XAI is a Double-Edged Sword

Leila Methnani

Virginia Dignum, Andreas Theodorou

UMEÅ UNIVERSITY

leila.methnani@umu.se

Status quo

- It is widely agreed that explainability for AI is no longer a ‘nice to have’.

UMEÅ UNIVERSITY

Still many questions around Article 14

- **"The AI system should be provided in a way that allows the overseer to understand its capabilities and limitations, detect and address issues, avoid over-reliance on the system, interpret its output, decide not to use it, or stop its operation."**
- Explainability and interpretability are indeed critical for such systems and mechanisms, perhaps *especially* with the human present
- Do these "oversees" need to be XAI experts, then?

Status quo

- It is widely agreed that explainability for AI is no longer a ‘nice to have’.
- There is no one-size-fits all solution when it comes to explainability.

Explainable AI: Beware of Inmates Running the Asylum

Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences

Tim Miller[*] and **Piers Howe**[†] and **Liz Sonenberg**[*]
[*]School of Computing and Information Systems
[†]Melbourne School of Psychological Sciences
University of Melbourne, Australia
{tmiller,pdhowe,l.sonenberg}@unimelb.edu.au

Abstract

In his seminal book *The Inmates are Running the Asylum: Why High-Tech Products Drive Us Crazy And How To Restore The Sanity* [2004, Sams Indianapolis, IN, USA], Alan Cooper argues that a major reason why software is often poorly designed (from a user perspective) is that programmers are in charge of design decisions, rather than interaction designers. As a result, programmers design software for themselves, rather than for their target audience; a phenomenon he refers to as the '*inmates running the asylum*'. This paper argues that explainable AI risks a similar fate. While the reemergence of explainable AI is positive, this paper argues most of us as AI researchers are building explanatory agents for ourselves, rather than for the intended users. But explainable AI is more likely to succeed if researchers and practitioners understand, adopt, implement, and improve models from the vast and valuable bodies of research in philosophy, psychology, and cognitive science; and if evaluation of these models is focused more on people than on technology. From a light scan of literature

portant in the 80s and 90s in expert systems particularly; see [Chandrasekaran *et al.*, 1989], [Swartout and Moore, 1993], and [Buchanan and Shortliffe, 1984]. High visibility of the term, sometimes abbreviated XAI, is seen in grant solicitations [DARPA, 2016] and in the popular press [Nott, 2017]. One area of explainable AI receiving attention is explicit *explanation*, on which we say more below.

While the title of the paper is deliberately tongue-in-cheek, the parallels with Cooper [2004] are real: leaving decisions about what constitutes a good explanation of complex decision-making models to the experts who understand these models the best is likely to result in failure in many cases. Instead, models should be built on an understanding of explanation, and should be evaluated using data from human behavioural studies.

In Section 2, we describe a simple scan of the 23 articles posted as 'Related Work' on the workshop web page. We looked at two attributes: whether the papers were built on research from philosophy, psychology, cognitive science, or human factors; and whether the reported evaluations involved human behavioural studies. The outcome of this scan supports the hypothesis that ideas from social sciences and human factors are not suf-

Status quo

- It is widely agreed that explainability for AI is no longer a ‘nice to have’.
- There is no one-size-fits all solution when it comes to explainability.
- Many methods have been—and continue to be—developed (for practitioners).

Status quo

- It is widely agreed that explainability for AI is no longer a ‘nice to have’.
- There is no one-size-fits all solution when it comes to explainability.
- Many methods have been—and continue to be—developed (for practitioners).

Which explainability method to apply?
When? Why? And how?

Problem

- Choosing an appropriate explanation technique and then interpreting the explanation correctly requires a technical understanding.
- Different stakeholders may demand different types of explanations without knowing which technique will offer it.

Explanation:

"an interface between humans and a decision maker that is … both an accurate proxy of the decision maker and comprehensible to humans.

— R. Guidotti et al.

UMEÅ UNIVERSITY
leila.methnani@umu.se

Research Question(s)

RQ 1: *can facts (and potentially beliefs) about the data / model, the stakeholder, and various explanation techniques be utilised to argue for the most suitable explanation in the given context?*

RQ 2: *can we measure the transparency be afforded by presenting these contextual arguments to the users of the system?*

Research Question(s)

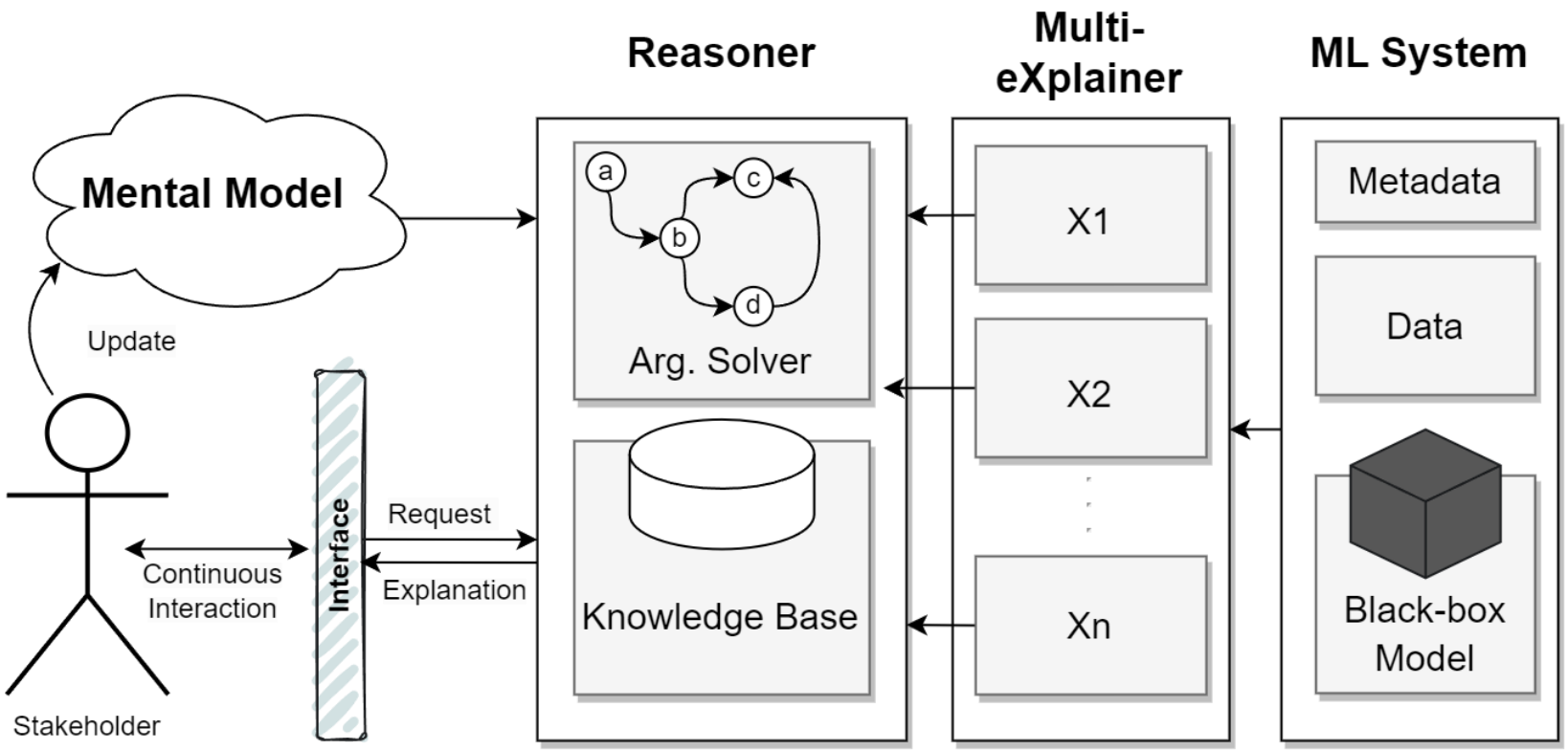
RQ 1: *can facts (and potentially beliefs) about the data / model, the stakeholder, and various explanation techniques be utilised to argue for the most suitable explanation in the given context?*

RQ 2: *can we measure the transparency be afforded by presenting these contextual arguments to the users of the system?*

Our proposal

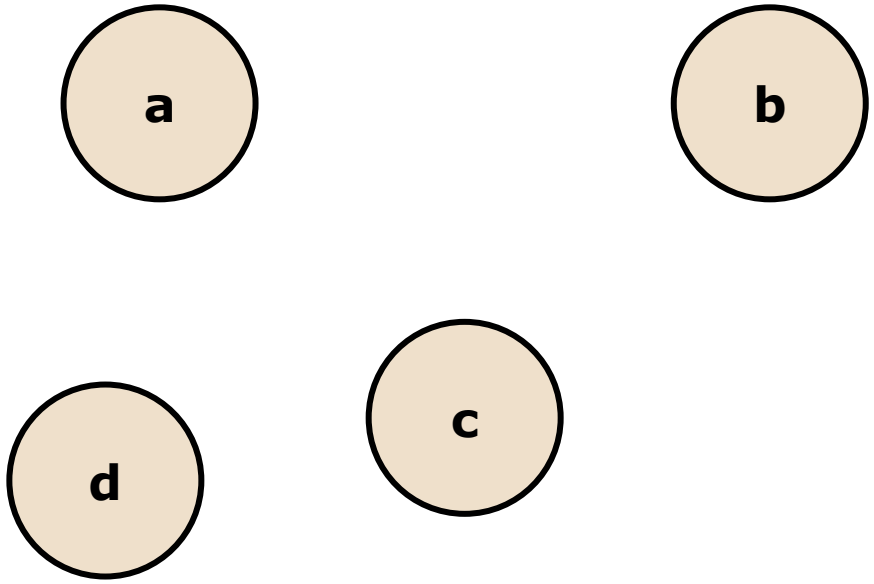
- Argumentation framework on top of system of multiple explainer methods.
- Takes into consideration various “facts” and “beliefs” about stakeholder (their mental model) and explanation techniques.
- Produces arguments and attacks on arguments.
- Solves for acceptable arguments that justify explanation choice.

