Tutorial: Approaches for Explainability in Autonomous Agents

Sergio Alvarez-Napagao (sergio.alvarez@bsc.es)
Sara Montese (sara.montese@bsc.es)
Victor Gimenez-Abalos (victor.gimenez@bsc.es)
Universitat Politècnica de Catalunya, Barcelona Supercomputing Center

Overview

FIRST BLOCK (1h45)

Introduction (15 min.)

Ante-hoc vs post-hoc explanations in agent settings (30 min.)

Hands-on: post-hoc explanation methods (30 min.)

Limitations of post-hoc explanations: moving towards causality (30 min.)

SECOND BLOCK (1h45)

Explanation methods with causal structure (15 min.)

What different explanation methods actually explain (45 min.)

Hands-on: contrastive and counterfactual explanations (30 min.)

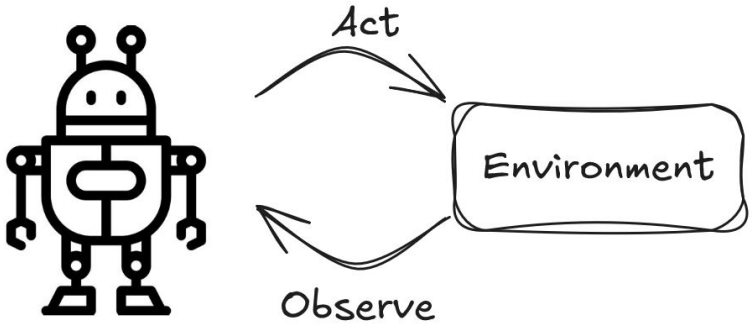
Discussion and conclusion (15 min.)

Introduction

Agent

An agent is anything that can be seen as sensing its environment through sensors and acting upon it via actuators (*AIMA, Russell&Norvig*)

Explainability in our context will take the presumption that what we are explaining is in the context of ‘an agent’ of the previous sort.



What is Explainability?

- **Causality**
- Decisions
- Autonomy
- Inquiry

- **Communication**
- Human-Computer Interaction
- Justification

What is Explainability?

- **Causality**
- Decisions
- Autonomy
- Inquiry

- **Communication**
- Human-Computer Interaction
- Justification

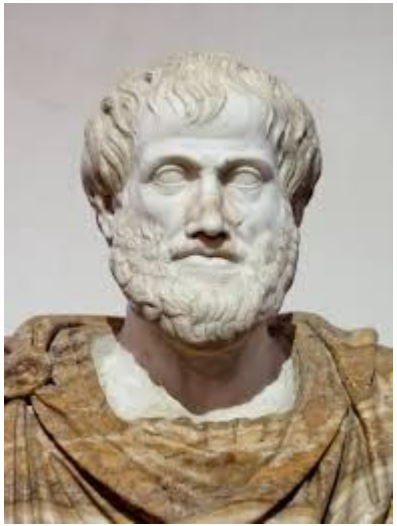
Scientific Inquiry

Social Enquiry

Two Families

Aristotelian

- Function
- Intention
- Teleologic



Galilean

- Cause
- Strict origin
- Mechanistic

Copernican revolution

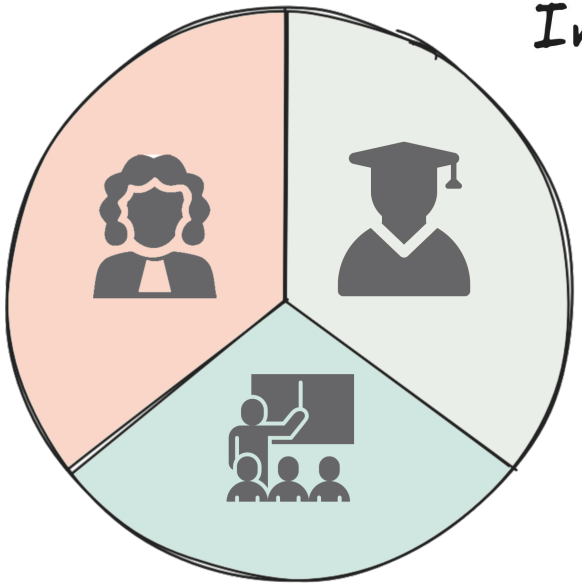
Explainable Artificial Intelligence (XAI)

- **XAI**: field of AI concerned with **producing explanations *about* artificial intelligence systems**
- **Explanation**: **communication** in which there is an **assignment of causal responsibility** / the product of such communication (*Explanation Social Sciences, Tim Miller*)
- **Explainability**: **score** to give a system regarding **how thorough** the explanations it provides are.
- In agents, it is often called **XAg** (Explainable Agency)

What do we want XAI for?

Justify

- Most prevalent
- For humans & Law
- Accountability



Improve/Control

- For machines & humans
- Bias control
- Model learns from it

Discover

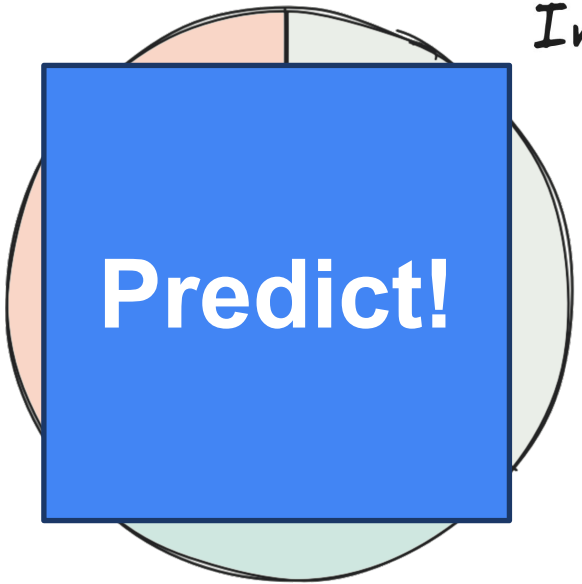
- For machines & humans
- Others learn from model

Adadi et al.

What do we want XAI for?

Justify

- Most prevalent
- For humans & Law
- Accountability



Improve/Control

- For machines & humans
- Bias control
- Model learns from it

Discover

- For machines & humans
- Others learn from model

Adadi et al.

What is XAI useful for?

Predictability as the key use of XAI

Downstream tasks are varied:

- Censor/Sanction a system (Justify)
- Improve/Debug unwanted behaviour (Control)
- Extract information about the task (Discover)

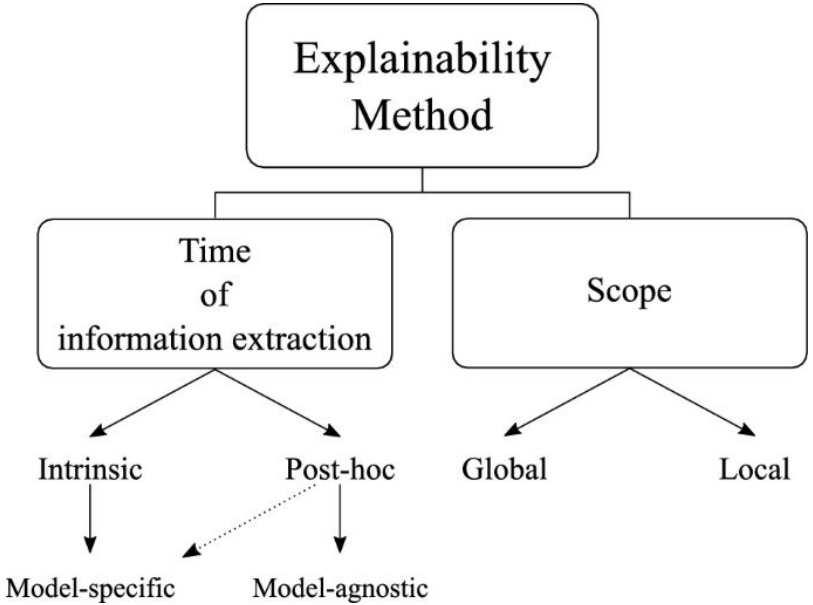
Humans don’t use explanations **only** in court!

Types of XAI

Stripes and colours of XAI

- Ante-hoc vs Post-hoc
- High-level surveys & taxonomies
- Feature Attribution
- Adoption of the technology

XAI Taxonomy



Puiutta E. et al.

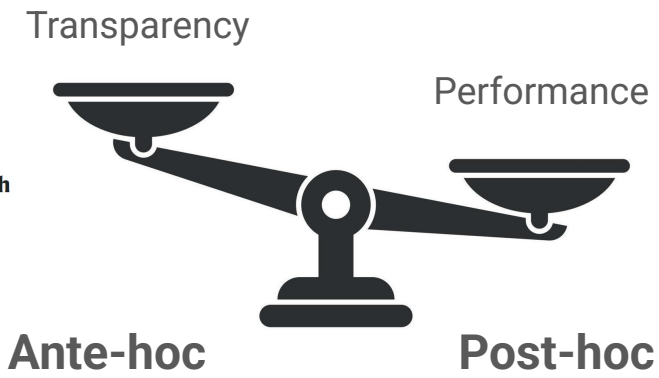
XAI Taxonomy

Perspective | Published: 13 May 2019

Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead

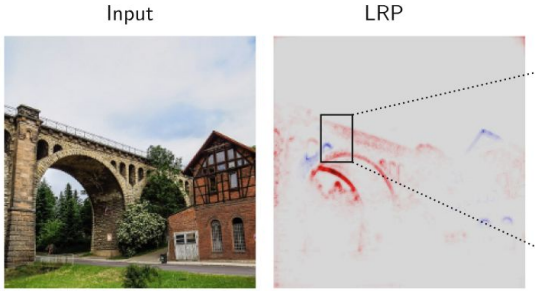
Cynthia Rudin

Nature Machine Intelligence **1**, 206–215 (2019) | Cite this article



How should it be?

