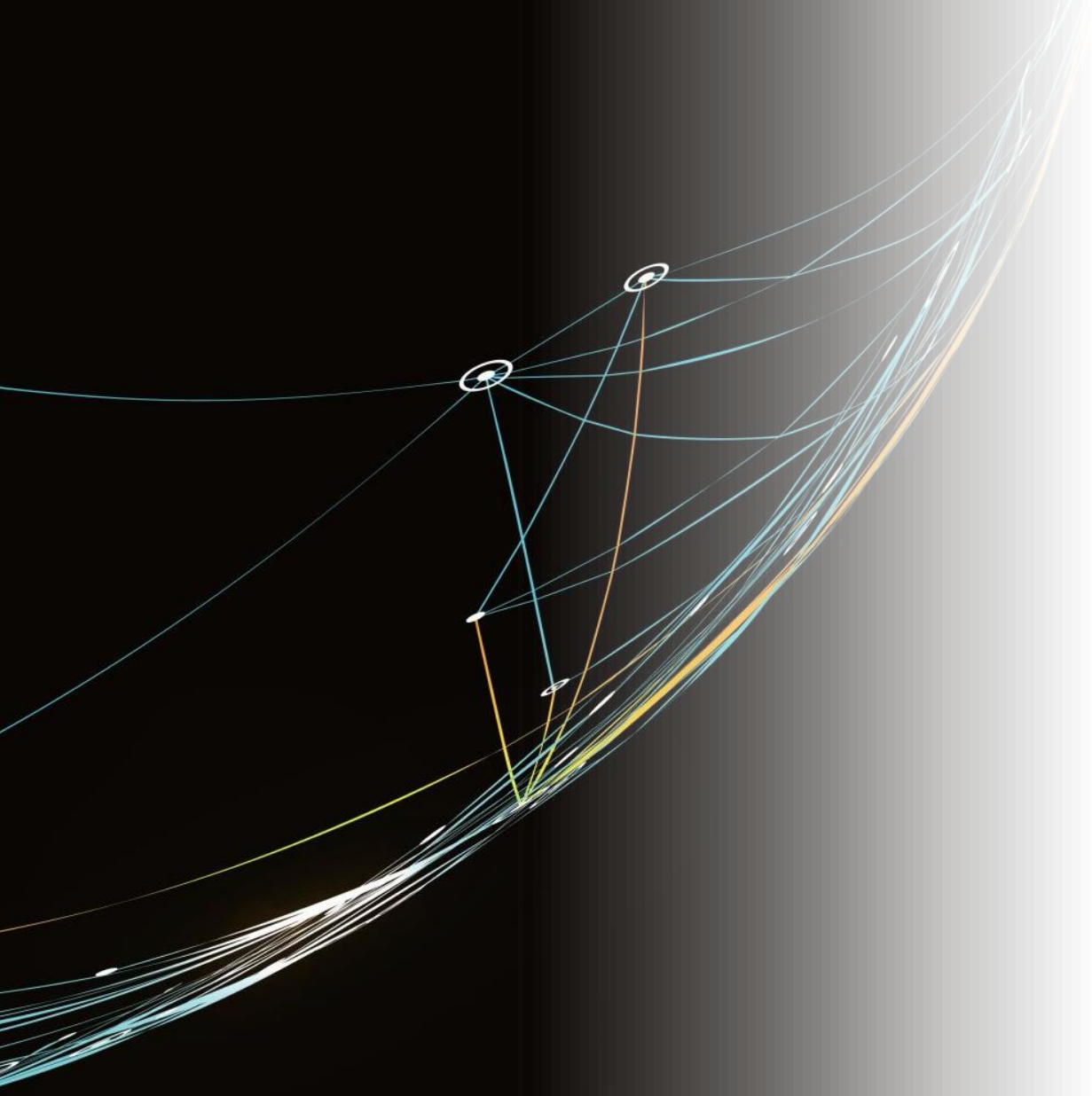


A Multi-Agent Reinforcement Learning Model for the Pension Ecosystem

Fatih Ozhamaratli
Dr. Paolo Barucca

University College London

Motivation

So what is the story?

