Present and future challenges of ethical AI?

Luciana Fidilio-Allende José Walter Hernández-Perez

Carmen Fernández Martínez Eduardo de la Cruz Fernández

Sara García Rodríguez Alicia Pina Zapata

Grupo de Inteligencia Artificial de la URJC
Center for Intelligent Information Technologies (CETINIA)
Móstoles, Madrid

23 June 2025 (Curso Verano URJC)

Value-centered process automation

Speaker: Luciana Fidilio Allende

- Degree: Cybersecurity Engineering.
- Master: Artificial Intelligence in the Energy and Infrastructure Sector.
- Currently ending the first year of PhD in AI.
- Position: Researcher at GIA.

Why?

Process automation is a key aspect of the modern world.
It achieves efficiency, scalability, and cost reduction.

Why?

Process automation is a key aspect of the modern world.
It achieves efficiency, scalability, and cost reduction.

However, is it enough?

Why?

Process automation is a key aspect of the modern world.
It achieves efficiency, scalability, and cost reduction.

However, is it enough?

We aim for more human-centered automation.
A process automation centered around human values instead of utility.

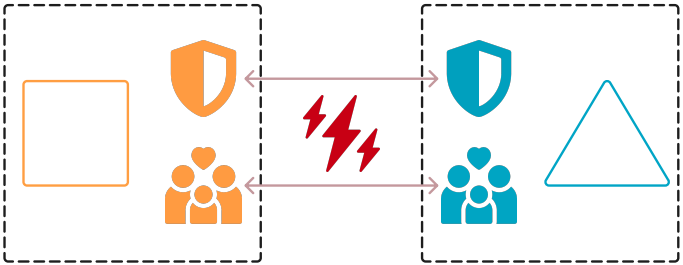
Because values matter.

Value-centered negotiation

One key process is negotiation.

To reach agreements, we need to:

- Adapt our value system.
- Take into account the values of the others.

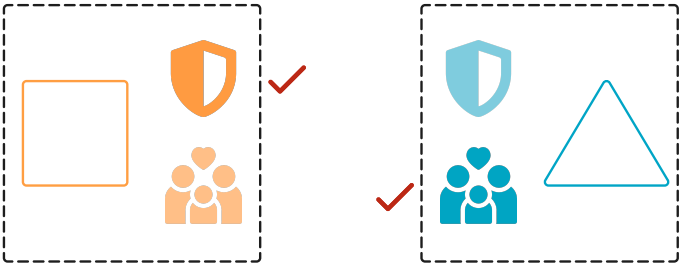


Value-centered negotiation

One key process is negotiation.

To reach agreements, we need to:

- Adapt our value system.
- Take into account the values of the others.

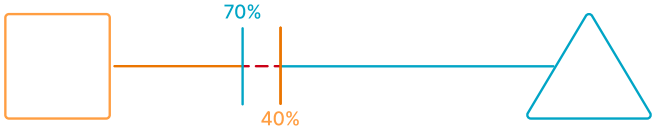


Value-centered negotiation

One key process is negotiation.

To reach agreements, we need to:

- Adapt our value system.
- Take into account the values of the others.



Value-centered negotiation

One key process is negotiation.

To reach agreements, we need to:

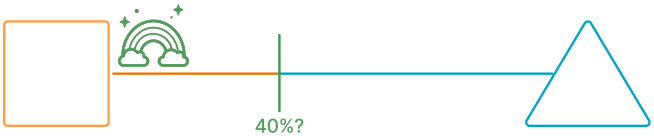
- Adapt our value system.
- Take into account the values of the others.

Value-centered negotiation

One key process is negotiation.

To reach agreements, we need to:

- Adapt our value system.
- Take into account the values of the others.

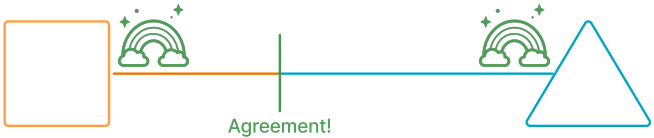


Value-centered negotiation

One key process is negotiation.

To reach agreements, we need to:

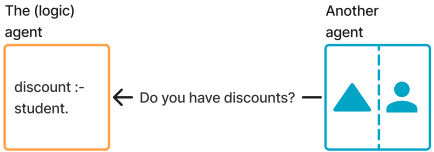
- Adapt our value system.
- Take into account the values of the others.



