Computational Argumentation for Fair and Explainable AI Decision-making

Elfia Bezou-Vrakatseli, Madeleine Waller, Andreas Xydis

June 23th

Conference on Fairness, Accountability, and Transparency (FAccT 2025)

Athens, Greece

About us

SAFE & TRUSTED AI

UKRI CENTRE FOR DOCTORAL TRAINING

KING'S College LONDON

The Alan Turing Institute

UNIVERSITY OF LINCOLN

Online Handbook of Argumentation for AI

Thanks to Daphne Odekerken for contributions to the slides!

What is argumentation?

https://www.menti.com/al1t4176xes4

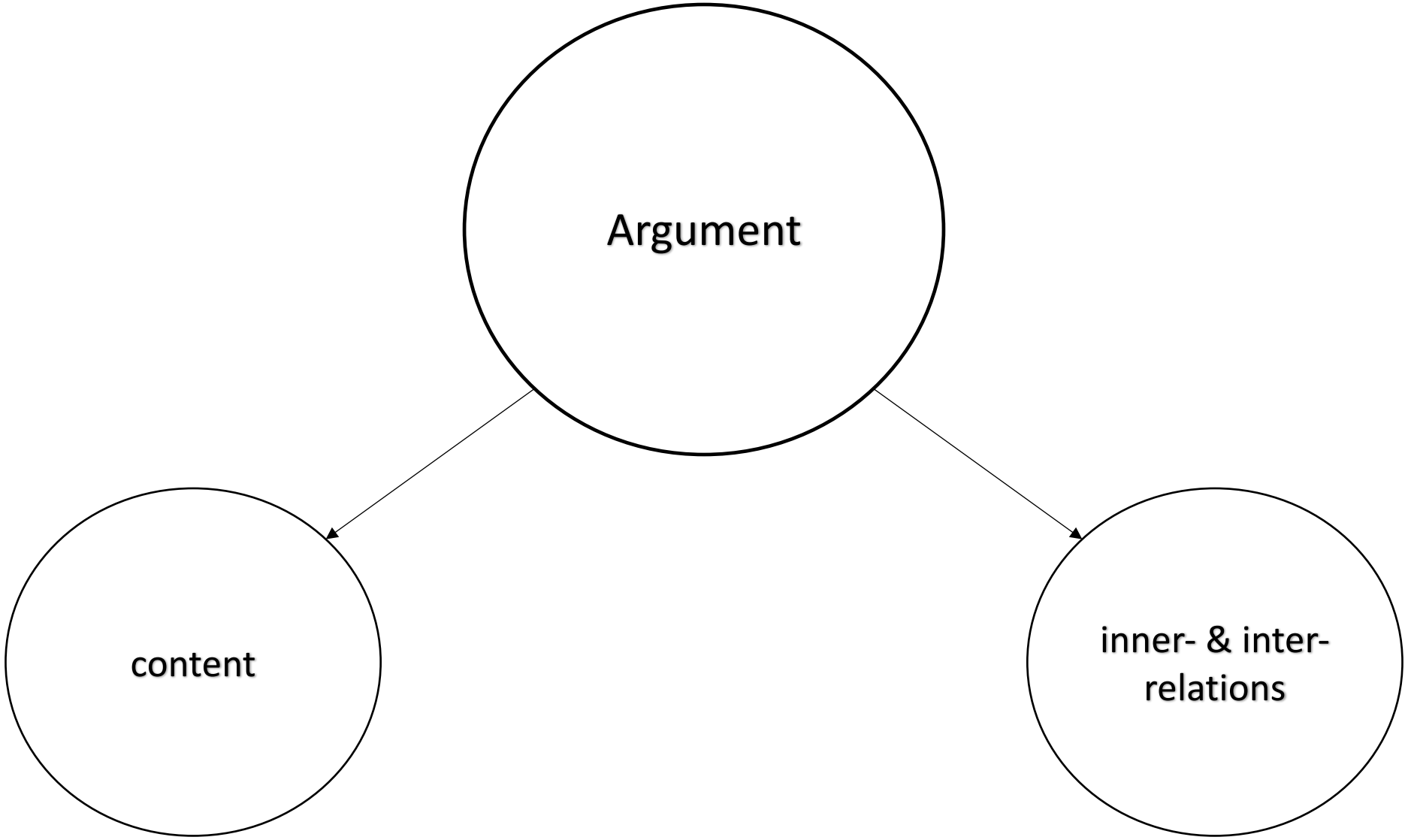
Mentimeter

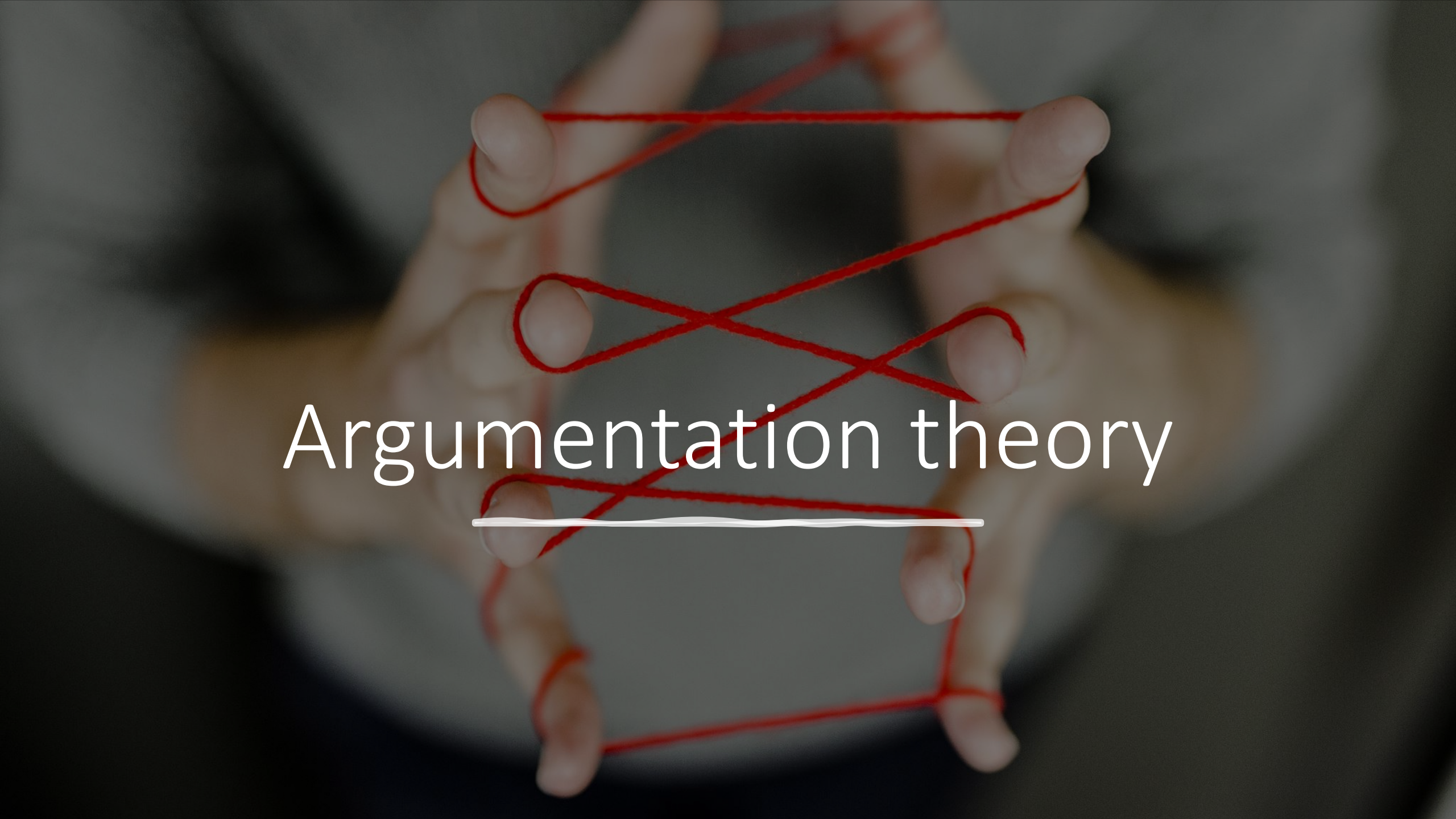
Which words come to mind when you think of argumentation?

3 27

Spin the wheel for a topic

- 2min to prepare
  - Arguments for
  - Arguments against
- 2min to debate

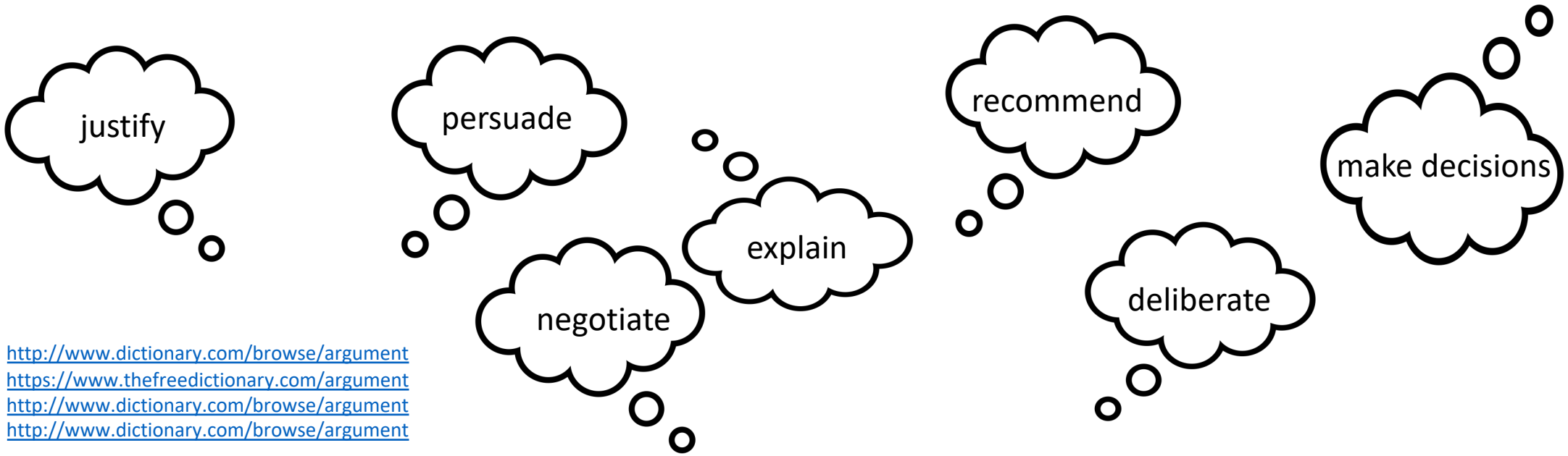

Argument
content
inner- & inter-
relations



Argumentation theory

What is an argument? Why do we argue?

- "A statement, reason, or fact for or against a point" [1]
- "A course of reasoning aimed at demonstrating truth or falsehood" [2]
- "A discussion involving differing points of view" [3]
- "An address or composition intended to convince or persuade" [4]



1. http://www.dictionary.com/browse/argument
2. https://www.thefreedictionary.com/argument
3. http://www.dictionary.com/browse/argument
4. http://www.dictionary.com/browse/argument

Internal Reasoning

- Information processing
- Reasoning about beliefs, goals, intentions

I will go left because it is the fastest route.

p: Left is the fastest route.
c: I will go left.
r: I want to take the fastest route.