- Modelling financial systems in a way that captures non-stationary dynamics
- Creating virtual doubles/potential alternatives of economies without relying on granular microdata
- Shift from DB to DC, Diversifying Income Trajectories and new modes of work, Frequent Crisis caused by so called Black Swan events
- Investigating a unified framework for implementing simulation and model training of financial systems in a computationally efficient and implemention-wise not complex methodology

Traditional Approach to Pension Savings

- (Merton, 1971).
- Traditional econometric models found to be restrictive due to certain assumptions.
- Models that consider factors such as labour income fluctuations, asset return fluctuations, and central bank decisions (Campanale et al., 2015; Campbell & Viceira, 2002; Cocco et al., 2005).

The Complexity of Pension Savings

- Long-term strategic investment to evade financial crises and black swan events, as evident from 2008 financial crisis (Impavido & Tower, 2009) and 2022 pensions leveraged gilt crisis (England, n.d.).
- Norwegian Sovereign Wealth fund's counter-cyclical investment strategy (Papaioannou & Rentsendorj, 2015).
- Labour income is counter-cyclically impacted by business cycle effects (Guvenen et al., 2012).
- The cascading effects of investment decisions and supply chain shocks lead to non-stationary market dynamics, contravening the assumptions of traditional pension models (Cont & Wagalath, 2016; Pichler & Farmer, 2022).

The Promise of Agent-Based Models

- Agent-Based Models (ABMs) using Multi-Agent Reinforcement Learning (MARL) can overcome limitations of traditional econometric models.
- ABMs respond to paradigm shifts and black swan events and accommodate heterogeneous profiles among pension investors (Asano et al., 2019).
- ABMs can generate synthetic and heterogeneous income trajectories for creating inclusive savings strategies.