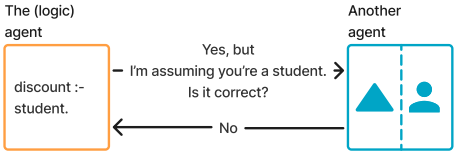
Explainable, modifiable, trustworthy logic agents

Explainable, modifiable, trustworthy logic agents

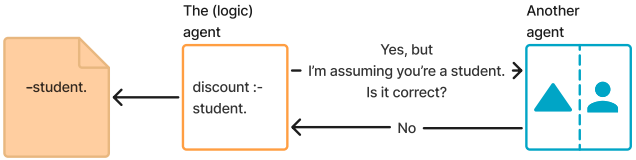
Explainable, modifiable, trustworthy logic agents



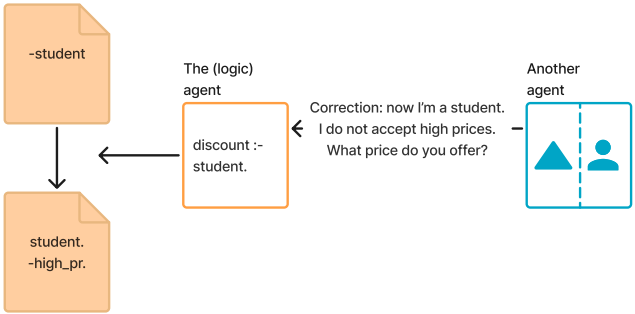
Explainable, modifiable, trustworthy logic agents



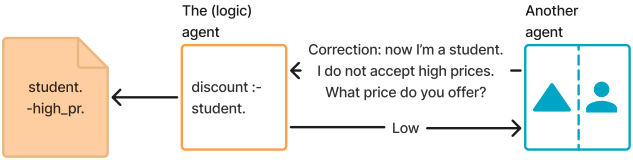
Explainable, modifiable, trustworthy logic agents



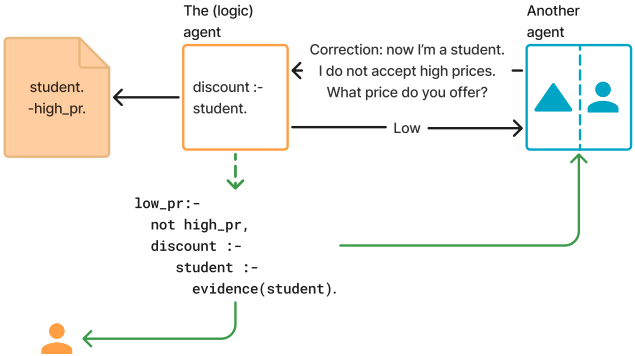
Explainable, modifiable, trustworthy logic agents



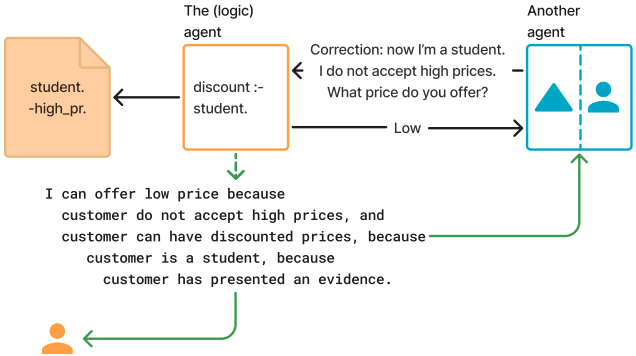
Explainable, modifiable, trustworthy logic agents



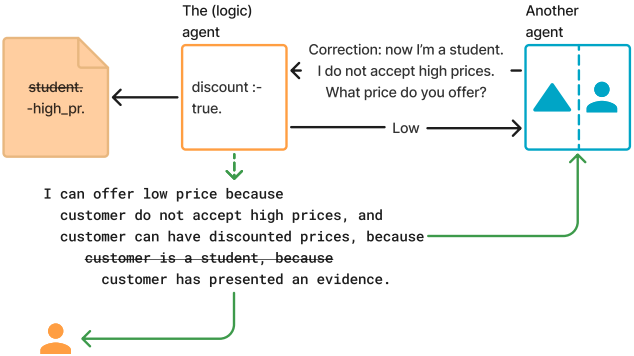
Explainable, modifiable, trustworthy logic agents



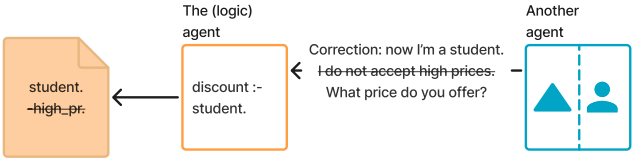
Explainable, modifiable, trustworthy logic agents