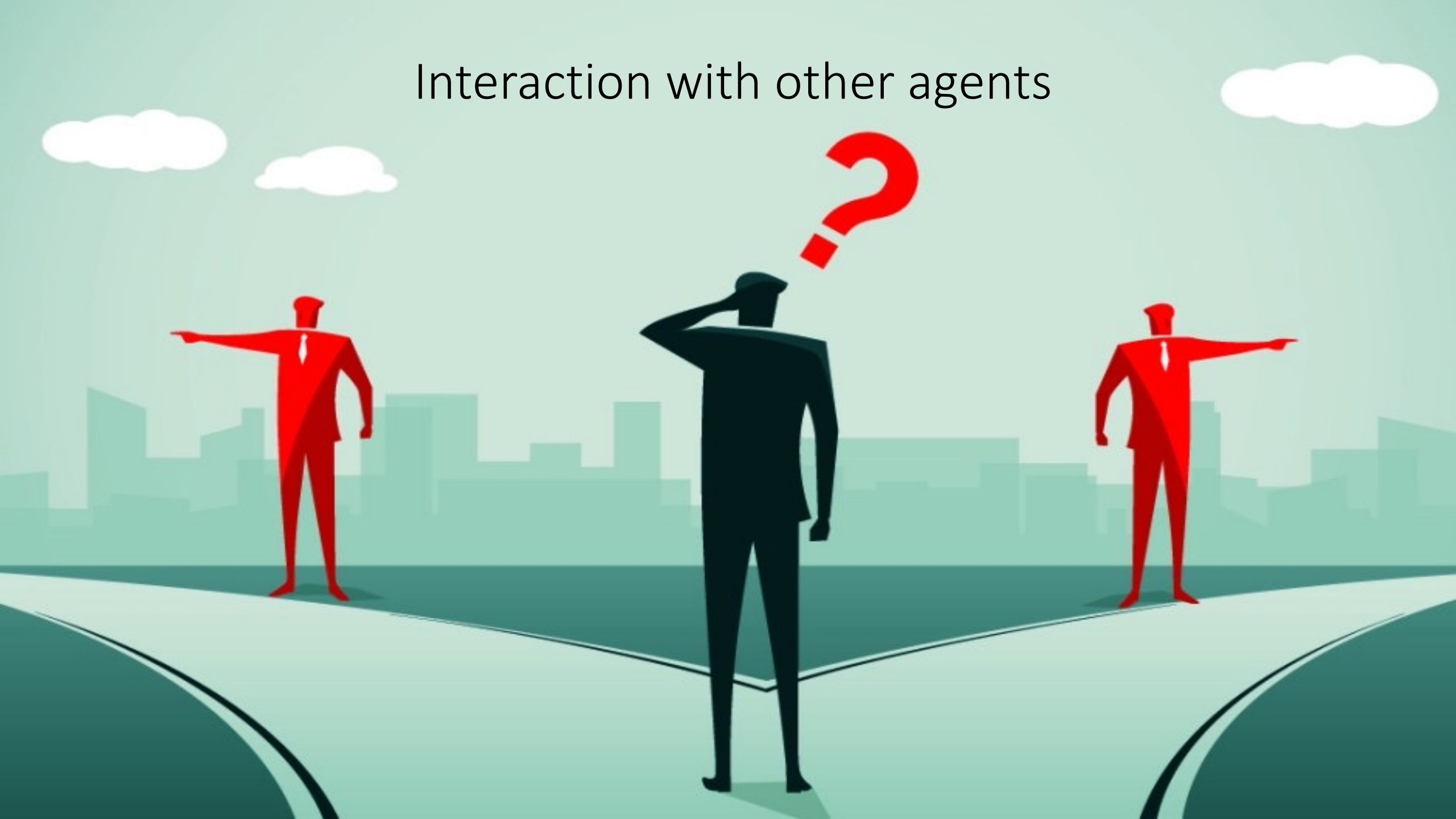
Commonsense reasoning: defeasible

- Inconsistent information
- Knowledge often uncertain or incomplete:
  - conclusions under certain assumptions
  - retract conclusions once learn an assumption is unwarranted

→Non monotonic logic

Actually, today, I will go right because there is an obstacle on the left.

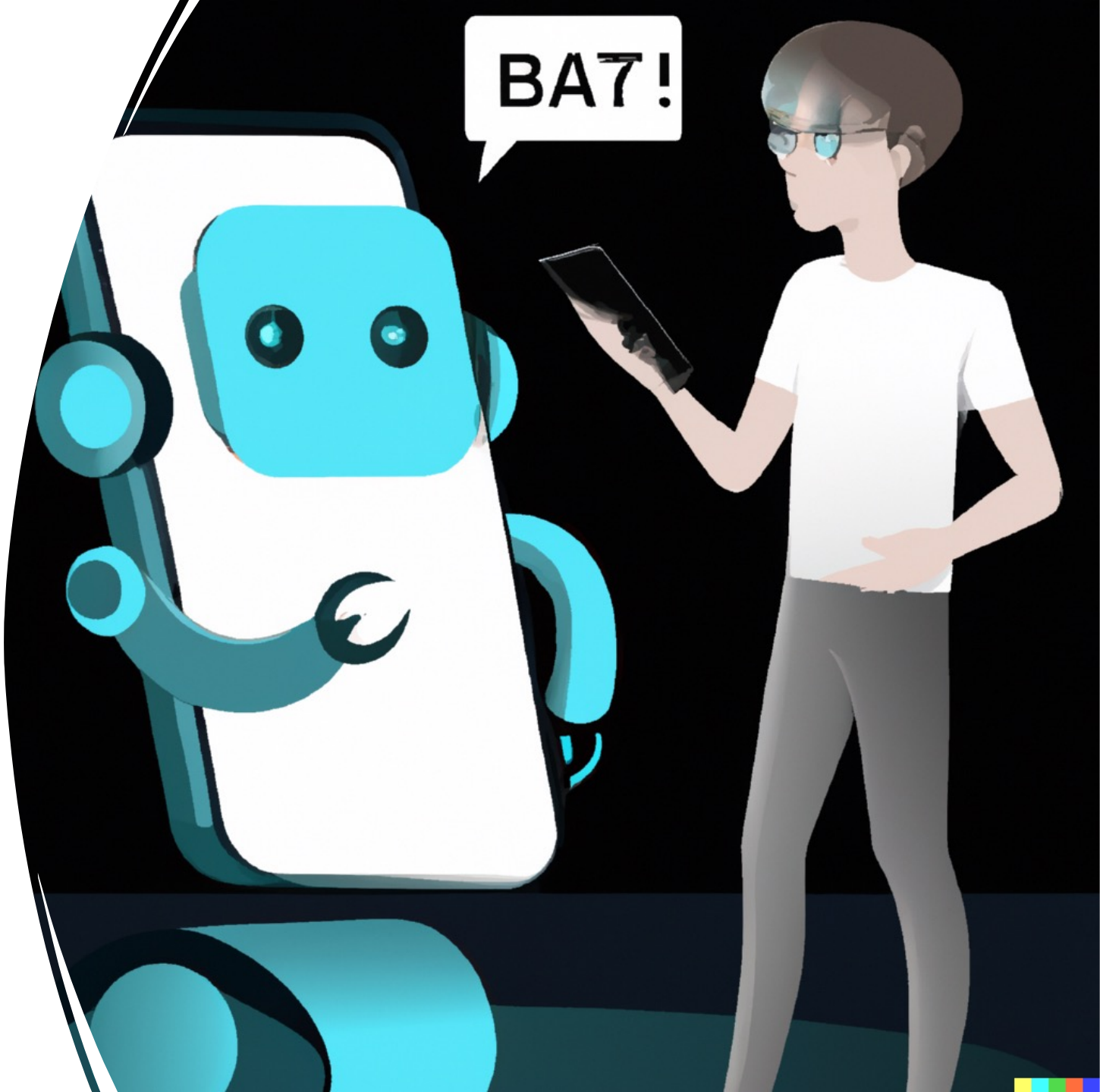
Interaction with other agents

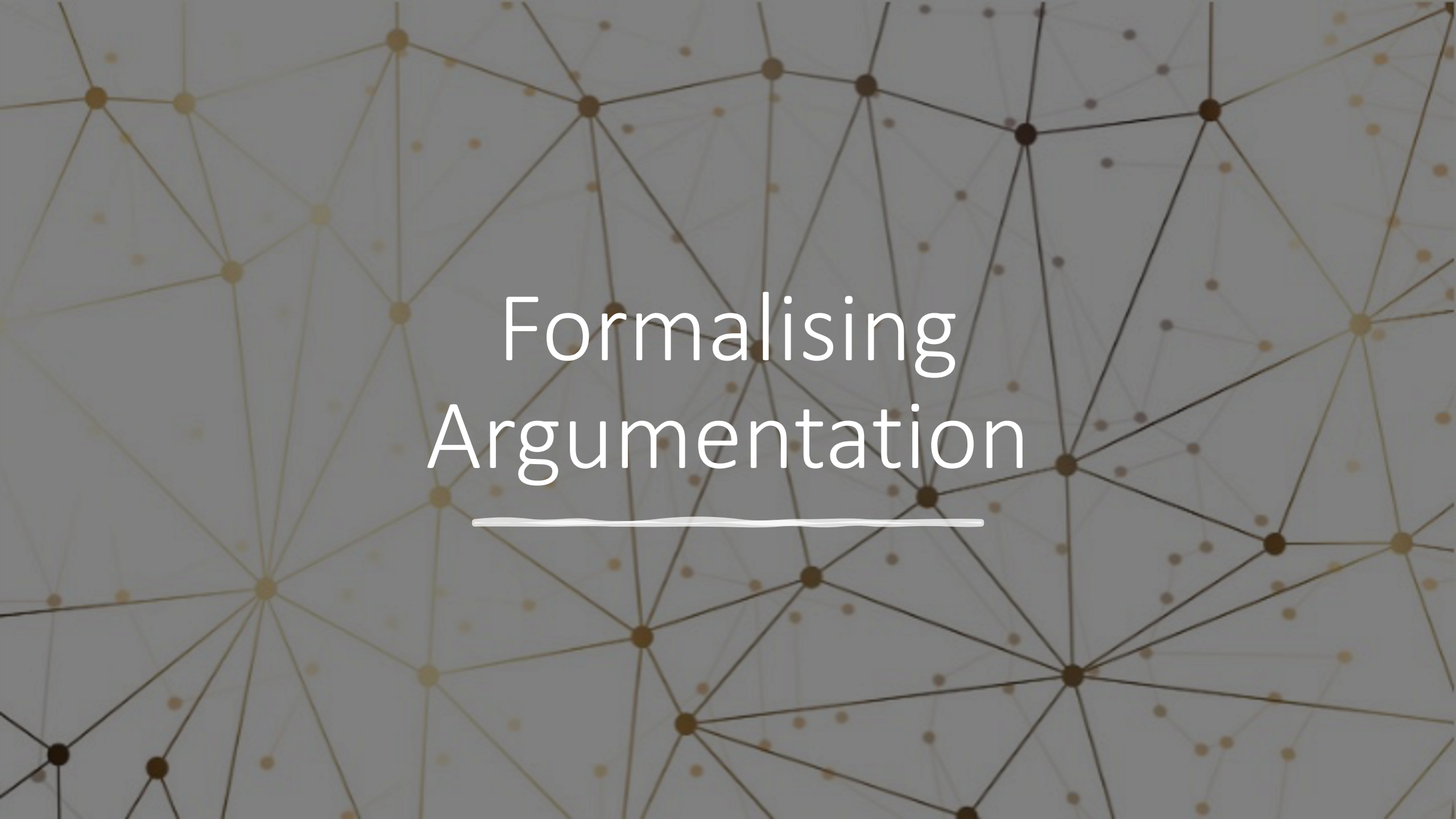


Dialogue

- Tool of interaction & communication
- Enables understanding of both parties involved
  - Information
  - Reasoning exploration

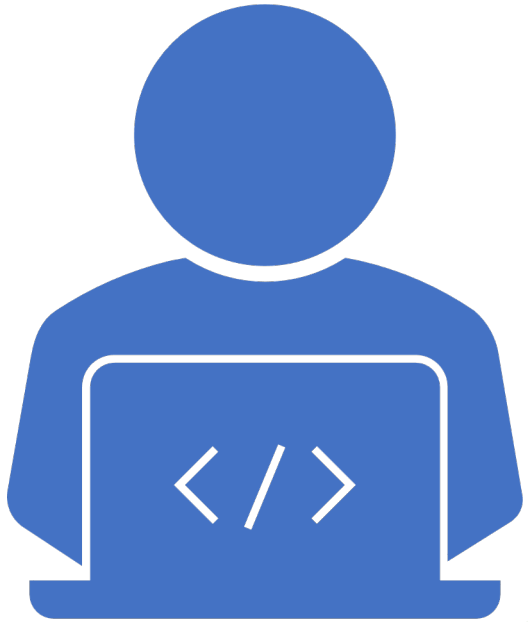
☞ Joint reasoning




Formalising
Argumentation

What is computational argumentation?

- Formalisation of argumentation theory
- Used to support human-computer interactions and computer-computer interactions
- Applications include:
  - providing reasoning and explaining decision-making
  - natural language processing and generation tasks



Abstract Argumentation

Disregards the internal structure of arguments and focusses on acceptability conditions that allow certain sets of arguments to co-exist in a rational manner based on a **given attack relationship between arguments**.

(P. M. Dung, 1995)

Should I go right or left?

Argument 1 (A1)

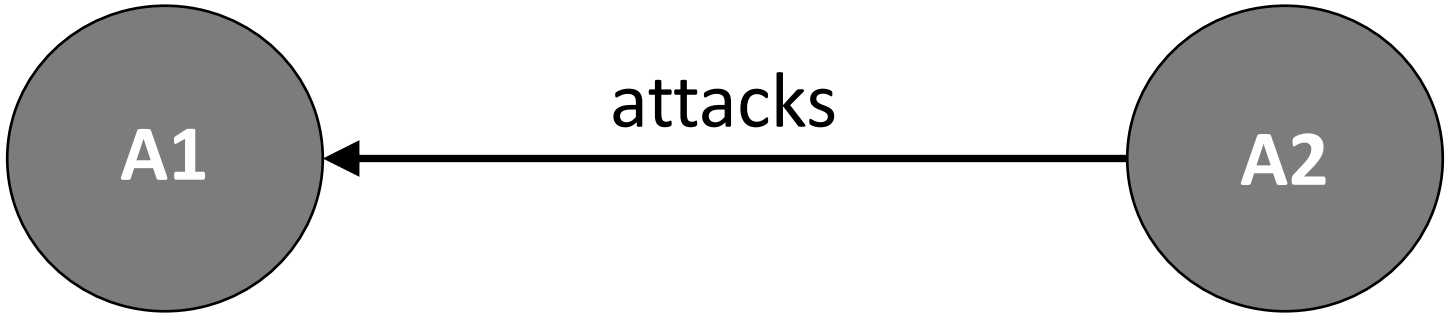
Going left is the fastest route,

therefore I should go left

Argument 2 (A2)

Today there is an obstacle to the left,

therefore I should go right



[1] Phan Minh Dung (1995). "On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming, and n–person games". *Artificial Intelligence*. **77** (2): 321–357.