Implementation Challenges of MARL

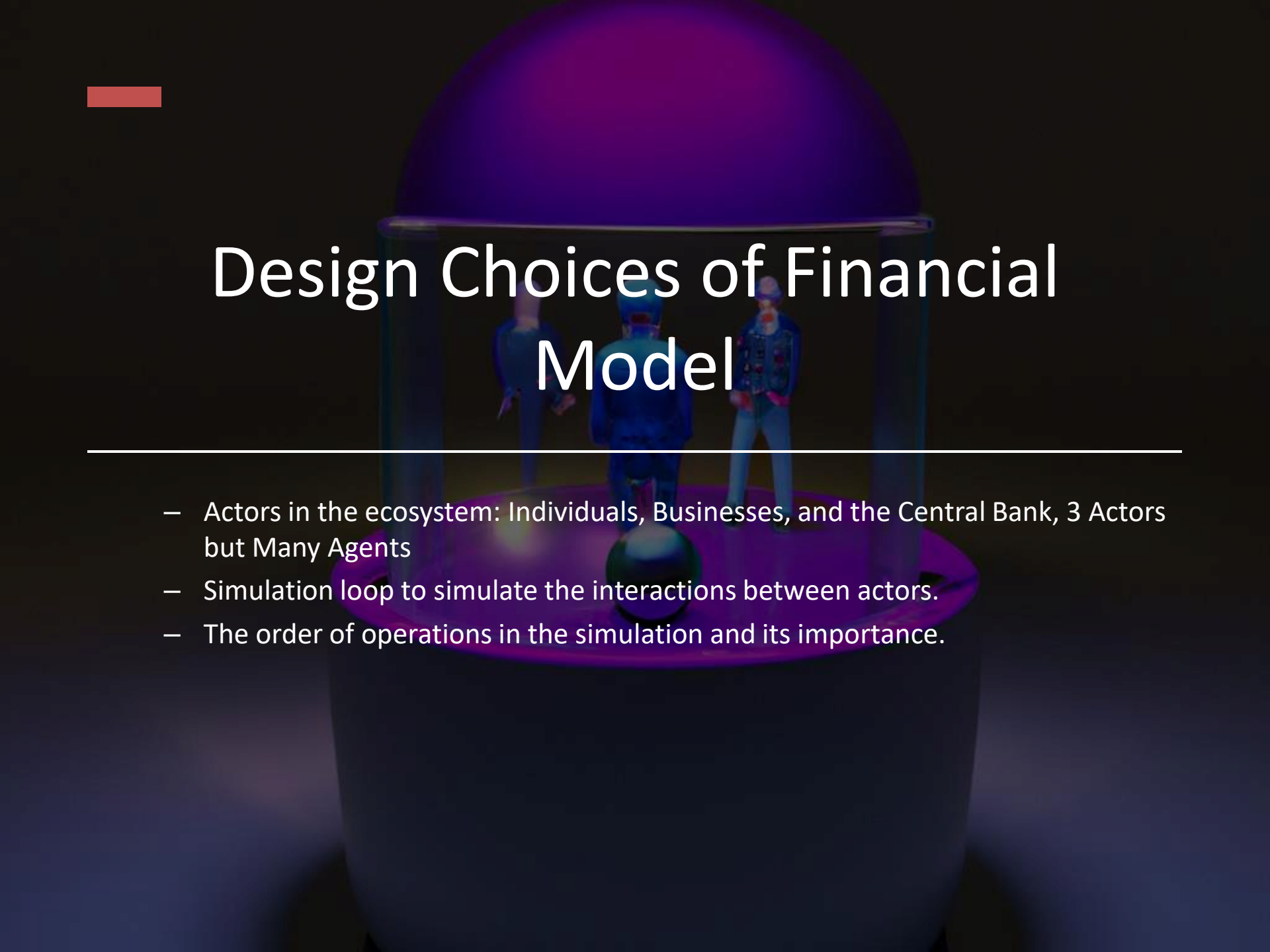
- MARL challenges: non-stationarity of the environment, combinatorial complexity (Samvelyan et al., 2019; Yang & Wang, 2021).
- Recent advancements in software architectures and reinforcement learning algorithms are promising solutions (Frostig et al., 2018; Harris et al., 2020; Yu et al., 2022).

Opportunities with Recent Advancements

- Bridge between the gap between mathematical formulations and GPU-accelerated Just-in-Time (JIT) executed codes (Frostig et al., 2018), using APIs similar to widely used NumPy API (Harris et al., 2020).
- Easy formulation of mathematical expressions without the need for factoring code for batches or distributed execution.

