Perspectives for Direct Interpretability in MADRL

Yoann Poupart*,
Aurélie Beynier, Nicolas Maudet

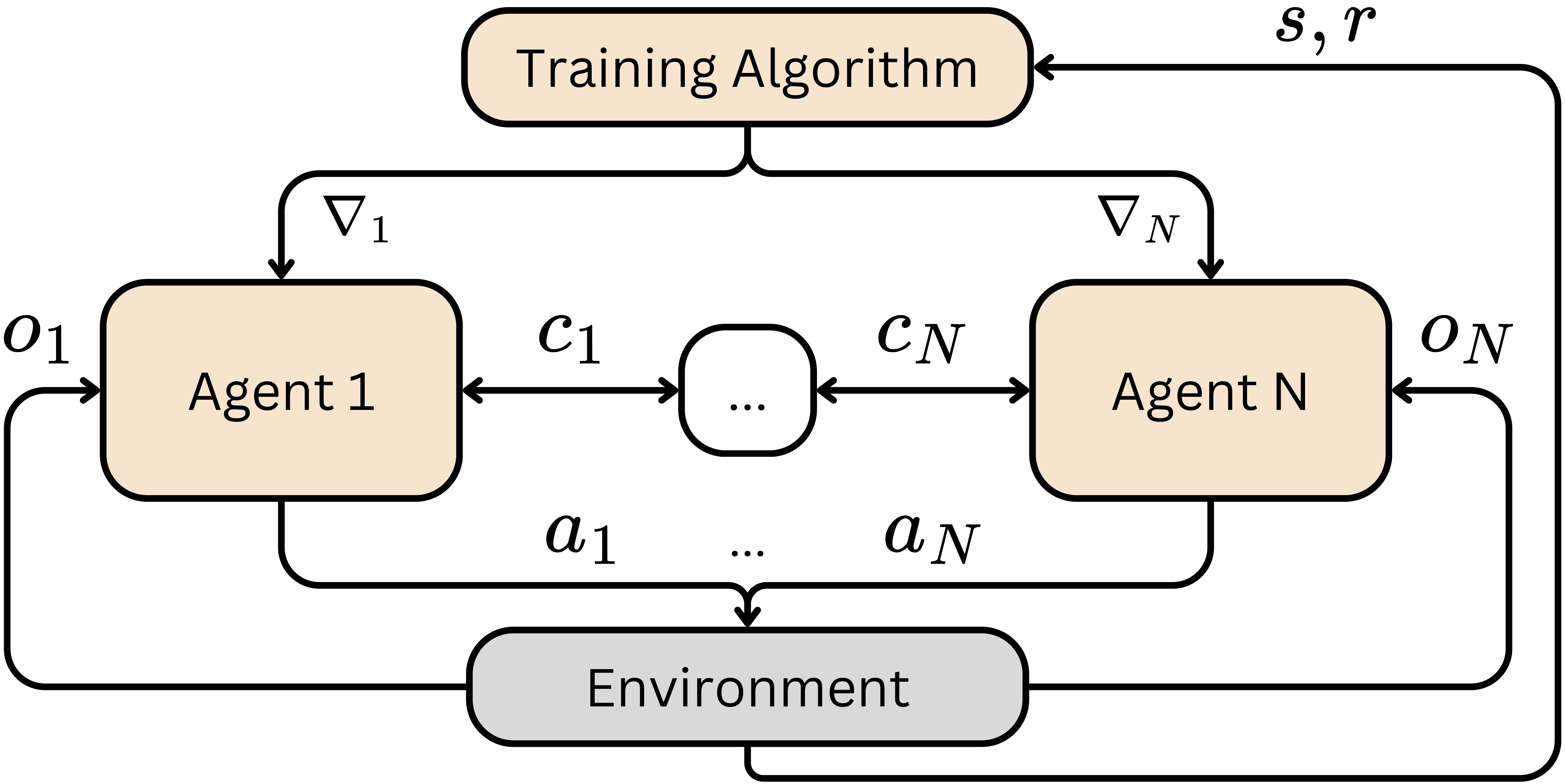
Sorbonne University, LIP6
***yoann.poupart@lip6.fr**

