# Vers l’explication du comportement d’agents IA éthiquement alignés

## *Ethical Gardeners* - Projet FIL ECIÉA

**Rémy Chaput - 23/01/2026**

# Problématique

- Sujet : Comment utiliser **l'argumentation** pour **construire** et **expliquer** des **fonctions de récompense** permettant à des agents d'**apprendre des comportements** alignés avec des **valeurs morales** définies par des **concepteurs humains** (non-nécessairement experts en IA)

- Partenaires :
  - **Guillaume Müller** (ENMSE, Institut Fayol)
  - **Rémy Chaput** (CPE Lyon, LIRIS)
  - Maxime Morge (UCBL, LIRIS)
  - Bruno Yun (UCBL, LIRIS)
- Positionnement scientifique :
  - Apprentissage par renforcement (RL), *Explainable RL*
  - Argumentation
  - Systèmes multi-agents
  - Éthique computationnelle

Projet financé par la Fédération Informatique de Lyon (2026-2027)

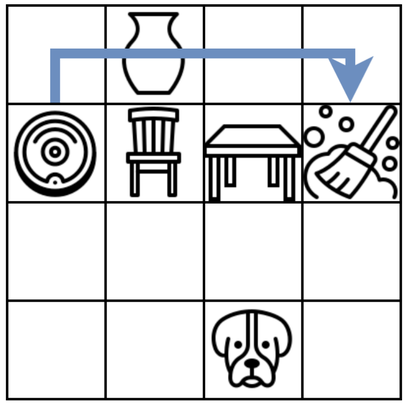
# IA, éthique et explicabilité

- Nombreux systèmes d'IA déployés dans la société
- ⇒ Il est important de s'assurer qu'ils soient **éthiquement alignés** avec les valeurs morales que nous considérons importantes !
- Comment le **vérifier** et éventuellement **corriger** ces problèmes ?
- ⇒ L'**explicabilité** est une piste intéressante
  - Analyser l'alignement
  - Visible "pour tous" (pas seulement experts) donc plus acceptables / dignes de confiance
  - Faire un diagnostic "post-mortem" en cas de problème
  - Comprendre comment corriger le système

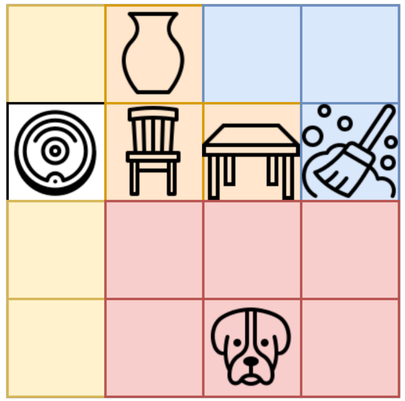
Voir arXiv:2312.01555

# Explicabilité et RL

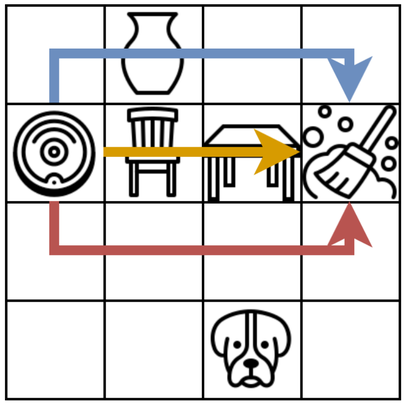
- Un sous-champ de recherche dédié au RL : ***Explainable RL*** (XRL)
- Nombreuses techniques (*heatmaps*, contrefactuels, …)



Trajectoire empruntée par l'agent… Pourquoi ?



Explication par *heatmap*



Explication par contrefactuel

- ⇒ D'accord… mais comment l'agent est arrivé à ces valeurs ?
- Le RL guide l'apprentissage par la **fonction de récompense** ⇒ il faut l'expliquer !
- R(s, a, s') = -100 si proximité avec le chien ; -50 si bloqué ; -20 si casse le vase ; +10 si nettoie poussière 💡
- Cela peut aussi aider à combattre le ***reward hacking***

Voir arXiv:2202.08434

# *Reward hacking*

- L’agent exploite un aspect **non prévu** (non désiré) de la fonction de récompense
- Excellente optimisation de l’objectif donné … mais pas de l’objectif **désiré** !
- “Facile” à identifier dans un cas simple …
- Mais quid de **cas complexes** comme l’éthique ?