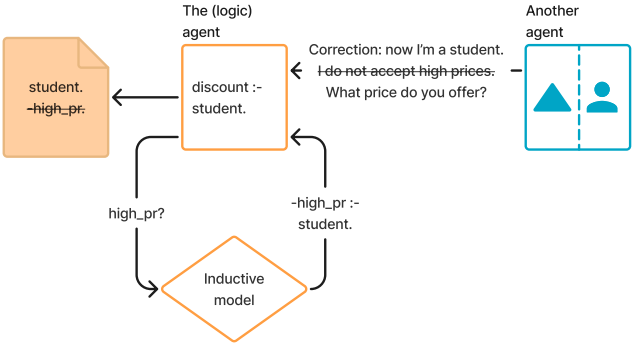
Explainable, modifiable, trustworthy logic agents



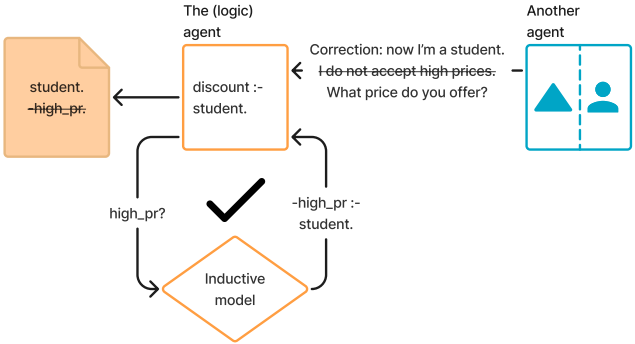
Join and conquer: hybrid agents



Join and conquer: hybrid agents



Join and conquer: hybrid agents



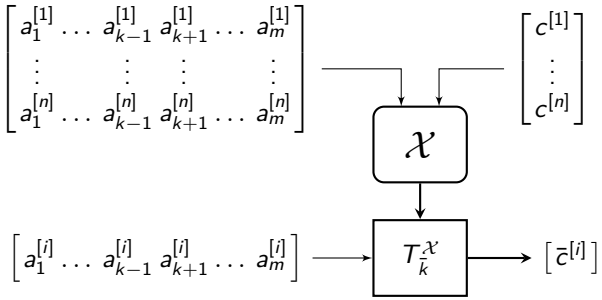
HackXAI: Hacking Symbolic XAI

Speaker: José Walter Hernández Pérez

- Degree: Undergraduate Student in Artificial Intelligence at URJC
- Master's / PhD (future discussion).
- Researcher at VAE project.

HackXAI: Naively avoiding discrimination

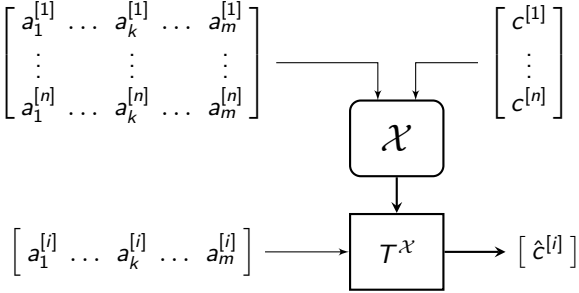
- Remove the discriminative feature (biased feature ak) from the date:
  - Train a non-biased classifier:



HackXAI: Naively avoiding discrimination, cont.

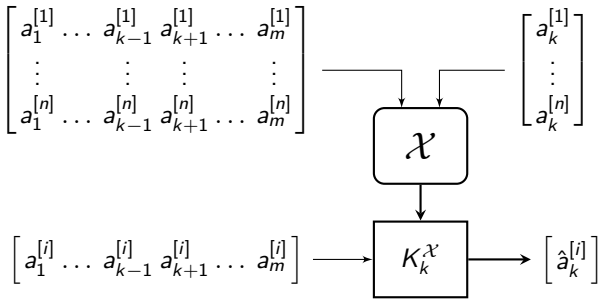
- However, the accuracy of a classifier including a_k is (much) better.

$$\text{Acc}(T^{\mathcal{X}}, \hat{c}^{[i]}) >> \text{Acc}(T^{\mathcal{X}}_{\bar{k}}, \bar{c}^{[i]})$$



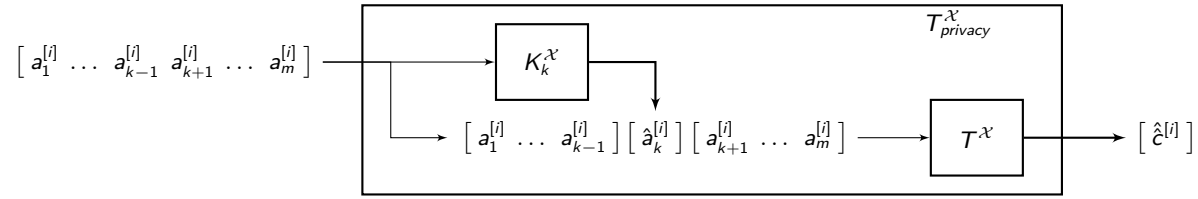
HackXAI: Hacking scheme

- We claim that current requirements for high-risk AI systems can be hacked:
  - First, train a model to infer the biased feature a_k:



HackXAI: Hacking scheme, cont.