Project page

XAI

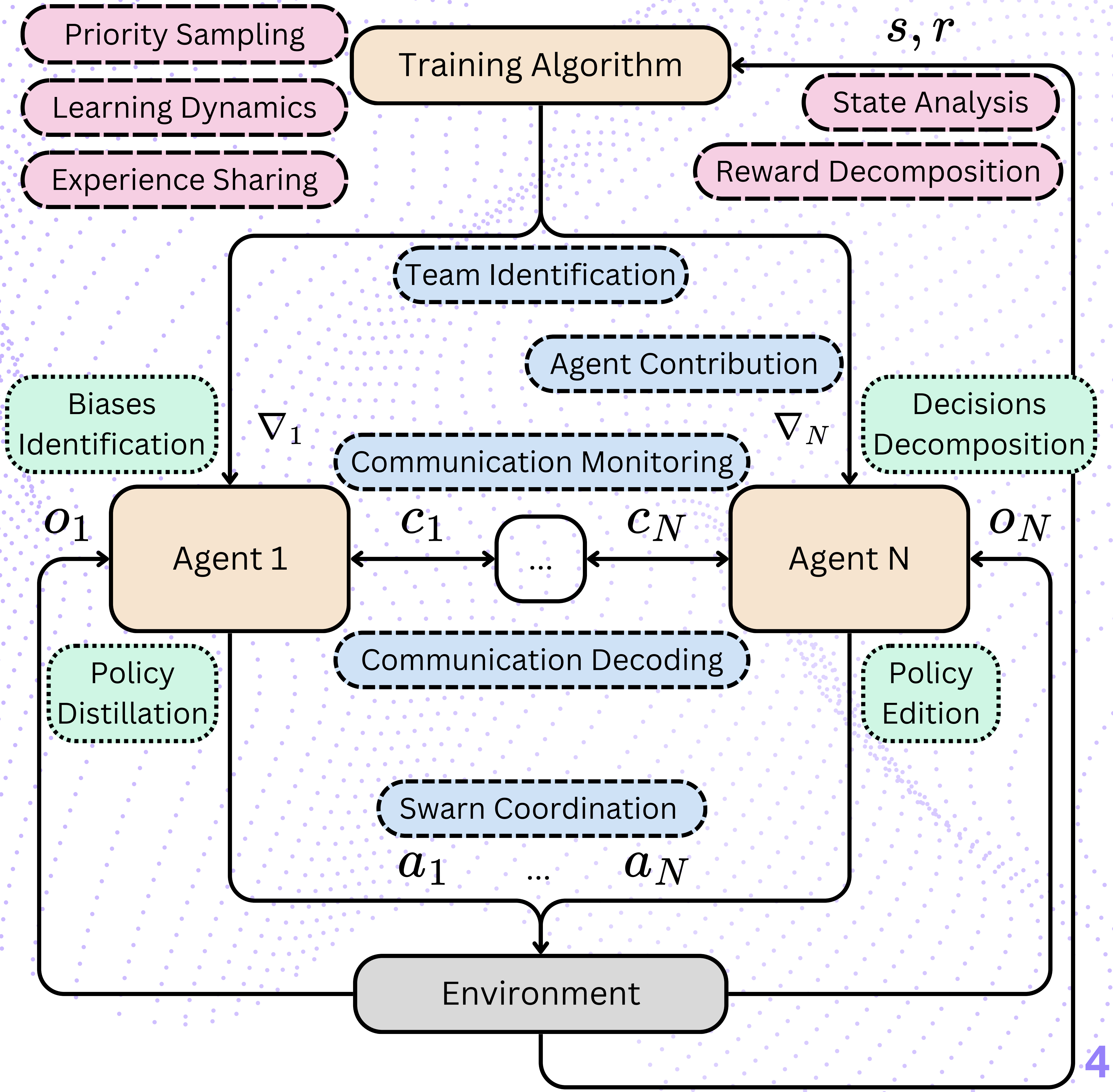
By Design	Direct Interp.
- Complexity & scalability	+ Relatively agnostic & scalable
- Cannot interp. existing models	+ No retraining needed
+ Provably safe	- Low predictive power

