Interactive Explanations for Contestable AI

Francesca Toni

Computational Logic and Argumentation Group

IMPERIAL

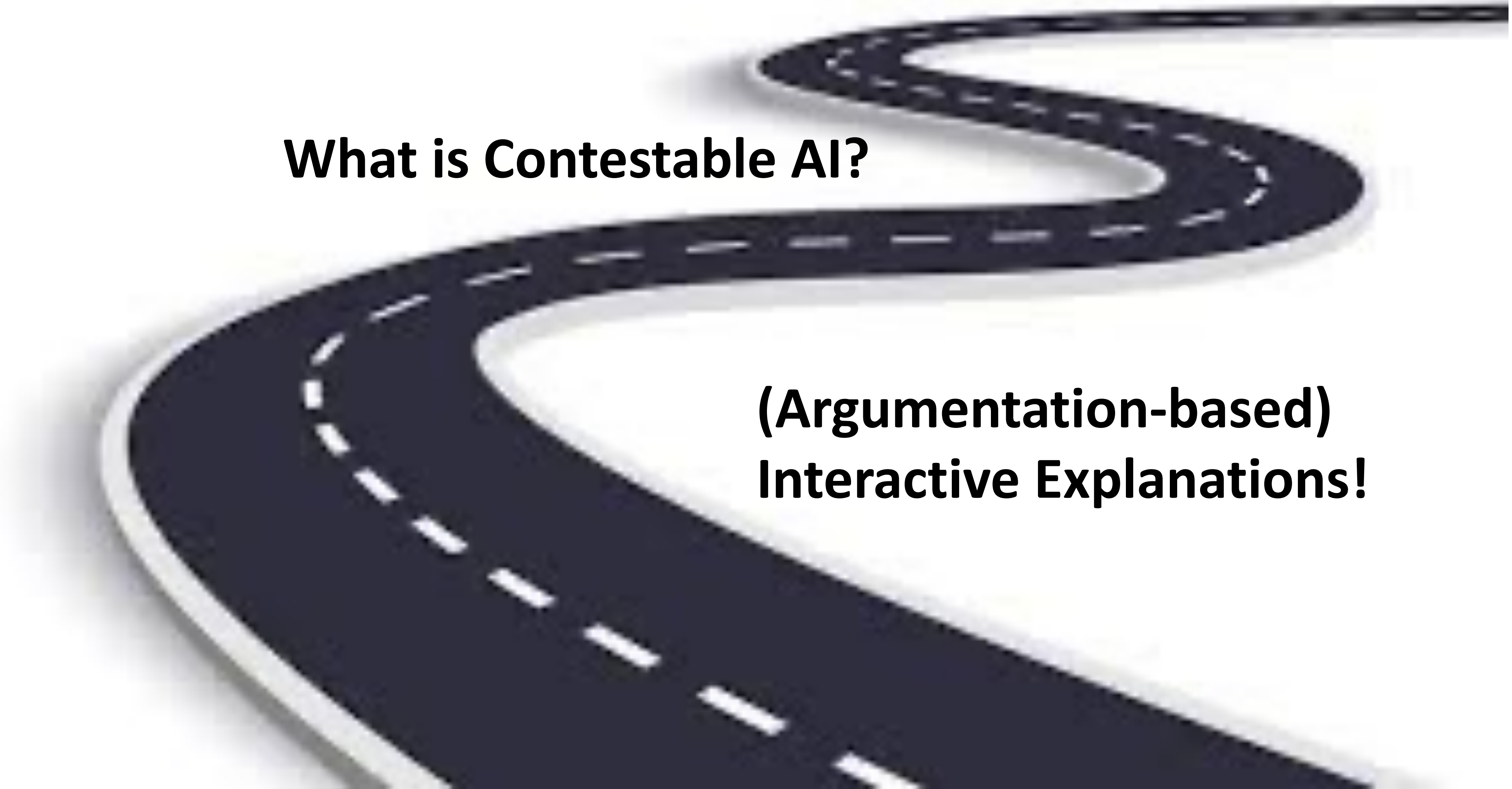
Department of Computing

Centre for Explainable AI

CILC2024

27 June 2024

Plan for this talk



Why Contestable AI?

What is Contestable AI?

(Argumentation-based)
Interactive Explanations!

What can AI do today?

Play (and win) games

AlphaGo

DeepMind

Answer queries

ChatGPT

Debate

Project Debater

IBM

Recognise speech

Hey Siri, call Mum

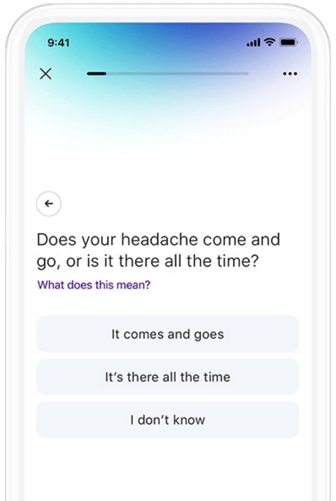
Recognise faces

facebook

Translate across languages

Google Translate

ciao Hi



Does your headache come and go, or is it there all the time?

What does this mean?

It comes and goes

It's there all the time

I don't know

Detect & Diagnose Diseases

babylon

Vacuum clean

dyson

Drive vehicles

WAYMO

Provide legal assistance

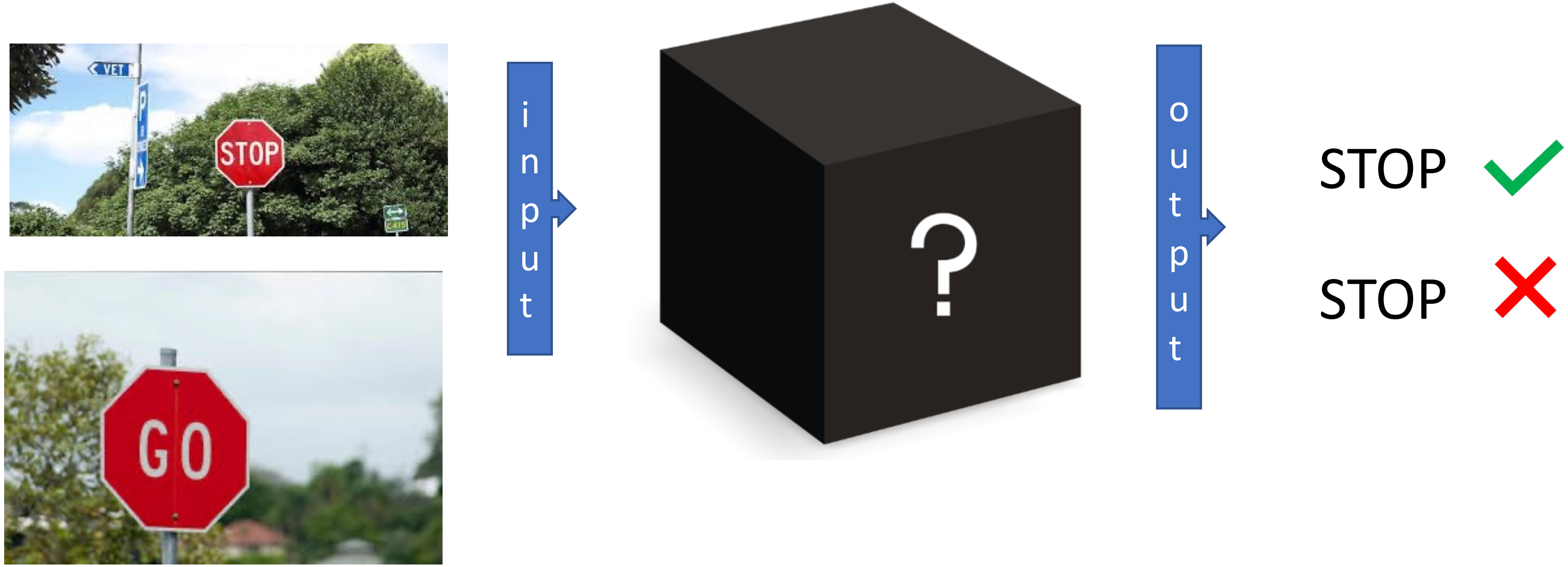
Thomson Reuters

Westlaw Edge UK

Can we trust AI?

