainer.

The interactive process



1. user queries the system with an input x, which goes to classifier
2. classifier produces output y, and sends pair (x, y) to explainer.
3. explainer produces explanation E for (x, y), using information from history of inputs/outputs $((x_0, y_0), \ldots, (x_n, y_n))$, and sends (x, y) and E back to user, who may then stop or further query the system

Setup: system

Post-hoc explanation. System is composed of:

1. a classifier $\mathbb{C} : X \to Y$

Setup: system

Post-hoc explanation. System is composed of:

1. a classifier $\mathbb{C} : X \to Y$
2. an explanation method $\mathbb{E} : Seq(X \times Y) \to Seq(\mathcal{E})$ mapping from a sequence of input-output pairs $(x_i, y_i)_{i \in [n]}$ to a sequence of explanations $(E_i)_{i \in [n]}$

Setup: system

Post-hoc explanation. System is composed of:

1. a classifier $\mathbb{C} : X \to Y$
2. an explanation method $\mathbb{E} : Seq(X \times Y) \to Seq(\mathcal{E})$ mapping from a sequence of input-output pairs $(x_i, y_i)_{i \in [n]}$ to a sequence of explanations $(E_i)_{i \in [n]}$
   - sequences for generality, instead of sets, since order might matter

Inference

- Core idea: a sequence of explanations $(E_i)_{i \in [n]}$ "commits" to some model behaviour.[5]

[5] Guilherme Paulino-Passos and Francesca Toni. "On Interactive Explanations as Non-Monotonic Reasoning". In: *Workshop on Explainable Artificial Intelligence (XAI) at IJCAI 2022*. 2022.

Inference

- Core idea: a sequence of explanations $(E_i)_{i \in [n]}$ "commits" to some model behaviour.[5]
  - We model this by an entailment relation $\models$ between $Seq(\mathcal{E})$ and $X \times Y$

[5] Paulino-Passos and Toni, "On Interactive Explanations as Non-Monotonic Reasoning".

Inference

- Core idea: a sequence of explanations $(E_i)_{i\in[n]}$ "commits" to some model behaviour.[5]
  - We model this by an entailment relation $\models$ between $Seq(\mathcal{E})$ and $X \times Y$
  - $(E_i)_{i\in[n]} \models (x, y)$ means that $(E_i)_{i\in[n]}$ "commits" to the outcome y, given the input x

[5] Paulino-Passos and Toni, "On Interactive Explanations as Non-Monotonic Reasoning".

Inference

- Core idea: a sequence of explanations $(E_i)_{i\in[n]}$ "commits" to some model behaviour.[5]
  - We model this by an entailment relation $\models$ between $Seq(\mathcal{E})$ and $X \times Y$
  - $(E_i)_{i\in[n]} \models (x, y)$ means that $(E_i)_{i\in[n]}$ "commits" to the outcome y, given the input x
- This entailment relation we keep abstract and application-dependent

[5] Paulino-Passos and Toni, "On Interactive Explanations as Non-Monotonic Reasoning".

Inference

- Core idea: a sequence of explanations $(E_i)_{i\in[n]}$ "commits" to some model behaviour.[5]
  - We model this by an entailment relation $\models$ between $Seq(\mathcal{E})$ and $X \times Y$
  - $(E_i)_{i\in[n]} \models (x, y)$ means that $(E_i)_{i\in[n]}$ "commits" to the outcome y, given the input x
- This entailment relation we keep abstract and application-dependent
- It should captures how a user would interpret the explanation or plausible inference