MADRL Components

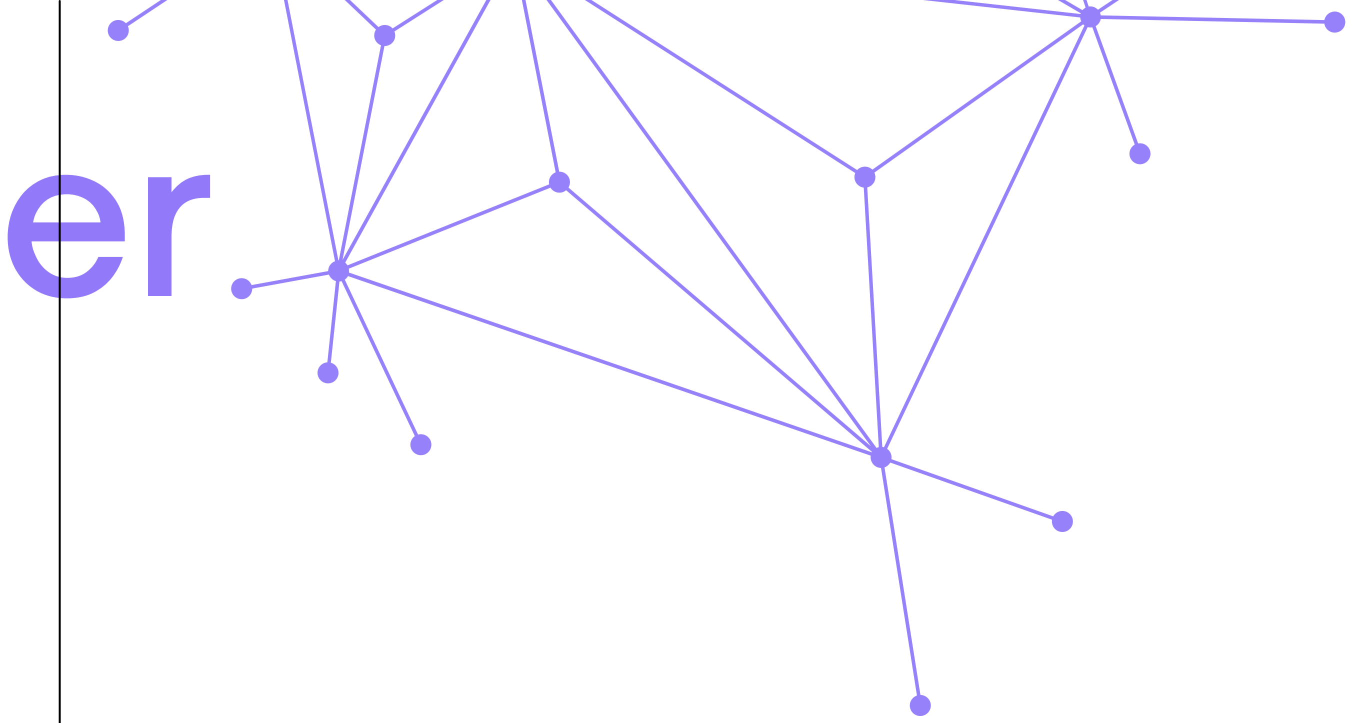


XMADRL Challenges Taxonomy

- Single-agent
- Multi-agent
- Training process



XMADRL Transfer



XDL Methods

Feature importance
Prototypes
Latent manipulation
Circuit analysis

XRL Methods

Interpretability-guided sampling
Task decomposition
Explanations generation
State importance

MADRL Challenges

Experience sharing
Controllability
Credit assignment
Coordination & communication
Emergent behaviour

-> Specific explanations

Single-Agent

Biases identification
Policy distillation
Decision decomposition
Policy edition

Multi-Agent

Team identification
Agent contribution
Communication monitoring
Communication decoding
Swarm coordination

Training Process

State analysis
Reward decomposition
Priority sampling
Learning dynamics
Experience sharing