AI=

Black-box AI systems



input

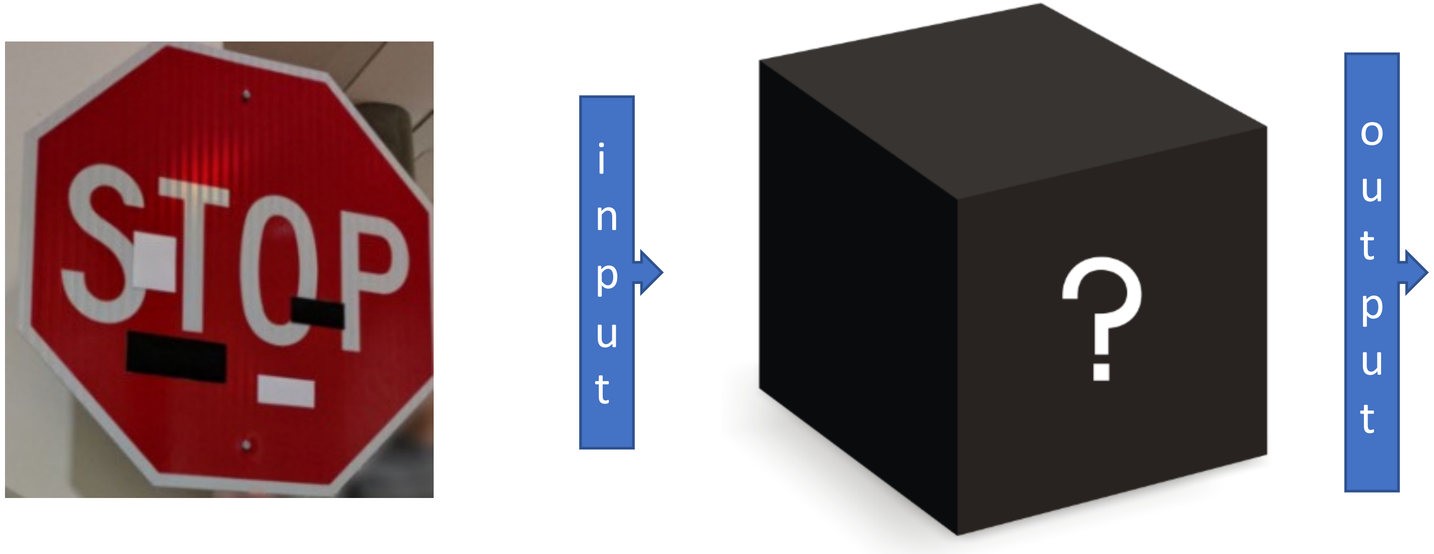
output

STOP ✓

STOP ✗

may learn (unwanted) **artifacts**

Black-box AI systems

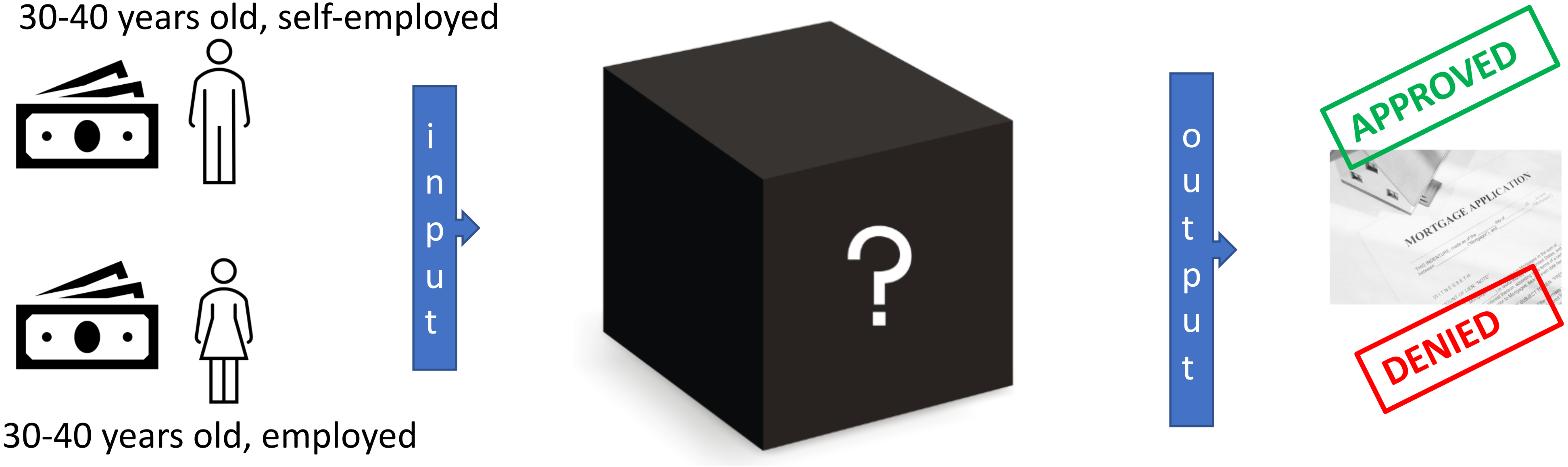


speed limit 45 ✗

may be (maliciously) **manipulated**

Eykholt et al CVPR2018

Black-box AI systems

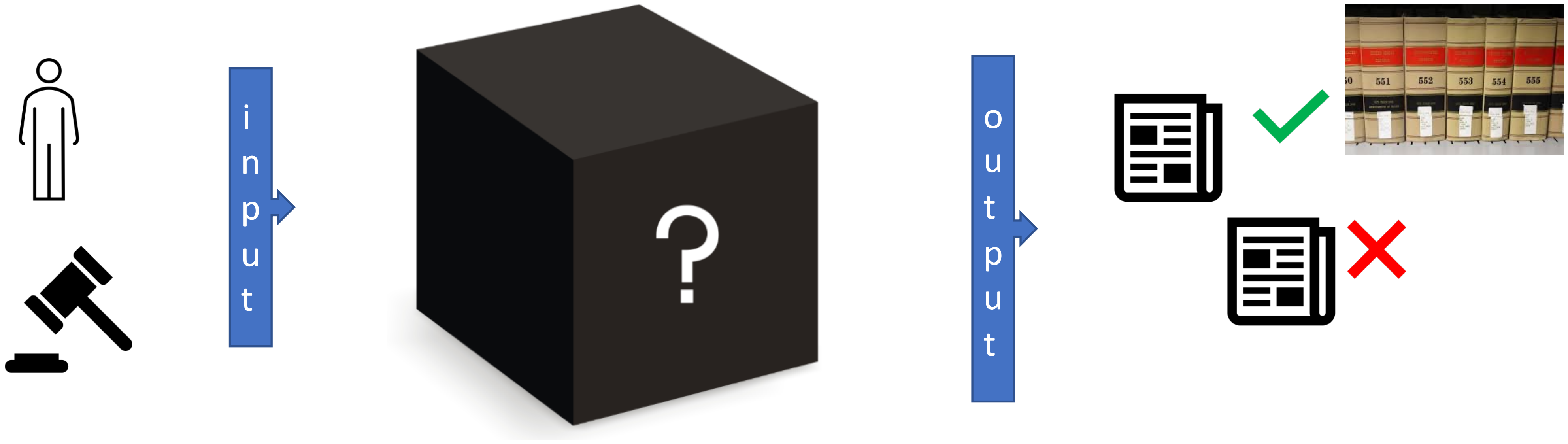


30-40 years old, self-employed

30-40 years old, employed

may be **biased**

Black-box AI systems



may **hallucinate**

Two US lawyers fined for submitting fake court citations from ChatGPT

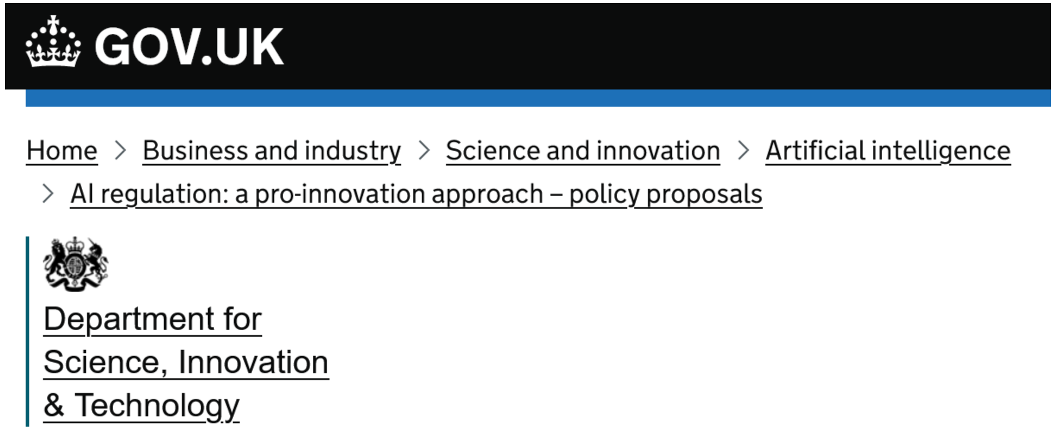
The Guardian

We need **contestable** AI

The **EU General Data Protection Regulation (GDPR)** is the most important change in data privacy regulation in 20 years - we're here to make sure you're prepared.

The Organization for Economic Cooperation and Development

OECD.AI Policy Observatory



GOV.UK

Home > Business and industry > Science and innovation > Artificial intelligence > AI regulation: a pro-innovation approach – policy proposals

Department for Science, Innovation & Technology

GDPR Article 22 (2018)

1. The data subject shall have the right not to be subject to *a decision based solely on automated processing*, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.

2. Paragraph 1 shall not apply if the decision:

(a) is necessary for entering into, or performance of, a contract between the data subject and a data controller;

(b) is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject’s rights and freedoms and legitimate interests; or

(c) is based on the data subject’s explicit consent.

3. In the cases referred to in points (a) and (c) of paragraph 2, the data controller shall implement suitable measures to safeguard the data subject’s rights and freedoms and legitimate interests, at least the right *to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision*.

4. Decisions referred to in paragraph 2 shall not be based on special categories of personal data referred to in Article 9(1), unless point (a) or (g) of Article 9(2) applies and suitable measures to safeguard the data subject’s rights and freedoms and legitimate interests are in place.

OECD.AI
Policy Observatory

Transparency and explainability (Principle 1.3)
This principle is about *transparency and responsible disclosure* around AI systems to ensure that people understand when they are engaging with them and can challenge outcomes

UK pro-innovation framework

Our framework is underpinned by 5 principles to guide and inform the responsible development and use of AI in all sectors of the economy:

- Safety, security and robustness
- Appropriate transparency and explainability
- Fairness
- Accountability and governance
- Contestability and redress

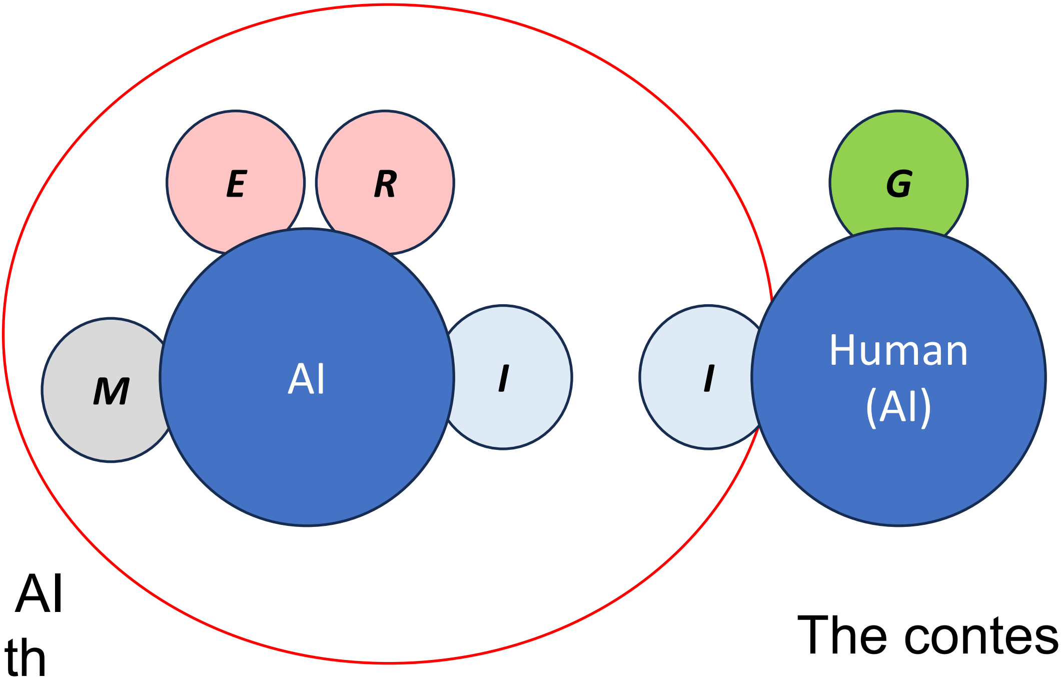
What is Contestable AI?