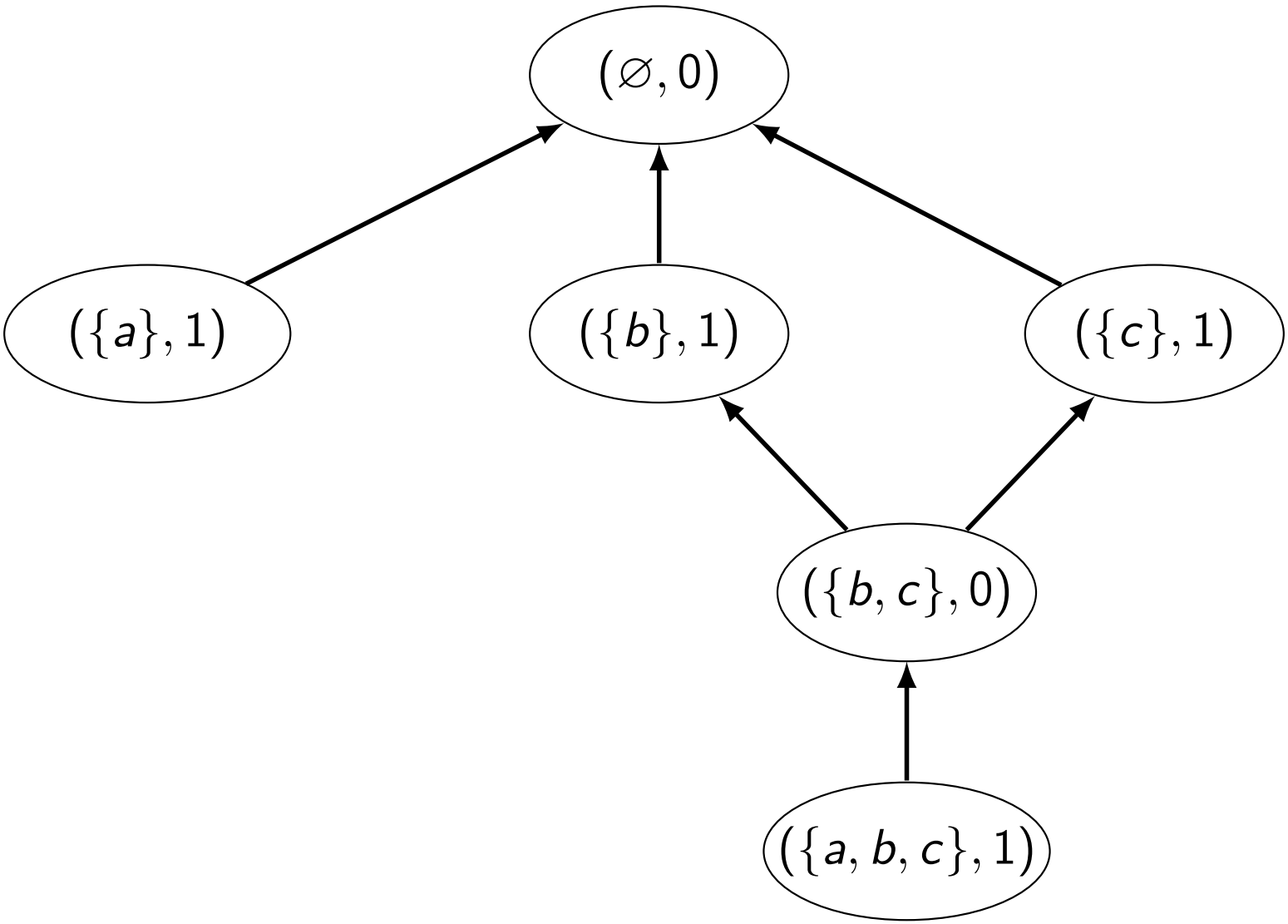
[5] Paulino-Passos and Toni, "On Interactive Explanations as Non-Monotonic Reasoning".

Arbitrated dispute trees

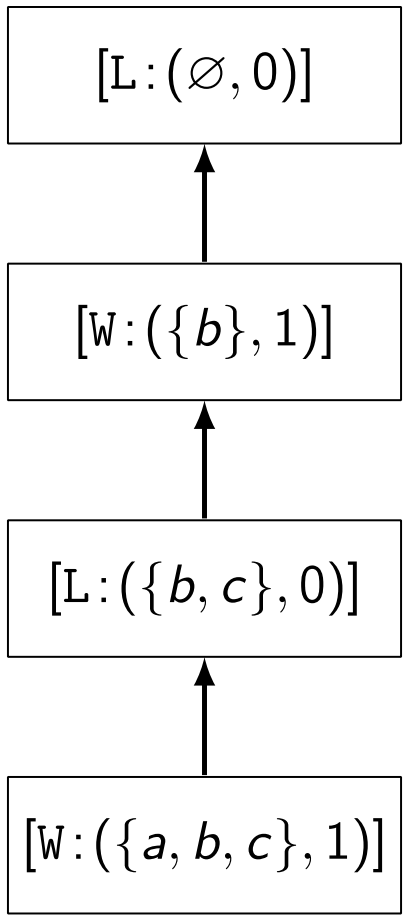
- a way of presenting and explaining the behaviour of AA-CBR as a more succinct structure than the entire argumentation framework

ADT example

Assume a casebase that results in the following mined AF:



ADT example



ADT explanation for input $x_1 = \{a, b, c, d\}$.