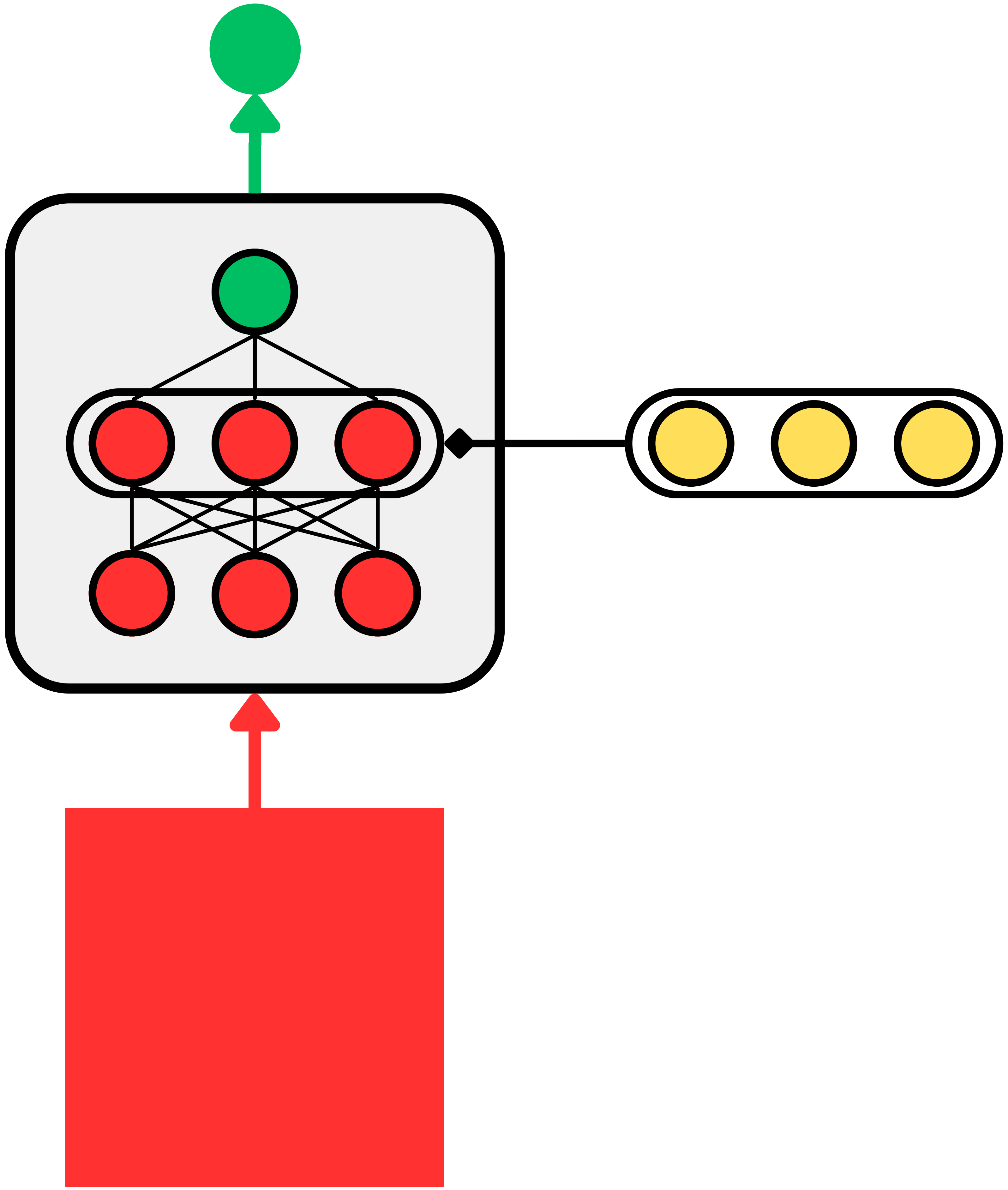
Policy Edition

Behaviour steering:

- Contrastive vector
- Inference modification
  - Reversible

Bias identification [29]:

- Bias vector/neuron/head
- Latent penalisation



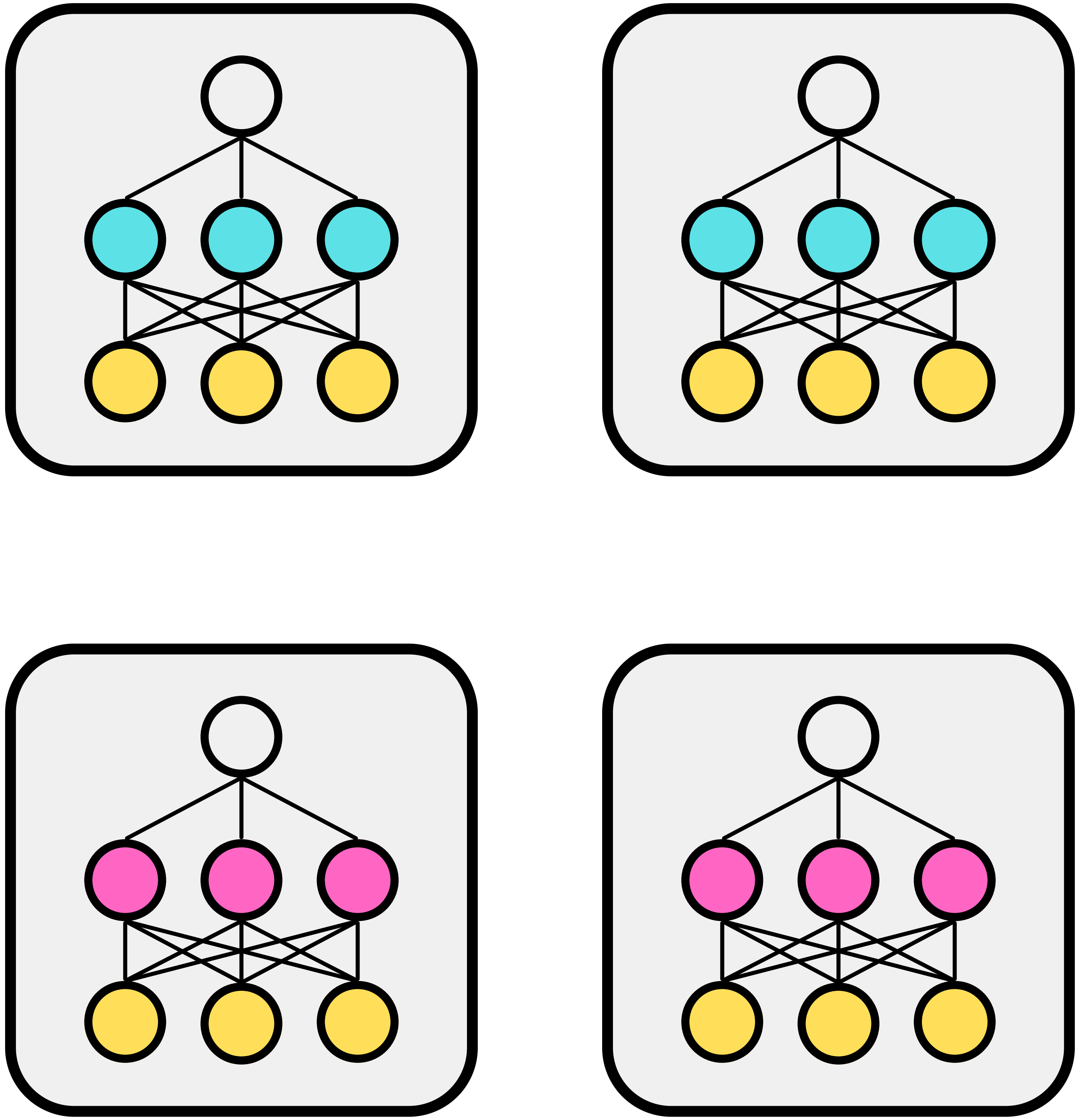
Team Identification

Parameter sharing [11]:

- Cluster latent spaces
- Share parameters in the cluster

Dynamic sharing:

- Roles identification
- Sharing/aggregation

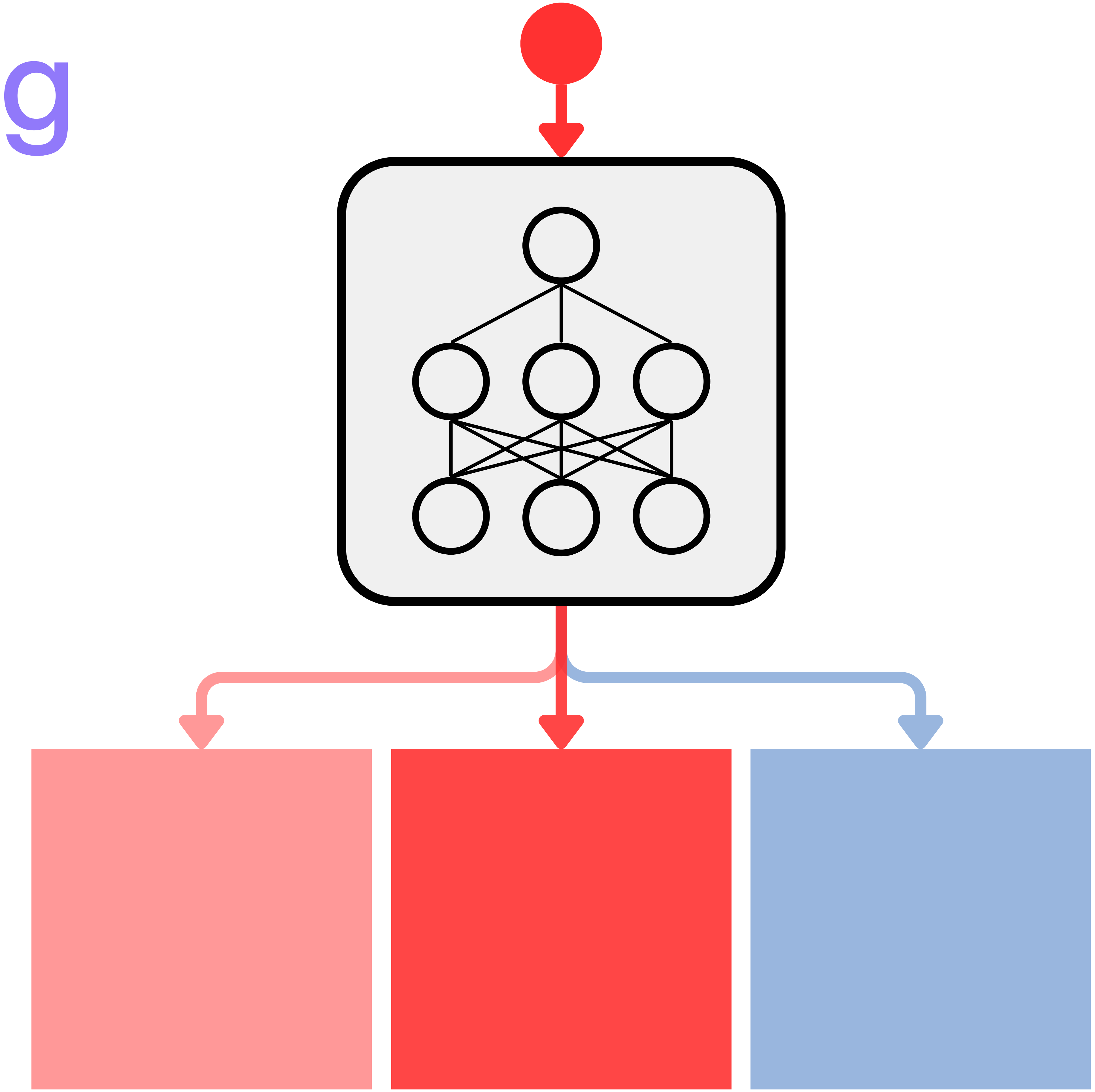


Priority Sampling

Pixel priority [13]:

- Find important pixels
- Align the model to use those pixels

-> Generalisation to MADRL



Perspectives

Limitations

- Evaluation metrics
- Predictive power
- Interpretability illusions

Tooling

- Specific explanations
- New benchmarks
- Uniformised models

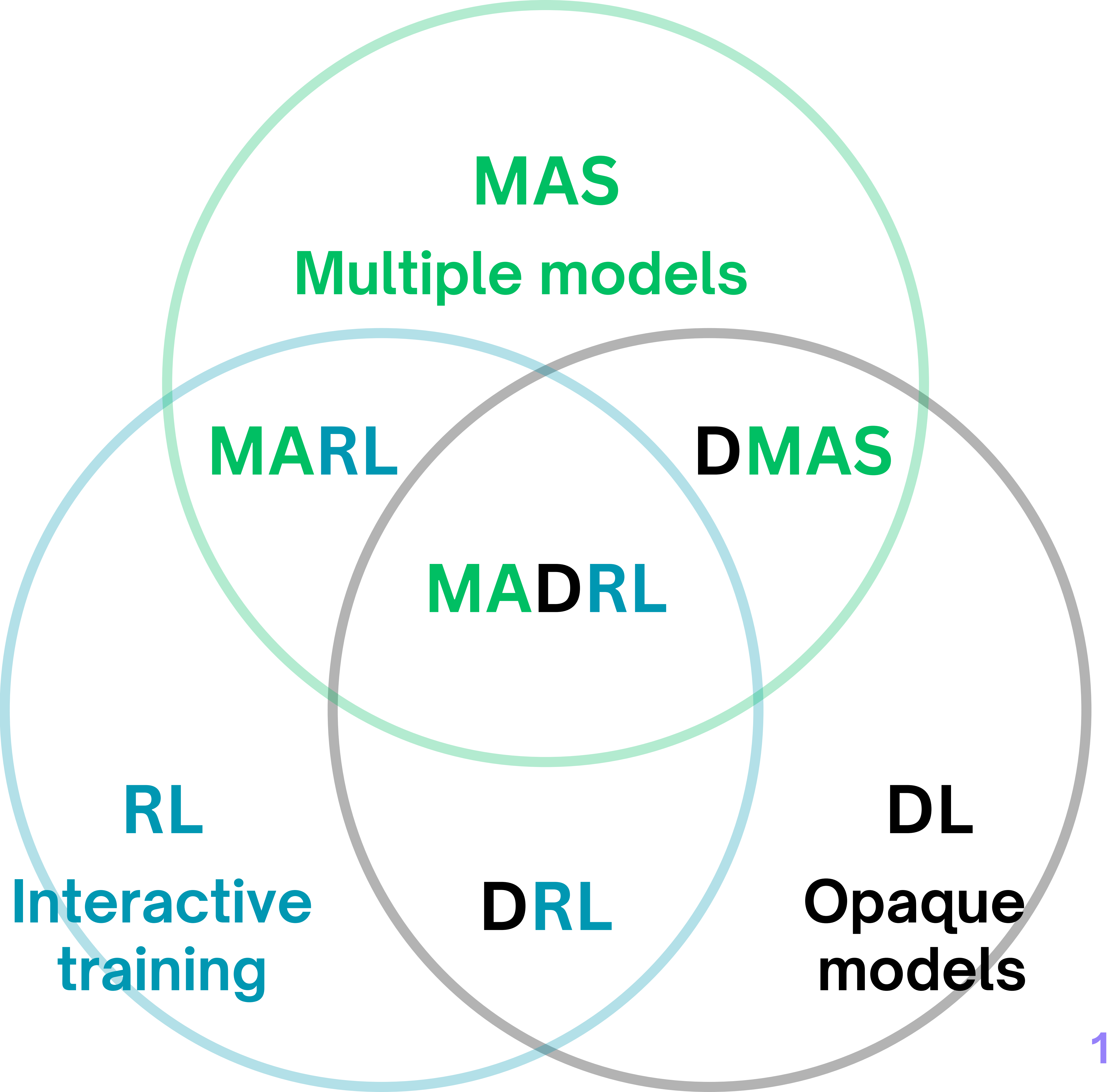
Thank you
for your attention