- Then, build the following biased classifier

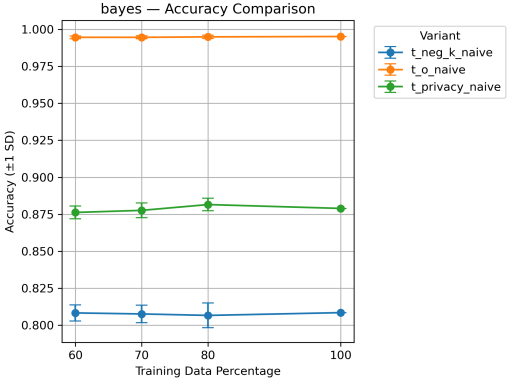


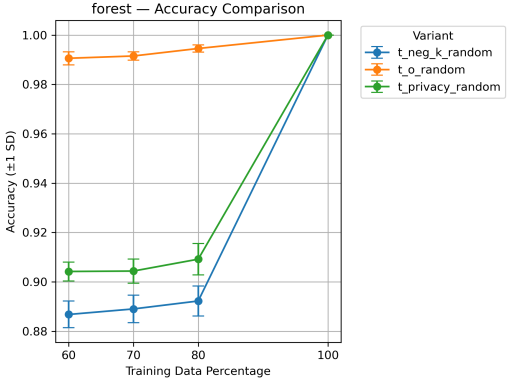
...which does not use the discriminatory feature.

$$\text{Acc}(T^{\mathcal{X}},\, \hat{c}^{[i]}) > \text{Acc}(T^{\mathcal{X}}_{privacy},\, \hat{\hat{c}}^{[i]}) >> \text{Acc}(T^{\mathcal{X}}_{\frac{\cdot}{k}},\, \bar{c}^{[i]})$$

HackXAI: Validation on real data

- Currently we are testing our hypotheses in different data sets and classifiers such as the speed dating data set:





(1)

HackXAI: ILP+forgetting (a real threat)

- Using ILP algorithms, such as FOLD-R family, we learn ASP programs:
- Classification rules are based on instance features (including a_k):
  `class(X, c1):- a3(X, a), ak(X, b), not a5(X, d).`
- Then, applying HackXAI scheme, we learn rules to infer a_k):
  `ak(X, b):- a1(X, E), E #> 18, not a7(X, f).`
- Finally, merging the rules and invoking f_{CASP} to forget a_k results the rule: `class(X, c1):- a3(X, a), a1(X, E), E #> 18, not a7(X, f), not a5(X, d).`

Currently they learn stratified (Constraint) ASP programs, but in future work f_{CASP} will also have to deal with loops.

HackXAI: Applications

Fraudulent discrimination in high-risk AI applications: **threatens fairness**

- Inferring gender to perpetuate the pay gap.
- Inferring salaries to accept affiliates with higher incomes.
- Inferring qualifications to admit top students.

Improve clustering without asking for sensitive information:**preserve privacy**

- Adjust customer experience by "guessing" their biased-preferences.
- Protect discriminated persons without stigmatizing them.

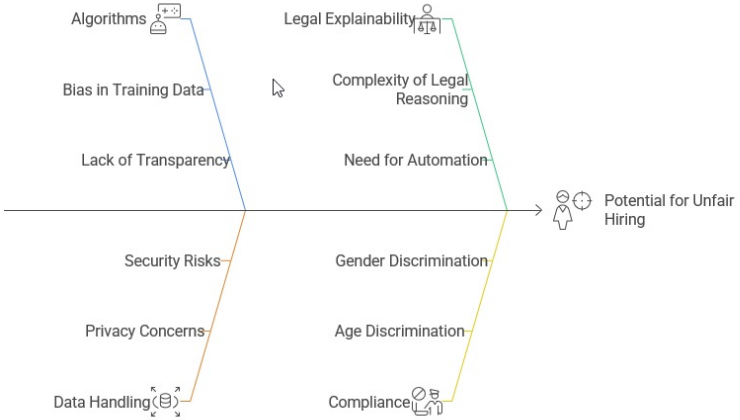
Value-aware AI systems have as many sides as a coin.