Has social media been good for humanity?

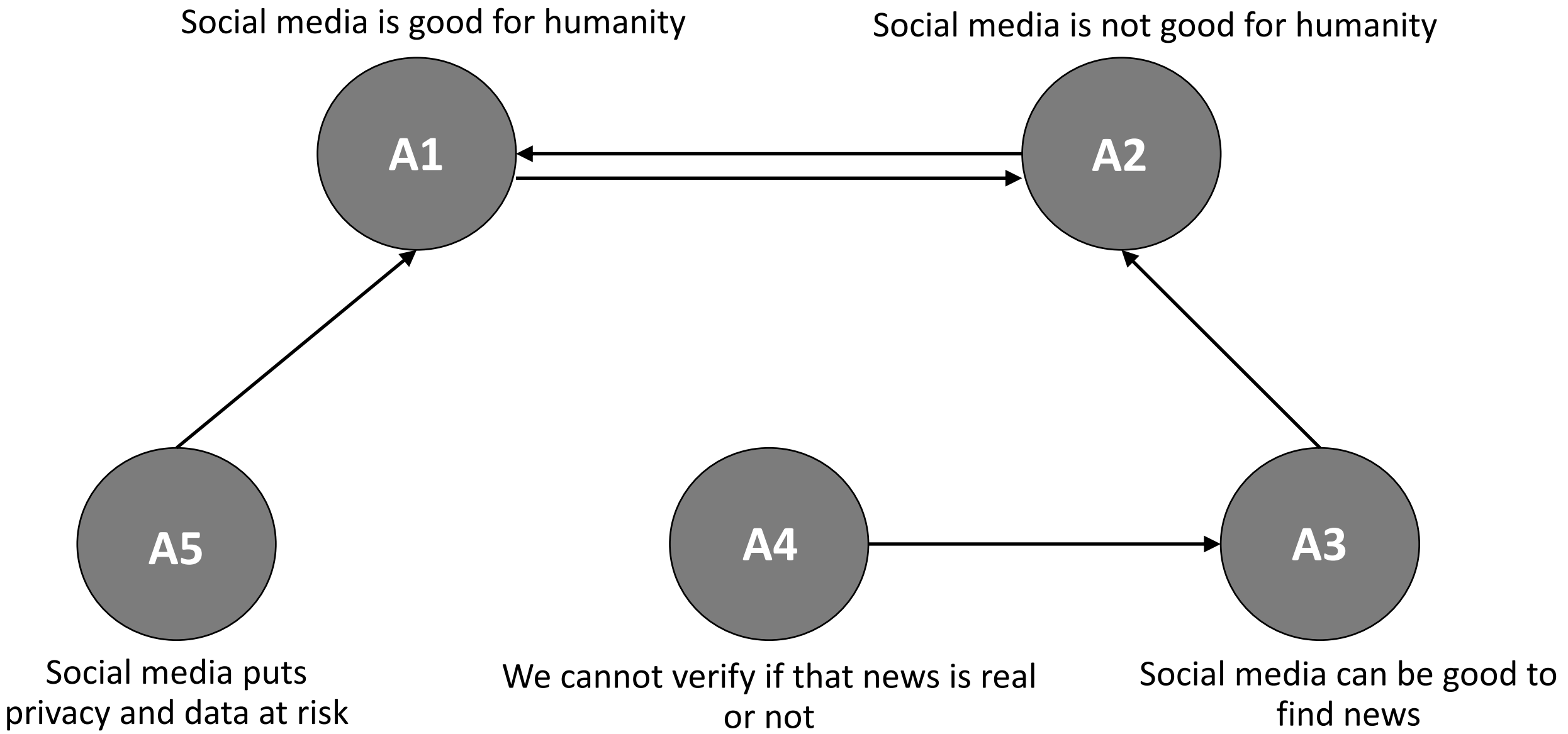
A1: Social media has been good for humanity

A2: Social media has not been good for humanity

A3: Social media can be good to find news

A4: We cannot verify if that news is real or not

A5: Social media puts privacy and data at risk


Social media is good for humanity
Social media is not good for humanity
A1
A2
A5
A4
A3
Social media puts privacy and data at risk
We cannot verify if that news is real or not
Social media can be good to find news

Label-based semantics

IN if all its attackers are out (or no attackers)

OUT if it has an attacker that is in

UNDEC if not all its attackers are out
and it does not have an attacker that is in

IN if all its attackers are out

OUT if it has an attacker that is in

UNDEC if not all its attackers are out
and it does not have an attacker that is in

Argument 1 (A1)

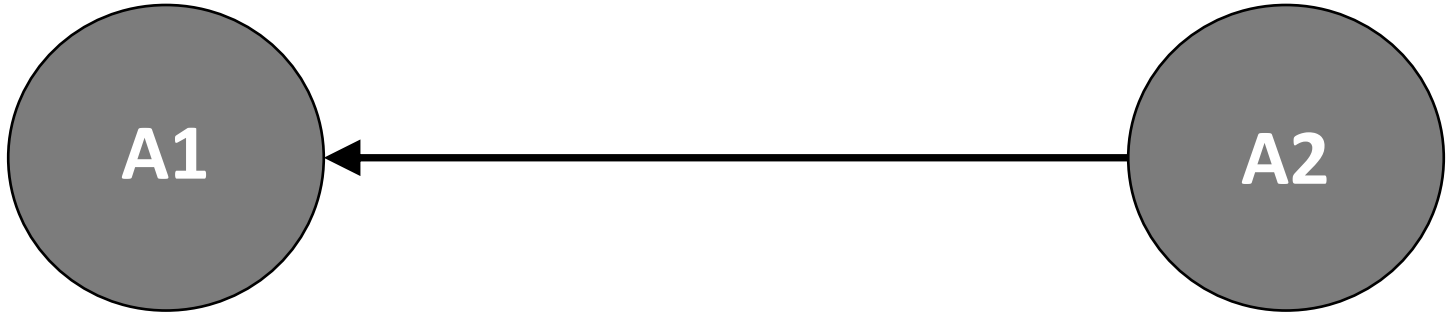
Going left is the fastest route,

therefore I should go left

Argument 2 (A2)

Today there is an obstacle to the left,

therefore I should go right



IN if all its attackers are out

OUT if it has an attacker that is in

UNDEC if not all its attackers are out
and it does not have an attacker that is in

Argument 1 (A1)

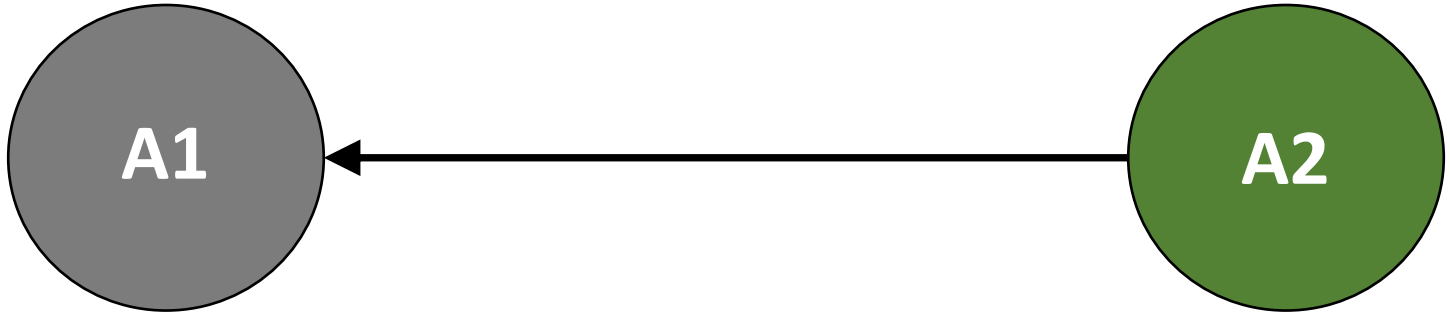
Going left is the fastest route,

therefore I should go left

Argument 2 (A2)

Today there is an obstacle to the left,

therefore I should go right



IN if all its attackers are out

OUT if it has an attacker that is in

UNDEC if not all its attackers are out
and it does not have an attacker that is in

Argument 1 (A1)

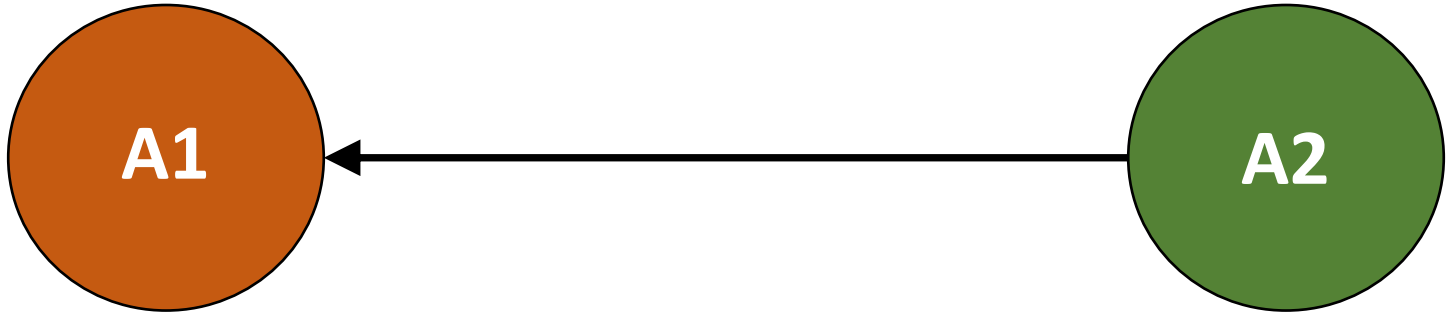
Going left is the fastest route,

therefore I should go left

Argument 2 (A2)

Today there is an obstacle to the left,

therefore I should go right



Has social media been good for humanity?

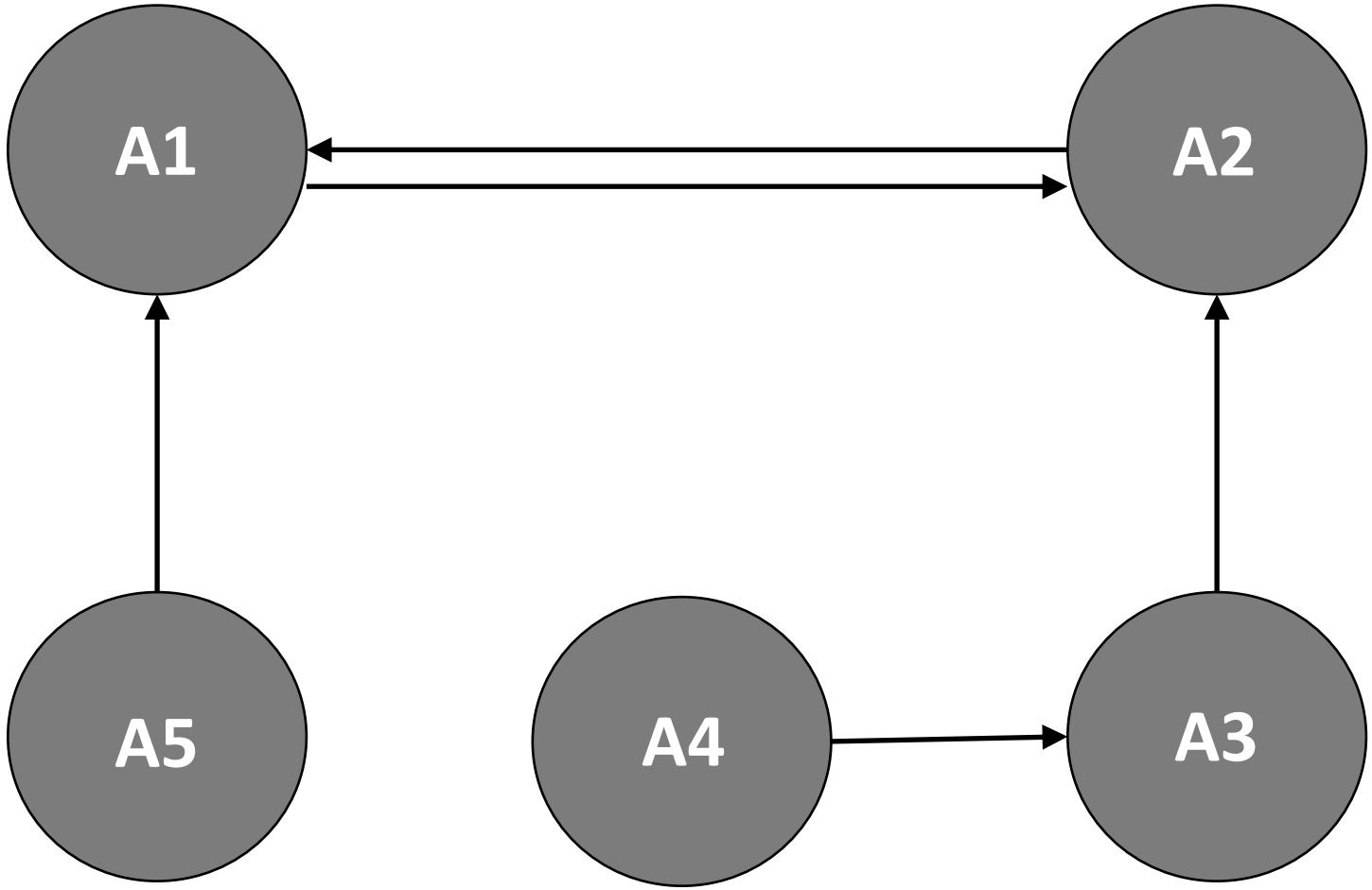
A1: Social media has been good for humanity