Our proposal



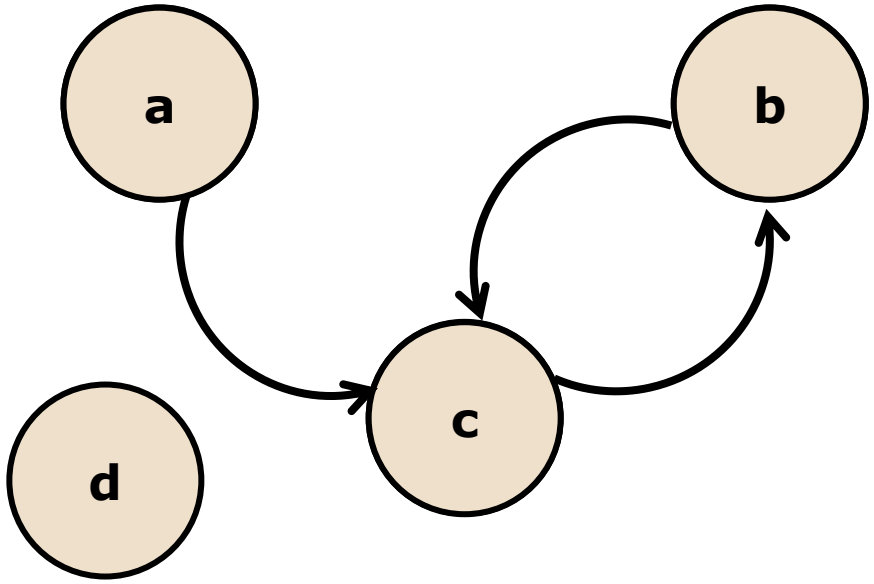
Abstract Argumentation

- Set of arguments
  - Ar = {a, b, c, d}



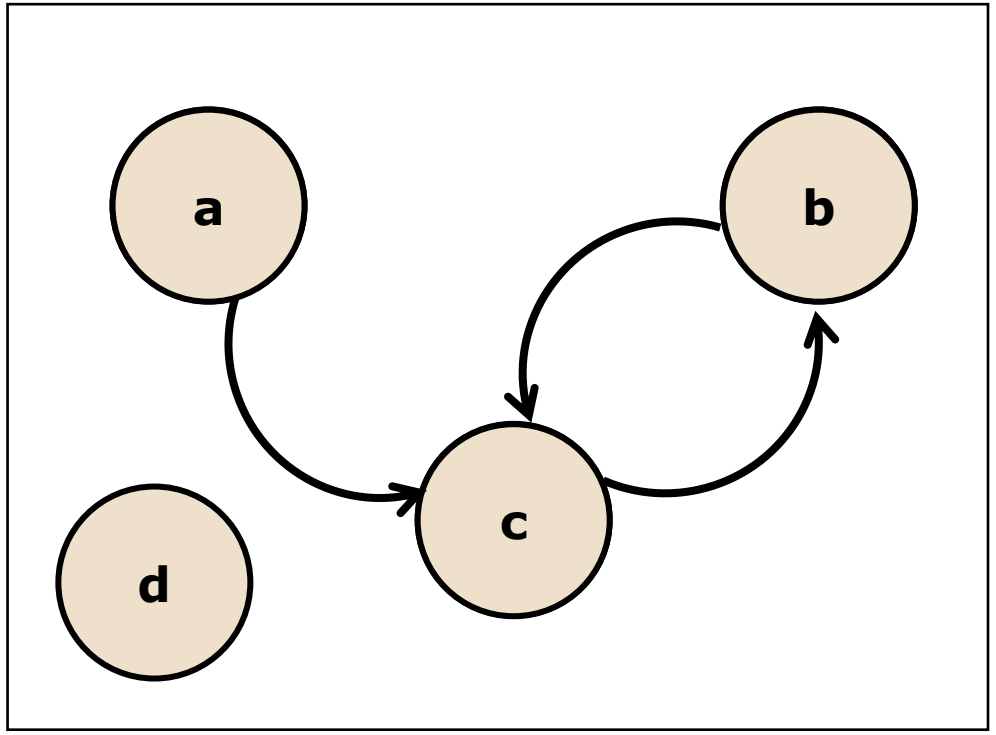
Abstract Argumentation

- Set of arguments
  - Ar = {a, b, c, d}
- Set of attack relations
  - R = { (a, c), (b, c) (c, b) }



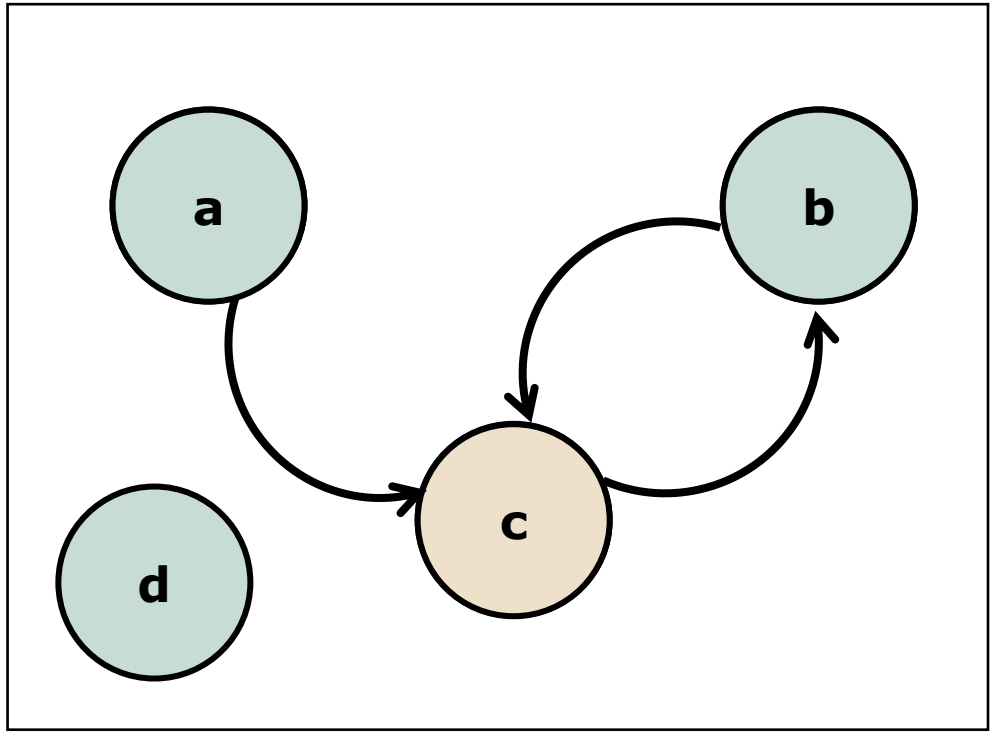
Abstract Argumentation

- Set of arguments
  - Ar = {a, b, c, d}
- Set of attack relations
  - R = { (a, c), (b, c) (c, b) }
- Framework
  - S = ⟨Ar , R⟩