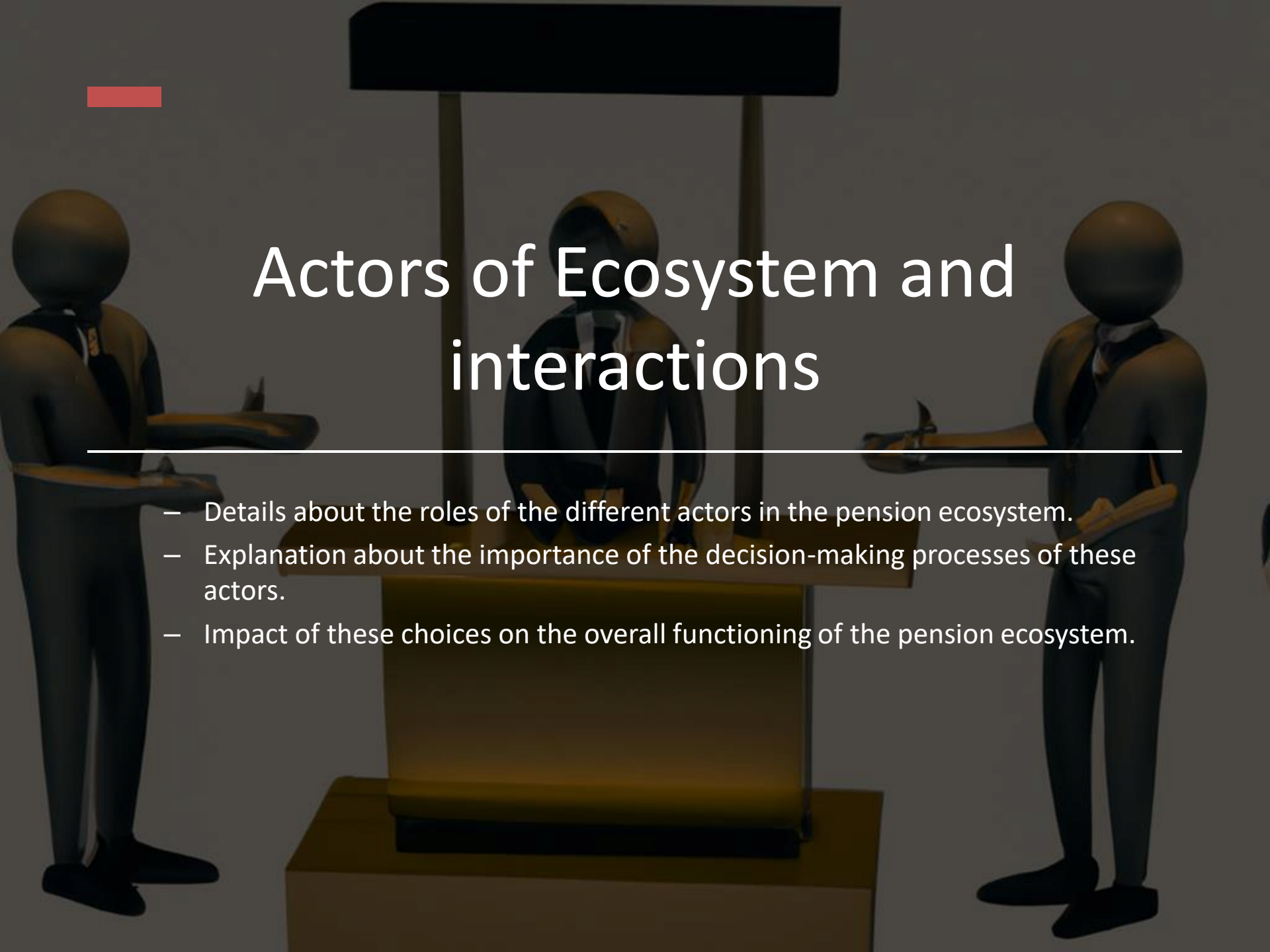
Introduction

- Introduction of a multi-agent reinforcement learning (MARL) model for the pension ecosystem.
- Model generates synthetic income trajectories.
- Proposed model is adaptable to various environmental changes.
- Traditional econometric approach limitations in pension savings problem.
- Introduction of agent-based model (ABM) of the pension ecosystem.
- Non-stationary nature of the model.