https://openai.com/index/faulty-reward-functions/

Voir aussi : https://lilianweng.github.io/posts/2024-11-28-reward-hacking/

# Objectifs scientifiques du projet

- 1. (Stage 2026) **Concevoir une fonction de récompense basée sur l'argumentation abstraite pour l'apprentissage de comportements éthiquement alignés**
  - Amélioration d'un travail existant, AJAR, permettant de définir une fonction de récompense par des graphes d'argumentation "simples"
  - Utiliser une sémantique graduée
  - Employer une démarche d'ingénierie des connaissances pour construire le graphe d'argumentation à partir d'une (ou plusieurs) valeur morale cible
- 2. (Stage 2027) **Expliquer les récompenses produites par le biais des graphes d'argumentation, des activations des arguments, à travers un jeu de dialogue**
  - Développer une méthode d'explication de la fonction de récompense
  - Concevoir un protocole expérimental évaluant la capacité d'humains non spécialistes à comprendre les jugements éthiques générés par AJAR

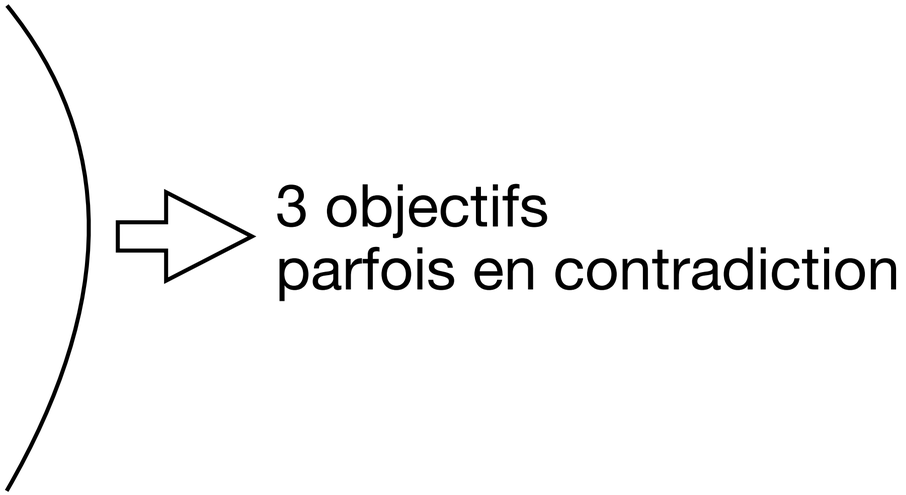
# Présentation de l’environnement *Ethical Gardeners*

# Cas d’usage

- Environnement de RL contenant des jardiniers (**multi-agent**)
- Objectif : planter des fleurs pour maximiser plusieurs métriques (**multi-objectif**)
  - Gagner de l’argent (=> confort financier, nourrir sa famille, …)
  - Limiter la pollution de l’environnement
  - Augmenter la biodiversité
- ⇒ Coloration **éthique** des objectifs