Arbitrated dispute trees (formal)

Definition

Let $AF_{\succeq}(D, x) = (Args, \rightsquigarrow)$. *An* ***arbitrated dispute tree*** *(ADT) is a tree* $\mathcal{DT}$ *such that:*

Arbitrated dispute trees (formal)

Definition

Let $AF_{\succeq}(D, x) = (Args, \rightsquigarrow)$. *An* ***arbitrated dispute tree*** *(ADT) is a tree* $\mathcal{DT}$ *such that:*

1. *every node of* $\mathcal{DT}$ *is of the form* $[\mathtt{N} : \alpha]$ *for* $N \in \{W, L\}$ *and* $\alpha \in Args$*, any such node being called an N-node labelled by argument* α*;*

Arbitrated dispute trees (formal)

Definition

Let $AF_{\succeq}(D, x) = (Args, \rightsquigarrow)$. *An* ***arbitrated dispute tree*** *(ADT) is a tree* $\mathcal{DT}$ *such that:*

1. *every node of* $\mathcal{DT}$ *is of the form* $[\mathtt{N} : \alpha]$ *for* $N \in \{W, L\}$ *and* $\alpha \in Args$*, any such node being called an N-node labelled by argument* α*;*
2. *the root of* $\mathcal{DT}$ *is labelled by the default argument* (δ_C, δ_o) *and is a W-node, if* $AA\text{-}CBR_{\succeq}(D, x) = \delta_o$*; and a L-node otherwise;*

Arbitrated dispute trees (formal)

Definition

Let $AF_{\succeq}(D, x) = (Args, \rightsquigarrow)$. *An* ***arbitrated dispute tree*** *(ADT) is a tree* $\mathcal{DT}$ *such that:*

1. *every node of* $\mathcal{DT}$ *is of the form* $[\mathtt{N} : \alpha]$ *for* $N \in \{W, L\}$ *and* $\alpha \in Args$*, any such node being called an N-node labelled by argument* α*;*
2. *the root of* $\mathcal{DT}$ *is labelled by the default argument* (δ_C, δ_o) *and is a W-node, if* $AA\text{-}CBR_{\succeq}(D, x) = \delta_o$*; and a L-node otherwise;*
3. *for every W-node n labelled* $\alpha \in Args$ *and for every* β *attacker of* α *in* $(Args, \rightsquigarrow)$*, there is a child m of n such that m is a L-node labelled by* β*;*

Arbitrated dispute trees (formal)

Definition (cont.)

Arbitrated dispute trees (formal)

Definition (cont.)

4. *for every L-node n labelled by $\alpha \in Args$, there is exactly one child which is a W-node labelled by some β attacker of α; and*

Arbitrated dispute trees (formal)

Definition (cont.)

4. *for every L-node n labelled by $\alpha \in Args$, there is exactly one child which is a W-node labelled by some β attacker of α; and*
5. *there are no other nodes in $\mathcal{DT}$.*

Arbitrated dispute trees (formal)

Definition (cont.)

4. *for every L-node n labelled by $\alpha \in Args$, there is exactly one child which is a W-node labelled by some β attacker of α; and*
5. *there are no other nodes in $\mathcal{DT}$.*

- $\mathcal{DT}$ is as an arbitrated dispute tree for (the prediction for) x

Arbitrated dispute trees (formal)

Definition (cont.)