Design Choices of Financial Model

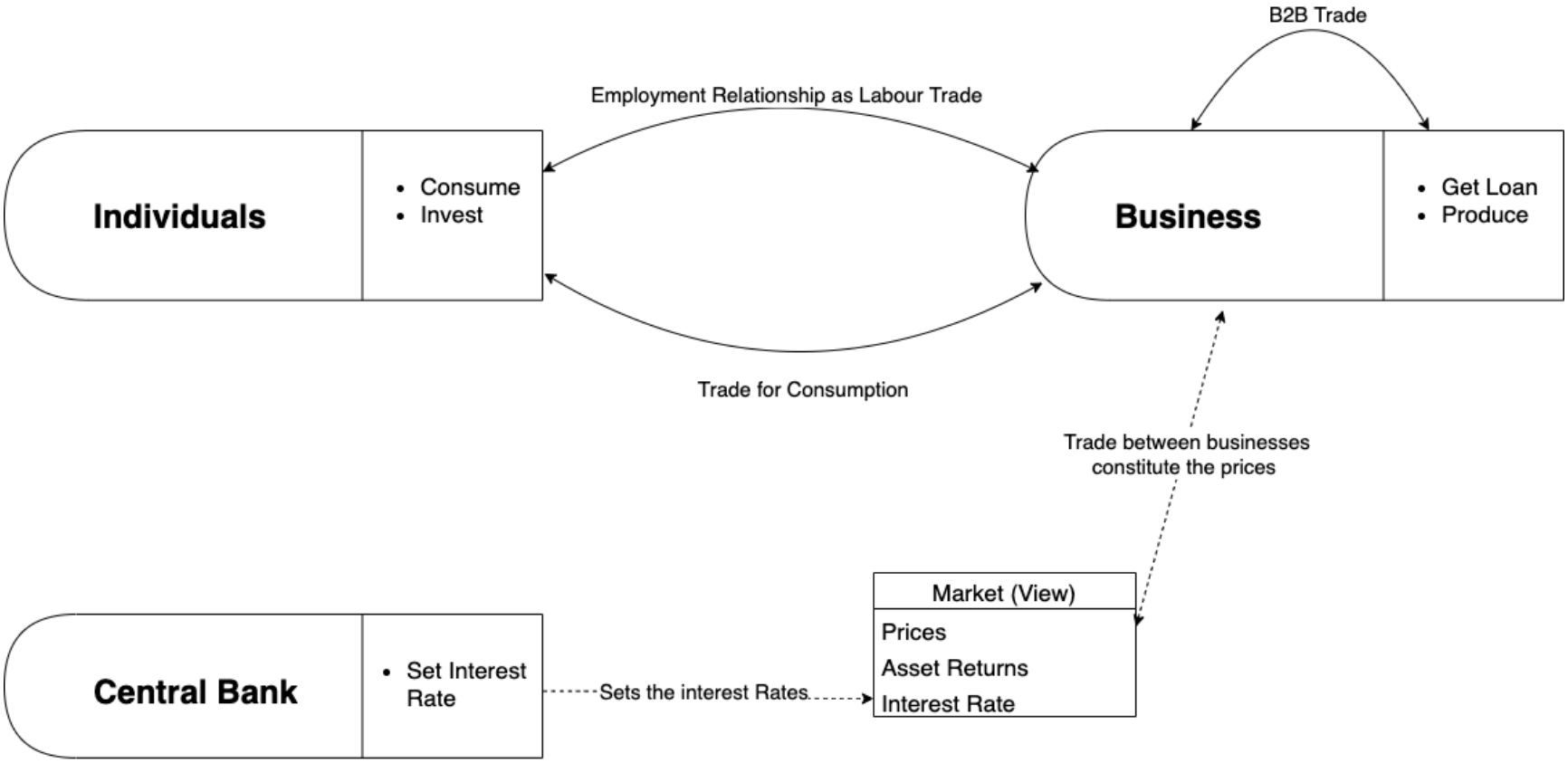
- Actors in the ecosystem: Individuals, Businesses, and the Central Bank, 3 Actors but Many Agents
- Simulation loop to simulate the interactions between actors.
- The order of operations in the simulation and its importance.