Leofante et al, arxiv 2024

A human may contest

1. The use of AI and/or the use of (private) data by AI
   1. with a human (regulatory challenge)
2. The outputs of AI
   1. with a human (regulatory challenge)
   2. directly with the AI (regulatory and/or computational challenge)

What is Contestable AI - Computationally?



The contested AI is equipped with

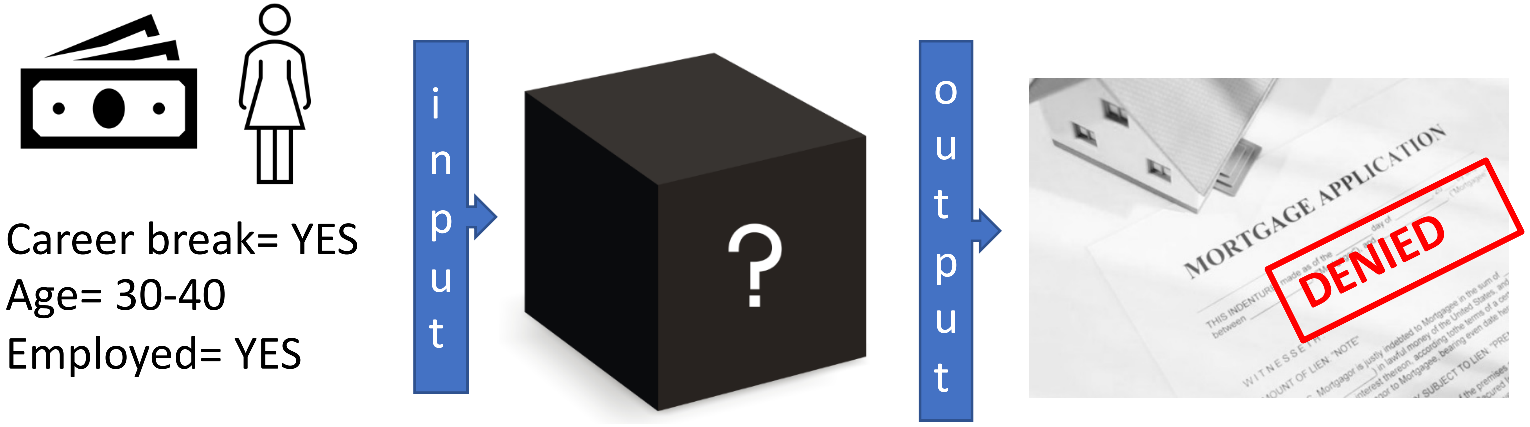
- a model (**M**)
- an explanation method (**E**)
- a redress method (**R**)

The contester (human or even AI) is equipped with

- a ground generator for contestations (**G**)

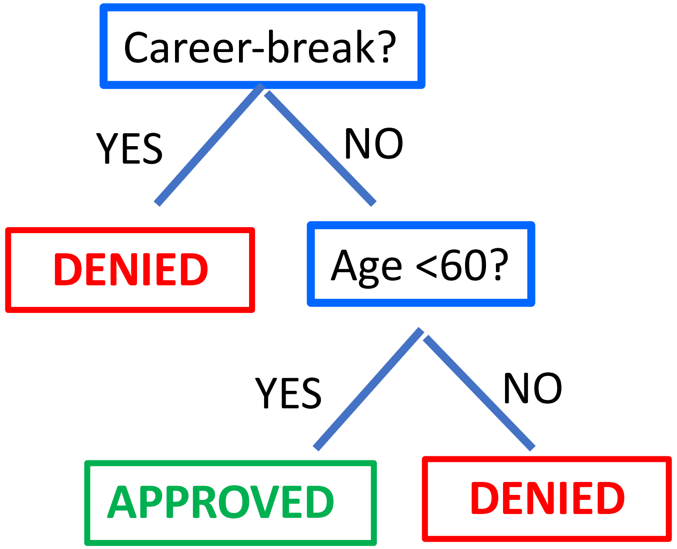
Both contested AI and contester are able to interact (**I**)

(some) XAI solutions for E



Career break= YES
Age= 30-40
Employed= YES

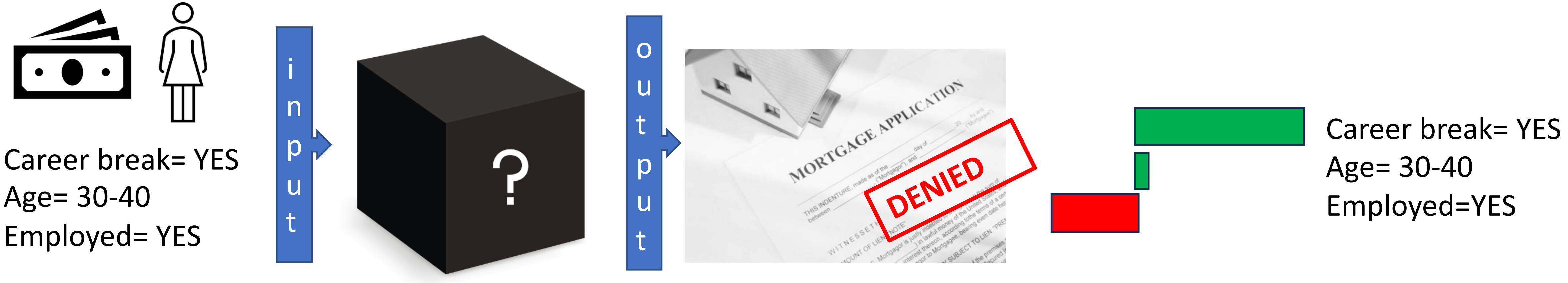
Transparent surrogate model



rule-based explanation

Career break= YES → DENIED

(some) XAI solutions for E



Career break= YES
Age= 30-40
Employed= YES

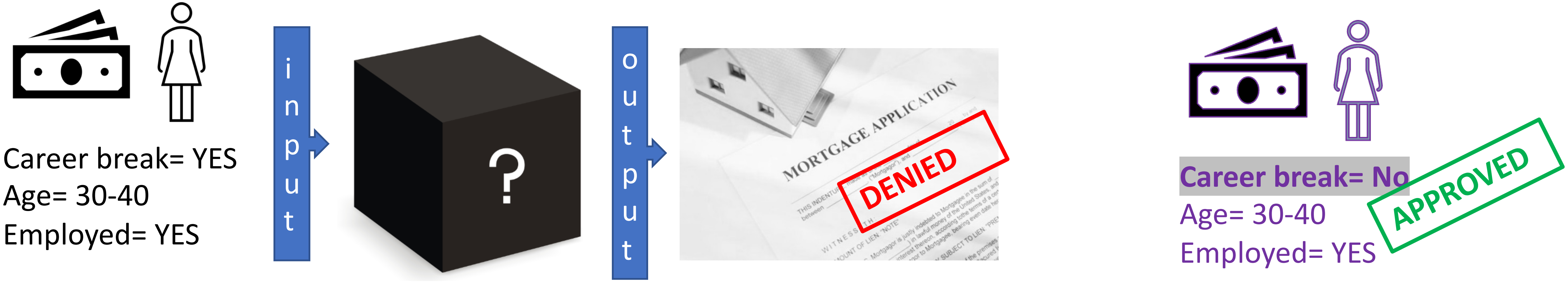
input

output

Career break= YES
Age= 30-40
Employed=YES

attribution-based explanation

(some) XAI solutions for E



Career break= YES
Age= 30-40
Employed= YES

input

?

output

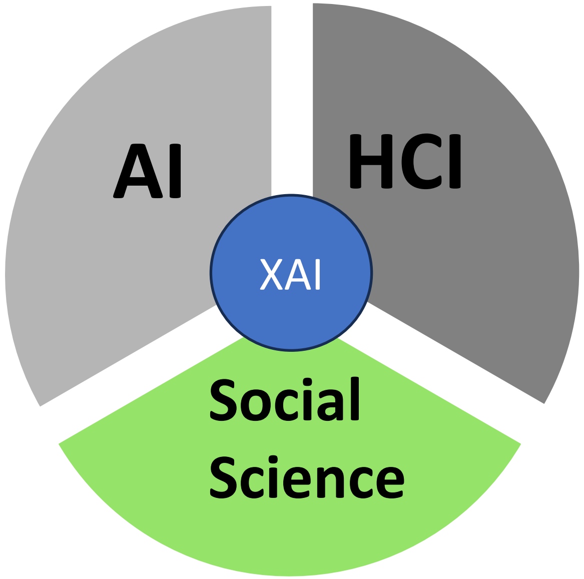
Career break= No
Age= 30-40
Employed= YES

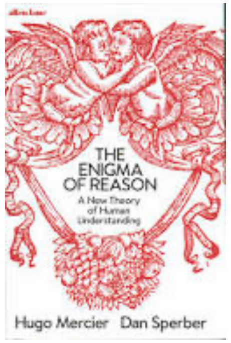
algorithmic recourse = counterfactual explanation
(a form of contestability)