Feature Relevance



Montavon et al., 2019.

Interpretability

The Mythos of Model Interpretability

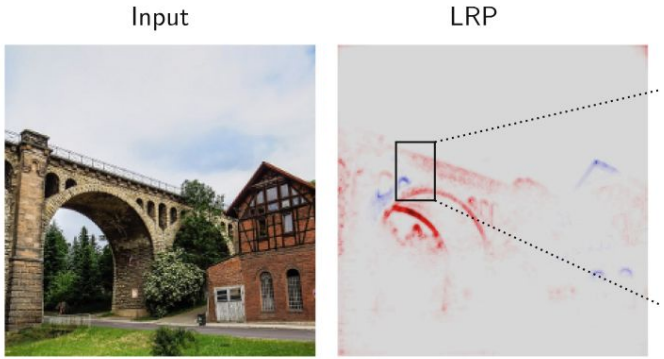
Zachary C. Lipton [1]

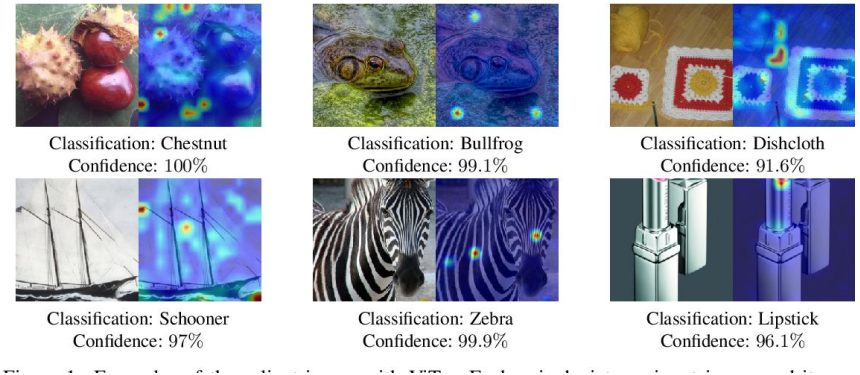
User studies

Fig. 2. Believability, acceptability, and comprehensibility scores for the three Cohorts and five explanations.

Winnikof et al., 2018

State of the Art in XAI (non-Agents)

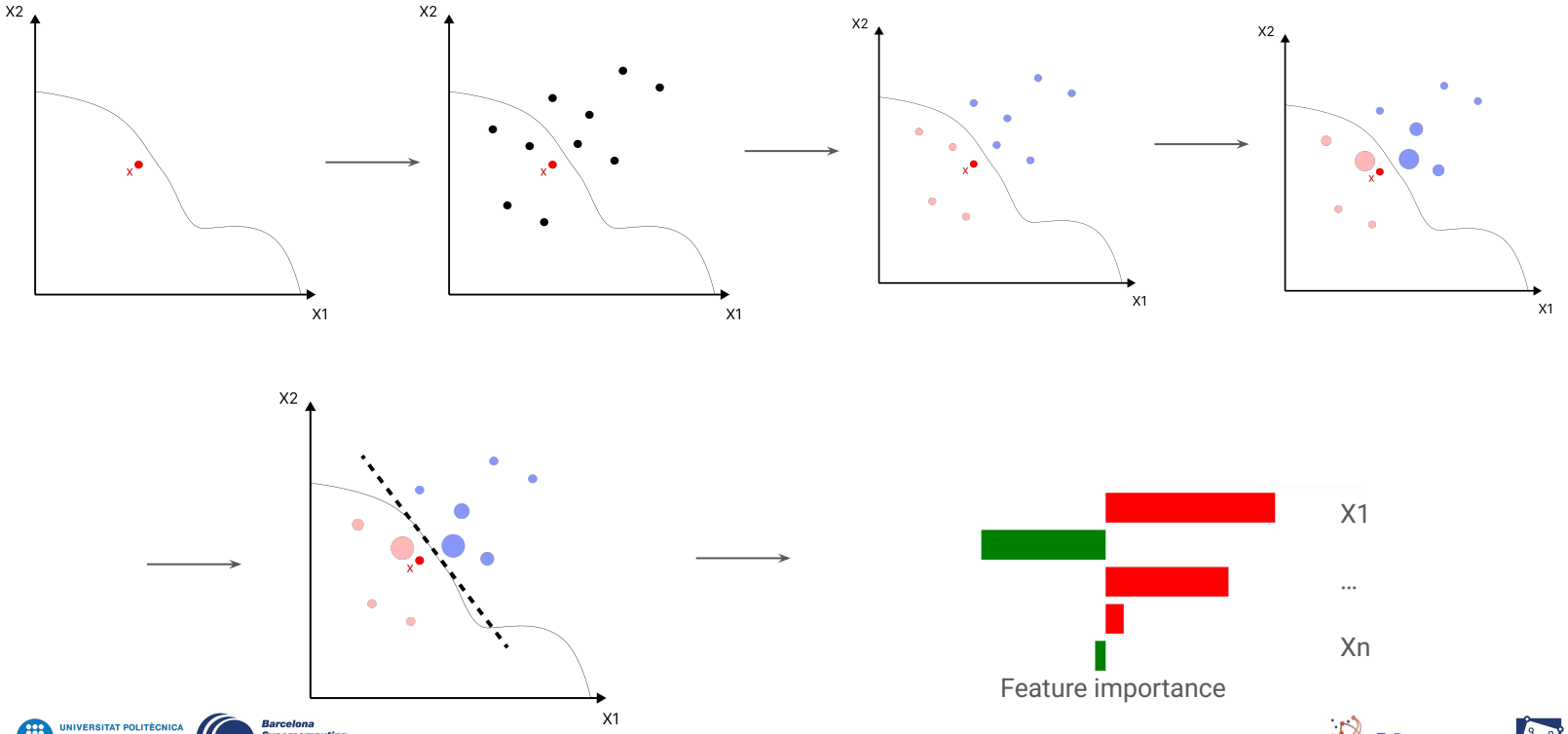




- SHAP
- Coefficients (for linear models)
- "Counterfactuals" (closest similar of another class)
- ...

Selvaraju et al.
Montavon et al.

LIME (Local Interpretable Model-agnostic Explanations)



SHAP (SHapley Additive exPlanations)

Shapley value: a player's contribution to the game payout, weighted and summed over all possible coalitions that the player could join

Shapley value of *player i*

Game

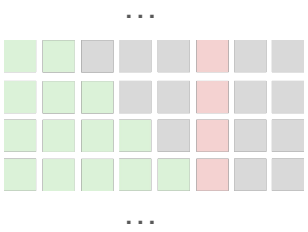
Weighting

$$\phi_i(f) = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f(S \cup \{i\}) - f(S)]$$

Contribution from adding *i*

Coalitions excluding *i*

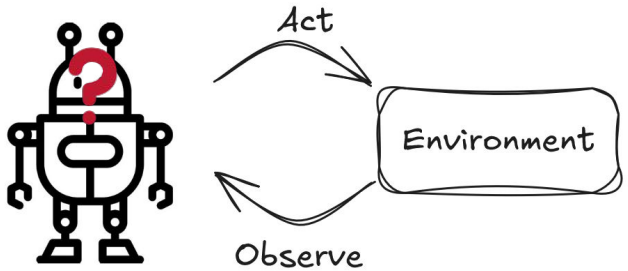
Feature set



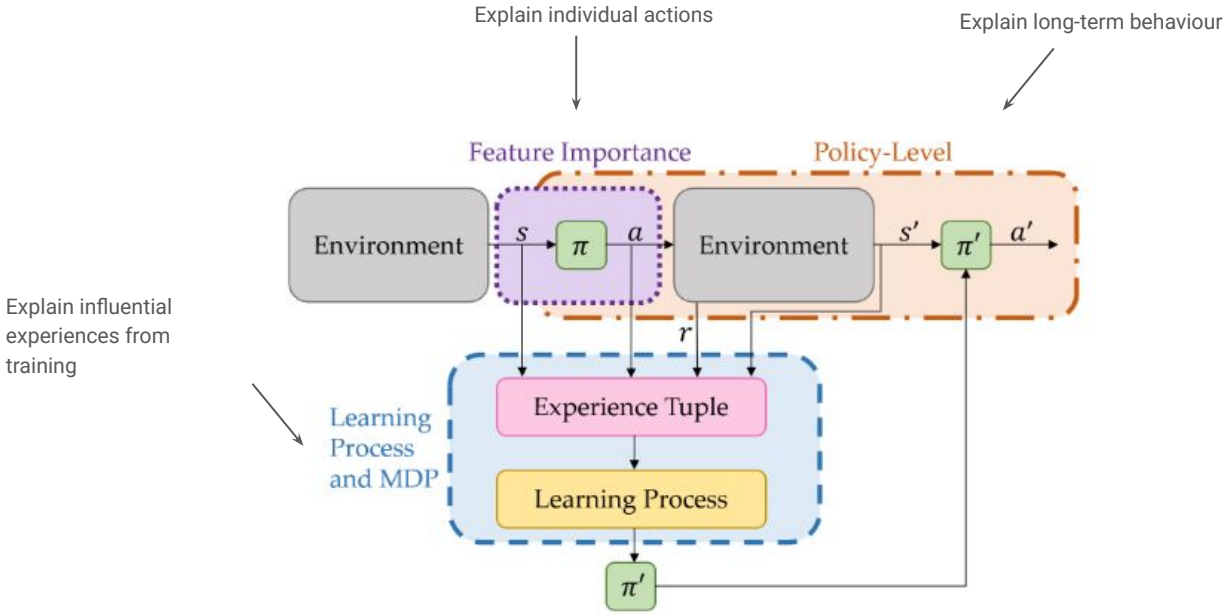
Approximations

KernelSHAP, (TreeSHAP, DeepSHAP...)