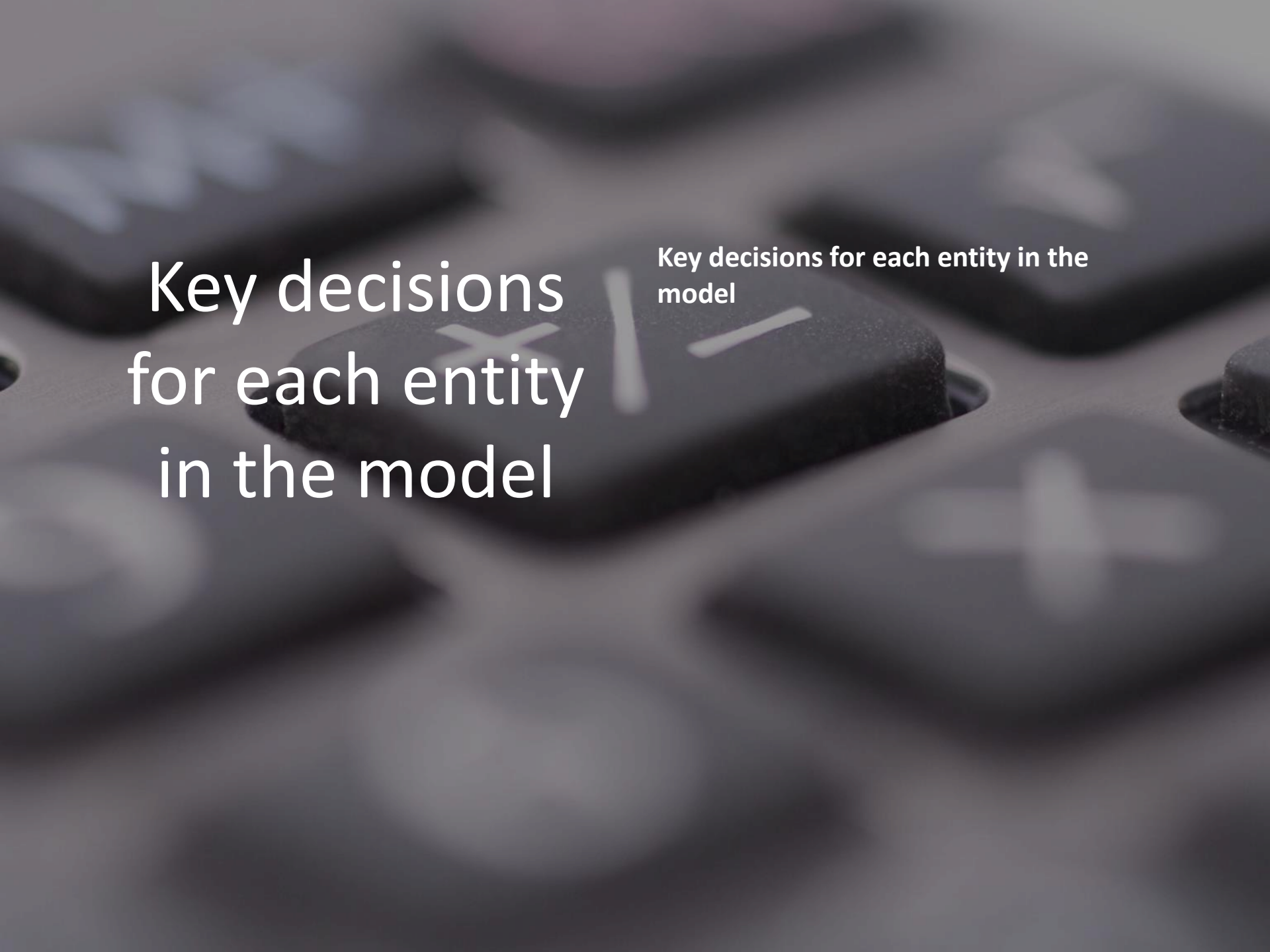
Actors of Ecosystem and interactions

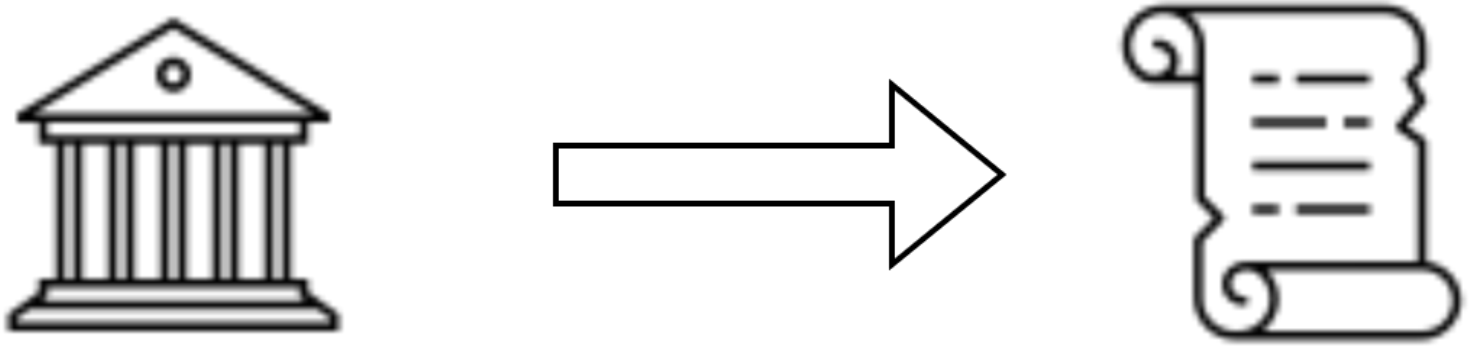
- Details about the roles of the different actors in the pension ecosystem.
- Explanation about the importance of the decision-making processes of these actors.
- Impact of these choices on the overall functioning of the pension ecosystem.