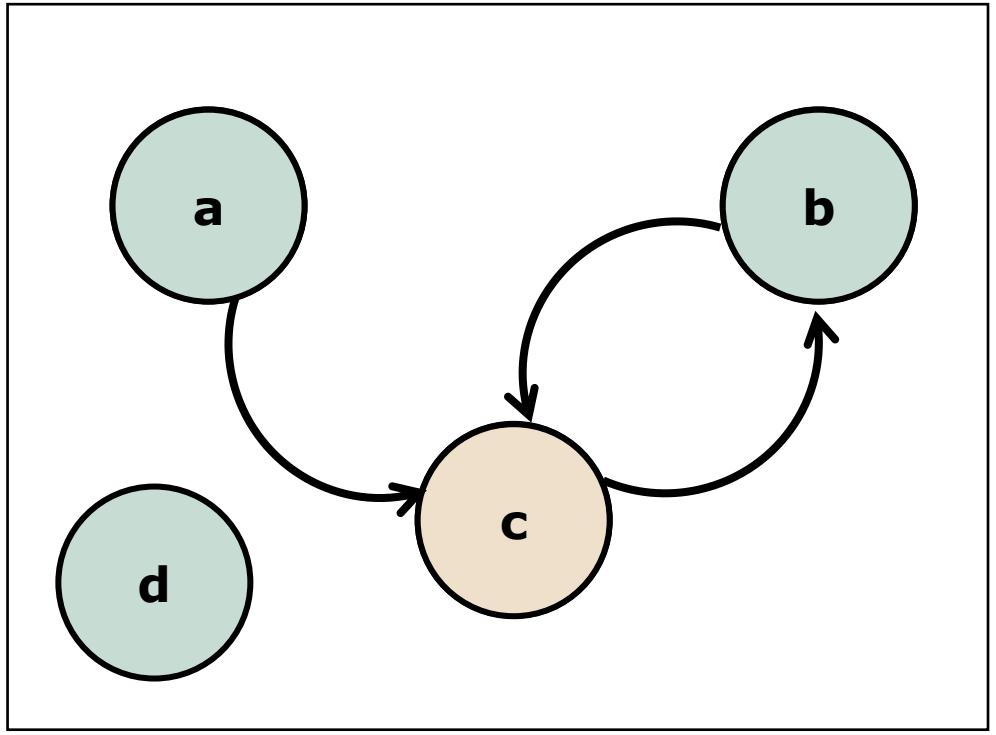
Abstract Argumentation

- Framework
  - S = ⟨Ar , R⟩
- Use semantics of acceptance
  - Sets of arguments computed are *extensions*
  - *Acceptable arguments*



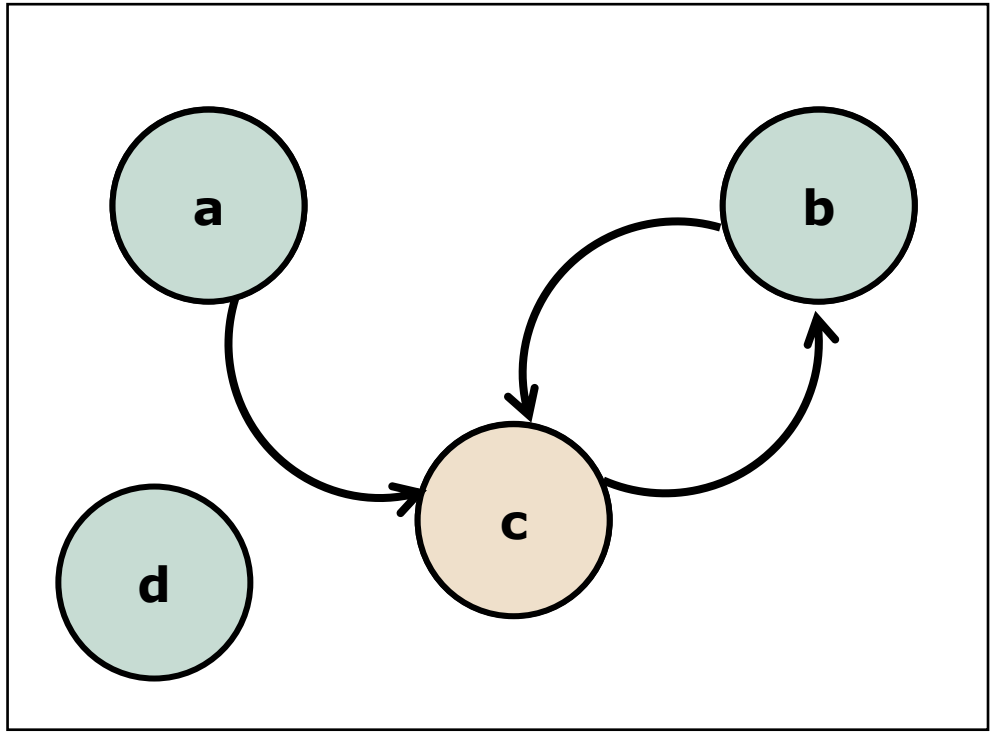
Abstract Argumentation

- Framework
  - S = ⟨Ar , R⟩
- Use semantics of acceptance
  - Sets of arguments computed are *extensions*
  - *Acceptable arguments*
  - *Conflict-free extensions*



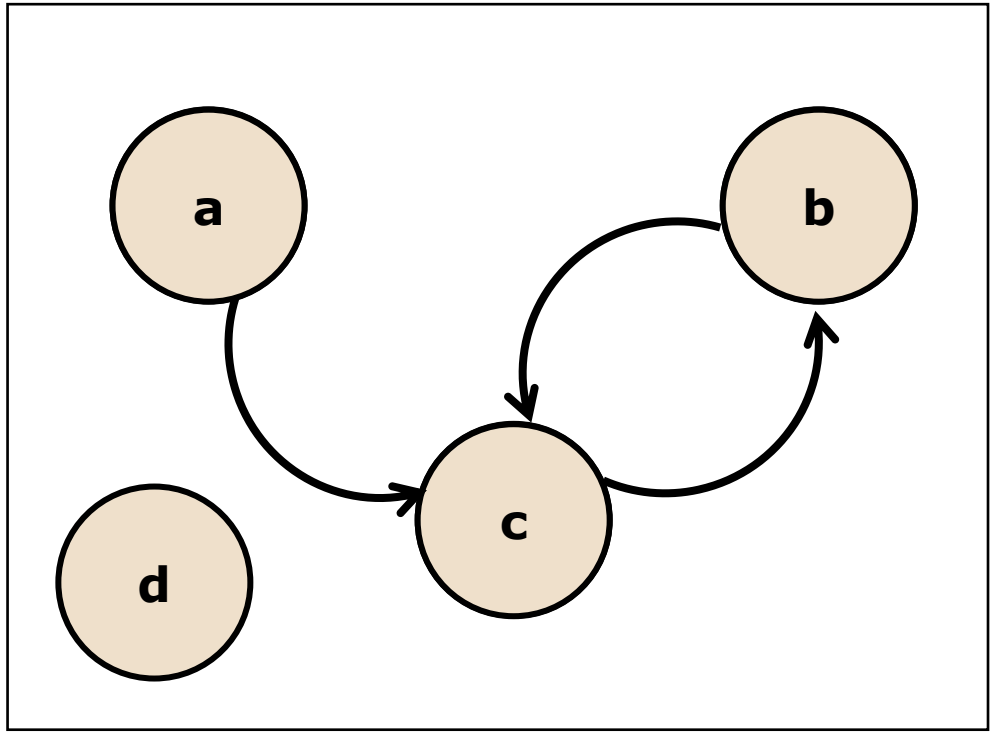
Abstract Argumentation

- Framework
  - S = ⟨Ar , R⟩
- Use semantics of acceptance
  - Sets of arguments computed are *extensions*
  - *Acceptable arguments*
  - *Conflict-free extensions*
  - *Admissible extensions*



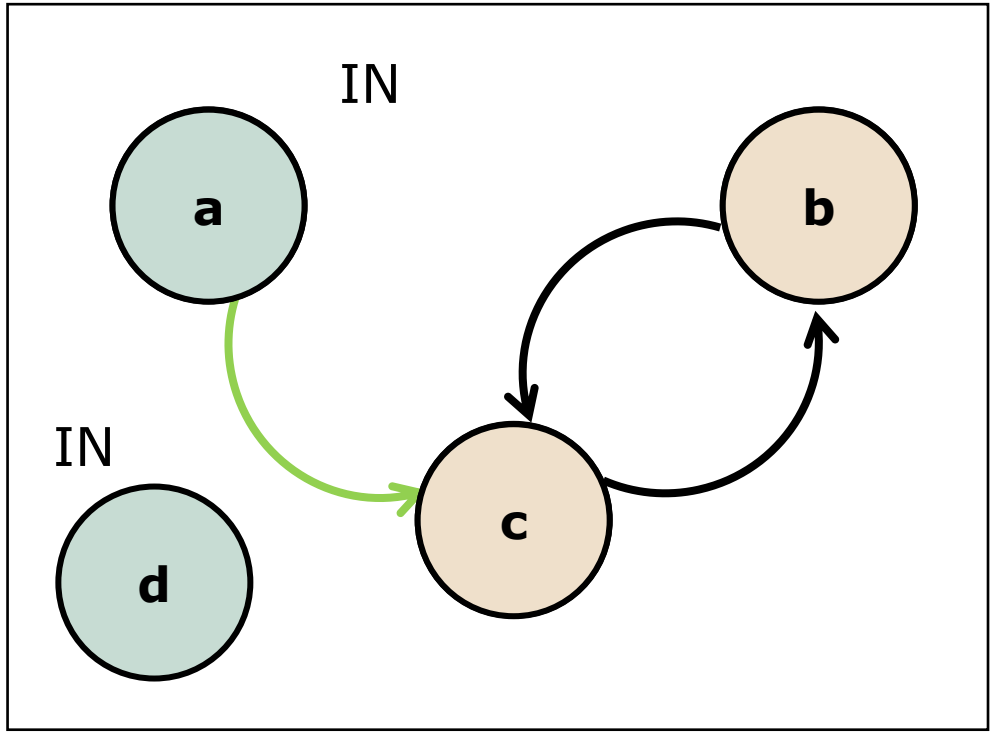
Abstract Argumentation

- Framework
  - $S = \langle Ar , R \rangle$
- Use semantics of acceptance
  - Labellings are expressive