Explanations need to be human-oriented

"Looking at how humans explain to each other can serve as a useful starting point for explanation in AI"

Explanation in artificial intelligence: Insights from the social sciences. Miller; AJ Journal 2019





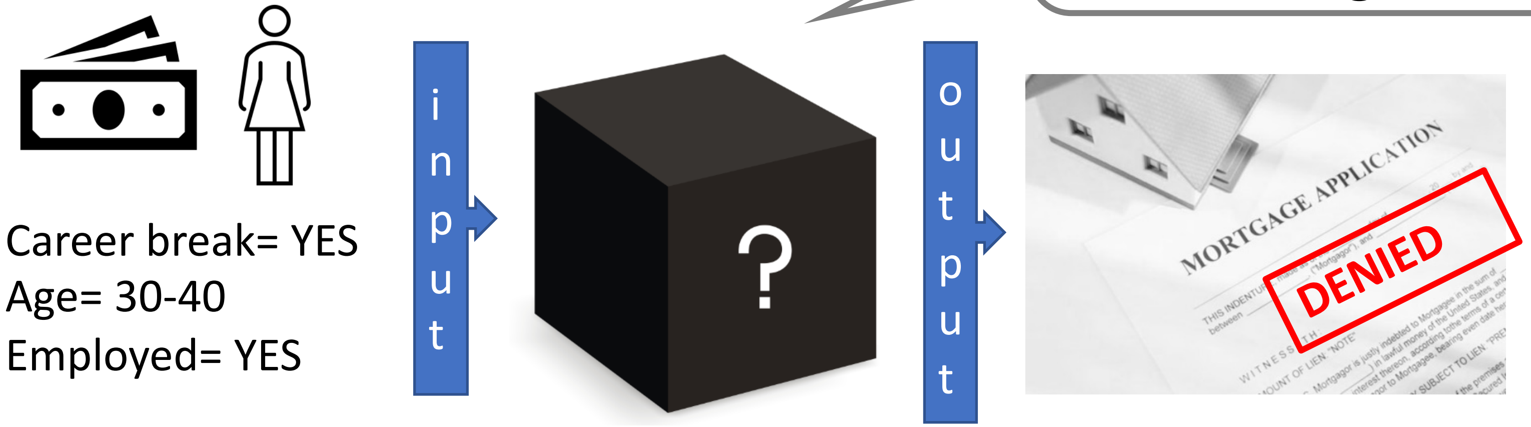
Interactionist view of reasoning

(for explanation)

Interactive explanations?

Why was my mortgage application denied?

Had you taken **no career break** the mortgage would have been granted



Career break= YES
Age= 30-40
Employed= YES

Career break= No
Age= 30-40
Employed= YES

APPROVED

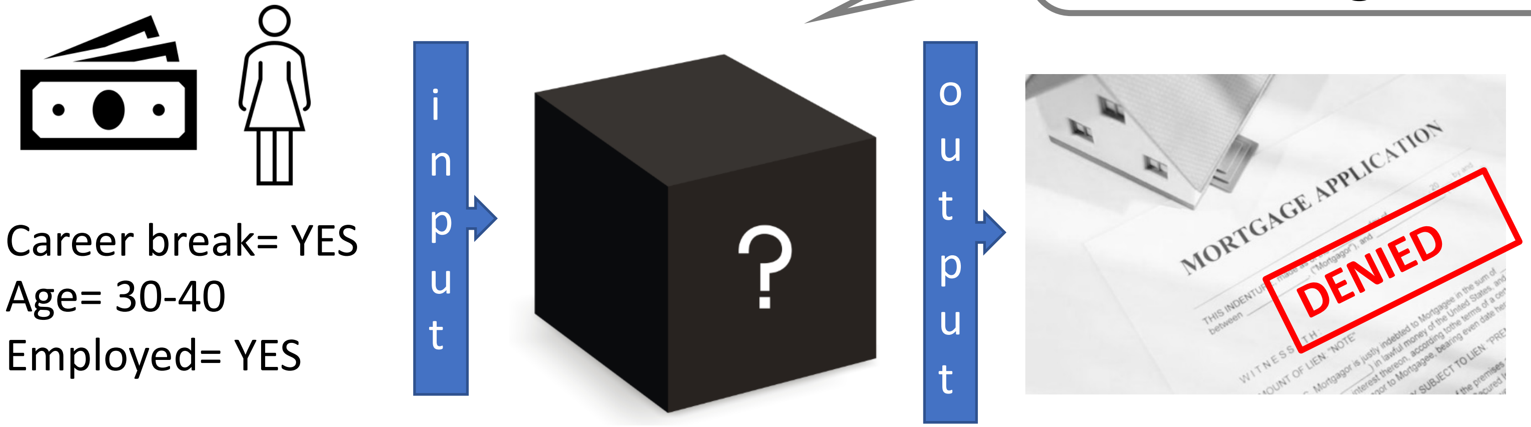
counterfactual explanation

(One-shot) info-seeking!

Interactive explanations?

Why was my mortgage application denied?

Had you taken **no career break** the mortgage would have been granted



Career break= YES
Age= 30-40
Employed= YES

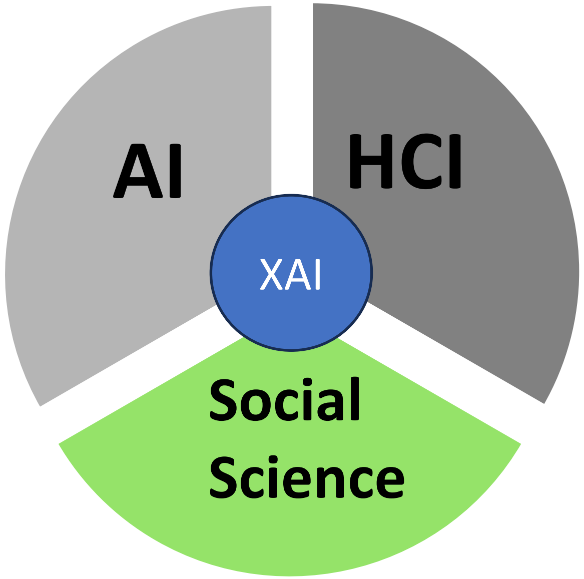
But my career breaks were due to maternity leaves, are you not gender biased?

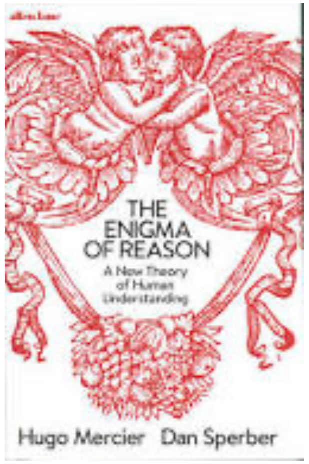
Contestability!

Explanations need to be human-oriented

“Looking at how humans explain to each other can serve as a useful starting point for explanation in AI”

Explanation in artificial intelligence: Insights from the social sciences. Miller; AJ Journal 2019




THE ENIGMA OF REASON
A New Theory of Human Understanding
Hugo Mercier Dan Sperber

Interactionist view of reasoning (for explanation)

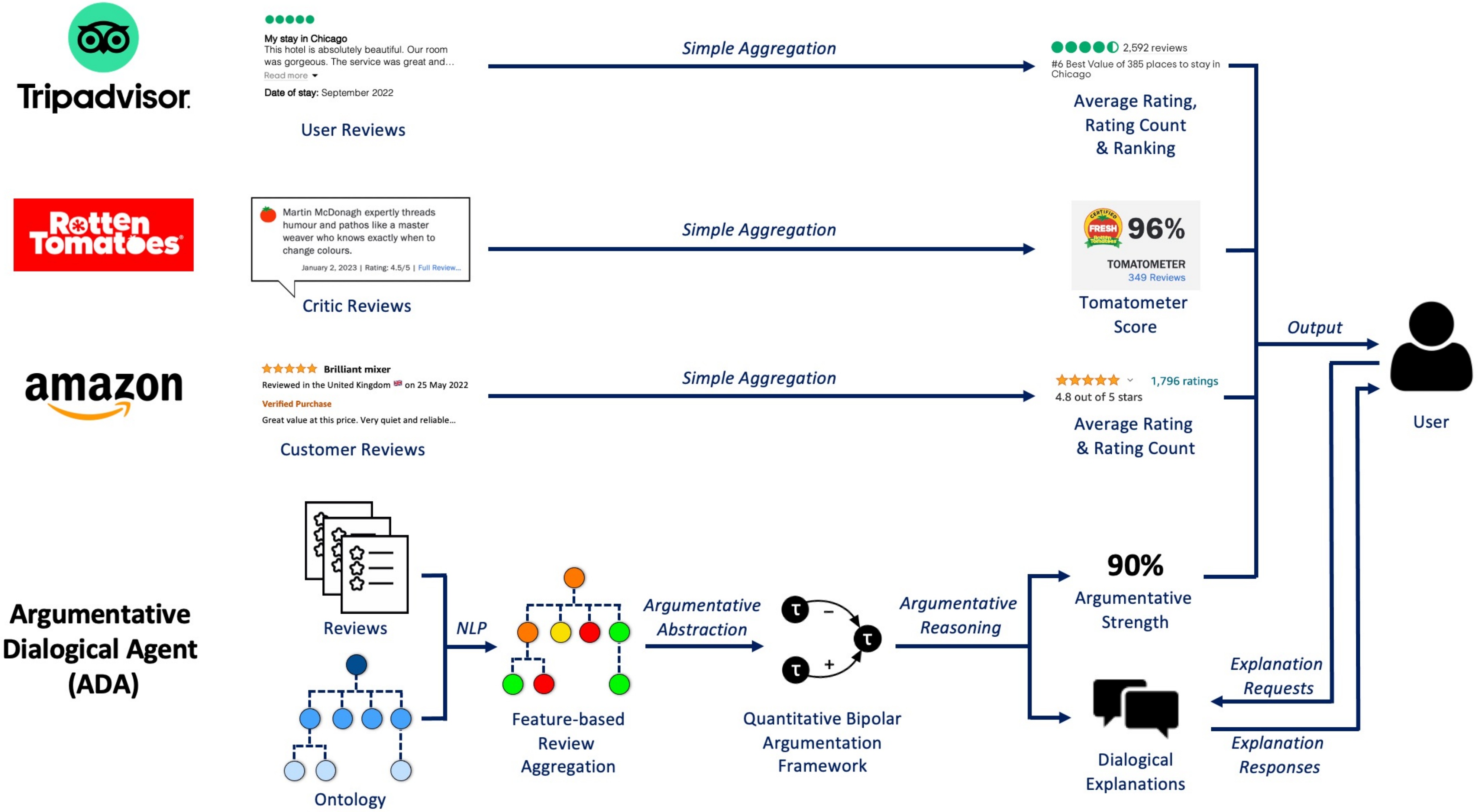
Argumentation-based Interactive Explanations