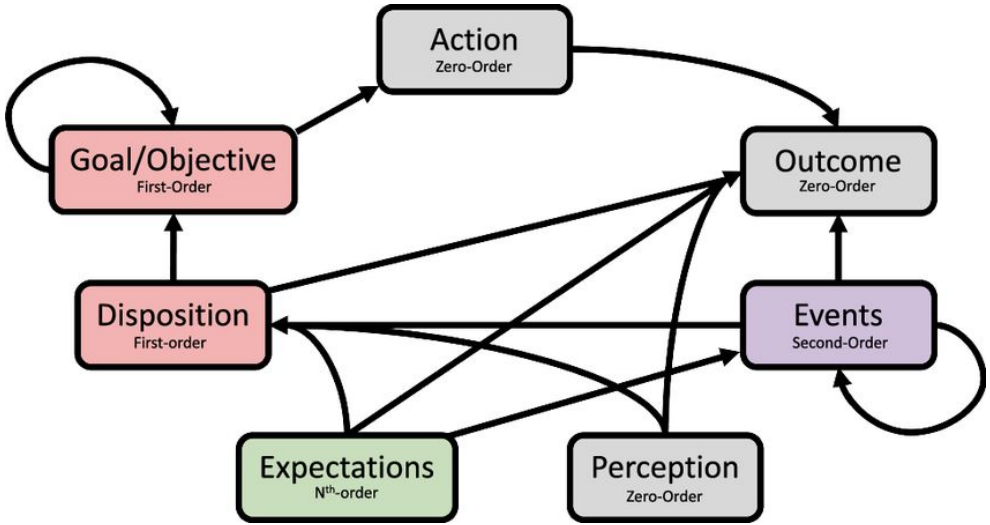
Is traditional XAI enough for Agents?



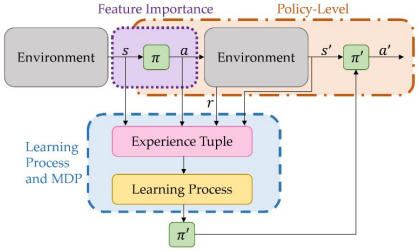
Taxonomy of XRL methods (Milani)

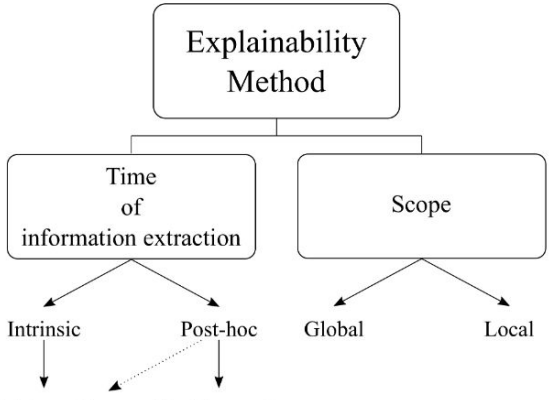


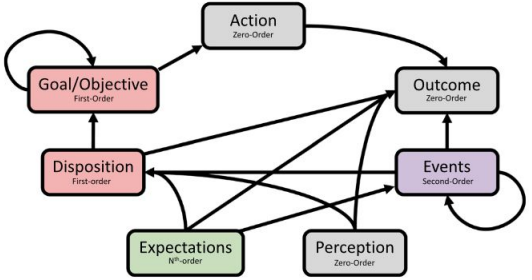
Taxonomy of XRL (Dazeley)



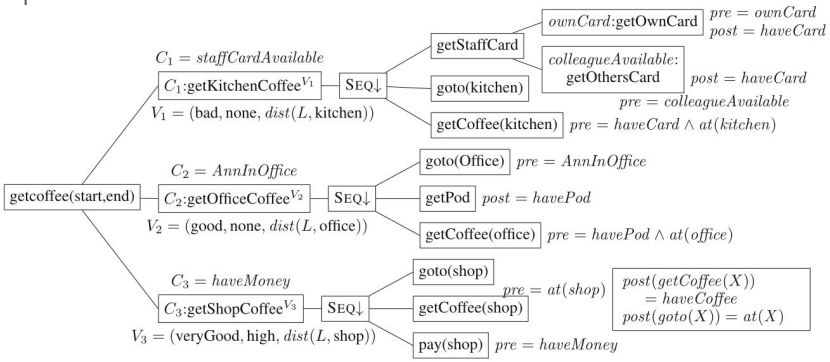
Agent XAI: How “Theory” looks like today



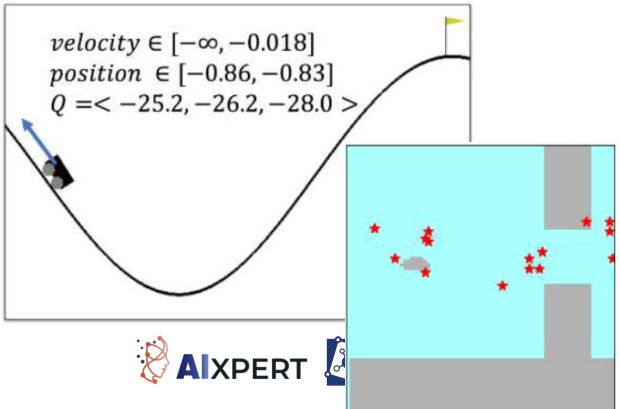




‘XAI-by-Design’



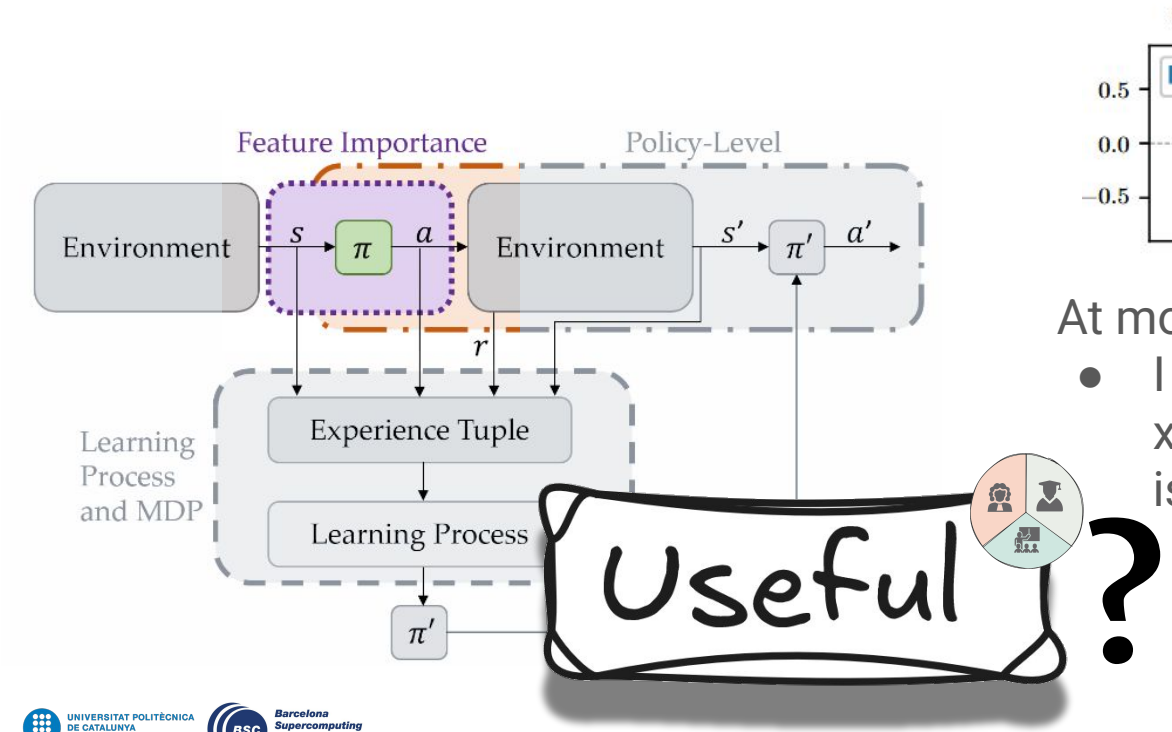
“General” Methods

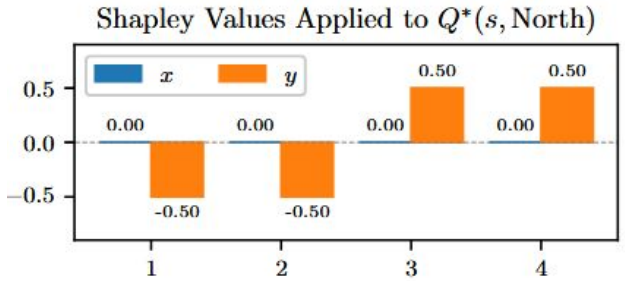


[Winnikof et al.](#)
[Puiutta & Veith](#)
[Dazeley et al.](#)
[Milani et al.](#)