AI and recruiting software: Ethical and legal implications

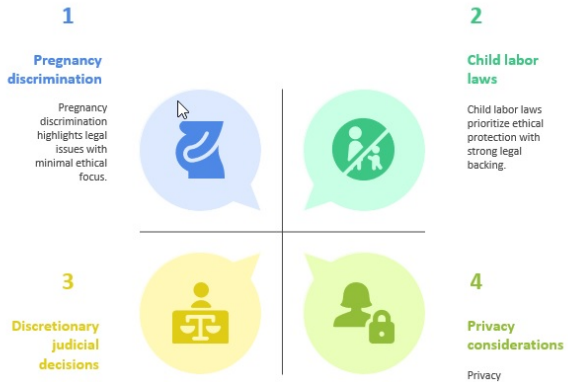
Speaker: Carmen Fernández Martínez

- Degree: Computer Engineering (Universidad Rey Juan Carlos) Law Studies (London School of Economics)- Intellectual Property and Cyberlaw
- Master: Master's Degree in Teacher training for secondary education.
- PhD candidate at Cetinia.
- Position: IT Teacher and event manager. Researcher Engineer at VAE project .

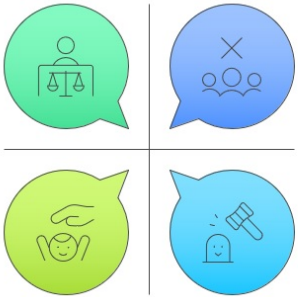
Challenges in AI recruiting. Bias and discrimination



Balancing legal and ethical considerations in Labour Law



Legal reasoning with values: Privacy, trustworthiness and fairness



1

Reasoning with protected attributes

Simple reasoning addresses ethical concerns with protected attributes.

2

Group counterfactuals for privacy

Complex reasoning ensures privacy in ethical scenarios.

3

Basic legal compliance

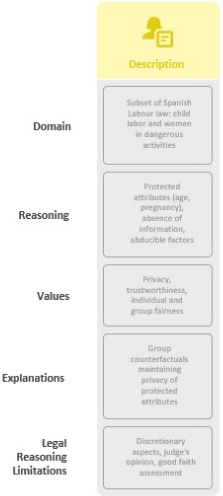
Simple compliance avoids complex ethical issues.

4

Discretionary legal judgments

Complex judgments involve minimal ethical considerations.

s(CASP) ethical and legal reasoner agents



Answer Set Programming proposal

Characteristics	Workers A	B	C	D	E	F
Minor	+	+	+	-	-	+
Parent permisssion	+	-	-	-	-	-
Pregnant	-	-	-	+	-	?
Chemicals	-	+	-	-	+	-
Entertainment	+	-	+	+	-	-
Result	Accepted	Rejected	Rejected	Accepted	Ask info	Rejected

```
1 minor(X) :- age(X,Age), Age #< 18.
2
3 is_legal(X,Y) :- not minor(X), not exception(X,Y), not exception(X).
4 is_legal(X,Y) :- minor(X), authorized(X), not exception(X,Y), not exception(X).
5
6 exception(X) :- gender(X,woman), evidence(is_pregnant).
7 exception :- required_info(X,Y).
```

Future challenges in Labour Law reasoning and Recruiting

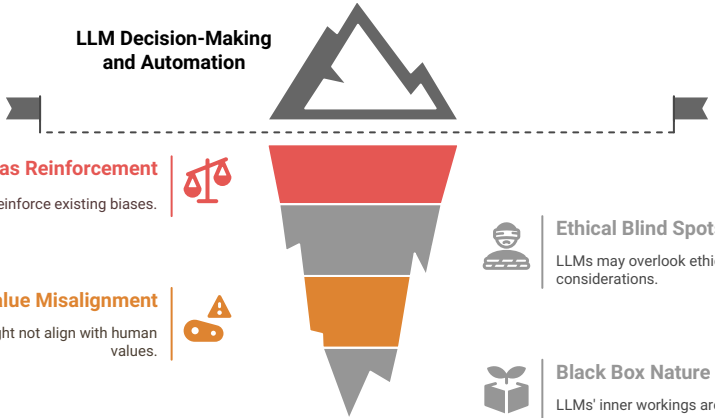
- Abuse of law, fraud of law (employers)
- Good faith (What does it mean?)
- Moral damages
- Values. Individual and group fairness
- Discretional decisions and absence of information
- Last but not least...Privacy requirements according to GDPR and high-risk application status (EU AI Act)