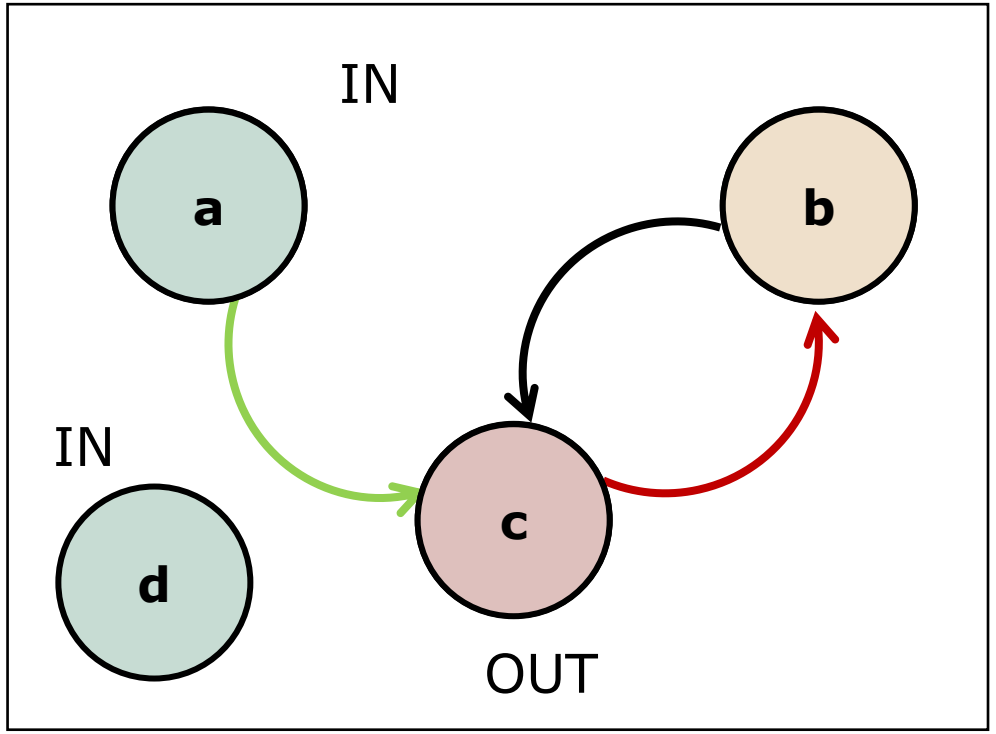
Abstract Argumentation

- Framework
  - S = ⟨Ar , R⟩
- Use semantics of acceptance
  - Labellings are expressive
  - IN means accepted



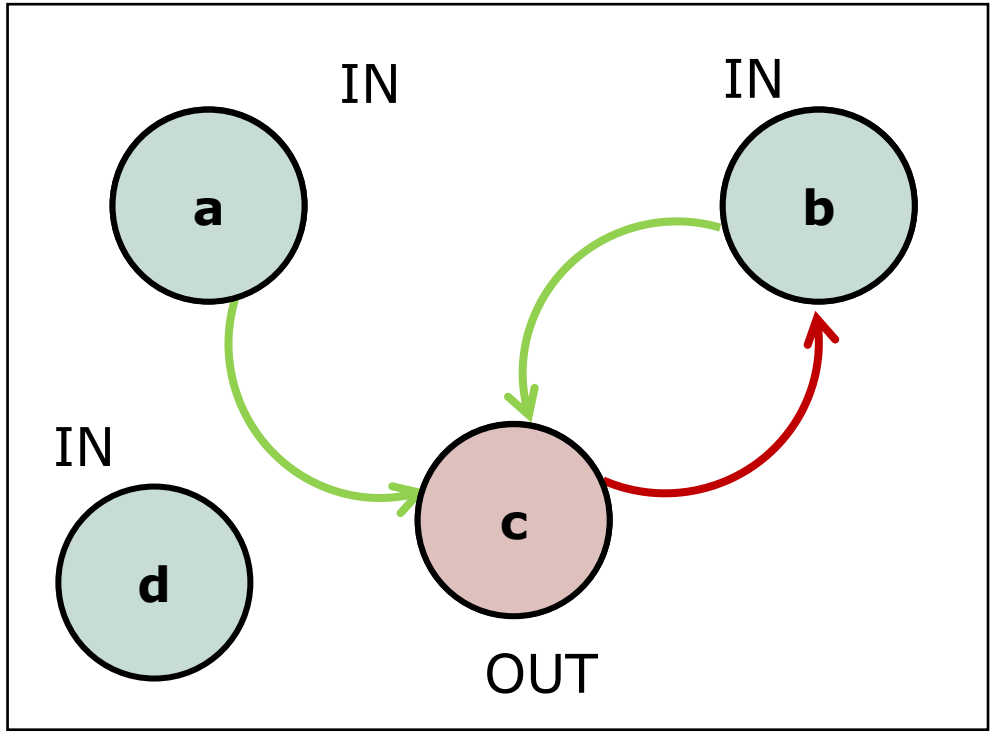
Abstract Argumentation

- Framework
  - S = ⟨Ar , R⟩
- Use semantics of acceptance
  - Labellings are expressive
  - IN means accepted
  - OUT means rejected



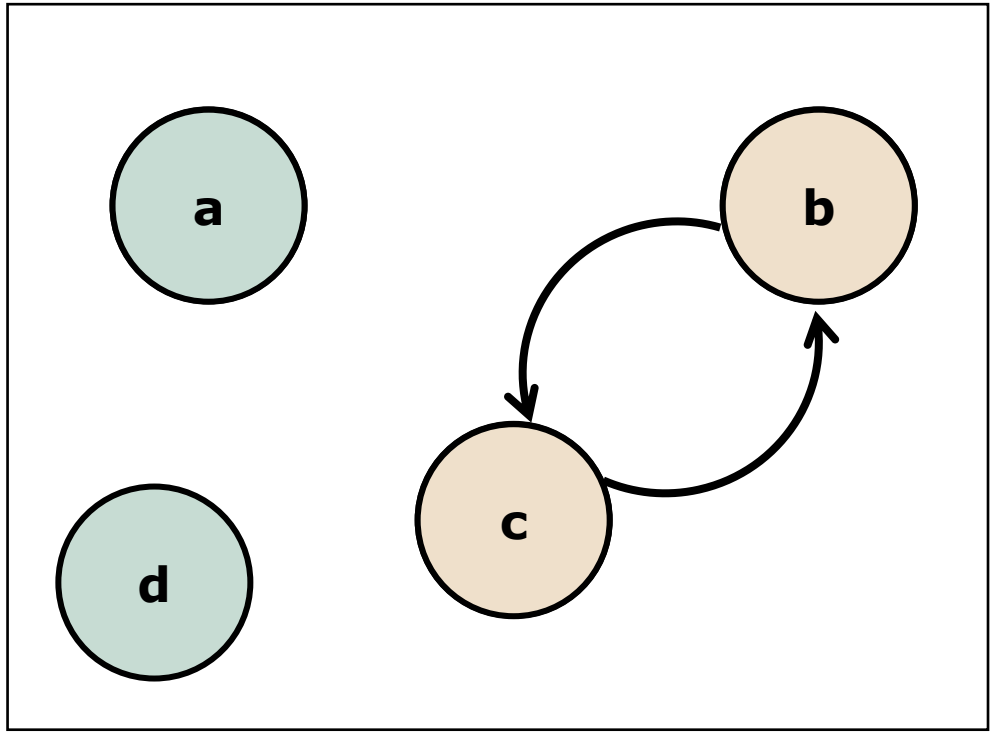
Abstract Argumentation

- Framework
  - S = ⟨Ar , R⟩
- Use semantics of acceptance
  - Labellings are expressive
  - IN means accepted
  - OUT means rejected
  - UNDEC means undecided



Abstract Argumentation

- Framework
  - $S = \langle Ar , R \rangle$
- Use semantics of acceptance
  - Labellings are expressive
  - IN means accepted
  - OUT means rejected
  - UNDEC means undecided



What about attack strength?