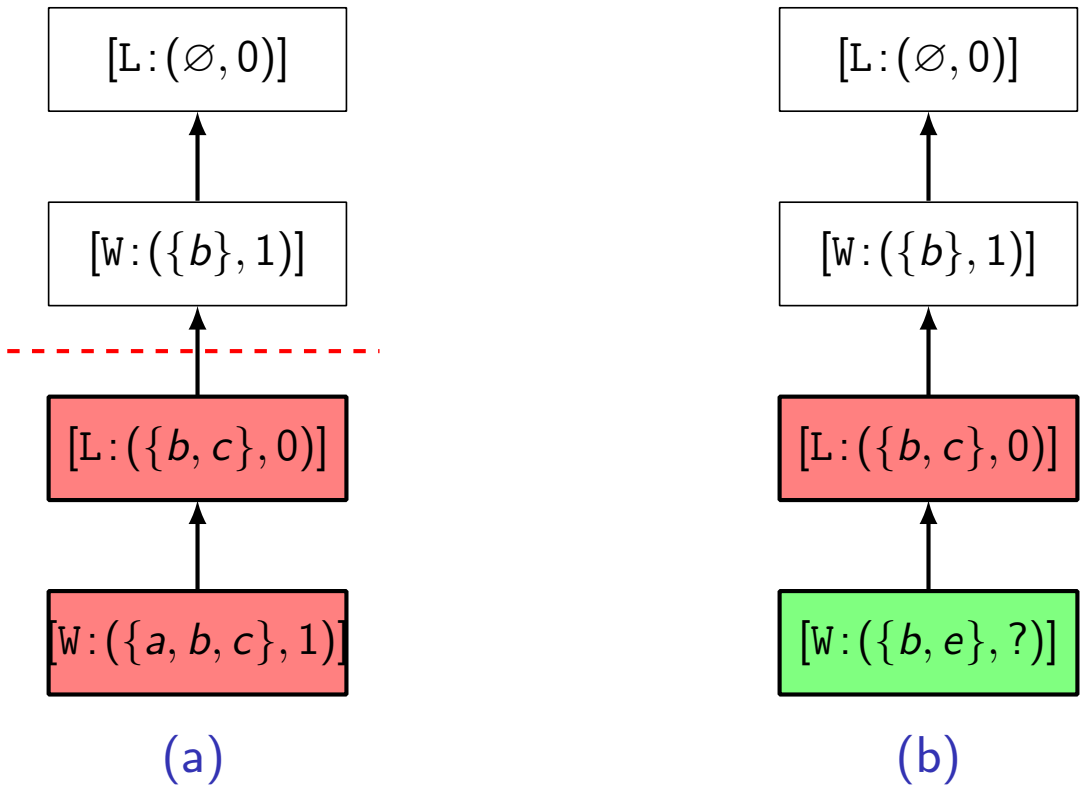
4. *for every L-node n labelled by $\alpha \in Args$, there is exactly one child which is a W-node labelled by some β attacker of α; and*
5. *there are no other nodes in $\mathcal{DT}$.*

- $\mathcal{DT}$ is as an arbitrated dispute tree for (the prediction for) x

Definition

*An **arbitrated dispute tree explanation** is a finite arbitrated dispute tree.*

Readaptation of ADTs



- Here is the process of adapting ADT for input $x_1 = \{a, b, c, d\}$ to the new input $x_2 = \{b, e\}$.

Readaptation of ADTs

- Theorem (informally): for an ADT $\mathcal{DT}$ for x, and a new input x_2, remove nodes from $\mathcal{DT}$ which are irrelevant for x_2. If the leaves in the new tree are all W-nodes, then the prediction for x_2 is the same as the prediction for x. Also, an ADT for x_2 may be created from this adaptation.

ADTs as interactive explanations

- $\mathbb{E}_\bullet$ as a function that receives a new case x and its predicted outcome $AA\text{-}CBR_{\succeq}(D, x) = y$, and returns an ADT explanation $\mathcal{DT}$ for it

ADTs as interactive explanations

- $\mathbb{E}_\bullet$ as a function that receives a new case x and its predicted outcome $AA\text{-}CBR_{\succeq}(D, x) = y$, and returns an ADT explanation $\mathcal{DT}$ for it
- $\mathbb{E}(((x_i, y_i))_{i\in[n]}) = (\mathbb{E}_\bullet(x_i, y_i))_{i\in[n]}$