“the majority of what might look like causal attributions turn out to look like ***argumentative*** claim-backings”

Explaining in conversation: Towards an argument model. Antaki&Leudar. Journal of Social Psychology 1992

Interactive explanations as computational argumentation

ADA (AAMAS 2019)



Tripadvisor

My stay in Chicago
This hotel is absolutely beautiful. Our room was gorgeous. The service was great and…
Read more
Date of stay: September 2022

User Reviews

Simple Aggregation

2,592 reviews
#6 Best Value of 385 places to stay in Chicago

Average Rating, Rating Count & Ranking

Rotten Tomatoes

Martin McDonagh expertly threads humour and pathos like a master weaver who knows exactly when to change colours.
January 2, 2023 | Rating: 4.5/5 | Full Review…

Critic Reviews

Simple Aggregation

96%
TOMATOMETER
349 Reviews

Tomatometer Score

amazon

Brilliant mixer
Reviewed in the United Kingdom on 25 May 2022
Verified Purchase
Great value at this price. Very quiet and reliable…

Customer Reviews

Simple Aggregation

1,796 ratings
4.8 out of 5 stars

Average Rating & Rating Count

Output

User

Argumentative Dialogical Agent (ADA)

Reviews

Ontology

NLP

Feature-based Review Aggregation

Argumentative Abstraction

Quantitative Bipolar Argumentation Framework

Argumentative Reasoning

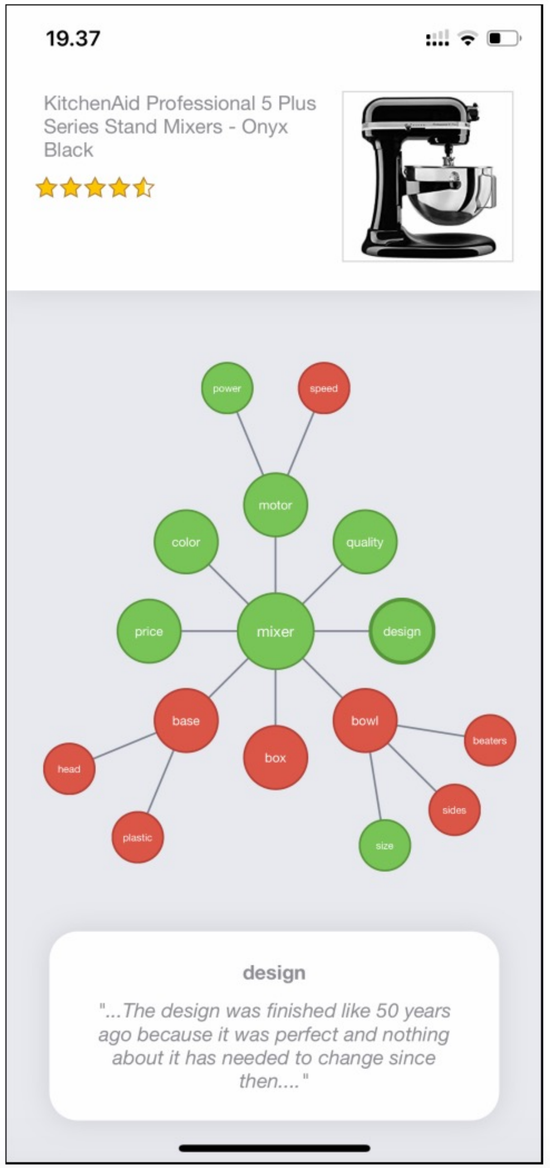
90%
Argumentative Strength

Dialogical Explanations

Explanation Requests

Explanation Responses

Interactive explanations as computational argumentation

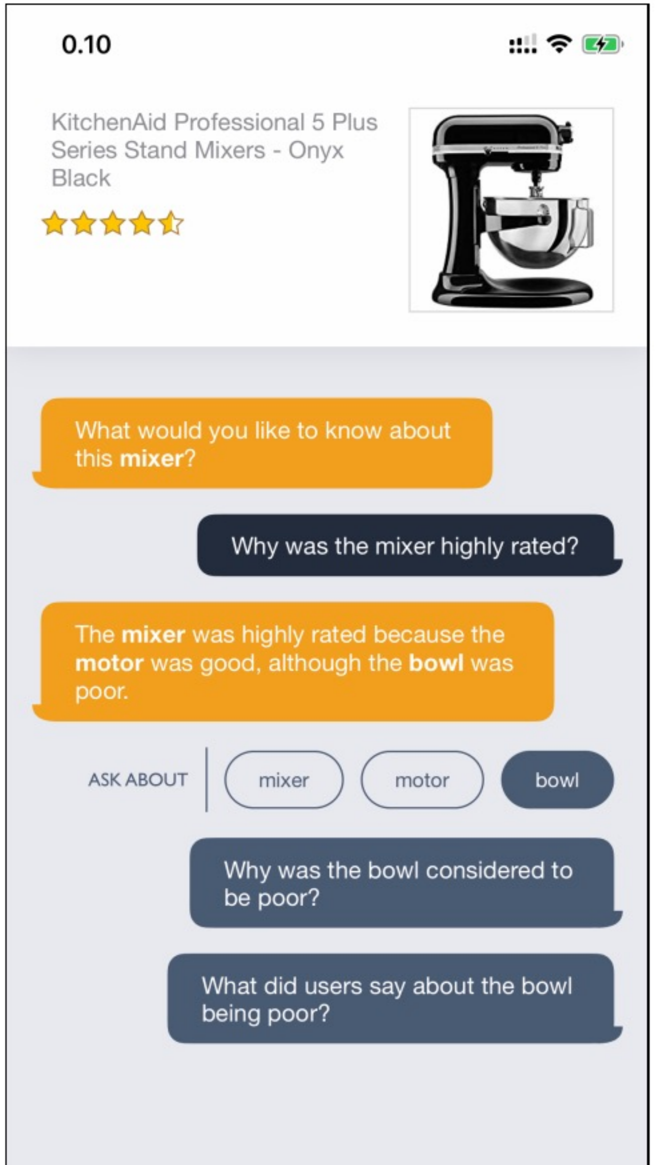


KitchenAid Professional 5 Plus Series Stand Mixers - Onyx Black

★★★★☆

design

"...The design was finished like 50 years ago because it was perfect and nothing about it has needed to change since then...."



KitchenAid Professional 5 Plus Series Stand Mixers - Onyx Black

★★★★☆

What would you like to know about this **mixer**?

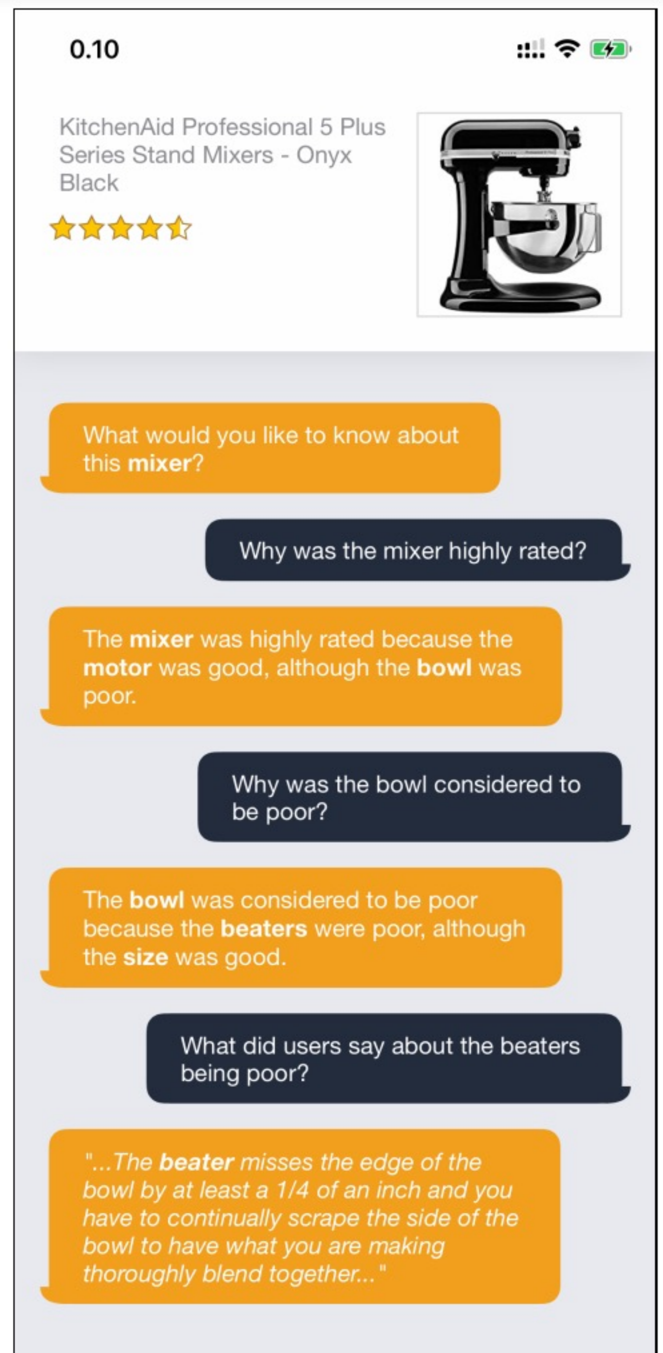
Why was the mixer highly rated?