AIXPERT

Agent XAI: How Practice looks like today





At most:

- I do ‘up’ when I am between x=[1,3] and y=[-2,3] and there is no screw at x,y=(1,4)

Hands-On

Post-hoc XAI

- Toy RL Example
- LIME/SHAP Methods
- Surprise at the end

Block 4: Limitations of the example

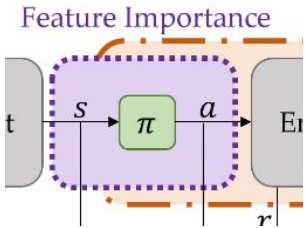
Impossibility theorems for Feature attribution

- Why do we do Feature Relevance
- Reliability of XAI
- A cynical view on XAI milieus
- How to improve
  - View 1: Consider Aristotle
  - View 2: Consider Causality

Why is post-hoc feature relevance so adopted?

Low cost. Many libraries. Easy to implement for a use-case.*

For Mechanistic explanations (in vogue)



Acceptable by legislators: Industry can say they do XAI if they're using this

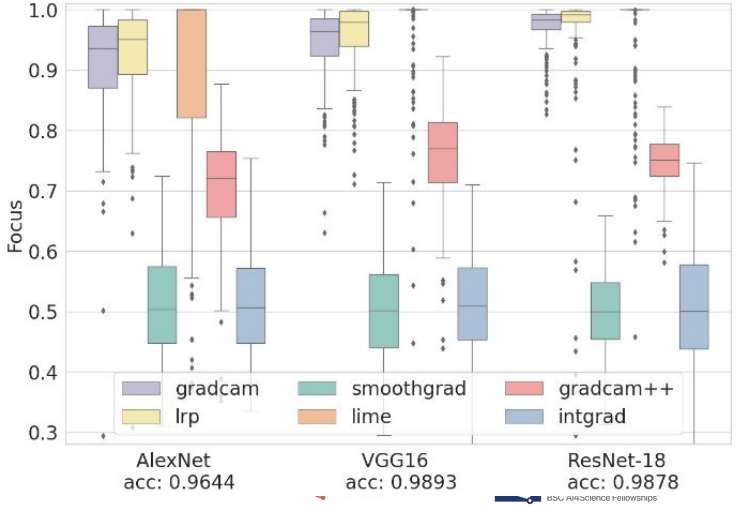
Well developed in XAI for ML, applies to XAg

*Meeting expectations? Whose?

Diatribe on Reliability

“Our main result is that for mildly rich model classes, it is **impossible** to conclude that the **user does better than random** guessing at inferring counterfactual model behaviour **using common feature attribution methods** without strong additional assumptions on the learning algorithm or data distribution.” ~ Bilodeau et al.

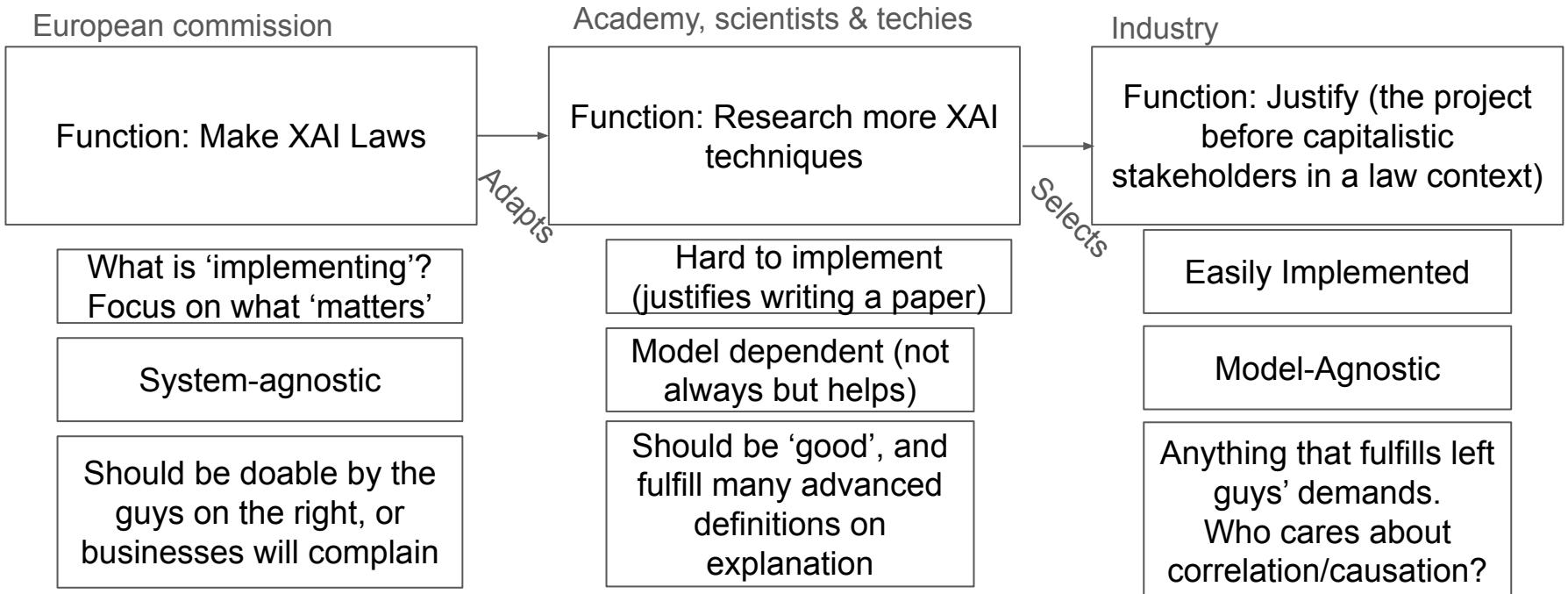
(d) Dogs vs. Cats (e) GradCAM



“Focus!” ~ A Arias-Duart et al.

Milieus of XAI: Why is it like it is today?

(a cynical view)



UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Barcelona Supercomputing Center Centro Nacional de Supercomputación

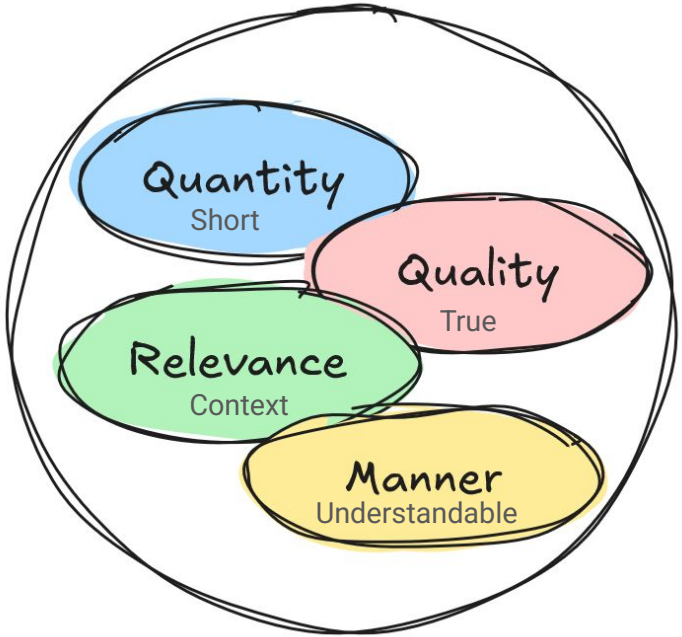
AIXPERT

AI4S BSC AI4Science Fellowships

How should it ACTUALLY be?

* Hopefully, not just to a few, but to society

What is the best Form?



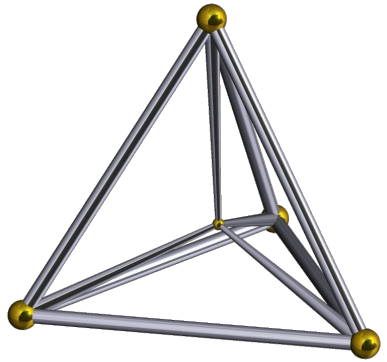
- Causal > Probabilities
- Contrastive/Counterfactual
- Selected
- Social

"Explanation in artificial intelligence" ~ T. Miller

Grice's Maxims

What does Understanding mean?