ADTs as interactive explanations

- $\mathbb{E}_\bullet$ as a function that receives a new case x and its predicted outcome $AA\text{-}CBR_{\succeq}(D, x) = y$, and returns an ADT explanation $\mathcal{DT}$ for it
- $\mathbb{E}(((x_i, y_i))_{i\in[n]}) = (\mathbb{E}_\bullet(x_i, y_i))_{i\in[n]}$
- $\models$ as: $(\mathcal{DT}_i)_{i\in[n]} \models (x, y)$ iff there is i such that $\mathcal{DT}_i$ can be readapted to x with outcome y

ADTs as interactive explanations

Theorem

Let $\mathbb{E}(((x_i, y_i))_{i\in[n]}) = (E_i)_{i\in[n]}$, *and* $(x, y) \in X \times Y$ *Then:*

ADTs as interactive explanations

Theorem

Let $\mathbb{E}(((x_i, y_i))_{i\in[n]}) = (E_i)_{i\in[n]}$, *and* $(x, y) \in X \times Y$ *Then:*

1. *(faithfulness) if* $(E_i)_{i\in[n]} \models (x, y)$, *then* $AA\text{-}CBR_{\succeq}(D, x) = y$.

ADTs as interactive explanations

Theorem

Let $\mathbb{E}(((x_i, y_i))_{i\in[n]}) = (E_i)_{i\in[n]}$, *and* $(x, y) \in X \times Y$ *Then:*

1. *(faithfulness) if* $(E_i)_{i\in[n]} \models (x, y)$, *then* $AA\text{-}CBR_{\succeq}(D, x) = y$.
2. $\models$ *is consistent;*

ADTs as interactive explanations

Theorem

Let $\mathbb{E}(((x_i, y_i))_{i\in[n]}) = (E_i)_{i\in[n]}$, *and* $(x, y) \in X \times Y$ *Then:*

1. *(faithfulness) if* $(E_i)_{i\in[n]} \models (x, y)$, *then* $AA\text{-}CBR_{\succeq}(D, x) = y$.
2. $\models$ *is consistent;*
3. $\models$ *is monotonic.*

Conclusion

- Interactive explanation process can be modelled as reasoning, subject to inference.

Conclusion

- Interactive explanation process can be modelled as reasoning, subject to inference.
- Arbitrated dispute trees in AA-CBR can be readapted to some other inputs, in a limited scope.

Conclusion

- Interactive explanation process can be modelled as reasoning, subject to inference.
- Arbitrated dispute trees in AA-CBR can be readapted to some other inputs, in a limited scope.
- Under this view, arbitrated dispute trees in AA-CBR are provably faithful and monotonic.

Conclusion

- Interactive explanation process can be modelled as reasoning, subject to inference.
- Arbitrated dispute trees in AA-CBR can be readapted to some other inputs, in a limited scope.
- Under this view, arbitrated dispute trees in AA-CBR are provably faithful and monotonic.
- Result is limited to inputs for which the tree can be readapted:

Conclusion

- Interactive explanation process can be modelled as reasoning, subject to inference.
- Arbitrated dispute trees in AA-CBR can be readapted to some other inputs, in a limited scope.
- Under this view, arbitrated dispute trees in AA-CBR are provably faithful and monotonic.
- Result is limited to inputs for which the tree can be readapted:
    - intuitively, only if, given an ADT, from all relevant arguments to the new case, the ones "said later" in the argumentative dispute are all winning.

Thank you!

Readaptation of ADTs (more formally)

Theorem

Let $\mathcal{DT}'$ be $\mathcal{DT}$ with all nodes labelled by $x1$ removed. For every leaf l of $\mathcal{DT}'$, let max_l be the node in the path from the root to l which is maximally relevant to $x2$ (that is, there is no node in the path greater than it such that it is also relevant to l).
If all max_l are W-nodes, then the predicted outcome for $x2$ is the same as the predicted outcome for x. Besides, let $\mathcal{DT}_{x2}$ be the tree constructed by the following process: start with the subtree of $\mathcal{DT}'$ containing only max_l and their ancestors, add all L-nodes which are children of max_l in $\mathcal{DT}'$, and, finally, for each L-node added in this way, add as a child a new W-node labelled by $x2$. Then $\mathcal{DT}_{x2}$ is an ADT for the prediction on $x2$.