The **mixer** was highly rated because the **motor** was good, although the **bowl** was poor.

ASK ABOUT | mixer | motor | bowl

Why was the bowl considered to be poor?

What did users say about the bowl being poor?



KitchenAid Professional 5 Plus Series Stand Mixers - Onyx Black

★★★★☆

What would you like to know about this **mixer**?

Why was the mixer highly rated?

The **mixer** was highly rated because the **motor** was good, although the **bowl** was poor.

Why was the bowl considered to be poor?

The **bowl** was considered to be poor because the **beaters** were poor, although the **size** was good.

What did users say about the beaters being poor?

*"...The **beater** misses the edge of the bowl by at least a 1/4 of an inch and you have to continually scrape the side of the bowl to have what you are making thoroughly blend together..."*

Interactive explanations as computational argumentation

DAX 2020 (arxiv)
SpArX 2023 (ECAI)

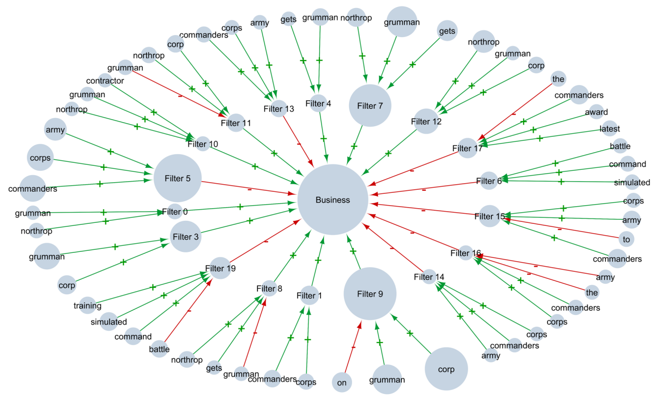
northrop grumman corp on Monday said it received a 10 year $ 408 million army contract to provide simulated battle command training support to army corps commanders the latest award in…

input

output

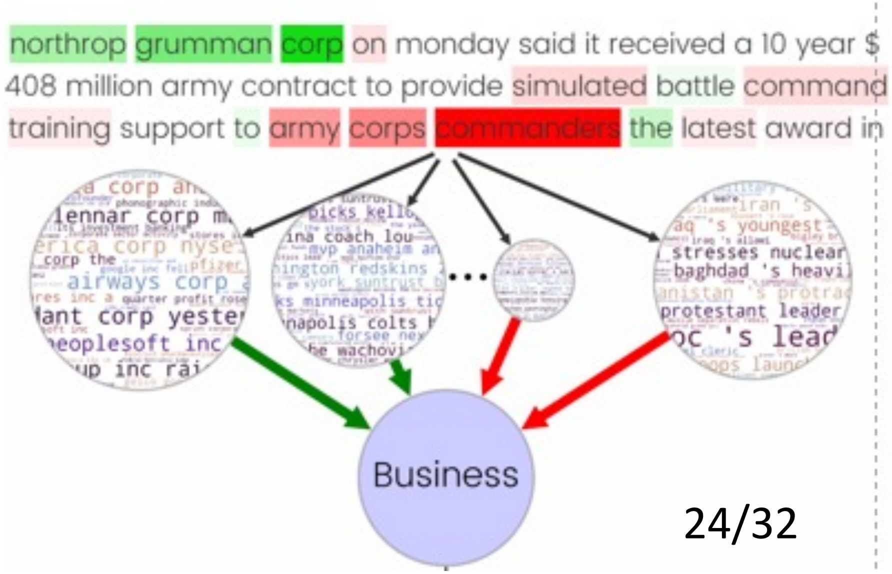
Business (65%)

Human-readable "debate"



Abstraction

= argumentation framework

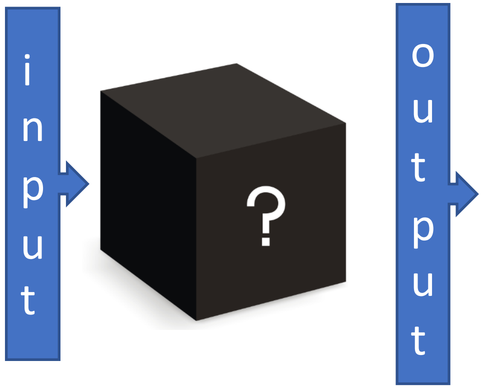


northrop grumman corp on monday said it received a 10 year $ 408 million army contract to provide simulated battle command training support to army corps commanders the latest award in

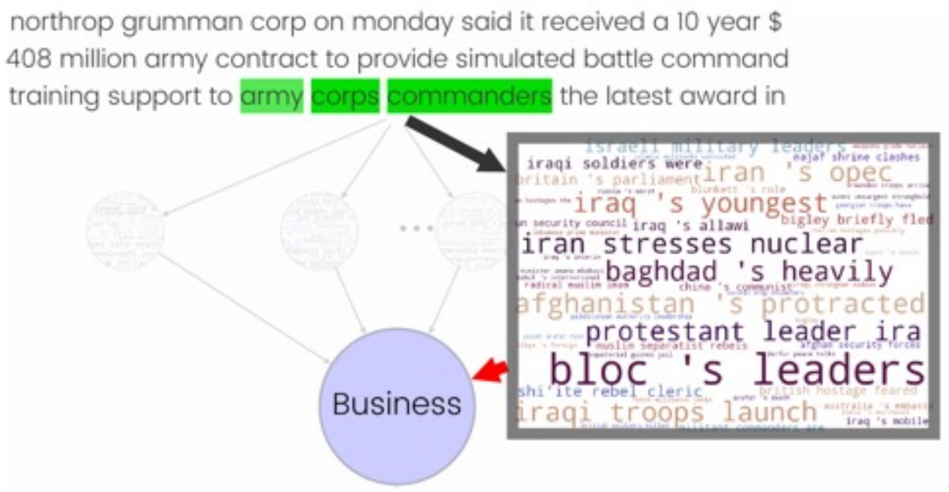
green =support
red =attack

Interactive explanations as computational argumentation

northrop grumman corp on Monday said it received a 10 year $ 408 million army contract to provide simulated battle command training support to army corps commanders the latest award in…

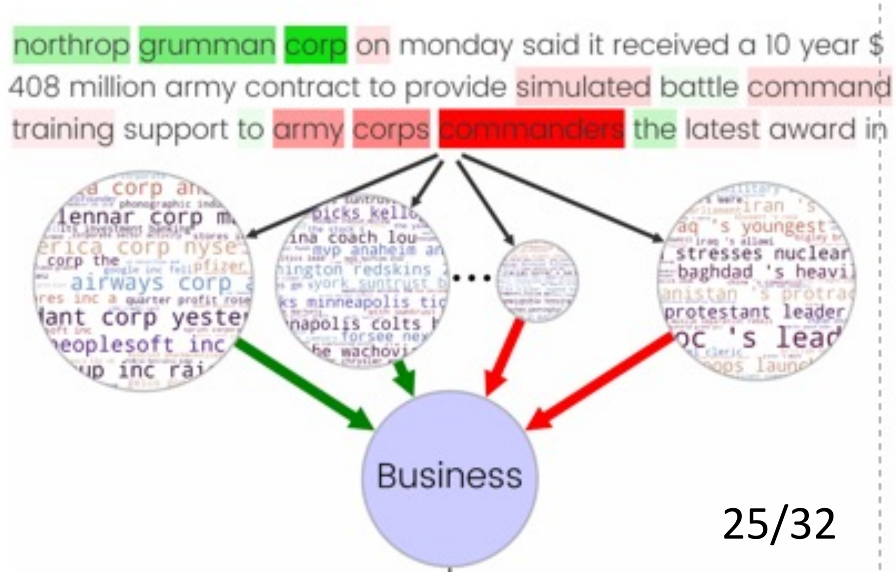


Business (65%)



northrop grumman corp on monday said it received a 10 year $ 408 million army contract to provide simulated battle command training support to army corps commanders the latest award in

Interactive (visual) Explanation

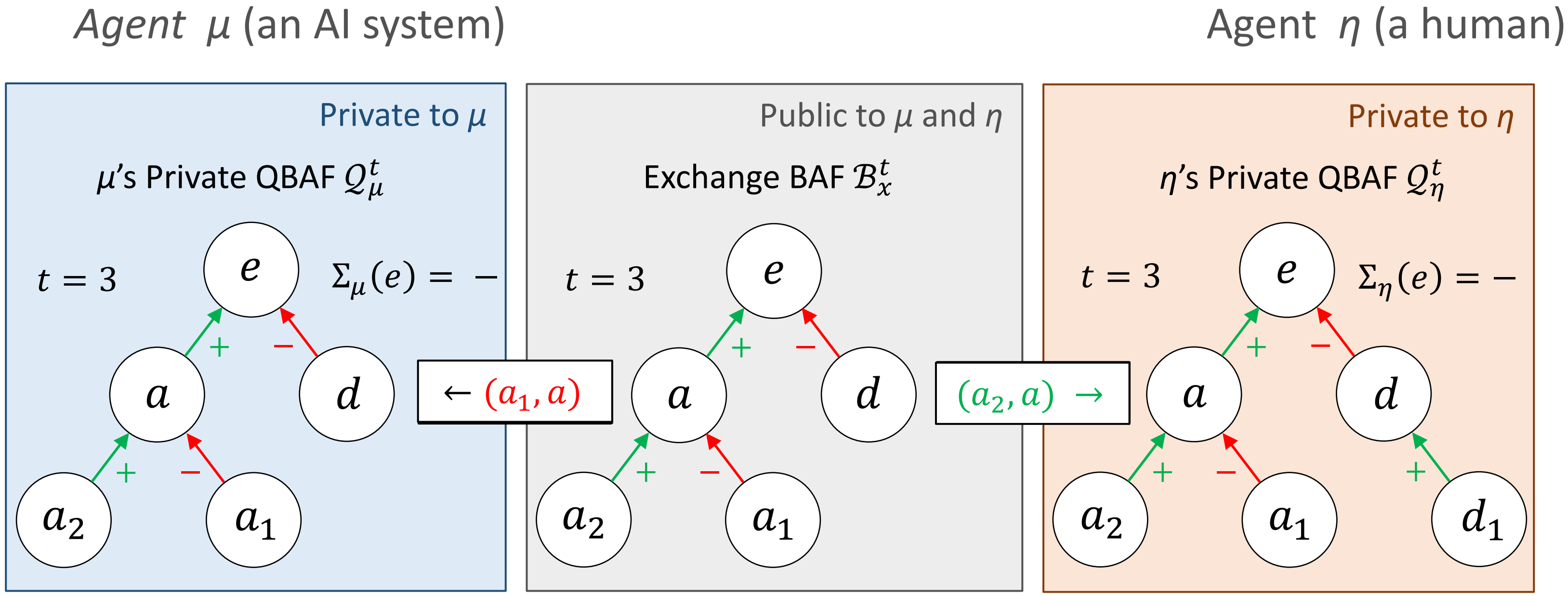


northrop grumman corp on monday said it received a 10 year $ 408 million army contract to provide simulated battle command training support to army corps commanders the latest award in