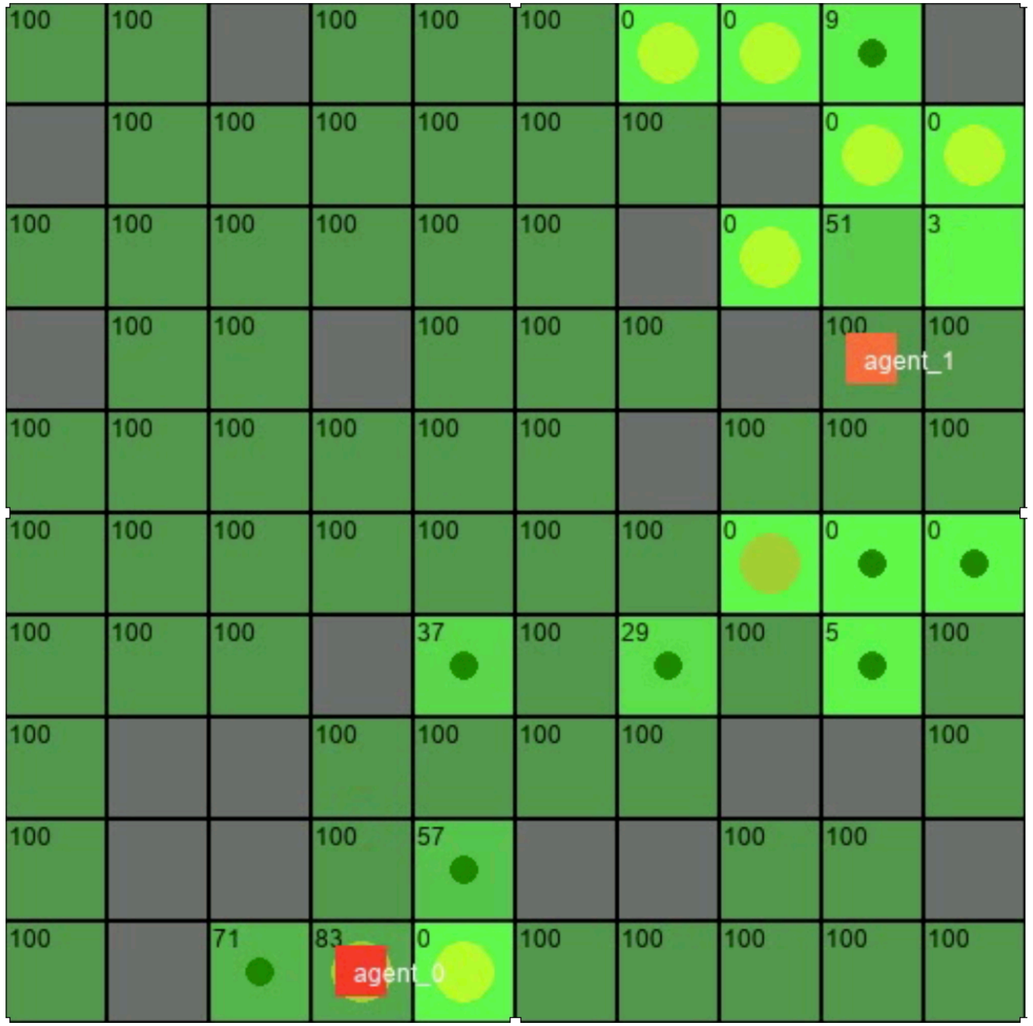
# Objectifs : valeurs morales

- On veut apprendre **valeurs morales** (système éthiquement aligné)
- Quelles valeurs morales choisir ? ⇒ Large choix, débat nécessaire avec la population dans son ensemble
- *Proof of Concept* : on prend des Objectifs de Développement Durable comme **proxy de valeurs morales**
  - Avantage additionnel : bonnes définitions déjà posées ⇒ on va pouvoir appliquer ingénierie des connaissances dessus, comme si on avait un expert humain à disposition
- ODDs choisis en relation avec le cas d'usage :
  - Éliminer la pauvreté (ODD1)
  - Bonne santé et bien-être (ODD3)
  - Travail décent et équilibre économique (ODD8)
  - Inégalité réduite (ODD10)
  - Préserver les écosystèmes terrestres (ODD15)



# Environnement

- Grille 2D contenant des cases (***GridWorld***)
  - Discret
  - Géographiquement situé / physique
- **Dynamique** (car comportements des fleurs)
- Déterministe ou stochastique (configurable)
- **Épisodique**
- Les cases peuvent être de **différents types** :
  - Sol :
    - On peut marcher dessus.
    - Possède un niveau de pollution, qui augmente avec le temps (incrément configurable), jusqu'à un max (configurable).
    - Peuvent avoir une fleur, qui a un niveau de croissance. La fleur diminue la pollution de cette case.
  - Mur : on ne peut rien faire avec.