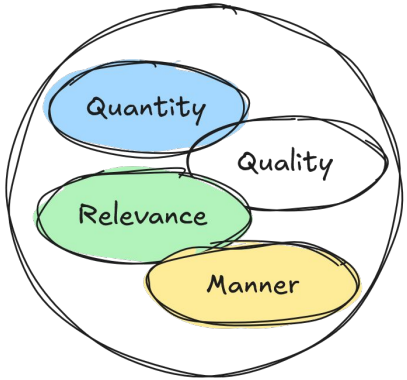
- **Understanding** (for humans) **means having a mental model**, to model/predict and reproduce events
  ~Craik, Johnson-Laird, Byrne
- Transfering a model to a human brain is not possible. What goes on?
  - **Defining boundaries of behaviour: what is vs what is not**
  - **Contrastive / Counterfactual knowledge**
- Wittgenstein: “Meaning is Use”
- Bilodeau: “... bespoke methods accounting for the downstream tasks fare better”



View 1: Consider Aristotle (Need for Intentions/Goals/Valuings etc.)

Why did you fill the pot with water?

- I filled the pot with water because it was empty, and because the hob was free, and because there is no fire alarm going on
- I filled the pot with water because I
  - need boiling water
  - intend to boil pasta
  - am cooking pasta carbonara



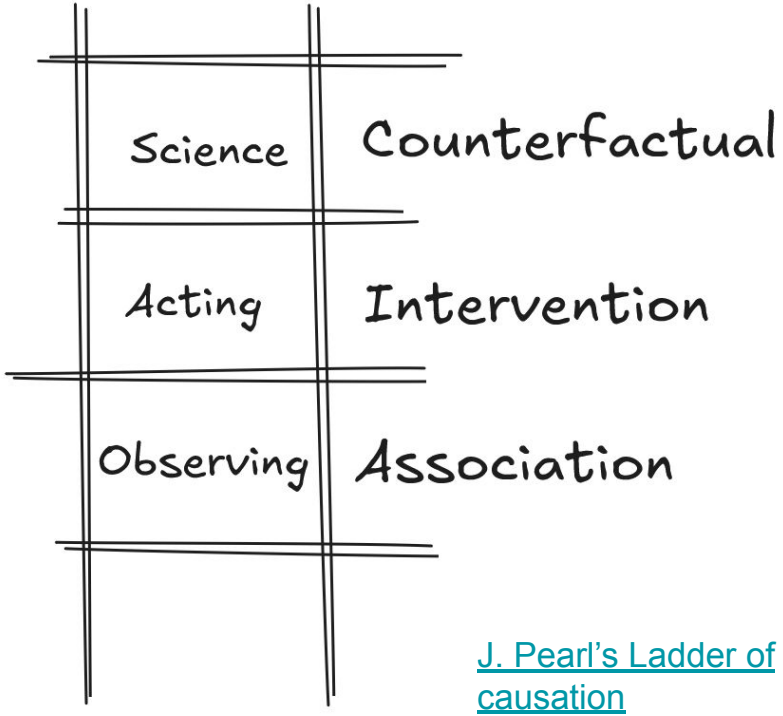
Are we OK to say a DL model has intentions?
Artificial things are created with a purpose.
‘Why that policy?’

View 2: Consider Causality

1. Correlation does not imply Causation
2. Then what is causation?
3. What can causation be used for?

Relevance of Agency:

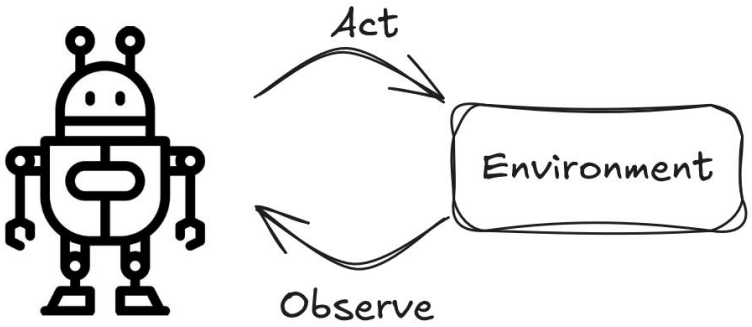
- Autonomous intervention
- Freedom of exploration
- Quality of knowledge



J. Pearl's Ladder of causation

View 2: Relevance of Agency for XAI

- XAI verification, causality, etc. can only be checked if **interventions** happen.*
- XAI as a task for intelligent beings.
- XAg as a requirement for clearing bigger field of XAI.



*E.g. deletion/insertion methods.

Coffee Break

Explanation with causal structure

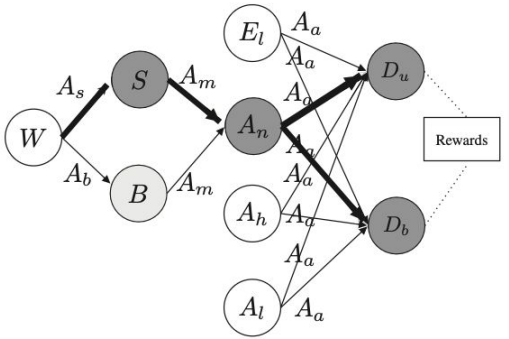
- Madumal
- CEMA
- IPGs

The need for structure

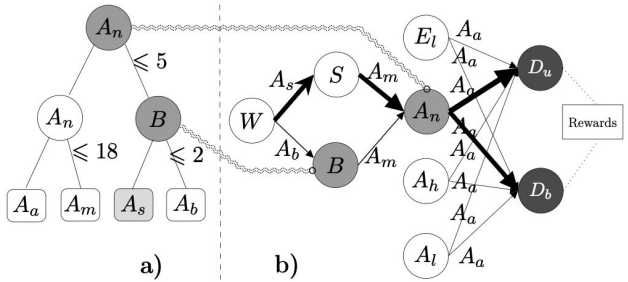
- Recap: good explanations should be
  - Short, true, causal, contrastive, user-centered, social, useful?
- These properties often require:
  - Causal structure: what could change under intervention?
  - Sequential structure: what future opportunity does an action enable?
  - Telic structure: what goal/desire does the action serve?
- References:
  - Miller, 2019
  - Madumal, 2020
  - Judea Pearl

XRL Through Causal Lens (Madumal)

- Model-free RL + causal explanation
  - State variables + action-labelled causal edges
- Action Influence Model:
  - Why action A?
  - Why not action B?
- Opportunity chains and distal explanations
  - A enables B
  - B causes C



State variables:
W - Worker number
S - Supply depot number
B - barracks number
E - enemay location
A_n - Ally unit number
A_h - Ally unit health
A_l - Ally unit location
D_u - Destoryed units
D_b - Destroyed buildings
Actions:
A_s - build supply depot
A_b - build barracks
A_m - train offensive unit
A_a - attack

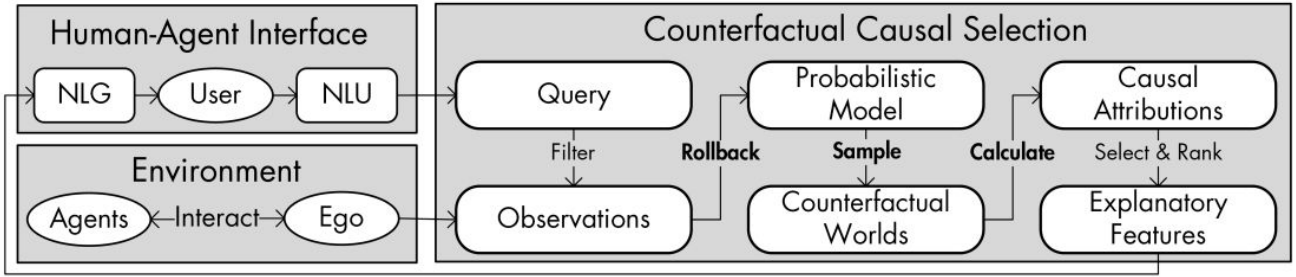


State variables:
W - Worker number
S - Supply depot number
B - barracks number
E - enemay location
A_n - Ally unit number
A_h - Ally unit health
A_l - Ally unit location
D_u - Destoryed units
D_b - Destroyed buildings

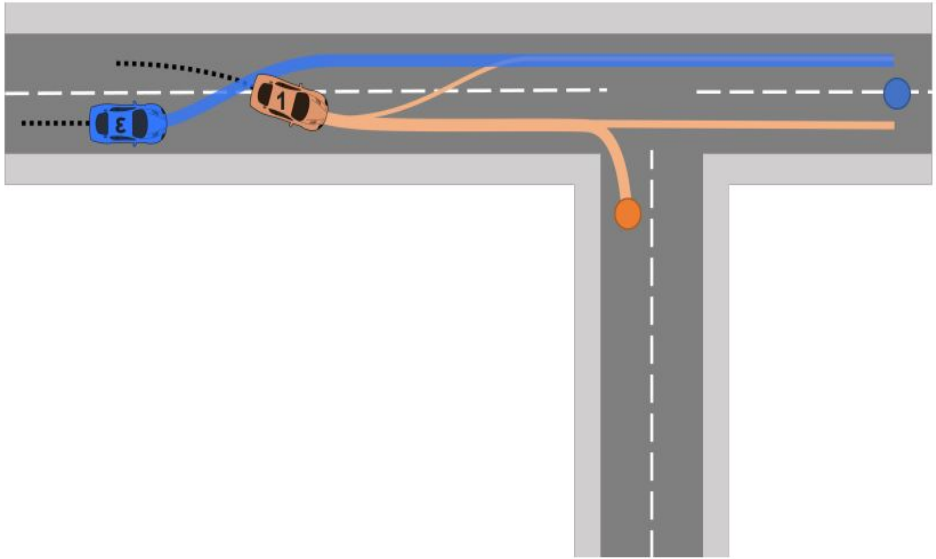
Actions:
A_s - build supply depot
A_b - build barracks
A_m - train marine
A_a - attack