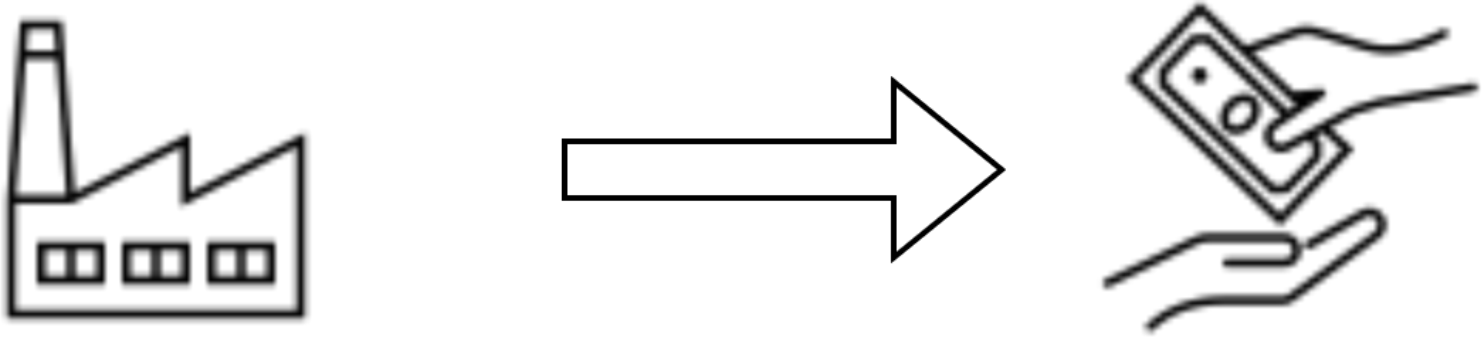
Agents and Environment Diagram

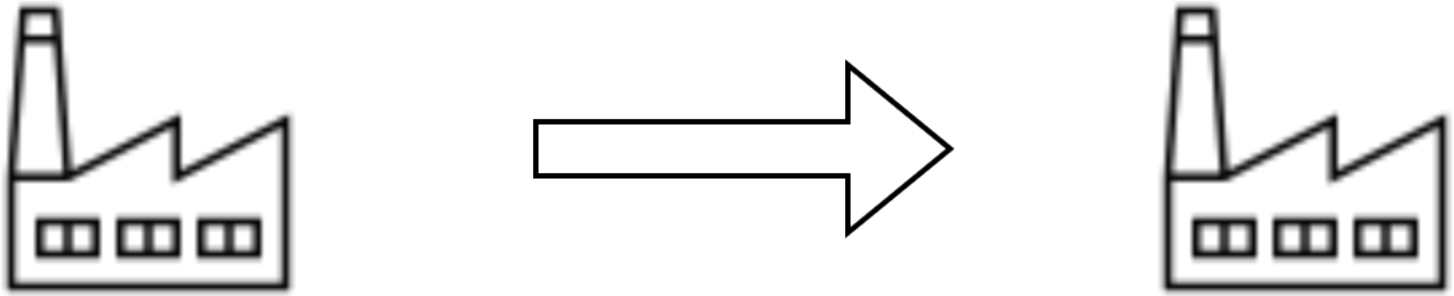


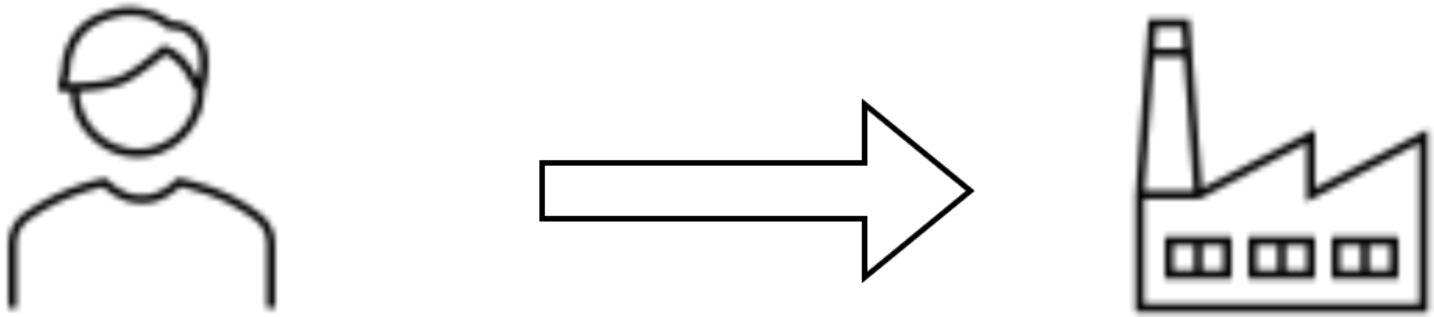