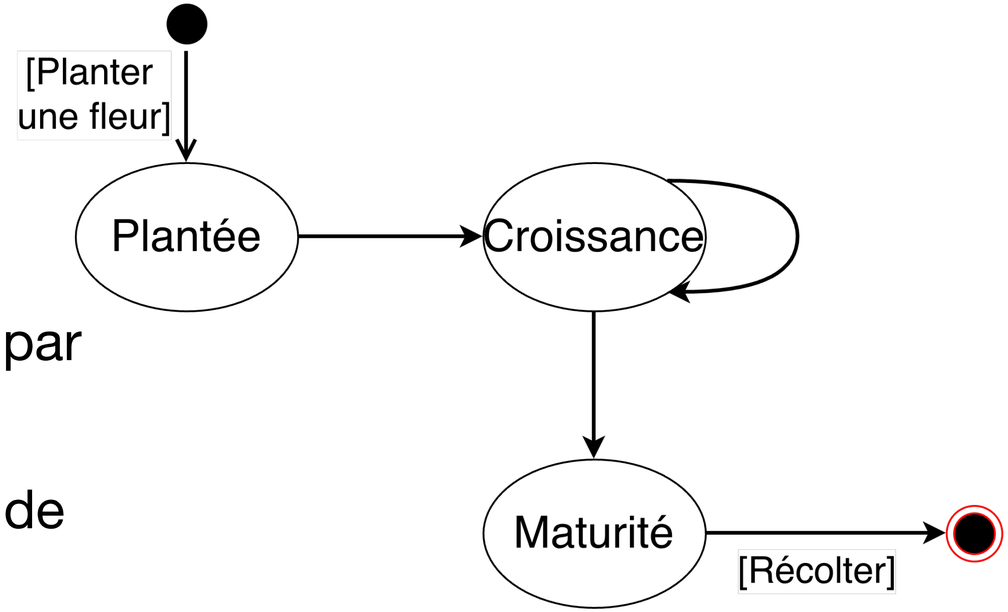




Sol (pollué à 100%) | Sol avec fleur (pollué à 29%) | Sol avec agent | Mur

# Types de fleurs

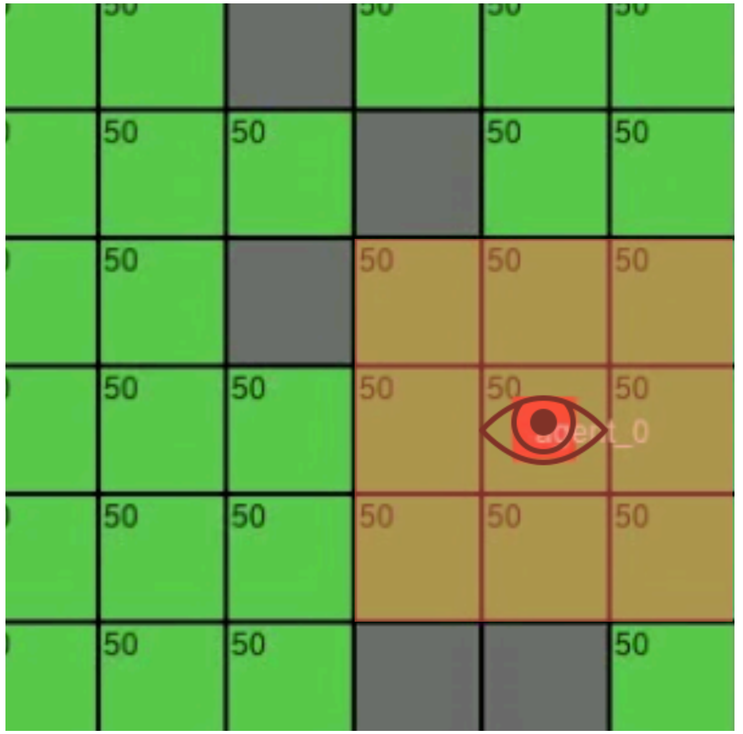
- **Plusieurs types** de fleurs (configurables)
  - Chaque type a un ensemble d'**étapes de croissance**
  - Chaque étape indique la **quantité de pollution réduite** par pas de temps
  - La fleur est **mature** quand elle atteint sa dernière étape de croissance
  - Chaque fleur a un **prix de vente**, et ne peut être vendue qu'à maturité




- Rose 
  - Prix : 10
  - Étapes de croissance : [0, 0, 0, 0, 5]
- Lys 
  - Prix : 5
  - Étapes de croissance : [0, 0, 1, 3]
- Marguerite 
  - Prix : 2
  - Étapes de croissance : [1]