A2: Social media has not been good for humanity

A3: Social media can be good to find news

A4: We cannot verify if that news is real or not

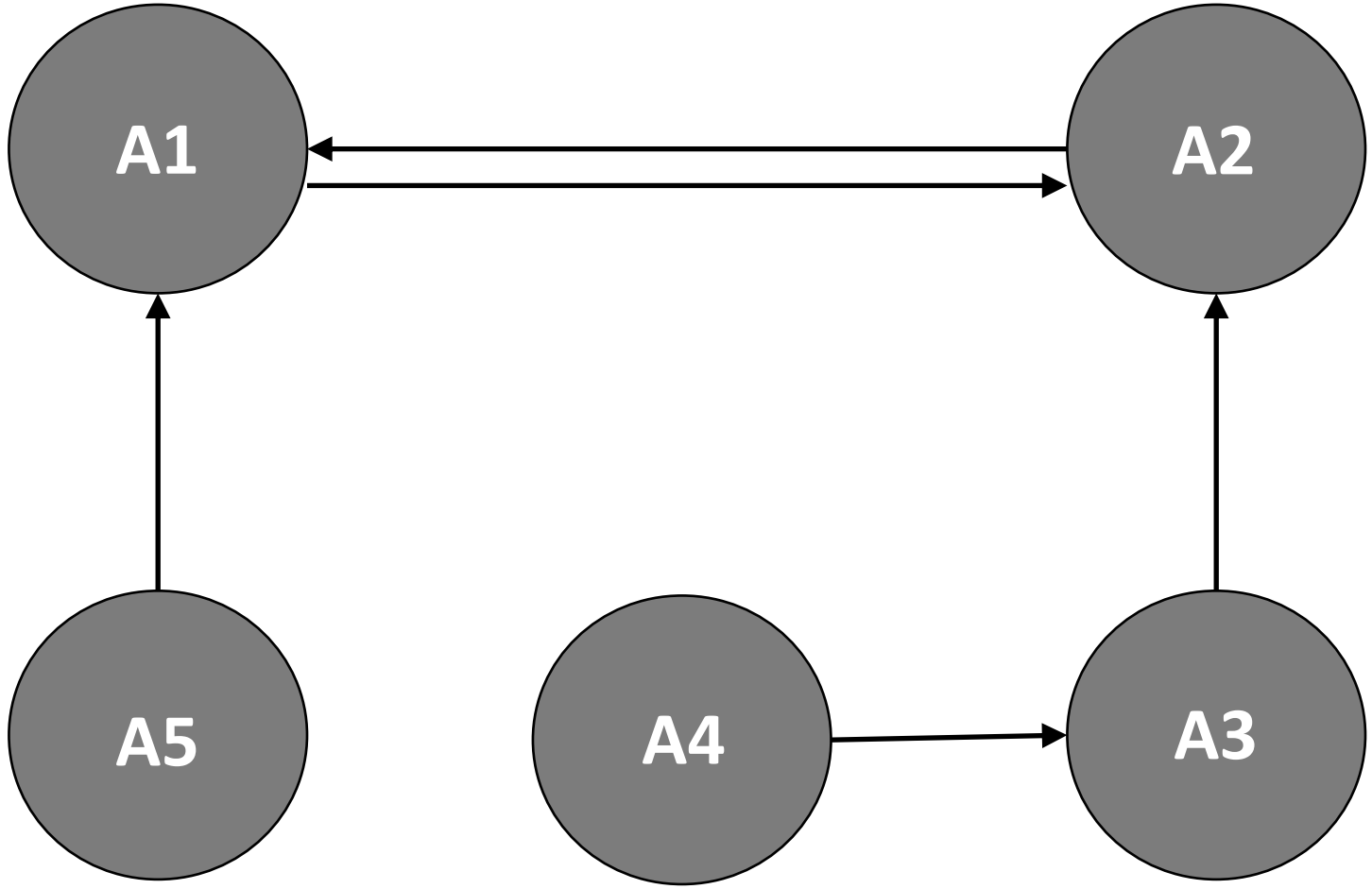
A5: Social media puts privacy and data at risk


A1
A2
A5
A4
A3

IN if all its attackers are out

OUT if it has an attacker that is in

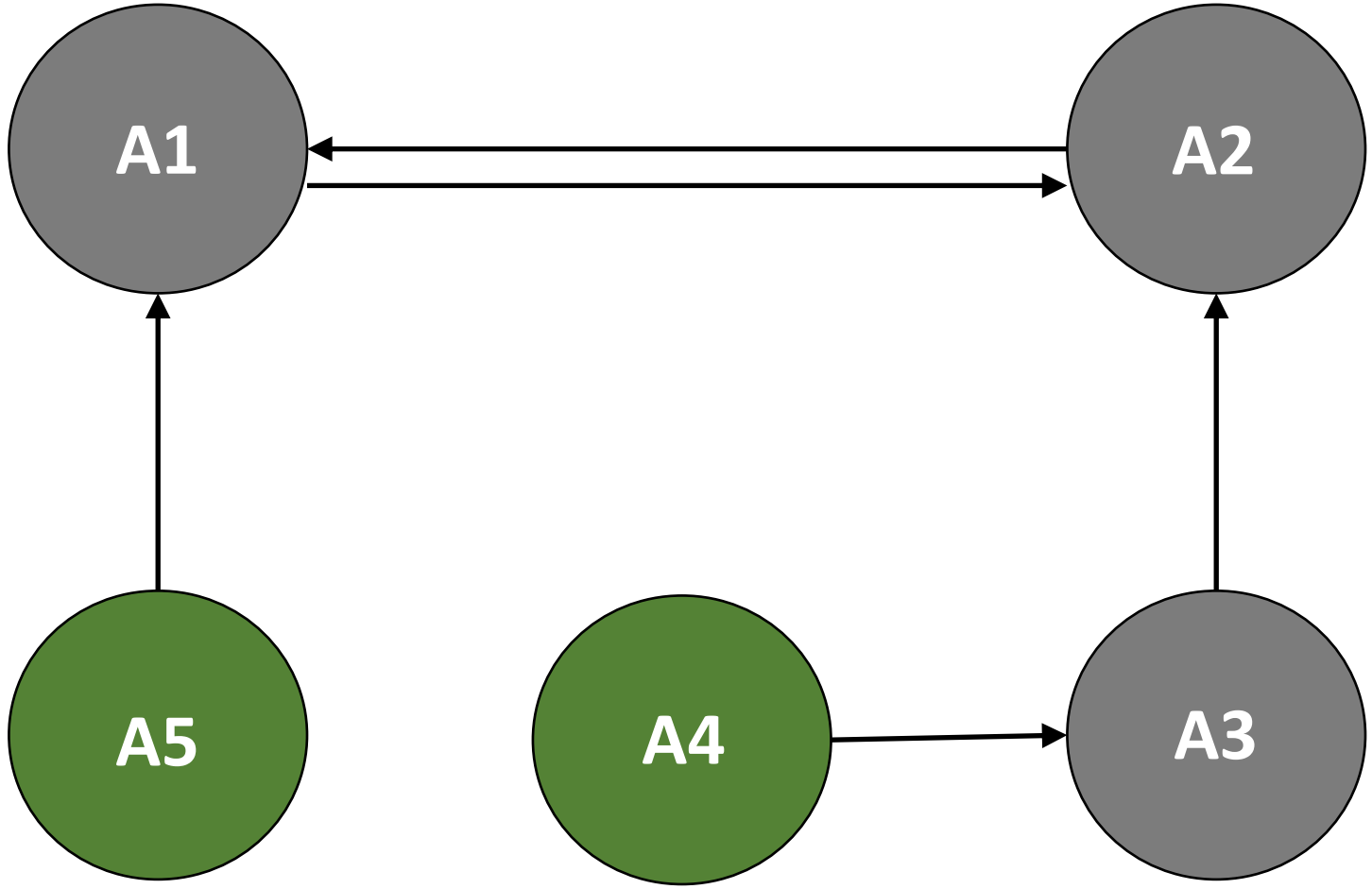
UNDEC if not all its attackers are out
and it does not have an attacker that is in



IN if all its attackers are out

OUT if it has an attacker that is in

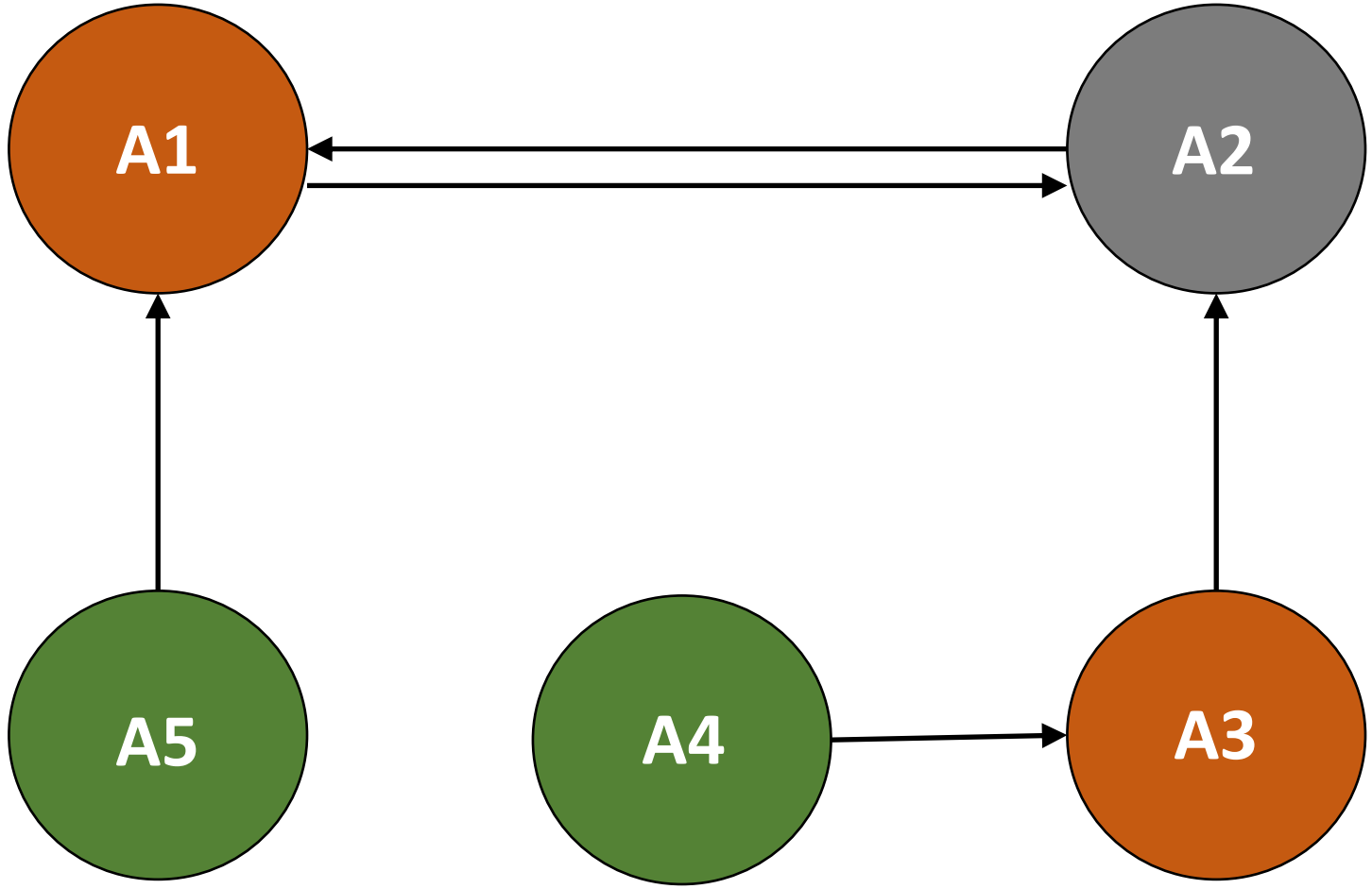
UNDEC if not all its attackers are out
and it does not have an attacker that is in



IN if all its attackers are out

OUT if it has an attacker that is in

UNDEC if not all its attackers are out
and it does not have an attacker that is in