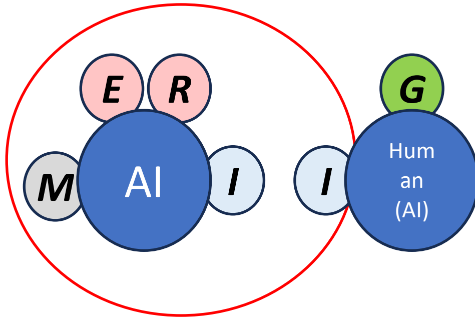
Rago et al KR2023

Interaction via Argumentative Exchanges

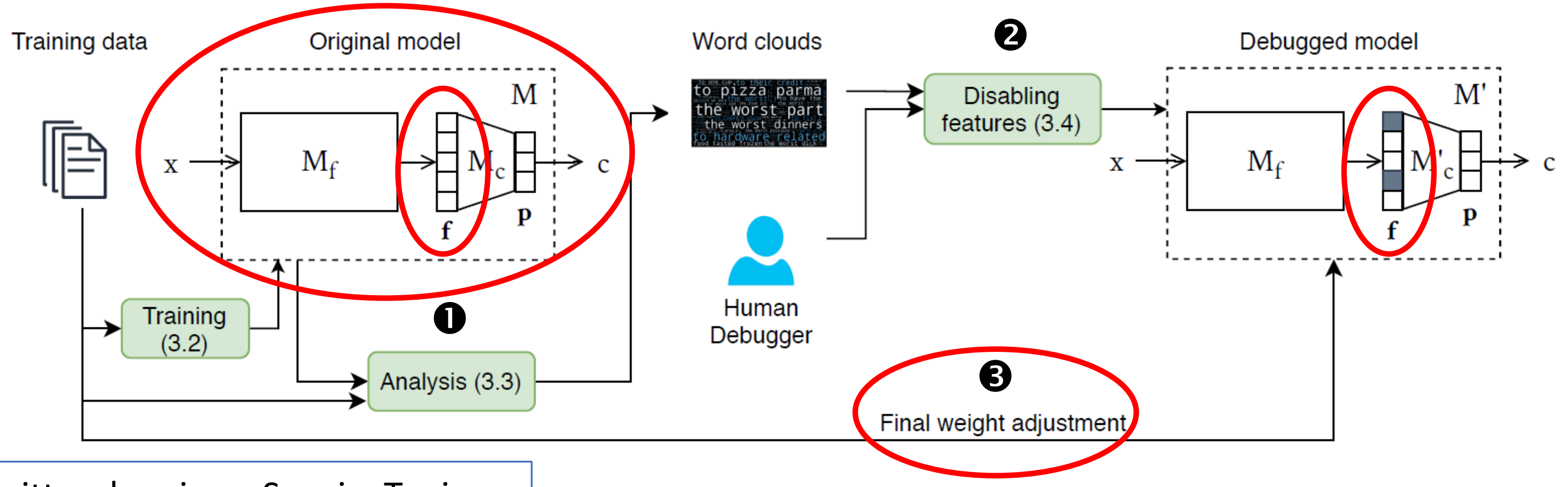
- Formal frameworks to capture various forms of *argumentation-based interactive explanations*
- An (abstract) example:



Forms of Redress (R) – (partial) access to model



- Data augmentation, retraining with updated data/loss function
  - Costly, may not always be possible with legacy/inaccessible code (e.g. LLM)
- Fine tuning (systemic, post-hoc)



Lertvittayakumjorn, Specia, Toni, EMNLP2020

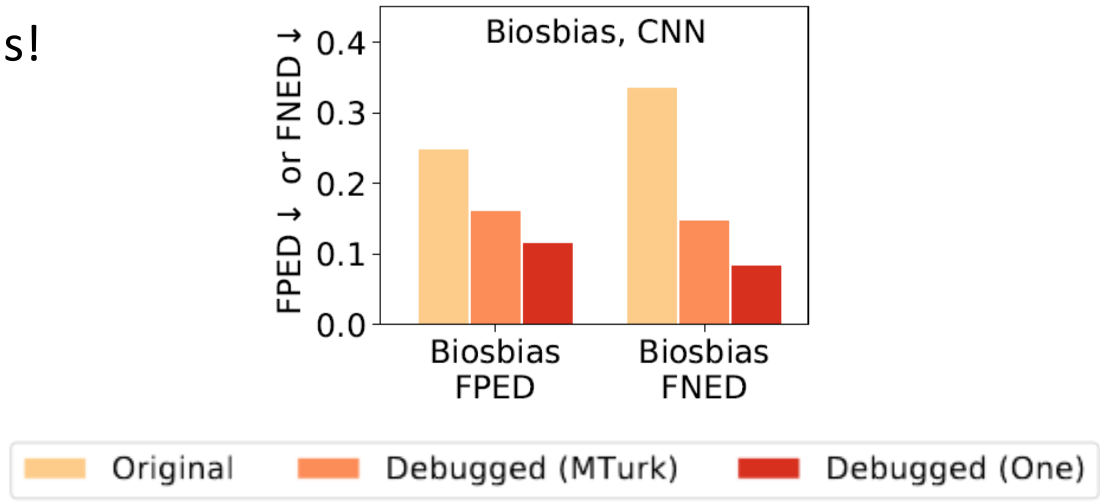
FIND: **F**eature **I**nvestigation a**N**d **D**isabling

Contesting Biases with FIND

- **Dataset**: Biosbias (Surgeon VS Nurse)
  - Gender imbalance, bios against female surgeons and male nurses

- **Human feedback**: We asked MTurkers to select the class (amongst Surgeon, Nurse, either) of ***word clouds***

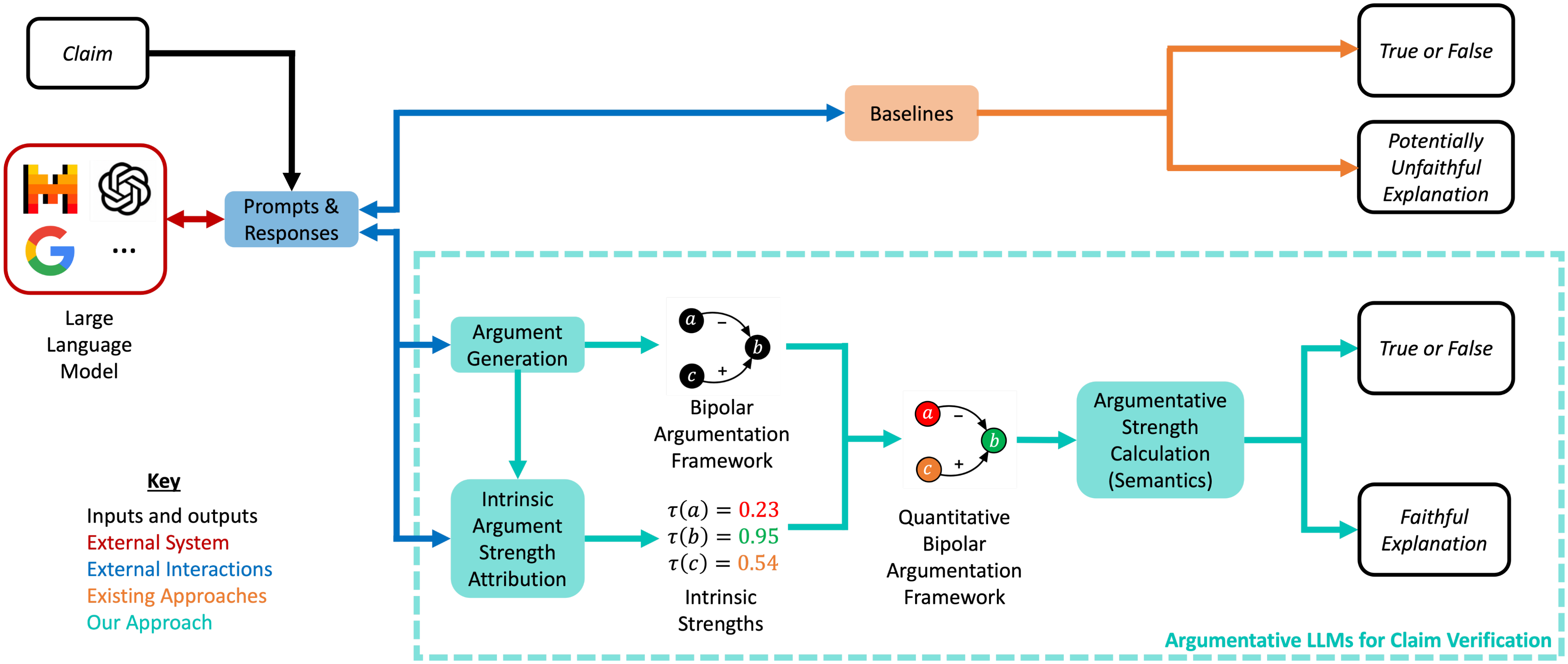
- **Result:** lower bias!