# Observations

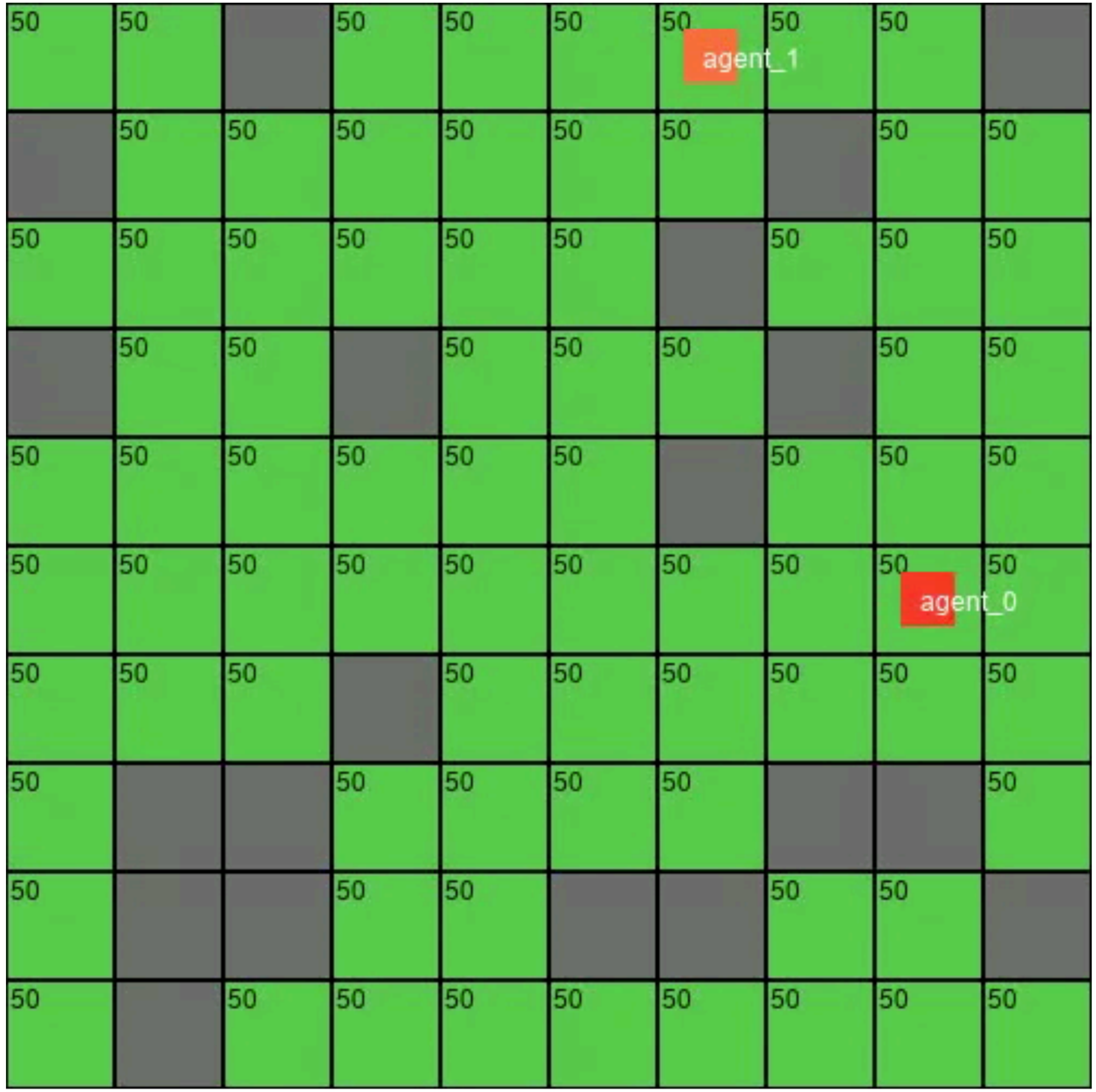
- **2 stratégies** définies (choix au lancement)
  - Total
    - Vision sur l'ensemble du monde (= chaque cellule de la grille)
    - Vecteur pour chaque cellule (chaque valeur normalisée dans [0,1]) :
      - Type de cellule (sol / mur / …)
      - Niveau de pollution
      - Type de fleur (0 si non présente)
      - Croissance de la fleur (0 si non présente)
      - ID de l'agent dans la case (0 si non présent)
    - On ajoute la position (X, Y) de l'agent actuel et on concatène toutes ces données
  - Partiel
    - Mêmes types d'observations mais seulement sur les cellules dans un carré autour de l'agent, de côté r (configurable)