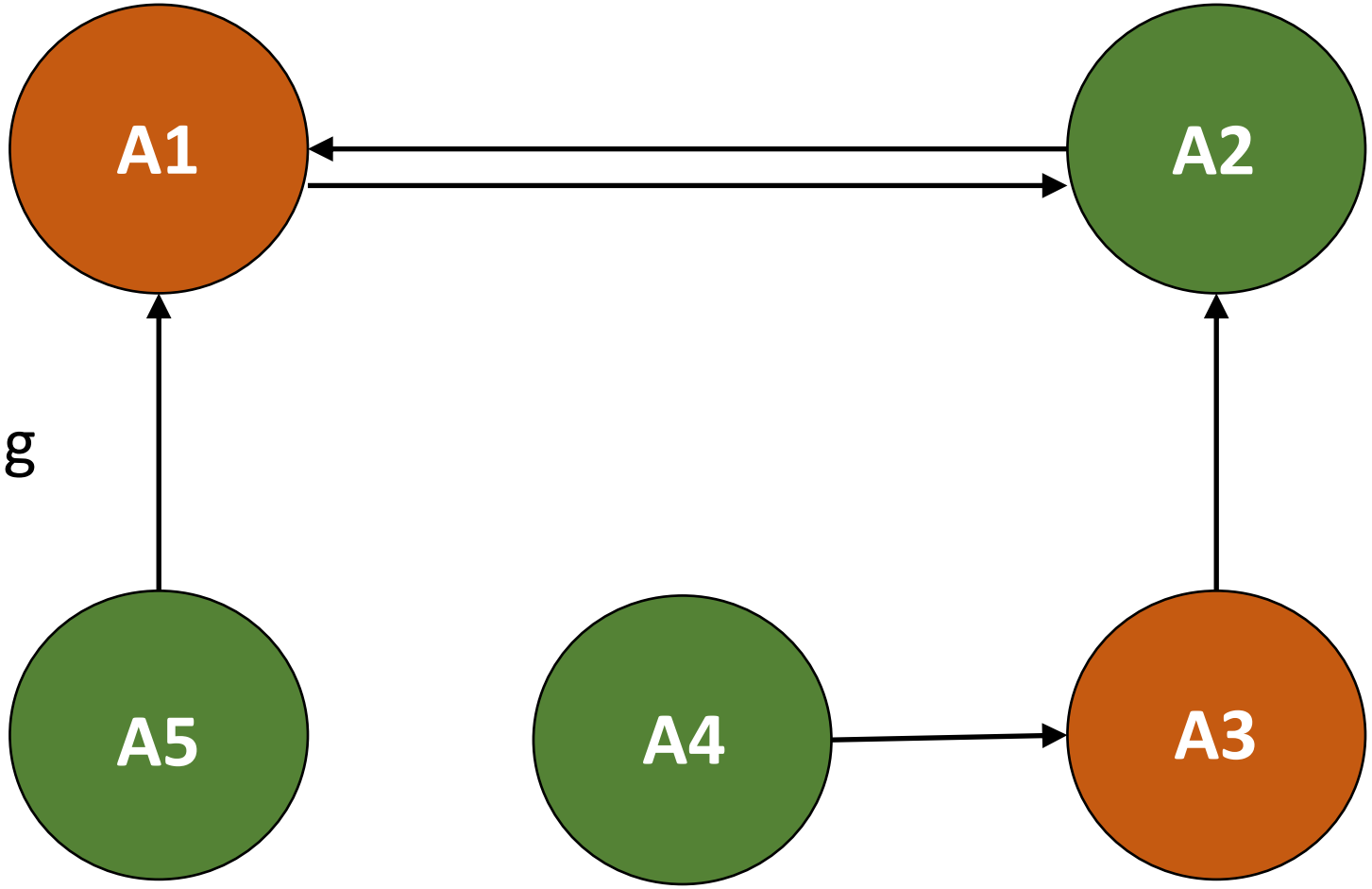


IN if all its attackers are out

OUT if it has an attacker that is in

UNDEC if not all its attackers are out
and it does not have an attacker that is in

We call this a complete labelling



Has social media been good for humanity?

A1: Social media has been good for humanity

A2: Social media has not been good for humanity

A3: Social media can be good to find news

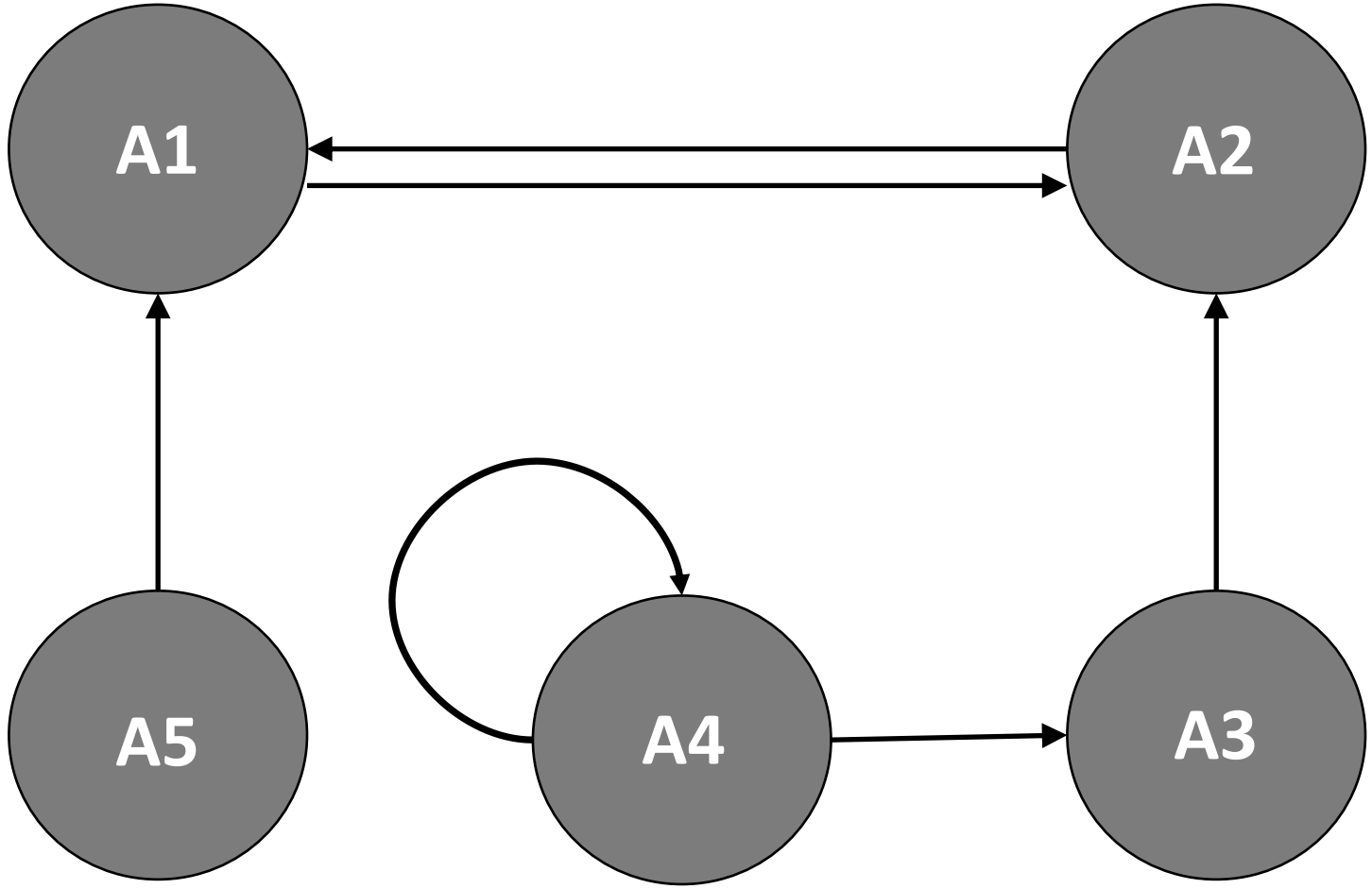
A4: I read on social media that we cannot verify whether news on social media is real or fake

A5: Social media puts privacy and data at risk

IN if all its attackers are out

OUT if it has an attacker that is in

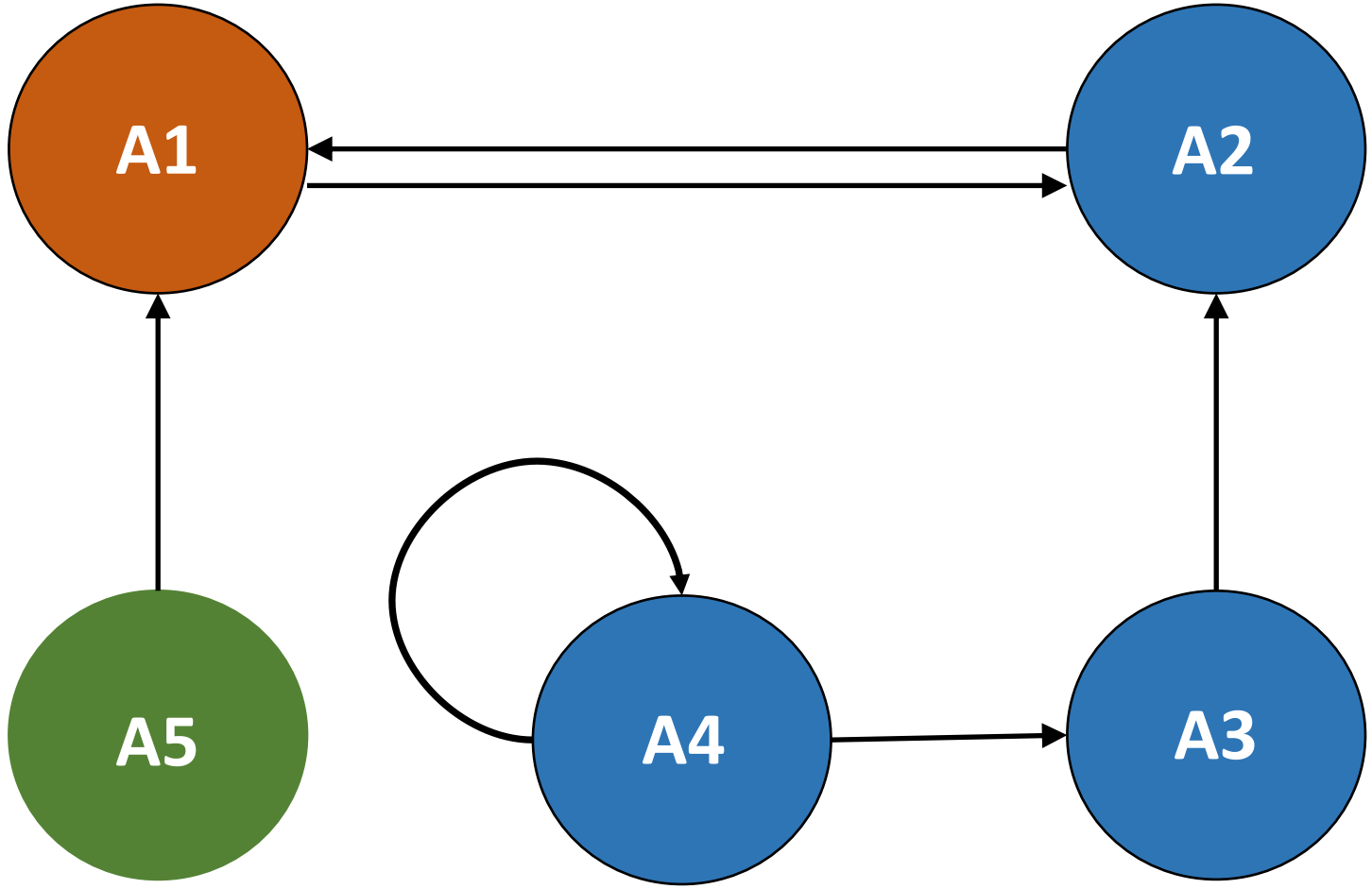
UNDEC if not all its attackers are out
and it does not have an attacker that is in



IN if all its attackers are out

OUT if it has an attacker that is in

UNDEC if not all its attackers are out
and it does not have an attacker that is in