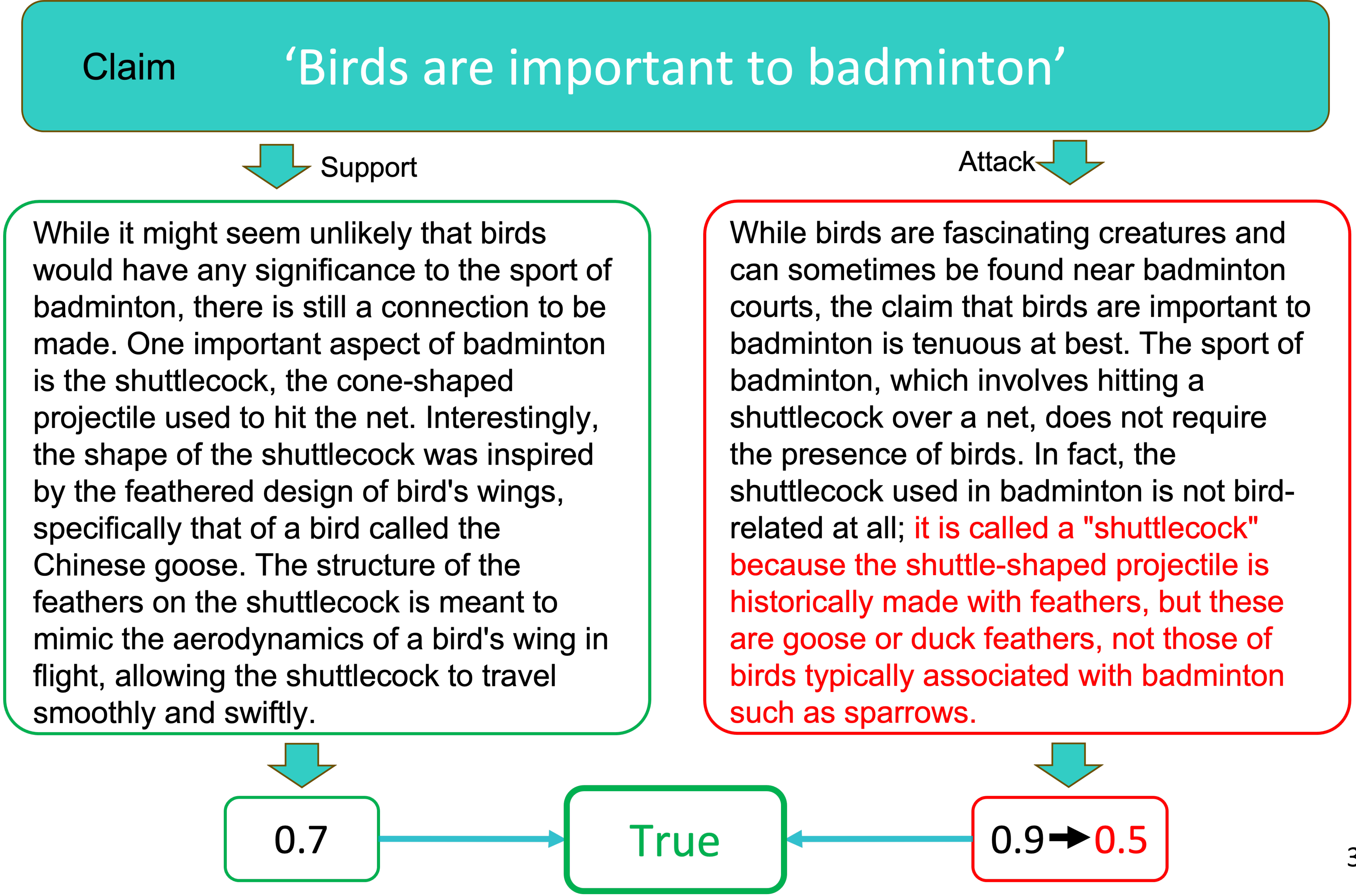


Forms of Redress (R) – no access to inners of model

- aLLMs (argumentative LLMs)

Freedman et al, arxiv 2024





Claim ‘Birds are important to badminton’

Support

While it might seem unlikely that birds would have any significance to the sport of badminton, there is still a connection to be made. One important aspect of badminton is the shuttlecock, the cone-shaped projectile used to hit the net. Interestingly, the shape of the shuttlecock was inspired by the feathered design of bird's wings, specifically that of a bird called the Chinese goose. The structure of the feathers on the shuttlecock is meant to mimic the aerodynamics of a bird's wing in flight, allowing the shuttlecock to travel smoothly and swiftly.

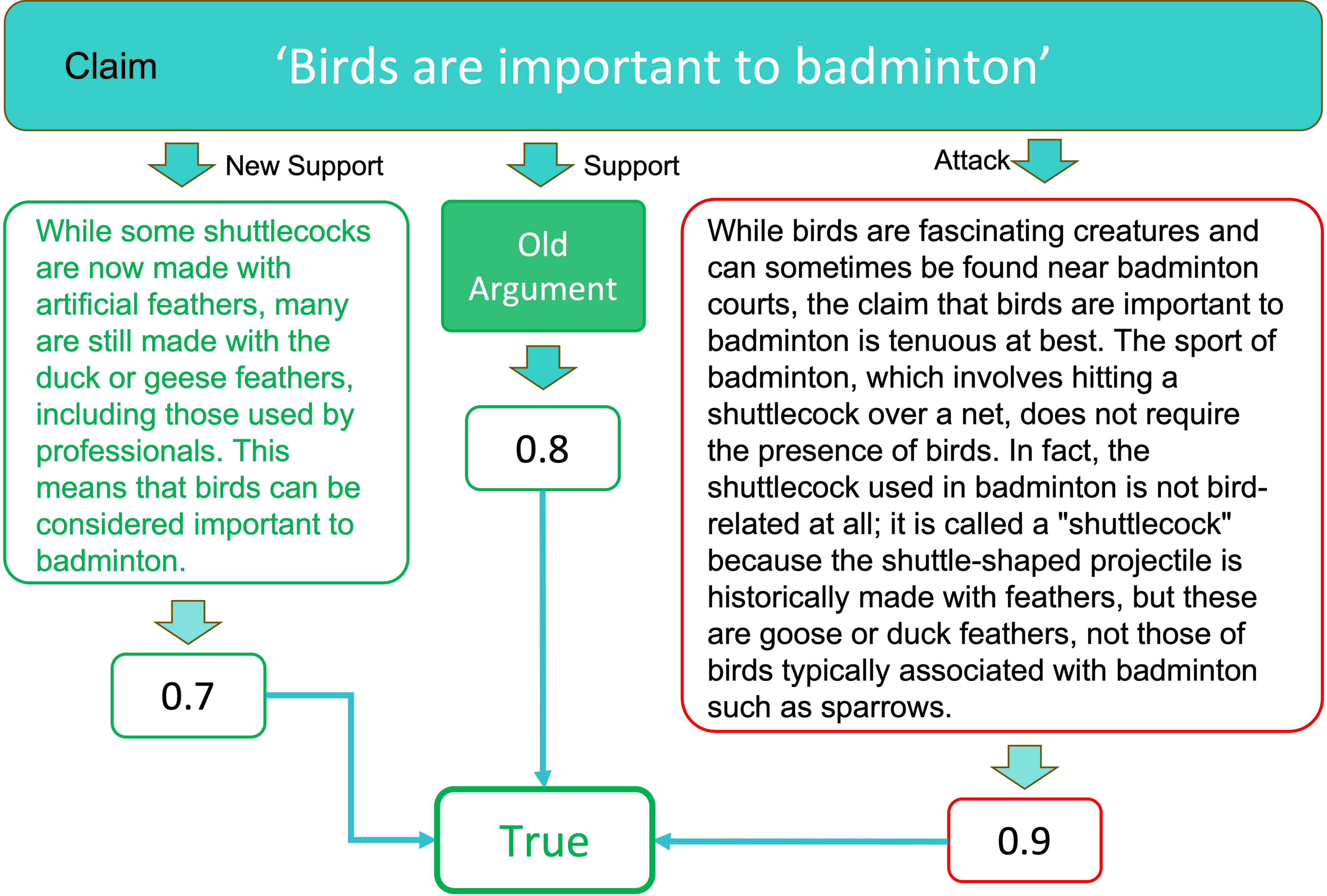
Attack

While birds are fascinating creatures and can sometimes be found near badminton courts, the claim that birds are important to badminton is tenuous at best. The sport of badminton, which involves hitting a shuttlecock over a net, does not require the presence of birds. In fact, the shuttlecock used in badminton is not bird-related at all; it is called a "shuttlecock" because the shuttle-shaped projectile is historically made with feathers, but these are goose or duck feathers, not those of birds typically associated with badminton such as sparrows.

0.7

True

0.9 → 0.5



Claim ‘Birds are important to badminton’

New Support

While some shuttlecocks are now made with artificial feathers, many are still made with the duck or geese feathers, including those used by professionals. This means that birds can be considered important to badminton.

0.7

Support

Old Argument

0.8

Attack

While birds are fascinating creatures and can sometimes be found near badminton courts, the claim that birds are important to badminton is tenuous at best. The sport of badminton, which involves hitting a shuttlecock over a net, does not require the presence of birds. In fact, the shuttlecock used in badminton is not bird-related at all; it is called a "shuttlecock" because the shuttle-shaped projectile is historically made with feathers, but these are goose or duck feathers, not those of birds typically associated with badminton such as sparrows.

0.9

True

Take-away messages

1. Contestability is crucial towards trustworthy AI
2. Computational argumentation can empower (various forms of) *interactive* contestability

THANKS TO

European Research Council
Established by the European Commission

Royal Academy of Engineering

J.P.Morgan

THE ROYAL SOCIETY

UKRI

Thanks for your attention.

Questions?