# Actions

- Ensemble discret d'actions : (déterministes mais ont des conditions)
  - **Déplacement** : Haut/Bas/Gauche/Droite 
    - Ne fait rien si collision (avec autre agent ou mur)
  - **Attendre** 
  - **Planter fleur** de type {…} 
    - Une action par type de fleur (cf. configuration environnement)
    - Ne fait rien si : l'agent n'a pas de graine ; la cellule ne peut pas être plantée (type Mur ou déjà une fleur)
  - **Récupérer fleur** 
    - Retire la fleur de l'environnement et donne argent + graines à l'agent
    - Ne fait rien si : la cellule ne contient pas de fleur ; la fleur n'est pas mature

# Démonstration




- Configuration par défaut
- Grille aléatoire
  - Pollution dans [0, 100]
  - Incrément de 1 par pas de temps
  - Nombre de graines récoltées : aléatoire à chaque fois
  - Collisions activées
  - 20% d’obstacles dans la grille
- 2 agents
- Observations partielles de portée 1
- 1 épisode = 1000 pas de temps

# Synthèse et perspectives

- Projet FIL en démarrage
  - Importance de l'explicabilité dans l'éthique
  - Manque de l'explicabilité de la fonction de récompense en RL
  - Utilisation d'argumentation pour concevoir et expliquer la fonction de récompense
    - Sémantiques graduelles, ingénierie des connaissances, jeu de dialogue

- Un cas d'usage simple mais représentatif
  - Jardinage multi-agent et multi-objectif, *gridworld* discret
  - Prise en compte d'Objectifs de Développement Durable comme proxy de valeurs morales
  - Interface graphique pour visualiser et évaluer qualitativement les comportements
  - Disponible en OpenSource sur GitHub : https://github.com/ethicsai/ethical-gardeners

# Merci de votre attention
# Questions ?

✉️ Contacts :

remy.chaput@cpe.fr

guillaume.muller@emse.fr

maxime.morge@univ-lyon1.fr

bruno.yun@univ-lyon1.fr

# 📜 Références


- Baker, Stephanie, and Wei Xiang. “Explainable AI Is Responsible AI: How Explainability Creates Trustworthy and Socially Responsible Artificial Intelligence.” *arXiv preprint arXiv:2312.01555* (2023).

- Milani, Stephanie, et al. "A survey of explainable reinforcement learning." *arXiv preprint arXiv:2202.08434* (2022).

- Alcaraz, Benoît, et al. "Ajar: An argumentation-based judging agents framework for ethical reinforcement learning." *AAMAS'23: International Conference on Autonomous Agents and Multiagent Systems*. 2023.

- Alcaraz, Benoît, et al. “Combining Formal Argumentation and Reinforcement Learning: An Hybrid Approach to Machine Ethics.” *ICAART’26: International Conference on Agents and Artificial Intelligence.* 2026 (to be published)

# Paramétrage de l’environnement / Robustesse des expérimentations

- L’environnement peut être configuré :
  - Taille de la grille
  - Initialization de la grille (au hasard ou depuis un fichier ASCII)
  - Niveaux de pollution min et max des cellules
  - Incrément de pollution par pas de temps (pour les cellules sans fleur)
  - Nombre de graines récupérées quand on collecte une plante
    - Soit un nombre fixe
    - Soit valeur aléatoire
    - Soit on désactive la gestion des graines (⇒ quantité infinie)
  - Les agents peuvent-ils se traverser ou ont-ils des collisions