CEMA: Causal Explanations for Multi-Agent Systems

- Sequential multi-agent systems
  - Roll back before the queried action
  - Simulate counterfactual futures
  - Rank causes by counterfactual effect size
- Explanation modes:
  - Mechanistic: factors caused by other agents or by the environment
  - Teleological: own goals or reward effects



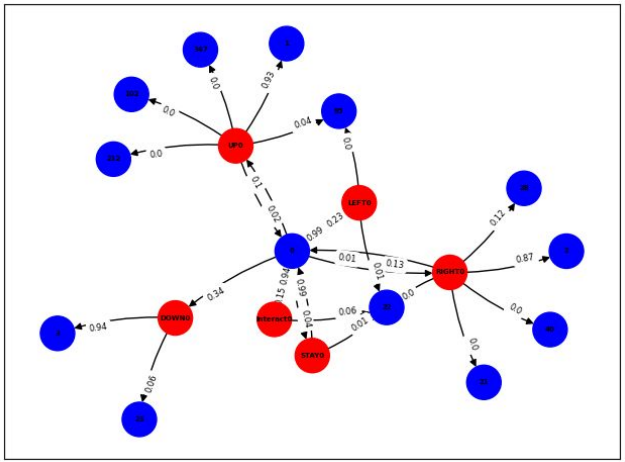
CEMA: Example



Scenario 1 (S1-A)

Why will you change lanes?
It will decrease the time to reach the goal.
Why does it decrease the time to the goal?
Because vehicle 1 will be slower than us.
Why will it be slower?
It will decelerate and turn right.
What if it hadn't changed lanes before?
We would have gone straight.

Intention-aware Policy Graphs



States (Discrete)
Actions

Edges:
$P(s'|a,s), P(a|s)$

Hypothesised Desire *d*: $\langle S_d, A_d \rangle$

Intention value $I_d(s)$: *probability to achieve* **d** *from* **s**

Standard questions:

- What will you do in state s?
- When do you do action a?
- Why not action a' in state s?

Telic questions:

- Why do you do action a in state s?
  - Action a increases the probability of achieving intention I_d

Hands-On

Contrastive/Causal XAI

- XRL through a causal lens
- CEMA
- IPGs

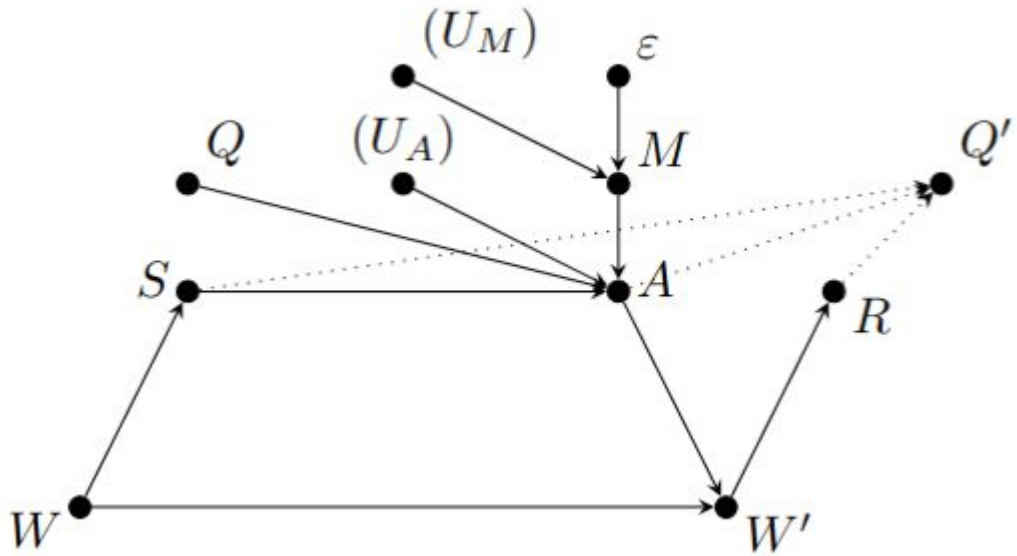
What XAg actually explain

- Explaining one Agent
- The goal behind a unification
- Why that policy? A Ladder of intentions

Explaining an Agent (Q-learning)

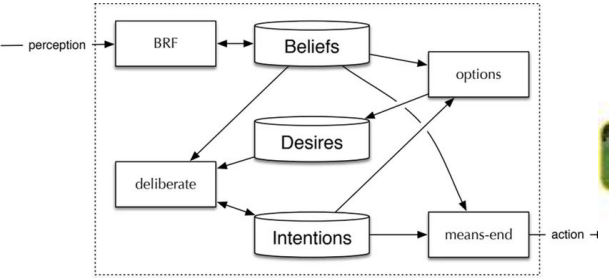
Q-learning

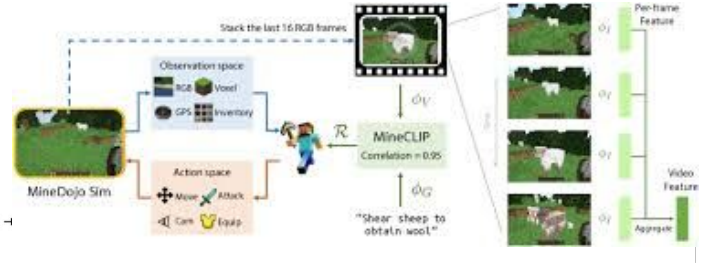
- Agent acts according to the ‘best known action value in a state’:
  argmax_a Q(s,a)
- Q is learnt from experiencing states (S), actions (A), and rewards from the environment (R)
- **BUT!** You can’t learn Q if you are only exploiting what you know! You have to sometimes explore alternatives, be curious, or you’ll converge to local minima
- Multiple Modes (M): Exploring/Exploiting. When exploring, pick A randomly!

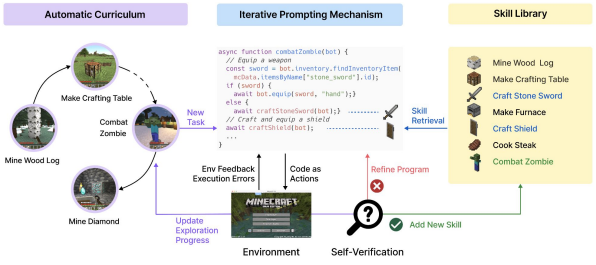


The gutpunch

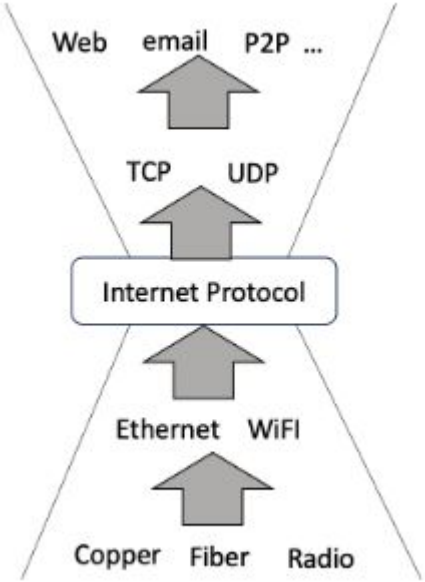
Now you know Q-learning. You can ‘ask’ XAI about this model’s particular decision, study how Q is created, or why this action was taken then.
You can measure XAI across 20 metrics, and run 50 user studies in 30 use-cases.
What about the 300 other, equally valid architectures?
And… what about people that **don’t work on AI?**

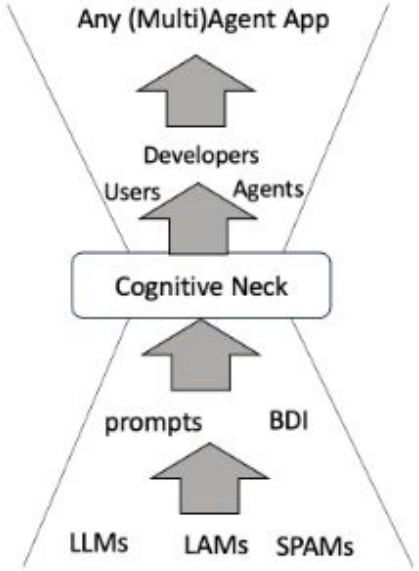


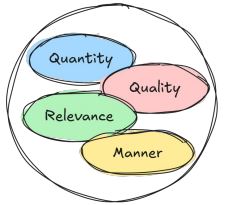




Why is a Unifying Model it important?