- E.g. GORGIAS: preference-based structured argumentation framework of Logic Programming with Priorities.
- *"one may prefer certain features over others; in scheduling, meeting some deadlines may be more important than meeting others; in legal reasoning, laws are subject to higher principles, like lex superior or lex posterior, which are themselves subject to 'higher order' principles."*

https://www.cs.ucy.ac.cy/~nkd/gorgias/

Kakas AC, Moraitis P, Spanoudakis NI, Cerutti F, Booth R. GORGIAS: Applying argumentation. Argument & Computation. 2019;10(1):55-81. doi:10.3233/AAC-181006

Transparency into XAI assumptions!

- Make assumptions about what informs XAI choices clear for any given stakeholder.
- Reasoning steps can be traced.
- Facts, beliefs, preferences, etc. can be explicitly set and then viewed by system users.
- Graph visualisations can be easily interpretable by humans and handled by machines.

Working example

- Housing sale price (California Housing Dataset).
- Buyer / seller is demanding an explanation to build trust in the model.
- The multi-explainer system consists of LIME, SHAP, and Counterfactual explanation techniques.
- Use contextual knowledge to argue for an “admissible” solution, based on what we/the system “knows”.

Local Interpretable Model-agnostic Explanation (LIME)

- Local: a single instance is explained faithfully (vs. global explainability).
- Model agnostic: can be applied to any model (vs. model-specific).
- Fit a "surrogate" interpretable model in the local perturbed neighbourhood of a single instance and use the new model as an explanation.

Image source: https://arize.com/glossary/local-interpretable-model-agnostic-explanations-lime

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why should i trust you?" Explaining the predictions of any classifier." *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016./

Local Interpretable Model-agnostic Explanation (LIME)

- Surrogate model (e.g. linear) is fit to a perturbed dataset around the local instance and used as the explanation.
- **Pro:** straight forward and intuitive. In fact, they are said to be "human-friendly" … whatever that means.
- **Con:** can't always trust the outcome due to inefficient sampling method in many implementations. Correlated features not accounted for.

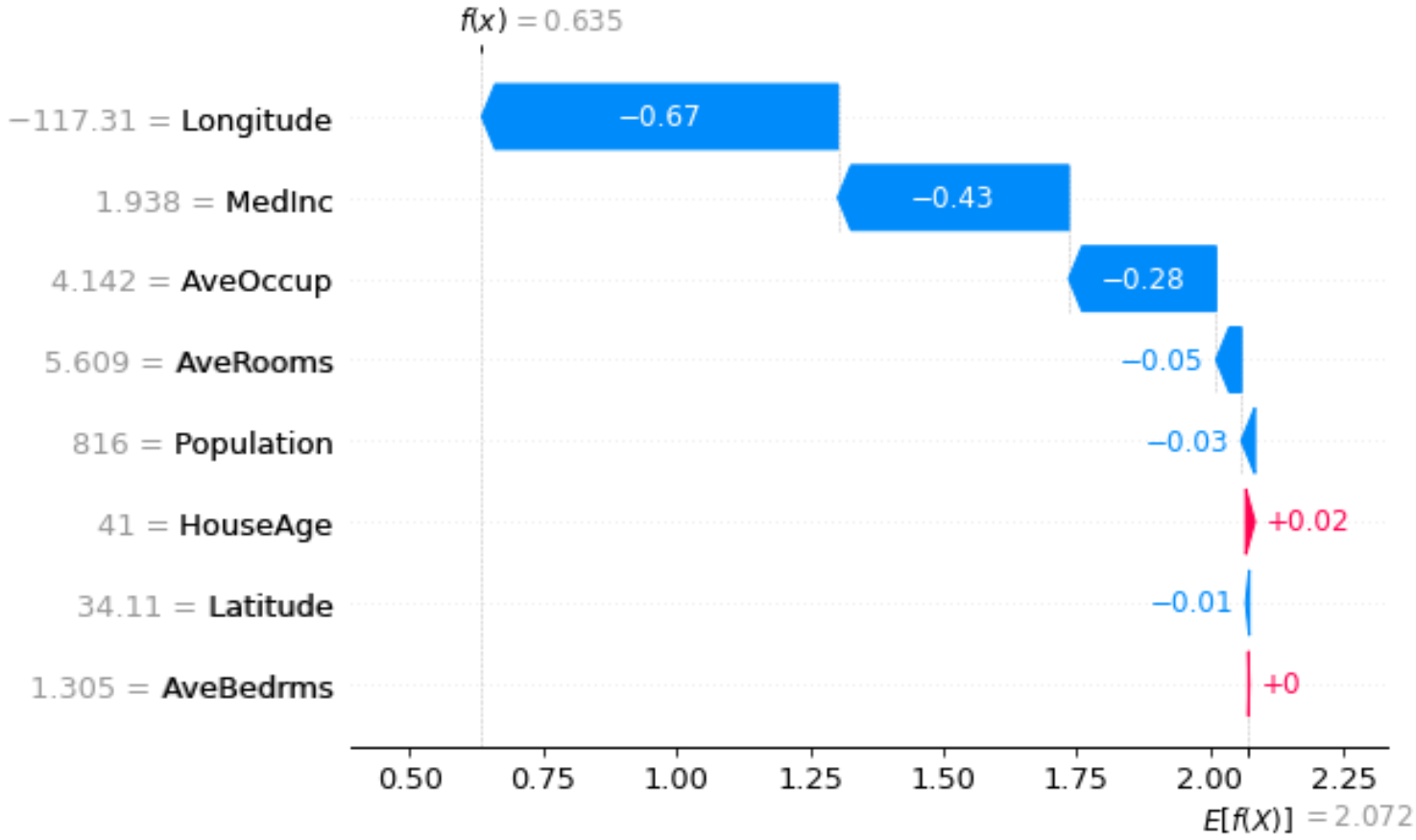
Molnar, C. (2022). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable (2nd ed.). christophm.github.io/interpretable-ml-book/

SHAP

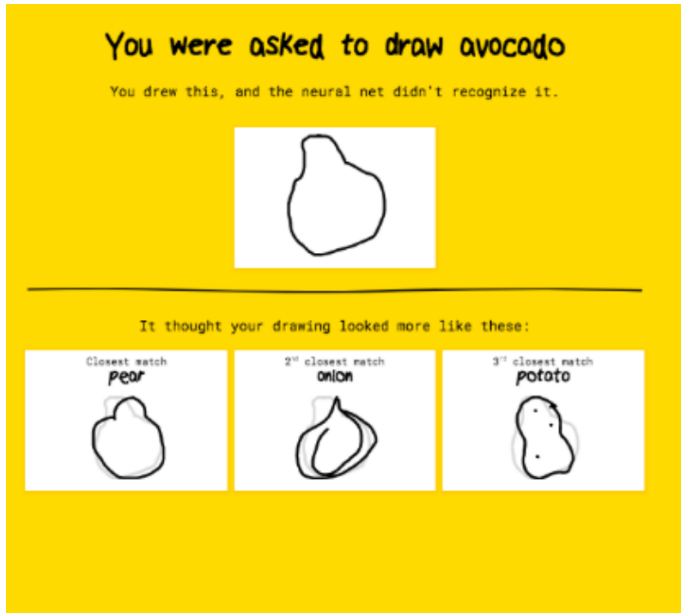
- **SH**apely **A**dditive ex**P**lanations
- Feature attribution method.
- **Pro:** prediction values are fairly distributed amongst feature values. Strong theoretical foundation in game theory.
- **Con:** can be manipulated to offer misleading explanations.

Slack, D. et al. "Fooling lime and shap: Adversarial attacks on post hoc explanation methods." In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, pp. 180-186 (2020).