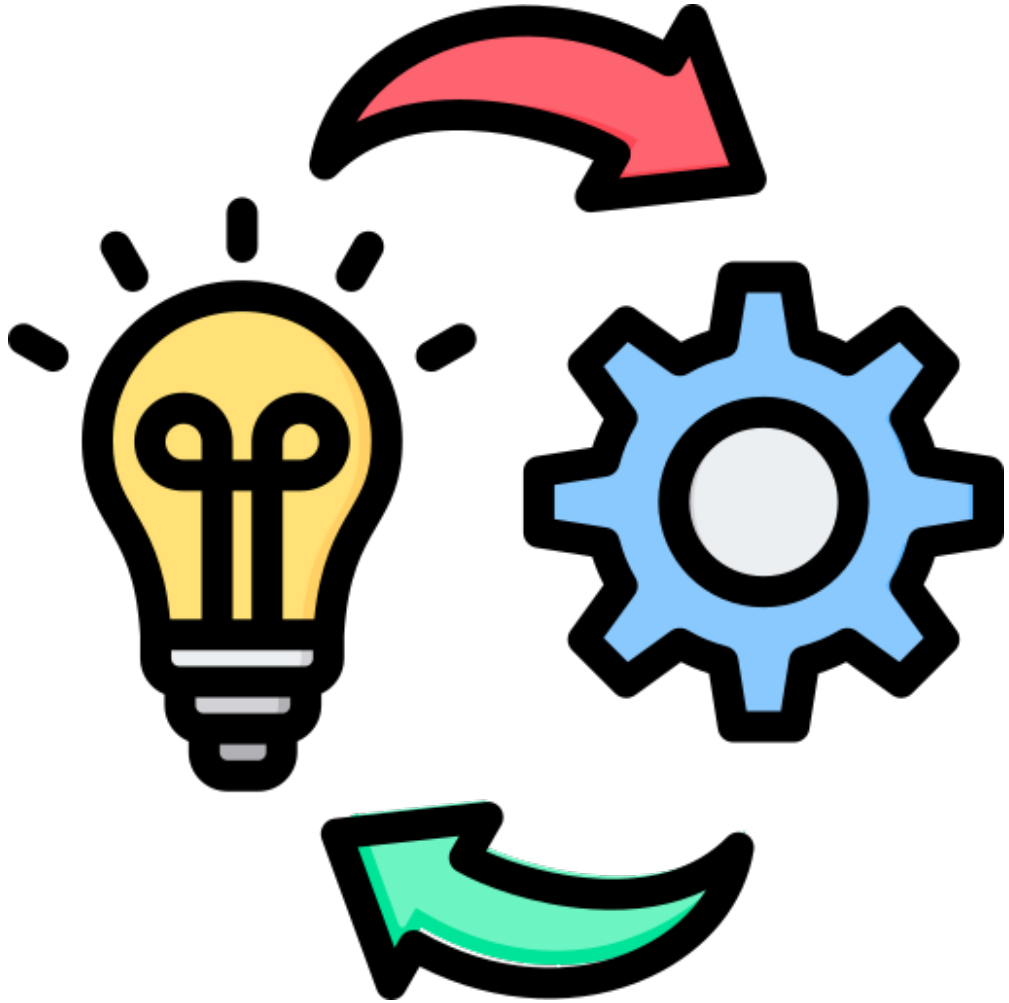


Other Labellings

- Grounded labelling – minimise the arguments that are IN
- Preferred labelling – maximise the arguments that are IN
- Stable labelling – no UNDEC arguments
- Semi-stable labelling – minimise the arguments that are UNDEC

Implementation

http://argteach.herokuapp.com



Bipolar argumentation frameworks (BAFs)

- Adds support relations to abstract argumentation frameworks
- Semantics defined differently to account for this:
  - An argument is accepted only if it is directly defended or supported by arguments that are themselves already accepted in a grounded manner.

Weighted argumentation frameworks (WAFs)

Adds numerical values to the abstract argumentation graph

Intrinsic weights assigned to arguments/attacks/supports representing their initial strength

Higher weights indicate stronger arguments/attacks/supports and therefore have more influence on the final acceptability calculated

Semantics used to calculate final weights of arguments based on the weights of incoming arguments/attacks/supports

Safe & Trusted AI

- Humans & AI Systems
  - Interaction & Communication
  - Human-AI Dialogue
  - Joint Reasoning
- Argumentation
  - Real-world Reasoning
  - Justification for its claims
  - Explainability & Transparency in Decision Making

Argumentation for XAI

Solving conflicts in multi-agent systems

Supporting human-computer interaction through transparent reasoning

Providing clear and intuitive justifications for AI decisions

G. Vilone & L. Rizzo. XAI and Argumentation Special Track. The 3rd World Conference on eXplainable Artificial Intelligence.

Types of argumentative explanations

Intrinsic

Explaining recommender systems built on argumentation

Post-hoc

(complete or approximate)

Explaining Bayesian networks using argumentation abstractions

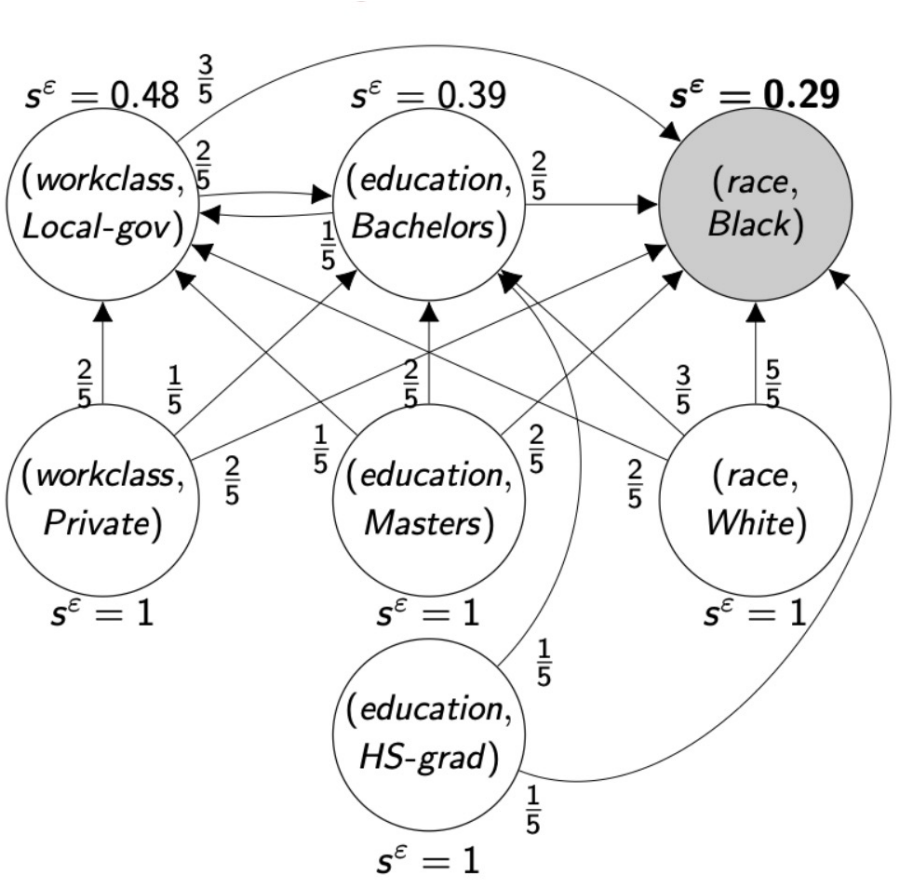
Approximating multi-layer perceptron with argumentation

K. Cyras et al. 2021. Argumentative XAI: A Survey. IJCAI 2021 Survey Track.

Bias detection

workclass	education	race	Classification
Local-gov	*Bachelors*	*Black*	−
Private	*Bachelors*	*White*	+
Local-gov	*HS-grad*	*White*	+
Local-gov	*Bachelors*	*White*	+
Private	*Masters*	*White*	+
Local-gov	*Masters*	*White*	+

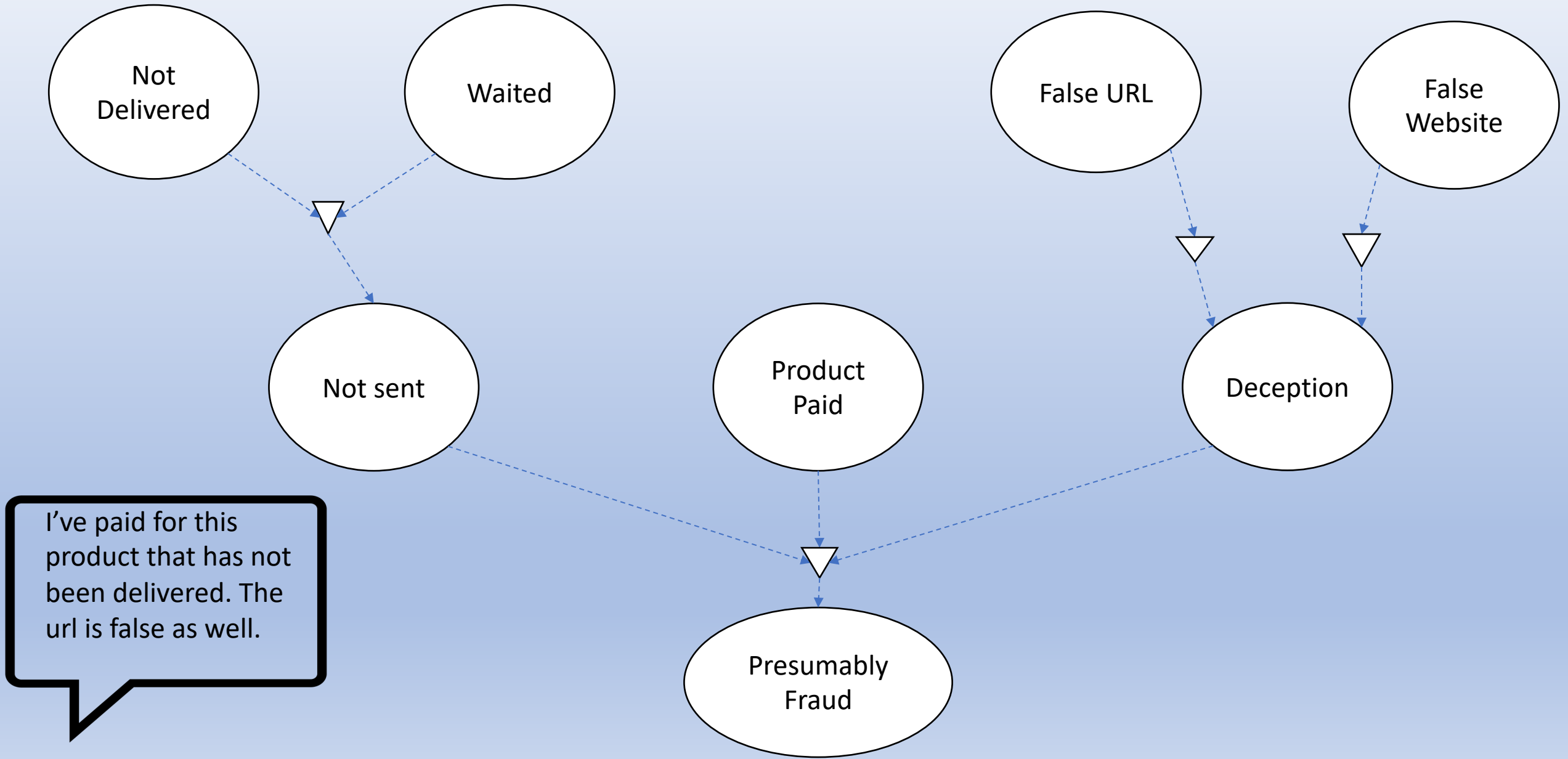
Smallest final weight(s) = attribute value(s) that contribute the most to the negative classification



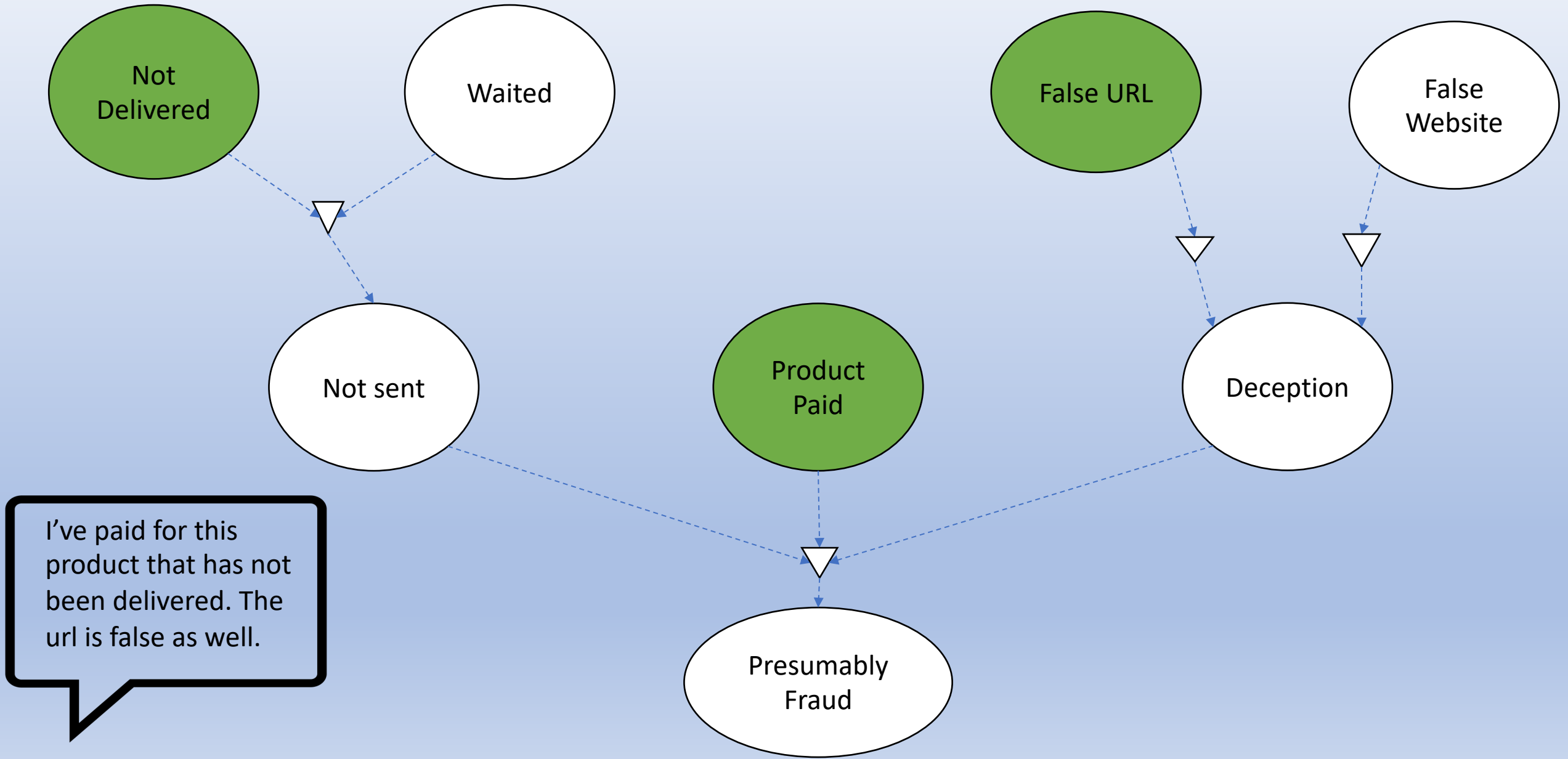
M. Waller et al. 2024. Identifying Reasons for Bias: An Argumentation-Based Approach. AAAI 2024.