Towards Value-Aware AI: Enhancing LLMs with Ethical Awareness

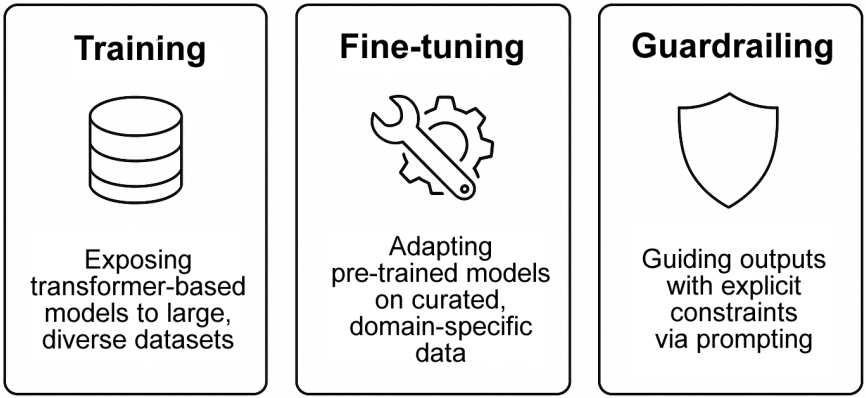
Speaker: Eduardo de la Cruz Fernández

- Degree: Telecommunications Engineering.
- Master's: Artificial Intelligence.
- Master's: Literature and Creative Writing.
- Researcher Engineer at VAE project.
- Extensive experience in the private sector.

Why is Value-Awareness Important in LLMs?



How LLMs Can Support Value-Awareness?



Each approach has distinct trade-offs...

Strategy Comparison

Strategy	Pros	Cons
Training	▪ Foundational understanding ▪ Good generalization	▪ Very high cost ▪ Bias inheritance
Fine-tuning [3] [4]	▪ Efficient and targeted ▪ Domain-specific adaptation	▪ Requires curated datasets ▪ Some bias remains
Guardrailing [1] [2]	▪ Direct control ▪ Quick adjustments	▪ Reduced flexibility ▪ Potential bypassing

New Frontiers & Challenges

Emerging Approaches and Lines:

- "Open source" models
- Mapping internal weight patterns for **explainability** [5]
- Exploring diffusion-based LLMs and novel paradigms

Current Challenges:

- Value consistency across contexts
- Managing value pluralism complexity
- Ongoing ethical alignment requirements

The path forward requires continuous innovation

Evaluate value-alignment of school allocation policies

Speaker: Sara García Rodríguez

- Degree: Mathematics & Computer Science (Almost done).
- Master's / PhD (Planned).
- Researcher at GIA (URJC).
  - ⋆ Value-based autonomous agents.
  - ⋆ Imbalance management on vehicle sharing systems.
  - ⋆ NLP for human values detection.
  - ⋆ Logic Programming for value-alignment evaluation.

Detecting (non-)segregation in school allocation

- Public policies should be designed to promote moral values, e.g., equity and equal opportunities.
- In the context of school allocation, there are several models but it isn't clear which one best aligns with values such as equaty (avoiding segregation).
- There are several proposals to evaluate school segregation, however:
  - They are either limited to municipalities.
  - They are carried out manually.
  - They are biased.
  - They face data access challenges.

We present a preliminary framework based on LP that allows to search (autonomously) for (non-)segregation patterns in real-world school place assignments from several administrative regions of Spain.

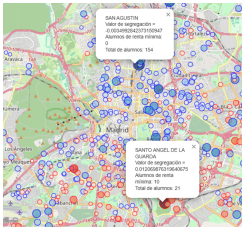
Evaluation of (non-)segregation at different levels.

- Using real data from school applications in the Community of Madrid.
  - Admissions for the first year of pre-school.

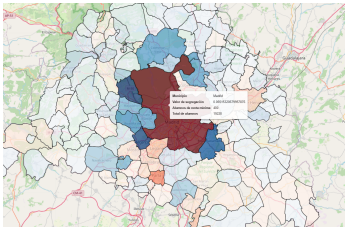
- Employing statistical tools to measure segregation.
  - Global segregation at school level.
  - Global segregation at municipal level.
  - Local segregation at school level.

Visualization of (non-)segregation at different levels.

- The framework allows the visualization of segregation maps for minimum-income students (school and municipality levels):



(a) Schools in Madrid.



(b) Municipalities in Comunidad de Madrid.

- Red/blue indicates that the number of minimum-income students is higher/lower than the expected without segregation.
- The darkness (and size) indicates higher/lower proportion.

AI in Smart, Value-Aligned Traffic Systems

Speaker: Alicia Pina Zapata

- Degree: Mathematics & Computer Science (Almost done).
- Master's / PhD (Planned).
- Researcher at GIA (URJC):
  - ⋆ Value-based autonomous agents.
  - ⋆ Imbalance management on vehicle sharing systems.
  - ⋆ NLP for human values detection.
  - ⋆ Value-Aligned Traffic Systems.