Proof of Theorem 5 I

- For every l, max_l exists: by regularity, the default case (thus the root) is relevant to $x2$. - Let us see that $\mathcal{DT}_{x2}$ is an ADT. Clearly the condition 1 is satisfied. Let us check conditions 3 and 4 for each node. For every branch, every node until max_l satisfies the conditions in Definition 1, since $\mathcal{DT}$ is an ADT. Now consider max_l. By assumption, every max_l is a W-node.
- Again, since $\mathcal{DT}$ is an ADT for x, for every max_l, each child of it is included in $\mathcal{DT}$ as a L-node child. Since those are also in $\mathcal{DT}_{x2}$, the condition is satisfied for each max_l
- Next, each of those attackers require exactly one W-node as a child, attacking it. This is satisfied by every added W-node labelled by $x2$, which is an attacker since the arguments which label such L-nodes are irrelevant (otherwise max_l would not be maximally

Proof of Theorem 5 II

relevant). Finally, since $x2$ has no attackers in $AF_{\succeq}(D, x2,)$, it satisfies the conditions of a leaf. Condition 5 is clearly satisfied.
- The last requirement to check is whether condition 2 is satisfied. Indeed it is, since, given that the other conditions are satisfied, then the set of arguments labelling W-nodes is in the grounded extension of $AF_{\succeq}(D, x2)$, which can be verified by induction[6].
- Therefore if the root is a W-node, it is in the grounded extension and thus the prediction if δ_o. Otherwise, it is a L-node and then it has as a child which is a W-node, that is, the default argument is attacked by the grounded extension (and thus not in it, since it is conflict-free) and so the prediction of $x2$ is the $\bar{\delta_o}$.

[6] Čyras et al., "Explanations by arbitrated argumentative dispute", Prop. 3.3.

References

Other notions of $\models$

- (Minimal) counterfactuals:
  - $\models$ as:
    - given original input
    - and the counterfactual as explanation
    - then every input between the original and the counterfactual has the same output

Formal definition of properties of $\vdash$, $\vdash_e$ - Non-monotonicity

Definition

- *Given a set S, a relation $\vdash'$ from $Seq(S)$ to S, and $s, s_i \in S$, for $i \in \mathbb{N}$, the relation $\vdash'$ is said to satisfy:*
  - *non-monotonicity iff there is $(s_i)_{i\in[n]}, s_{n+1}, s$ s.t. $(s_i)_{i\in[n]} \vdash' s$ and $(s_i)_{i\in[n+1]} \not\vdash' s$;*
  - *reflexivity iff for every $(s_i)_{i\in[n]}$ and i, $(s_i)_{i\in[n]} \vdash' s_i$;*
  - *cautious monotonicity iff for every $(s_i)_{i\in[n]}$, s_{n+1}, and s, if $(s_i)_{i\in[n]} \vdash' s_{n+1}$ and $(s_i)_{i\in[n]} \vdash' s$, then $(s_i)_{i\in[n+1]} \vdash' s$;*

References

- Cocarascu, Oana et al. "Data-Empowered Argumentation for Dialectically Explainable Predictions". In: *ECAI 2020 - 24th European Conference on Artificial Intelligence, Santiago de Compostela, Spain, 10-12 June 2020*. 2020.
- Paulino-Passos, Guilherme and Francesca Toni. "On Interactive Explanations as Non-Monotonic Reasoning". In: *Workshop on Explainable Artificial Intelligence (XAI) at IJCAI 2022*. 2022.
- Čyras, Kristijonas, Ken Satoh, and Francesca Toni. "Abstract Argumentation for Case-Based Reasoning". In: *Principles of Knowledge Representation and Reasoning: Proceedings of the Fifteenth International Conference, KR 2016, Cape Town, South Africa, April 25-29, 2016*. Ed. by Chitta Baral, James P. Delgrande, and Frank Wolter. AAAI Press, 2016, pp. 549–552. URL: http://www.aaai.org/ocs/index.php/KR/KR16/paper/view/12879.
- Čyras, Kristijonas et al. "Explanations by arbitrated argumentative