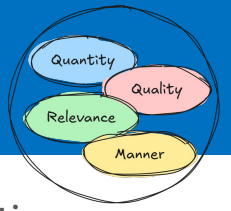
- How to do good XAI?
  How do we evaluate manner?
  - Cohorts with explainees

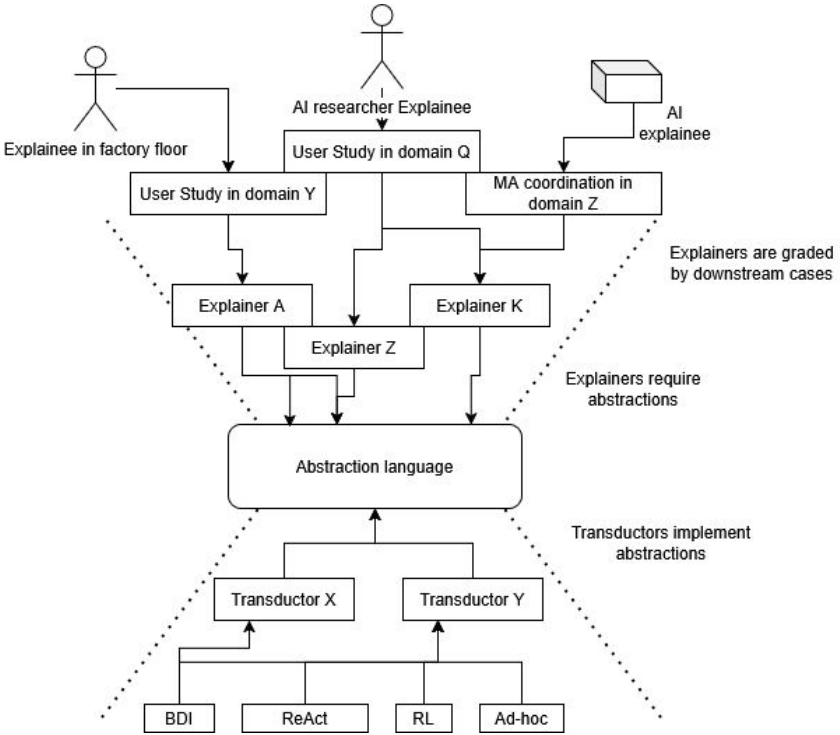
  How do we evaluate truthfulness for a particular case?

  Do the user studies extrapolate?

  Must we repeat every user study with every architecture, kind of possible explainee, situation and context?

Unifying XAI proposal

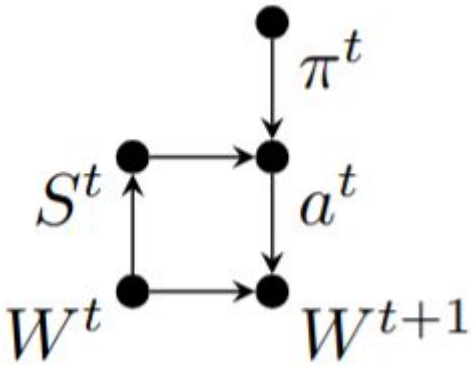




- Language decouples applications and their usage from substrate
- Transductors to language are **not necessarily** perfect, but must acknowledge their sources of errors and assumptions
- User studies become independent from architecture, permitting extrapolation
- **Explainability becomes a non-functional requirement enforceable by contract in advance**

A Ladder of Intentions

Key Idea: Why that policy?



- Why π? Reasons(π)
- Then… why Reason(π)? Reasons(Reason(π))
- And so on…

The agent can be ‘summarised’ into π. Splitting it to see its causes makes it more understandable.

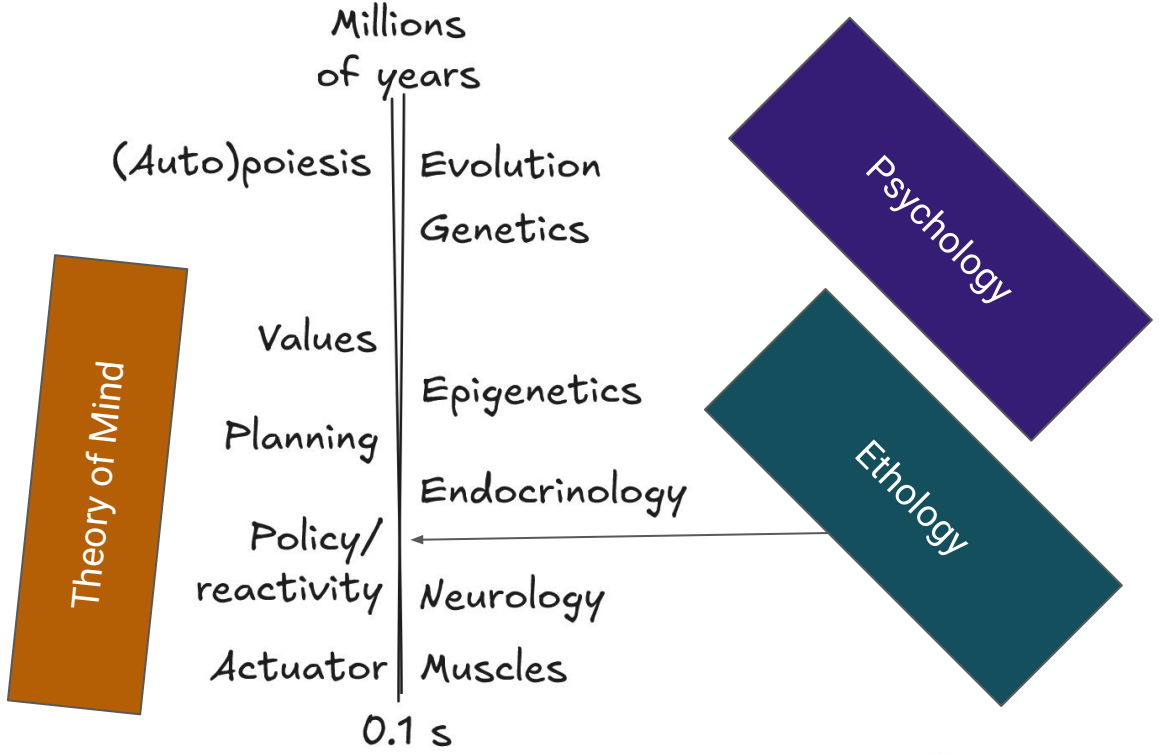
How to split?

Causality of action, Humans & AI

Why does Entity exist with some properties?

How does it arrive at a manner on how to do things?

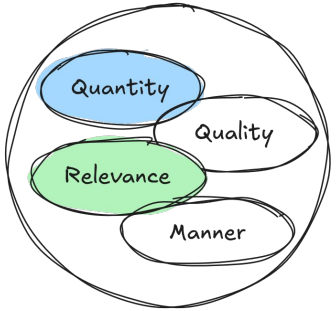
How does it finally act based on its manner?

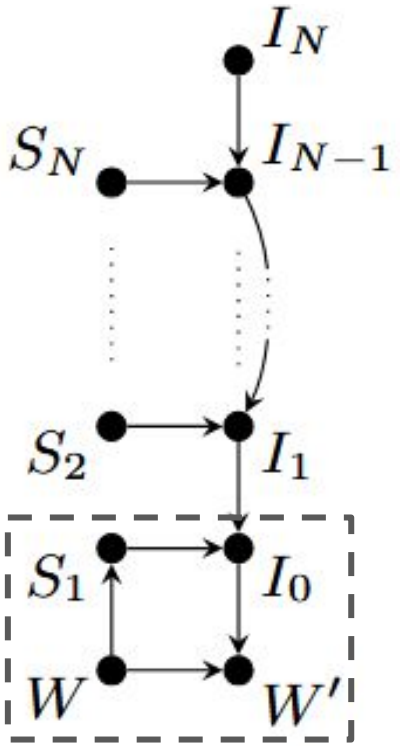


Ladder of Intentions

- Holistic, universal taxonomy of Behaviour
- Intention-centric
- Praxia vs Gnosia; declarative vs imperative
- **Applies to BDI, RL, LLM-agents, custom code…**
- Exhaustively finds **ALL** XAI, split in bits and rungs

Decouples how to make Interpretable XAI from truthful & feasible XAI





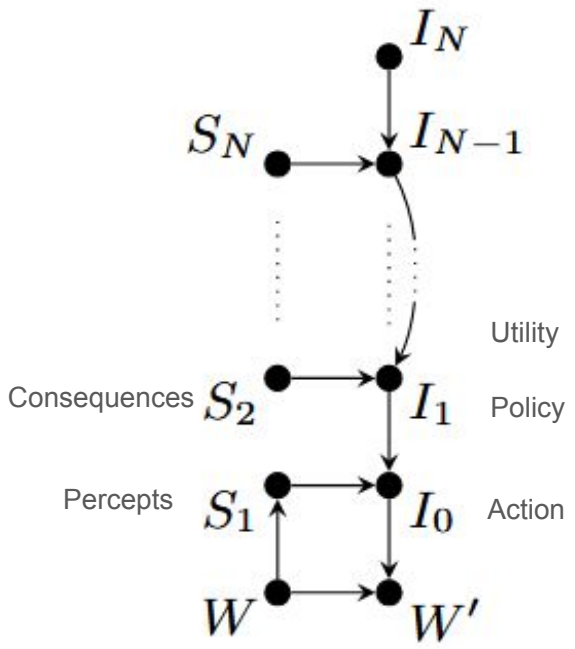
The Levels, Reification, Robustness

- Splitting into levels, separating statements (S) from code (I) using them
- How are ‘rungs’ separated?