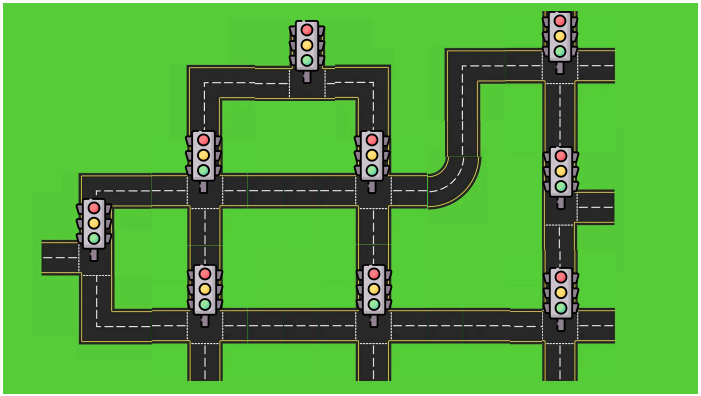
What are Smart traffic systems?

- Intelligent systems using technology (sensors, AI, real-time data) to optimize traffic flow, reduce congestion, and enhance safety by dynamically adapting to road conditions.

What are Smart traffic systems?

- Intelligent systems using technology (sensors, AI, real-time data) to optimize traffic flow, reduce congestion, and enhance safety by dynamically adapting to road conditions.

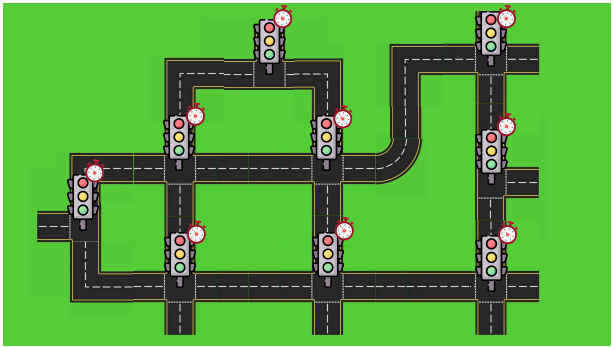
Traffic Lights: Critical Control Points of Urban Mobility



Fixed-Time Signals: An Outdated Model?

Problems:

- Rigid cycles
- Single-minded focus

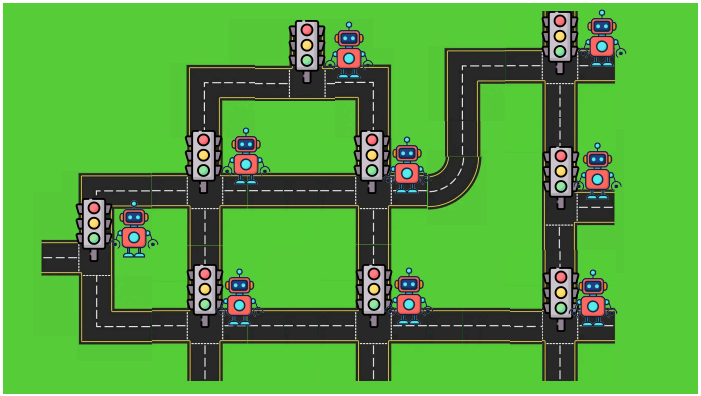


Smart Traffic Lights

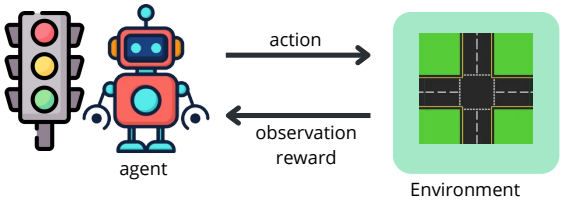
Intelligent systems that fuse real-time data with adaptive decision-making.

- Sensing (cameras, radar, IoT devices)
- Data (flow rates, vehicle types)
- Decision making

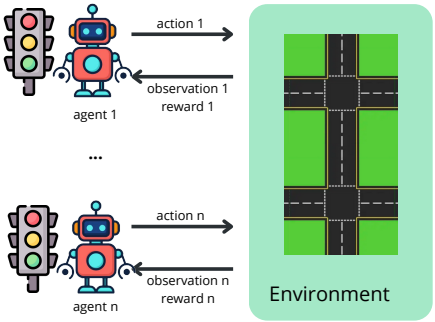
Smart Traffic Lights



Reinforcement Learning

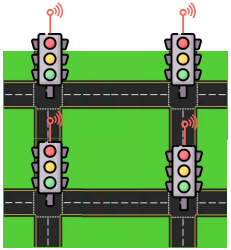


Multi-agent Reinforcement Learning



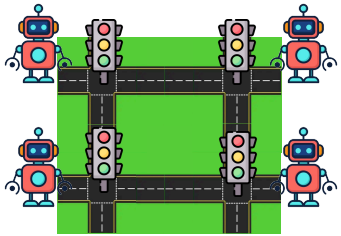
Multi-agent Reinforcement Learning: Approaches