SHAP


f(x) = 0.635
−117.31 = Longitude
1.938 = MedInc
4.142 = AveOccup
5.609 = AveRooms
816 = Population
41 = HouseAge
34.11 = Latitude
1.305 = AveBedrms
−0.67
−0.43
−0.28
−0.05
−0.03
+0.02
−0.01
+0
0.50
0.75
1.00
1.25
1.50
1.75
2.00
2.25
E[f(X)] = 2.072

Example-based explanations



Comparative

Normative

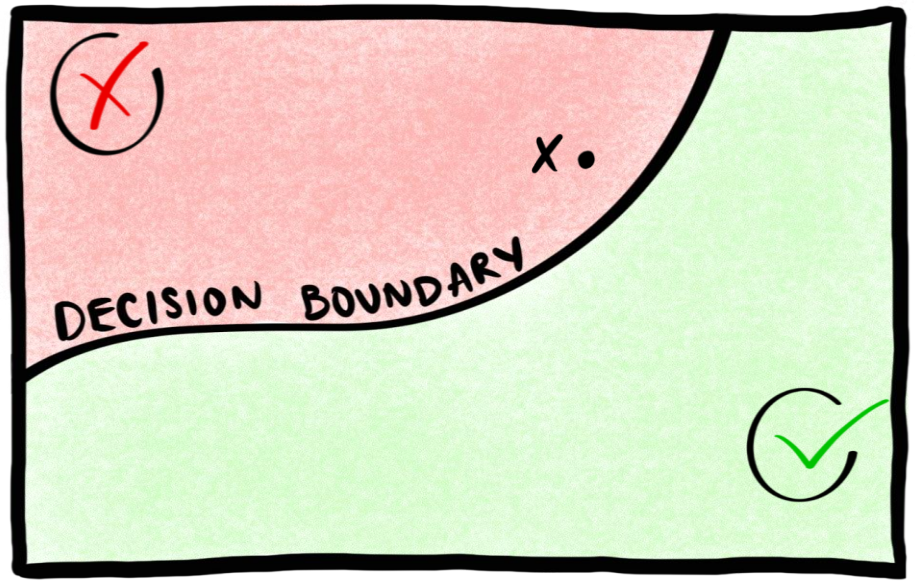
UMEÅ UNIVERSITY

Source: Cai et al. "The effects of example-based explanations in a machine learning interface." (2019)

Counterfactual explanations

- Counterfactuals are also example-based.
- Aim to answer why ‘not’ instead of ‘why’ questions.
- Perform minimal feature changes until an alternative prediction is made.

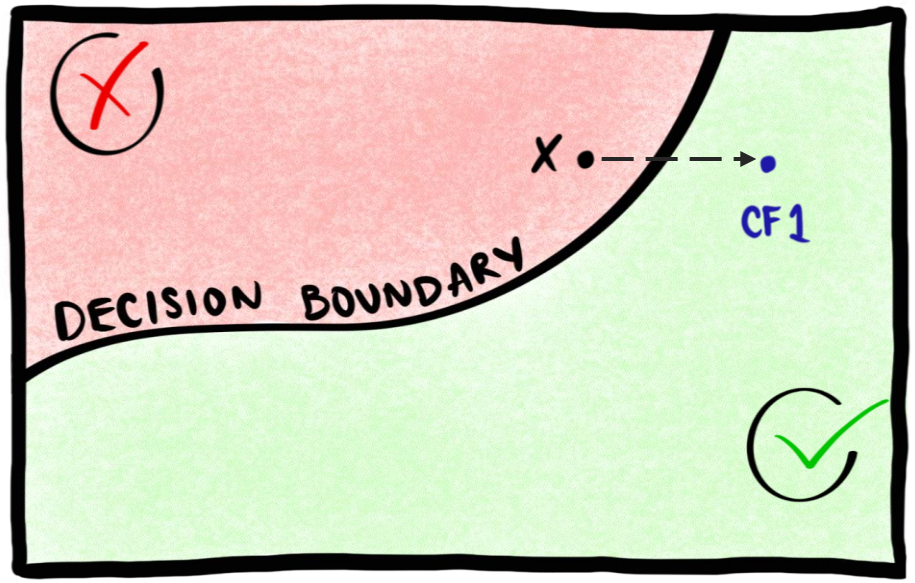
…	Education	Years Experience	…	…
…	…	…	…	…



Counterfactual explanations

- Counterfactuals are also example-based.
- Aim to answer why ‘not’ instead of ‘why’ questions.
- Perform minimal feature changes until an alternative prediction is made.

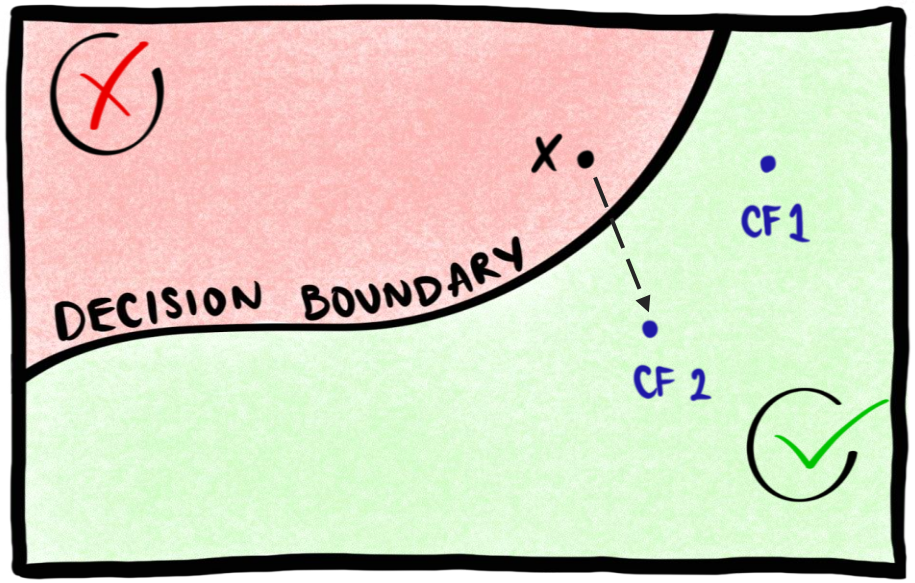
…	Education	Years Experience	…	…
…	…	…	…	…



Counterfactual explanations

- Counterfactuals are also example-based.
- Aim to answer why ‘not’ instead of ‘why’ questions.
- Perform minimal feature changes until an alternative prediction is made.

…	Education	Years Experience	Sex	…
…	…	…	…	…



Counterfactuals

- What feature values need to be changed and by how much in order to “flip” the prediction?
- Example-based method.
- **Pro:** data/model not required to generate explanation.
- **Con:** there can be several counterfactuals. Hard to avoid the Rashomon effect.

Molnar, C. (2022). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable (2nd ed.). christophm.github.io/interpretable-ml-book/

Some rules and preferences

- r1 = use explainer if it’s sparse.
- r2 = don’t use explainer if it’s not computationally cheap.
- r3 = use the explainer if it’s trustworthy

- Preference rules and attacks help us build our graph.
- Prefer computationally cheap over sparse.
- Prefer trustworthiness over computationally cheap.
- Populate KB with facts / beliefs.

Some rules and preferences

- r1 = use explainer if it’s sparse.
- r2 = don’t use explainer if it’s not computationally cheap.
- r3 = use the explainer if it’s **trustworthy (????)**

- Preference rules and attacks help us build our graph.
- Prefer computationally cheap over sparse.
- Prefer trustworthiness over computationally cheap.
- Populate KB with facts / beliefs.

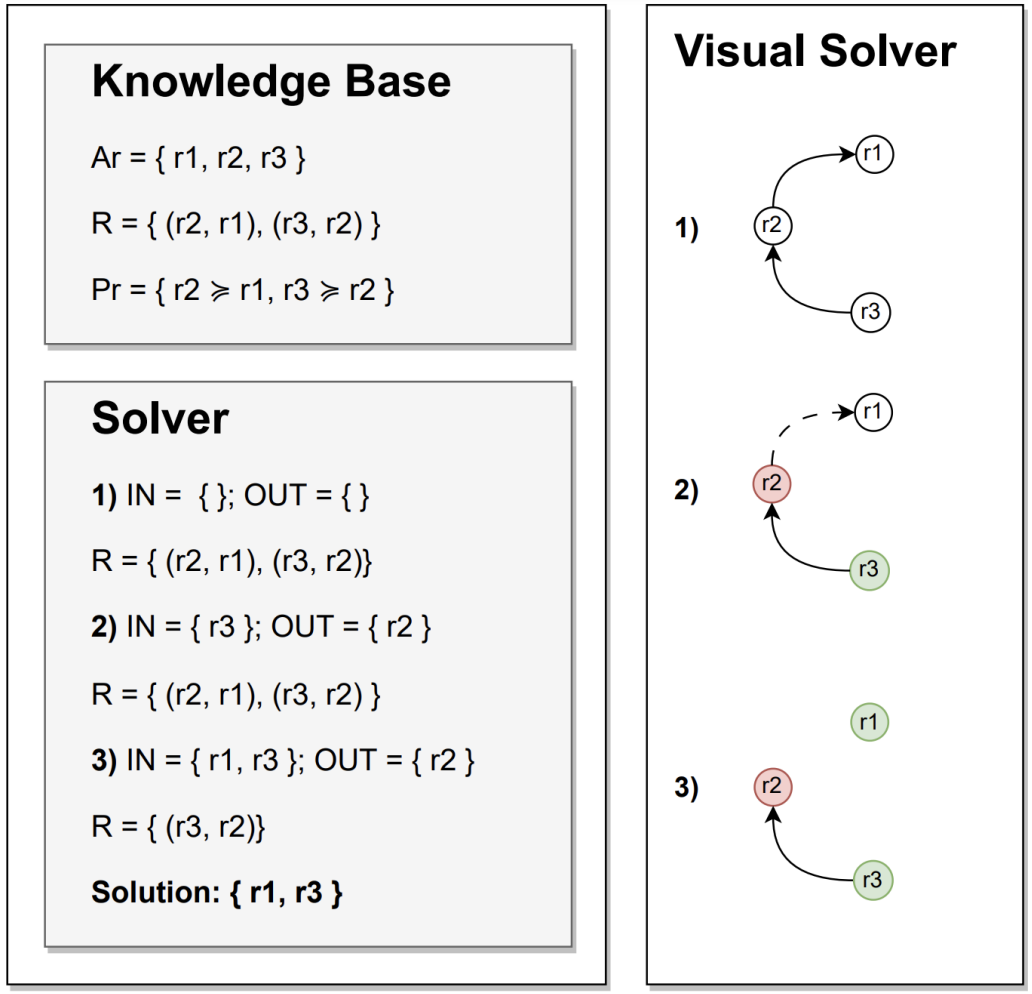
Some rules and preferences

- r1 = use explainer if it’s sparse.
- r2 = don’t use explainer if it’s not computationally cheap.
- r3 = use the explainer if it’s **trustworthy**
- Trustworthiness dependent on stability and susceptibility to adversarial attack.