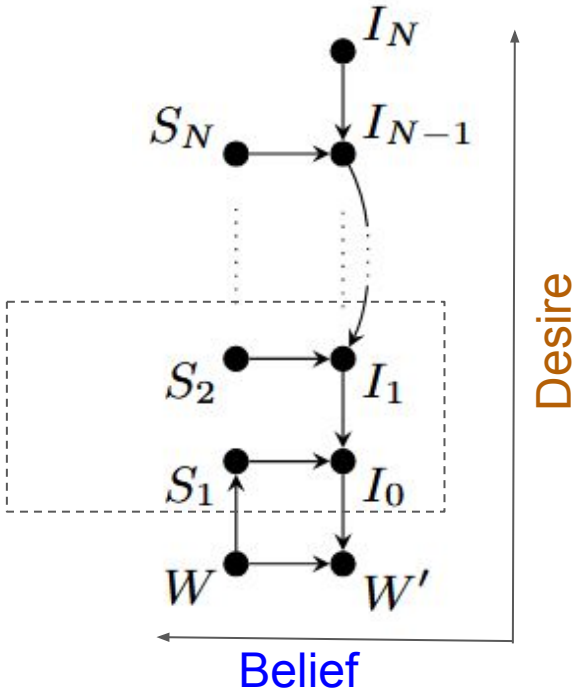
  Statement **Reification**:
  s_2=Consequence of [i_0] in [s_1] is [s_1]

  Elaborations of lower levels

  - Desirable s_1 ; Value of *a* in s_1…
  - Confidence in transition s_1 to s_1’, preferences between desires…

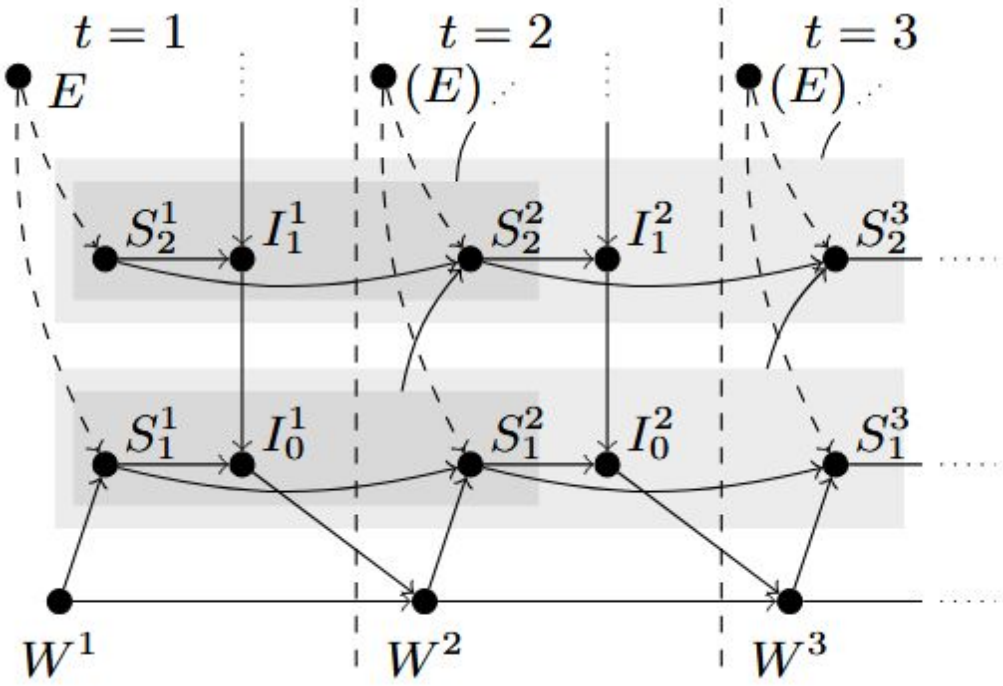


Folk-Psychology grounding + Example



- Why go right? Because [there is a wall on top] and [it's the best action]
- Why is it the best? Because [] [I am exploring random actions]
- Why are you exploring random actions? Because [I have an estimator of when it's better to exploit my knowledge vs explore] [and I use it to decide to explore now]
- Why do you use a value to decide to explore or exploit? Because [that's the intent of my creator]

Ladder Learning



- Learning a statement $\mathbf{s}_i$ implies learning something about $\mathbf{s}_{i-1}$,$\mathbf{i}_{i-2}$.
- They can be given a priori
- They can be **learnt** by experiencing the lower level
- An agent could be designed to **climb** the ladder

The 3 kinds of questions

Intentional

I to I

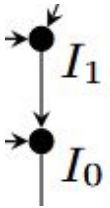
Goal oriented. Climbs up.

Concise, but not comprehensive.

May need clarification (Weakly-epistemic)

High predictability.

Why up?
Because I am going to the pot.



Weakly Epistemic

I_i to S_{i+1} (in the context of I_{i+1})

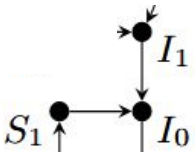
Why believe in a knowhow?

Declarative-relevance, supports a way of acting, bounded rationality.

How S came to be remains unclear (Strongly Epistemic)

Cognitively harder on explainee.

Why go up when going to the pot?
In this state, the pot is above me.



Strongly Epistemic

S to S, I (of a lower level)

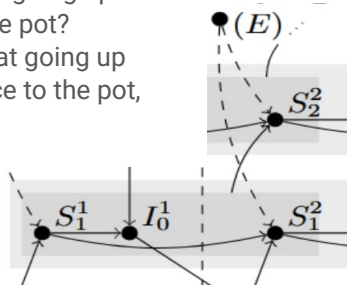
Why believe in a statement?

Learning explanations.

Understanding how E works is non-trivial

Aggregates tons of data, currently unfeasible in many domains.

Why did you think that going up results in you being closer to the pot?
I have experienced that going up decreases my distance to the pot, increasing reward.



Level 2 formalisation: general S_2

$$a >_s a'$$
$$a >_s^d a'$$
$$U_{\gamma(s^{t'})}(a^t, s^t)$$
$$P(\gamma(s^{t'})|a^t, s^t)$$
$$P(s^{t+1}|a^t, s^t)$$
$$SCM : \langle properties(s^{t+1}), \{a^t, properties(s^t), randomness\}, F\rangle$$

General idea: establish action preference/deliberation mechanism over action preference. Bottom can reduce to top.

From easiest (least informative) to hardest (most expressive):

- Easiest: action preference per state
  - Why a in s? Because it's the best. Why? Strongly epistemic question…
- Medium: probability about characteristics of future states
  - Why a in s? It results in a future state where x happens (tied to I_2, x is part of its intention)
- Hardest: causal model of state transitions
  - Why a in s? It likely changes the state such that… and then I will do a'... and I get x.

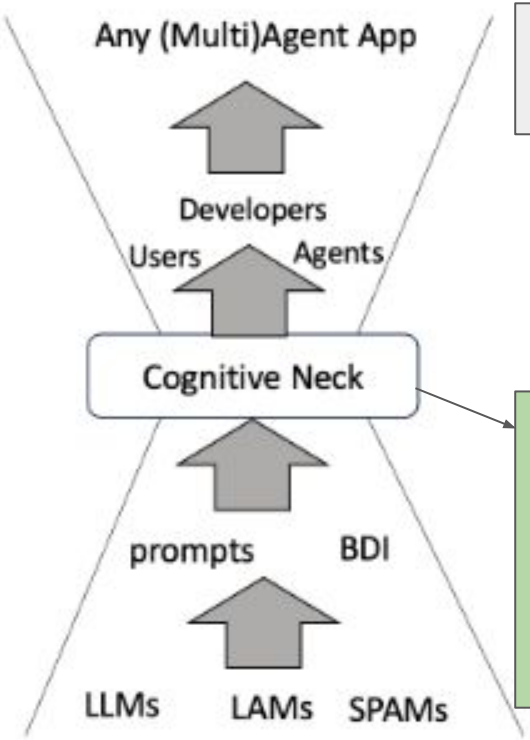
What this means

Process for 'Usage'

XAI of type *K* are good for factory floor robotics, followed by type *P*; according to user studies. I need to pick a substrate which uses formalisms X+Y (to use *K*) or Z (to use *P*)

Process for Agent Expert

I think I can do a new implementation of BDI2 such that performance is better and it implements Y with 60% reliability.



Process for Explainer Expert

I can invent a new type T of answering an XAI question; using formalism Z+Y. I will run a user study for AV to check if the type T is good for that scenario.

Agent BDI implements X(90%), Y(20%) and Z(40%); and has a good performance for this task

Agent Voyager implements…

…

Discussion & Conclusion

- Are you using XAg (or planning to) for anything beyond *justifying*?
- Is there any other XAg method that you liked?
- Can we explain the behaviour of LLM-agents?