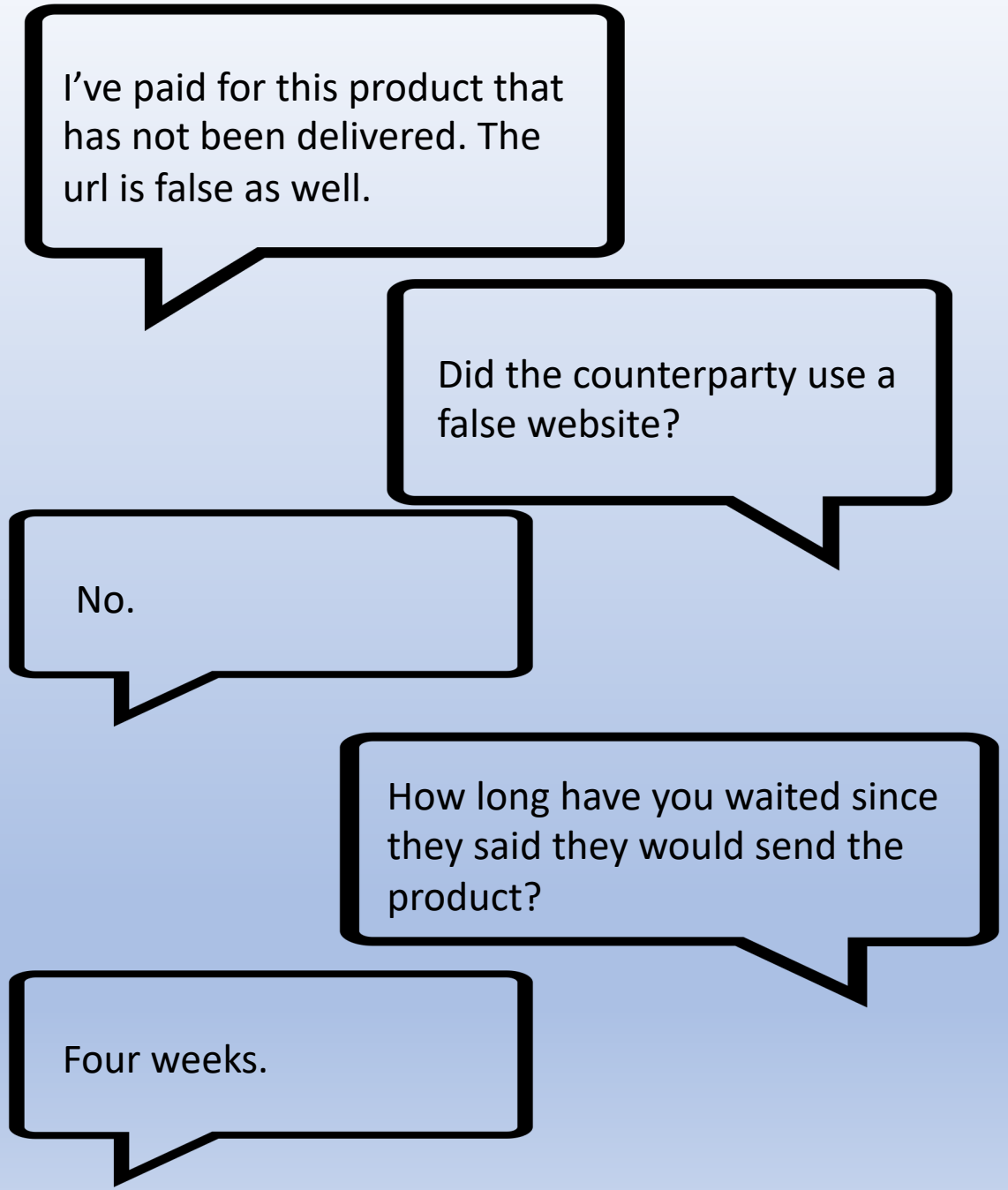

FROM THEORY TO PRACTICE:
ARGUMENTATION IN ACTION
POLICE
POLICE
POLICE

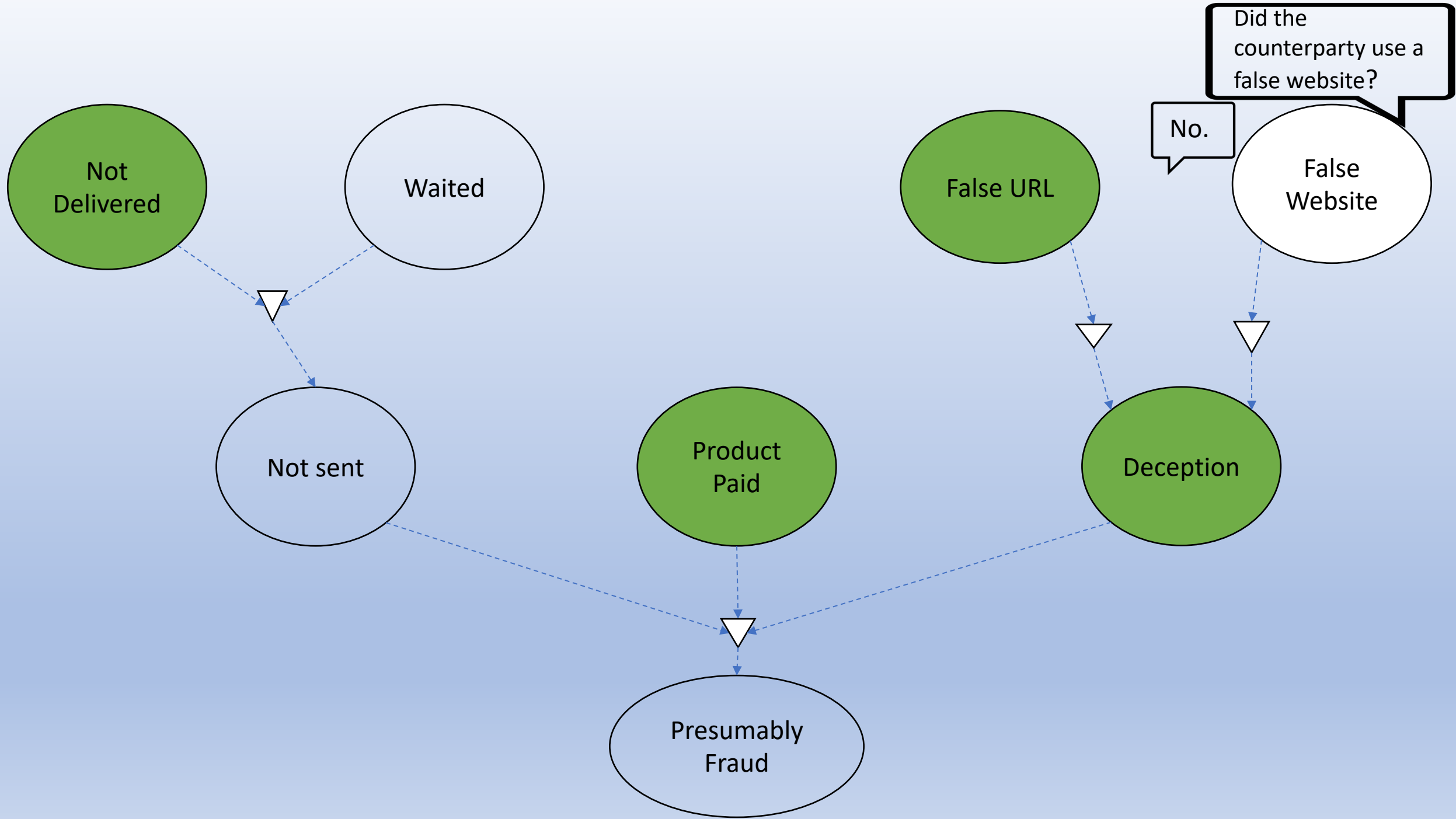
Application: Dutch Police

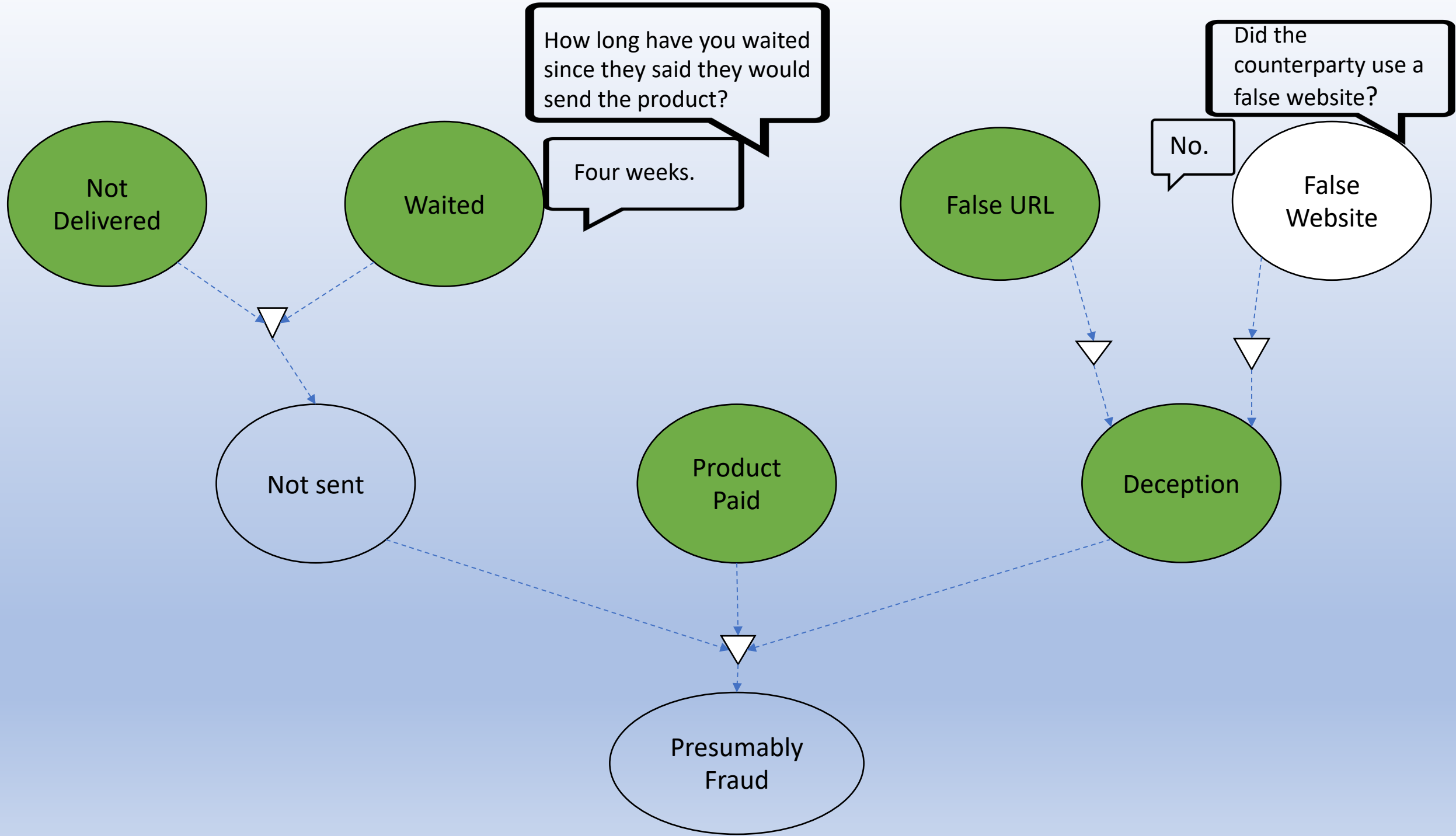


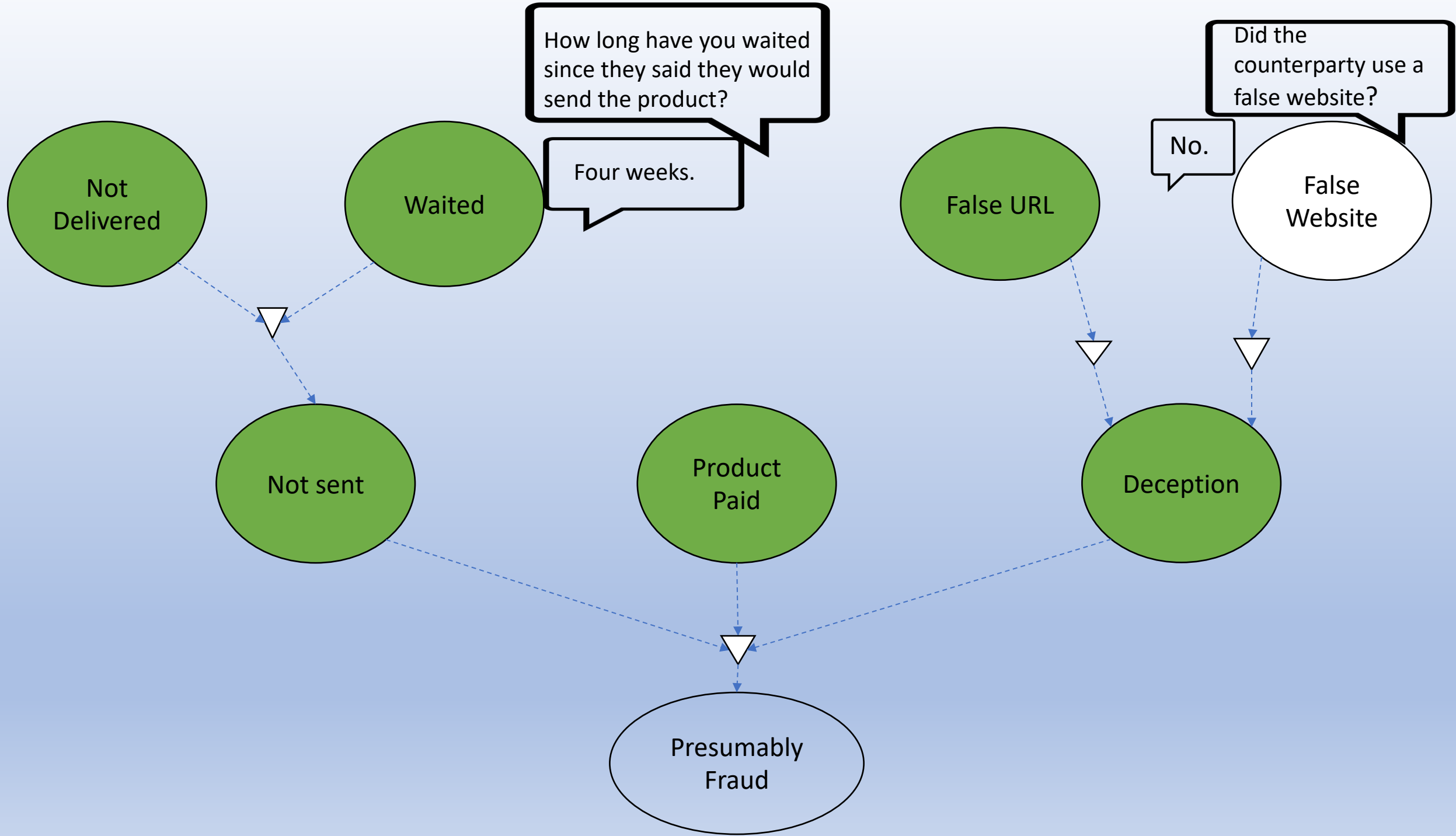
Application: Dutch Police

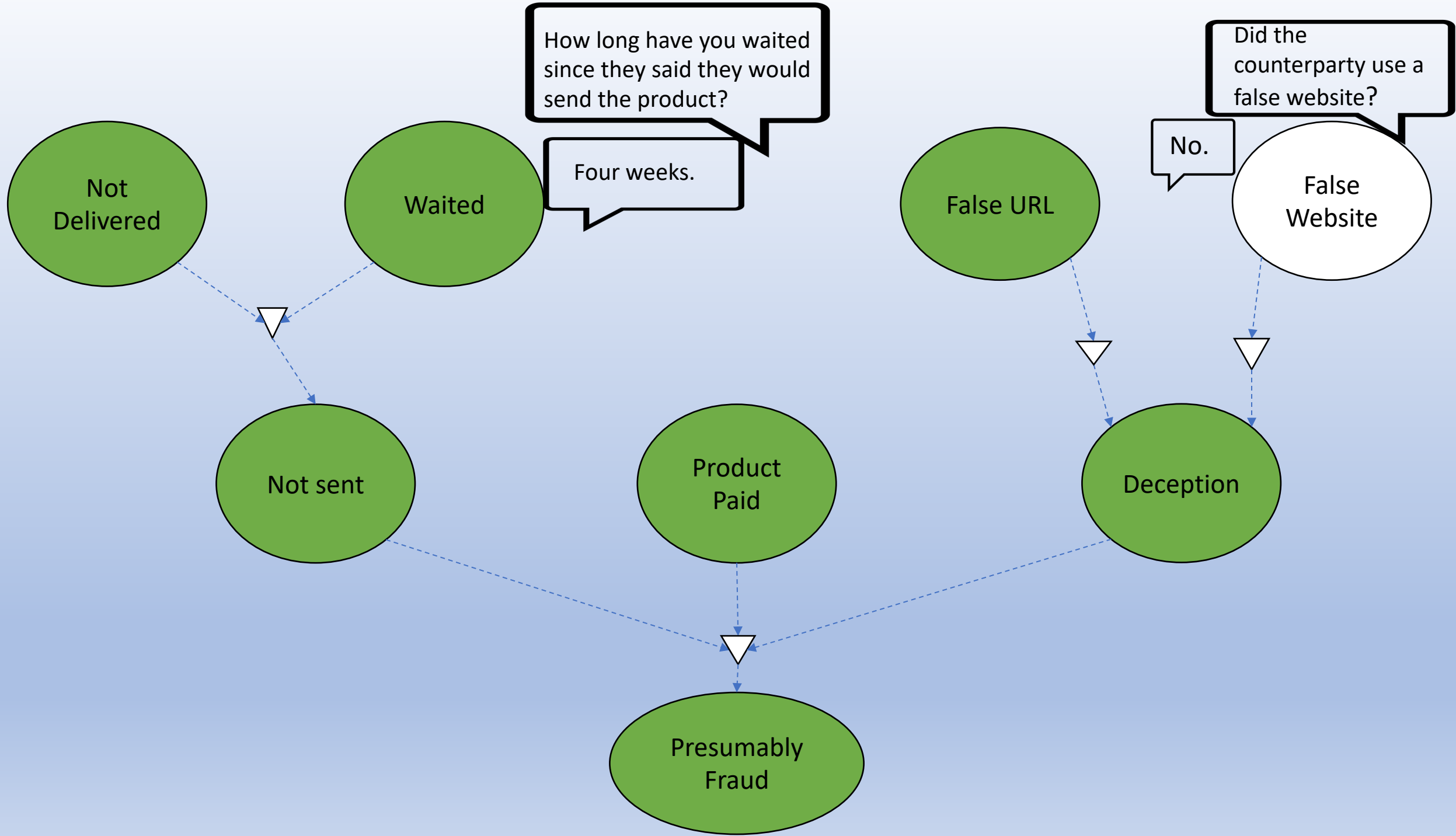


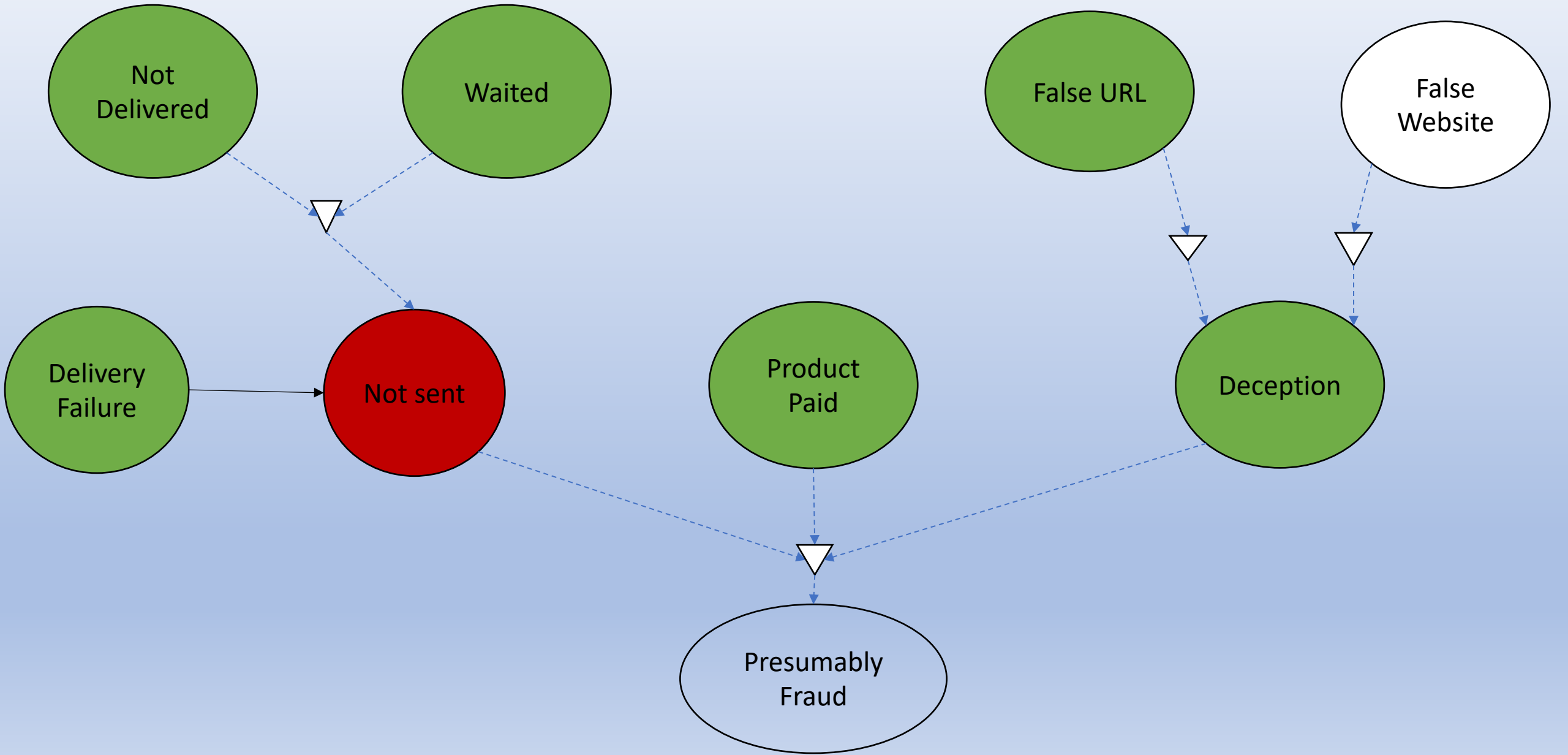

I’ve paid for this product that has not been delivered. The url is false as well.
Did the counterparty use a false website?
No.
How long have you waited since they said they would send the product?
Four weeks.


Not Delivered
Waited
Not sent
Product Paid
False URL
Did the counterparty use a false website?
No.
False Website
Deception
Presumably Fraud


How long have you waited since they said they would send the product?
Four weeks.
Did the counterparty use a false website?
No.
Not Delivered
Waited
False URL
False Website
Not sent
Product Paid
Deception
Presumably Fraud


How long have you waited since they said they would send the product?
Four weeks.
Did the counterparty use a false website?
No.
Not Delivered
Waited
False URL
False Website
Not sent
Product Paid
Deception
Presumably Fraud


How long have you waited since they said they would send the product?
Four weeks.
Did the counterparty use a false website?
No.
Not Delivered
Waited
False URL
False Website
Not sent
Product Paid
Deception
Presumably Fraud


Not Delivered
Waited
False URL
False Website
Delivery Failure
Not sent
Product Paid
Deception
Presumably Fraud

OHAAI

Online Handbook of Argumentation for AI

Trends in argumentation research

Theory	65,12%
Application	41,86%
Abstract Argumentation	55,81%
Structured Argumentation	37,21%
Argument Mining; NLP	16,28%
Dialogues	34,88%
Explainable/Responsible AI	25,58%
Logic	18,60%
Neural Networks	9,30%
Complexity	9,30%
Multi-Agent Systems	6,98%
Enthymemes	9,30%
Other	30,23%

Thank you!

Please provide us some feedback!!
https://forms.gle/fZsqyL5Tu6LkoCFF7