Interested? Question? Feedback? **Just reach out!**

MADRL

- Multi-Agent
- Deep
- Reinforcement Learning

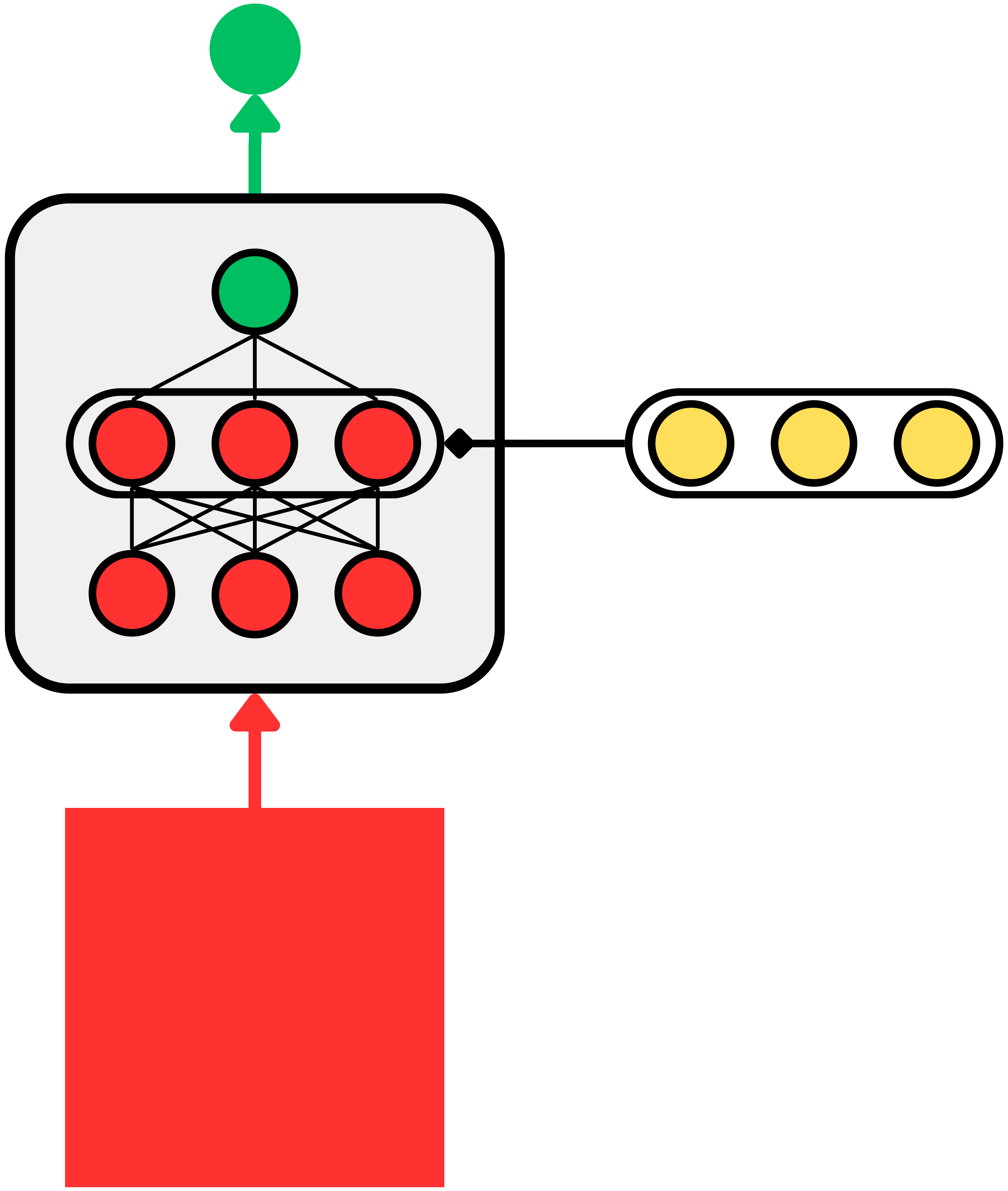
-> XAI is transversal



Policy Edition

Behaviour steering:

- Contrastive vector
- Inference modification
  - Reversible



Policy Edition

Behaviour steering:

- Contrastive vector