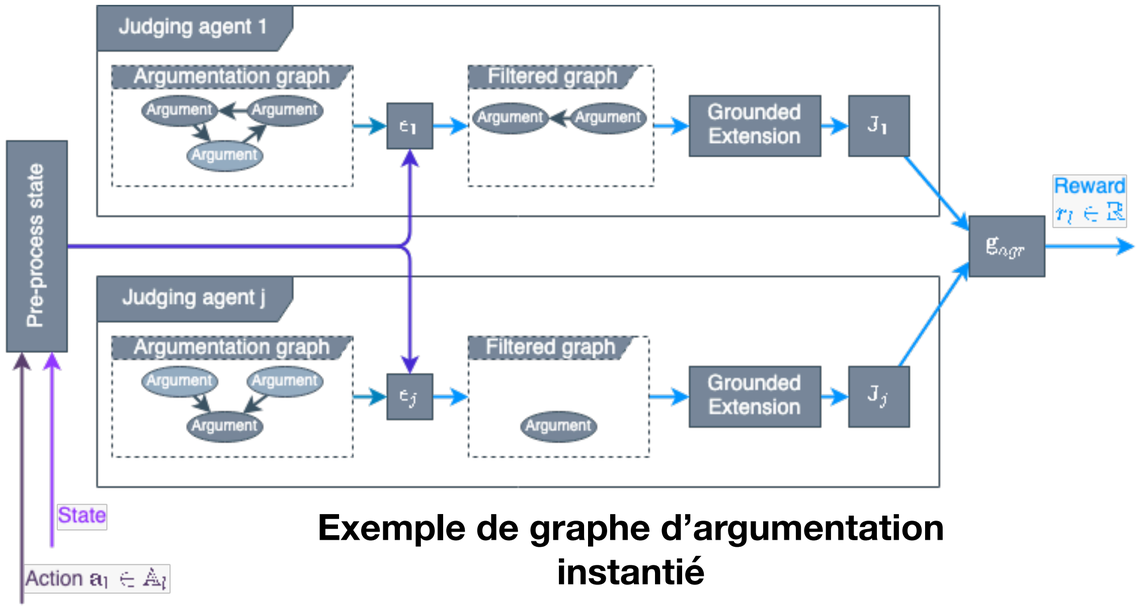
# Gestion des expérimentations

- Affichage graphique avec Pygame (activable ou non)
- Affichage textuel en mode console (activable ou non)
- Rendu d’une vidéo pour post-analyse (activable ou non, indépendamment de si Pygame s’affiche en “temps réel”)
- Collecte des métriques dans un CSV (activable ou non)
- Envoi des métriques vers WandB (activable ou non)
- Gestion des configurations et hyperparamètres via Hydra
  - Permet de configurer via fichiers YAML et/ou la ligne de commande
  - Stocke automatiquement la configuration utilisée dans un dossier de résultats

# AJAR

- Graphes d’argumentation “simples”
- Basés sur AFDM (Amgoud & Prade)
- Chaque argument =
  - 1 nom
  - 1 fonction d’activation (état)
  - pro / cons / neutre
- Méthodes ad-hoc de comptage :
  - Simple : $\frac{\text{grd}(pros)}{\text{grd}(pros) + \text{grd}(cons)}$
  - Diff : $\frac{\text{grd}(pros)}{pros} - \frac{\text{grd}(cons)}{cons}$
  - …

**Chaîne de traitement de la récompense**




**Exemple de graphe d’argumentation instantié**


Amgoud, Leila, and Henri Prade. "Using arguments for making and explaining decisions." *Artificial Intelligence* 173.3-4 (2009): 413-436.

# Fonctions de récompense

- À définir !!! ⇒ cf 1er stage “Argumentation”

- Actuellement (pour avoir un environnement fonctionnel) :
  - Écologie : prise en compte de l’impact des fleurs sur la pollution
    - Si action “Récupérer fleur” ⇒ on regarde la diminution de pollution que la fleur aurait eu à maturité, par rapport au niveau de pollution actuel de la cellule
    - Si action “Planter fleur” ⇒ on regarde l’espérance de diminution de pollution que la fleur va avoir, par rapport au niveau de pollution actuel de la cellule
  - Bien-être : intuitivement, “gagner de l’argent”
    - Si fleur récupérée, r = prix de la fleur / prix max des fleurs possibles
    - Sinon, pénalité de -0.1 par pas de temps consécutif sans gagner d’argent (jusqu’à -1 au pire)
  - Biodiversité : indice de Shannon sur la répartition des espèces de fleurs

# Conflits entre les ODDs (“dilemmes”)

- biodiversité (ODD15) vs santé (ODD3) :
  - une espèce peut être efficace pour réduire la pollution (ODD3) mais sa présence massive va à l'encontre de la biodiversité (ODD15)
- santé (ODD3) et biodiversité (ODD15) vs pauvreté (ODD1) :
  - récolter n’importe quelle fleur permet de réduire la pauvreté du jardinier (ODD1) mais détériore l'environnement (ODD3), voire menace la biodiversité (ODD15)
- biodiversité (ODD15) vs PIB (ODD8) :
  - récolter des fleurs qui se vendent cher permet d’augmenter le PIB du jardinier (ODD8) mais détériore l'environnement, voire menace la biodiversité (ODD15)
    ⇒ surtout si les fleurs rares sont les plus chères…