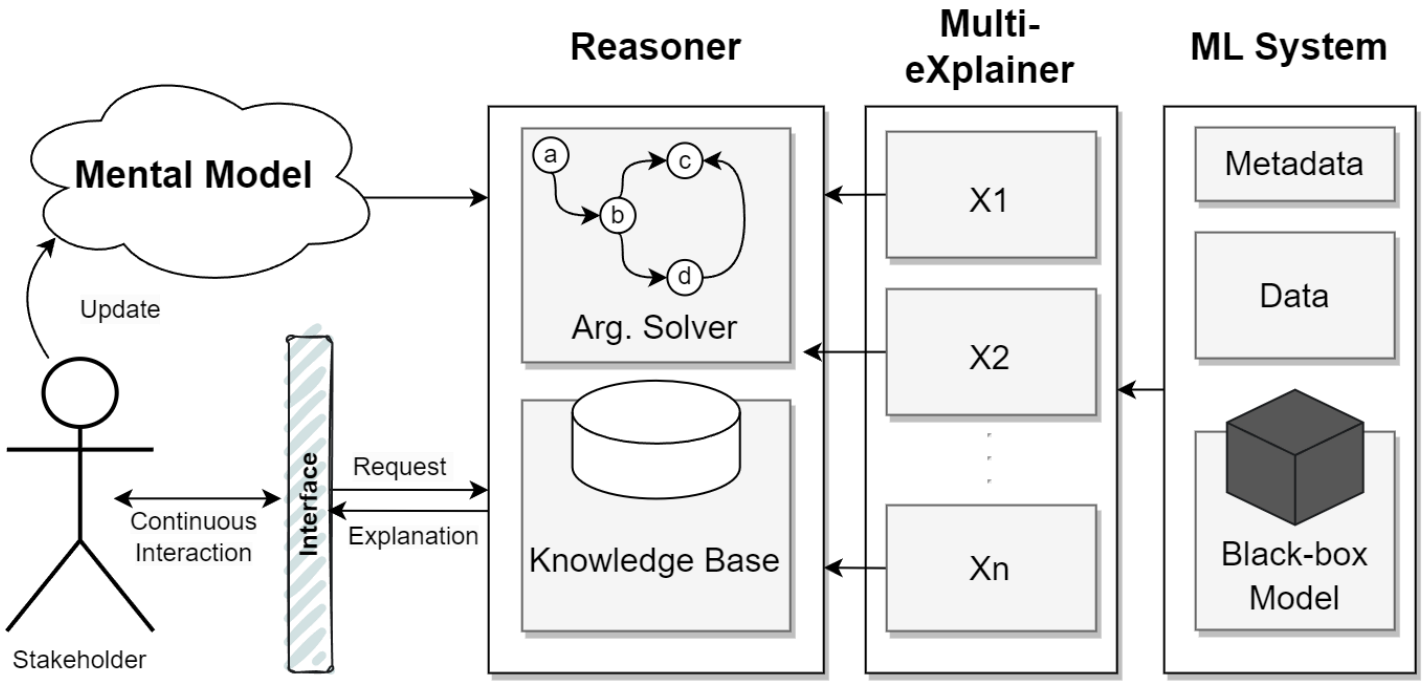
- Preference rules and attacks help us build our graph.
- Prefer computationally cheap over sparse.
- Prefer trustworthiness over computationally cheap.
- Populate KB with facts / beliefs.

Visualising solution steps

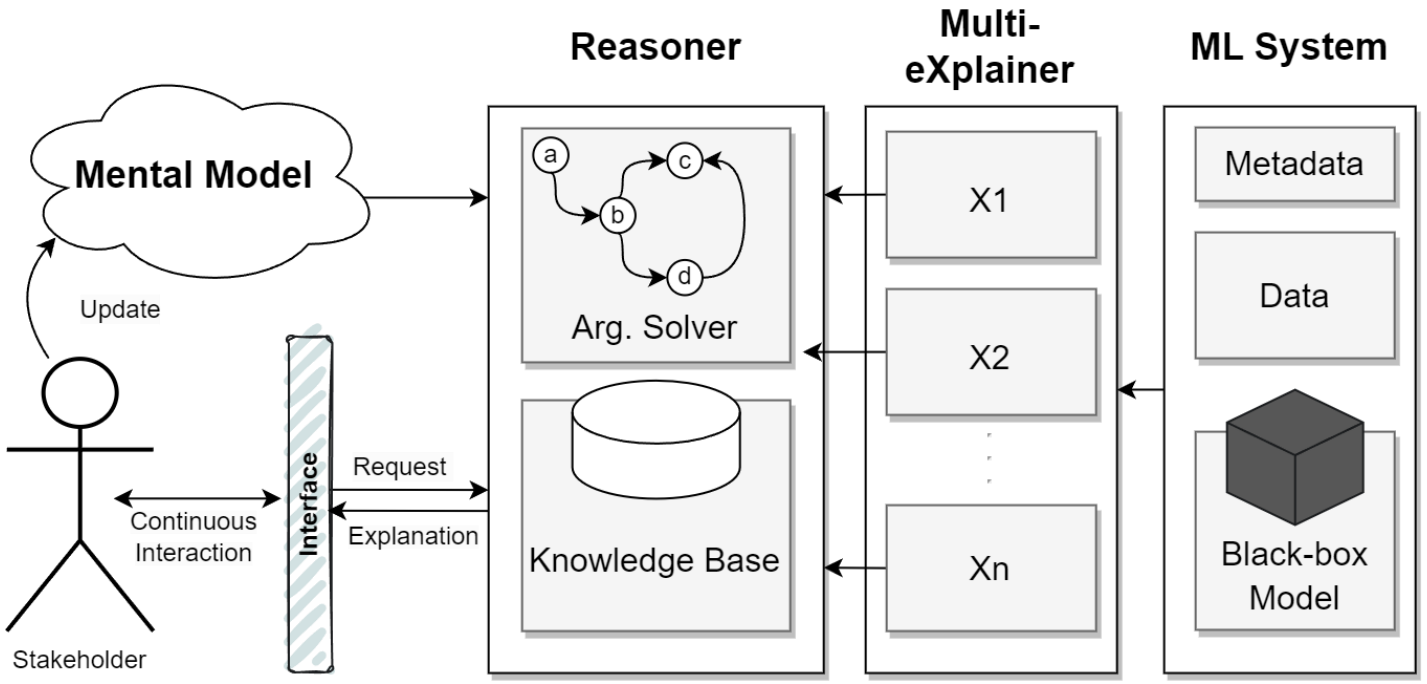
- r1 = use explainer if it’s sparse.
- r2 = don’t use explainer if it’s not computationally cheap
- r3 = use the explainer if it’s trustworthy
- Trustworthiness computed based on e.g. stability



RQ 1: *can facts (and potentially beliefs) about the data / model, the stakeholder, and various explanation techniques be utilised to argue for the most <u>suitable</u> explanation in the given context?*



RQ 1: *can facts (and potentially beliefs) about the data / model, the stakeholder, and various explanation techniques be utilised to argue for the most <u>suitable</u> explanation in the given context?*



Mental Model:

"any internal representation of the relations between a set of elements … [such as] expectations regarding use and consequences … used to guide the individual's interactions with the system or product in question."

—American Psychological Association.

Why care about XAI assumptions?

- What one considers “trustworthy” – is it definition biased? Incomplete? Culturally-determined?
- What one highlights as generally interpretable – is it correct? Robust to adversarial attack?
- We assign many properties to explanation “goodness”.