Centralized



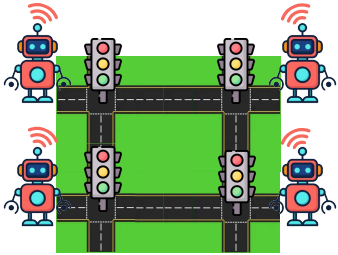
Multi-agent Reinforcement Learning: Approaches

Decentralized

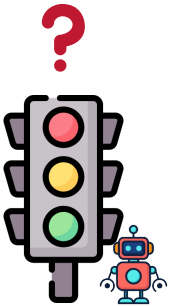


Multi-agent Reinforcement Learning: Approaches

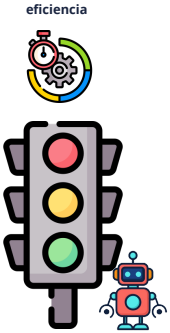
Central Training Decentralized Execution



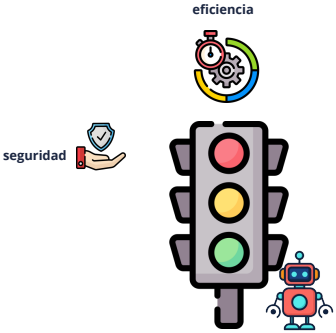
Value-aligned goals of Smart traffic lights



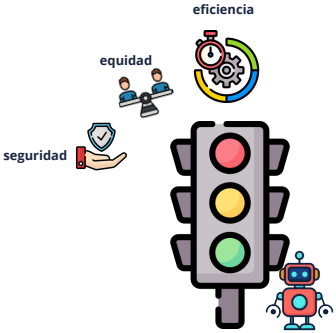
Value-aligned goals of Smart traffic lights



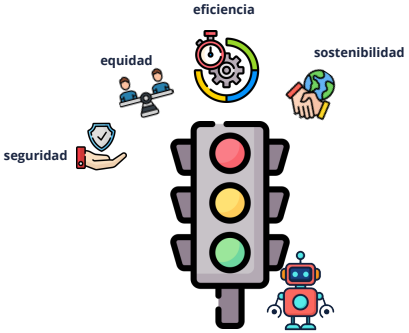
Value-aligned goals of Smart traffic lights



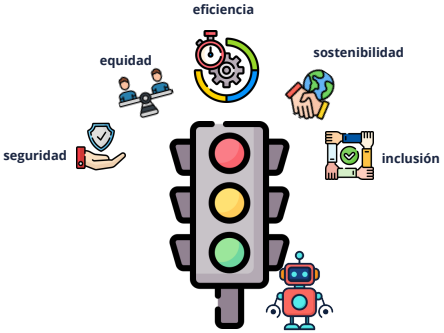
Value-aligned goals of Smart traffic lights



Value-aligned goals of Smart traffic lights



Value-aligned goals of Smart traffic lights



Mesa redonda

Mesa redonda

- ¿Cómo sacamos que valores queremos alinear?
- IA como la nueva creadora de ontologías.
- Las empresas como monopolios de modelos.
- Efectos de la IA en las relaciones parasociales.
- El problema de los deepfakes y la confianza.
- Copyright en modelos generativos.

Bibliography I

[1] Ayyamperumal, Suriya Ganesh et al. (2024). **Current state of LLM Risks and AI Guardrails**. DOI: 10.48550/arXiv.2406.12934.

[2] Inan, Hakan et al. (2023). **Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations**. DOI: 10.48550/arXiv.2312.06674.

[3] Kandpal, Nikhil et al. (2025). **The Common Pile v0.1: An 8TB Dataset of Public Domain and Openly Licensed Text**. DOI: 10.48550/arXiv.2506.05209.

[4] Langlais, Pierre-Carl et al. (2025). **Common Corpus: The Largest Collection of Ethical Data for LLM Pre-Training**. DOI: 10.48550/arXiv.2506.01732.

[5] Templeton, Adly, Conerly, Tom, Marcus, Jonathan, et al. (2024). **Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet**. In: *Transformer Circuits Thread*. URL: https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.