Attention is not Explanation

Sarthak Jain
Northeastern University
jain.sar@husky.neu.edu

Byron C. Wallace
Northeastern University
b.wallace@northeastern.edu

Abstract

Attention mechanisms have seen wide adoption in neural NLP models. In addition to improving predictive performance, these are often touted as affording transparency: models equipped with attention provide a distribution over attended-to input units, and this is often presented (at least implicitly) as communicating the relative importance of inputs. However, it is unclear what relationship exists between attention weights and model outputs. In this work we perform extensive experiments across a variety of NLP tasks that aim to assess the degree to which attention weights provide meaningful "explanations" for predictions. We find that they largely do not. For example, learned attention weights are frequently uncorrelated with gradient-based measures of feature importance, and one can identify very different attention distributions that nonetheless yield equivalent predictions. Our findings show that standard attention modules do not provide meaningful explanations and should not be treated as though they do. Code to reproduce all experiments is available at https://github.com/successar/AttentionExplanation.

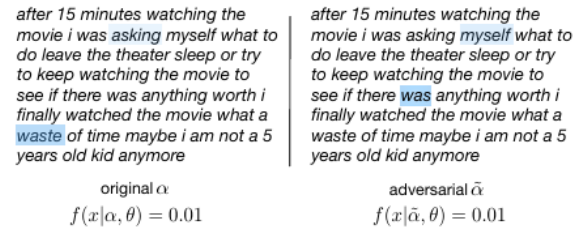


Figure 1: Heatmap of attention weights induced over a negative movie review. We show observed model attention (left) and an adversarially constructed set of attention weights (right). Despite being quite dissimilar, these both yield effectively the same prediction (0.01).

interpretability are common in the literature, e.g., (Xu et al., 2015; Choi et al., 2016; Lei et al., 2017; Martins and Astudillo, 2016; Xie et al., 2017; Mullenbach et al., 2018).[1]

Implicit in this is the assumption that the inputs (e.g., words) accorded high attention weights are responsible for model outputs. But as far as we are aware, this assumption has not been formally evaluated. Here we empirically investigate the relationship between attention weights, inputs, and outputs.

Assuming attention provides a faithful explanation for model predictions, we might expect

Learning to Deceive with Attention-Based Explanations

Danish Pruthi[†], **Mansi Gupta**[‡], **Bhuwan Dhingra**[†], **Graham Neubig**[†], **Zachary C. Lipton**[†]
[†]Carnegie Mellon University, Pittsburgh, USA
[‡]Twitter, New York, USA
ddanish@cs.cmu.edu, mansig@twitter.com,
{bdhingra, gneubig, zlipton}@cs.cmu.edu

Abstract

Attention mechanisms are ubiquitous components in neural architectures applied to natural language processing. In addition to yielding gains in predictive accuracy, attention weights are often claimed to confer *interpretability*, purportedly useful both for providing insights to practitioners and for explaining *why a model makes its decisions* to stakeholders. We call the latter use of attention mechanisms into question by demonstrating a simple method for training models to produce deceptive attention masks. Our method diminishes the to-

Attention	Biography	Label
Original	Ms. X practices medicine in Memphis, TN and is affiliated ... Ms. X speaks English and Spanish.	Physician
Ours	Ms. X practices medicine in Memphis , TN and is affiliated ... Ms. X speaks English and Spanish.	Physician

Table 1: Example of an occupation prediction task where attention-based explanation (highlighted) has been manipulated to whitewash problematic tokens.

sometimes thought of intuitively as indicating which tokens the model *focuses on* when making

Why is Attention Not So Interpretable?

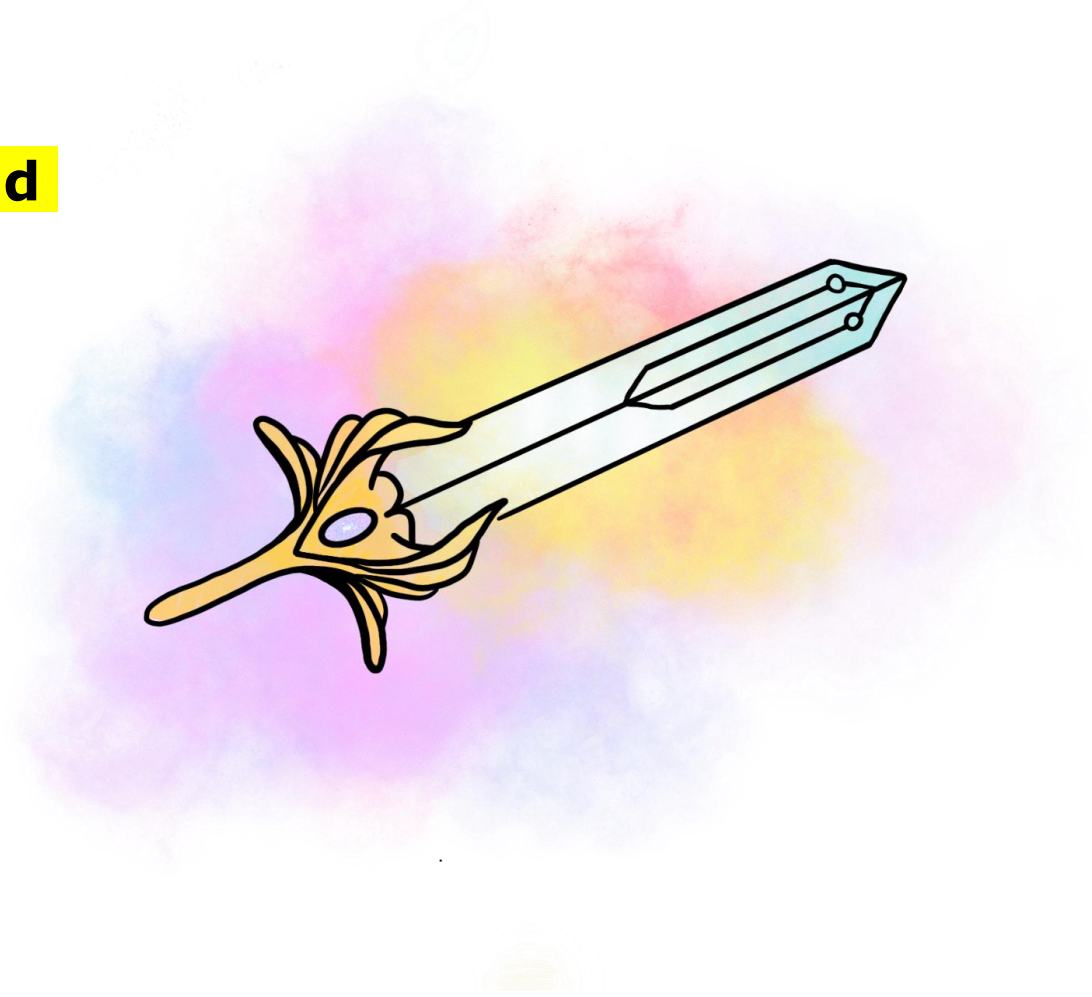
Bing Bai[1*], **Jian Liang**[2*], **Guanhua Zhang**[13], **Hao Li**[1], **Kun Bai**[1], **Fei Wang**[4]
[1]Tencent Inc., China, [2]Alibaba Inc., China,
[3]Harbin Institute of Technology, China, [4]Cornell University, USA
{icebai,guanhzhang,leehaoli,kunbai}@tencent.com,
xuelang.lj@alibaba-inc.com, few2001@med.cornell.edu

Abstract

Attention-based methods have played an important role in model interpretations, where the calculated attention weights are expected to highlight the critical parts of inputs (e.g., keywords in sentences). However, recent research points out that attention-as-importance interpretations often do not work as well as we expect. For example, learned attention weights sometimes highlight less meaningful tokens like "[SEP]", ",", and ".", and are frequently uncorrelated with other feature importance indicators like gradient-based measures. Finally, a debate on the effectiveness of attention-based interpretations has been raised. In this paper, we reveal that one root cause of this phenomenon can be ascribed to the *combinatorial shortcuts*, which stands for that in addition to the highlighted parts, the attention weights themselves may carry extra information which could be utilized by downstream models of attention layers. As a result, the attention weights are no longer pure importance indicators. We theoretically analyze the combinatorial shortcuts, design one intuitive experiment to demonstrate their existence, and propose two methods to mitigate this issue. Empirical studies on attention-based interpretation models are conducted, and the results show that the proposed methods can effectively improve the interpretability of attention mechanisms on a variety of datasets.

XAI is a double-edged sword

- **“The AI system should be provided in a way that allows the overseer to understand its capabilities and limitations”**
- We are selective over what we choose to explain.
- Different methods of explaining may lead to malicious use of XAI too!
- Explanations can be misleading and misinterpreted, even if all actors have good intentions.



Ways forward

- Interdisciplinary methods for impactful XAI methods.
- Human-centricity and context-specificity.
- Interactive and adaptive XAI for effective human-machine teaming.

UMEÅ UNIVERSITY

Thank you