Agents and Environment Diagram


Key decisions for each entity in the model



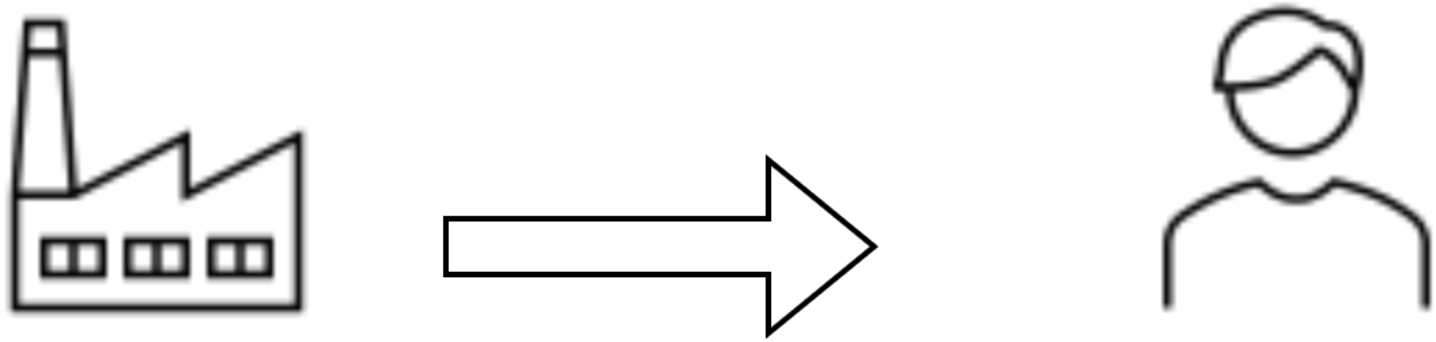


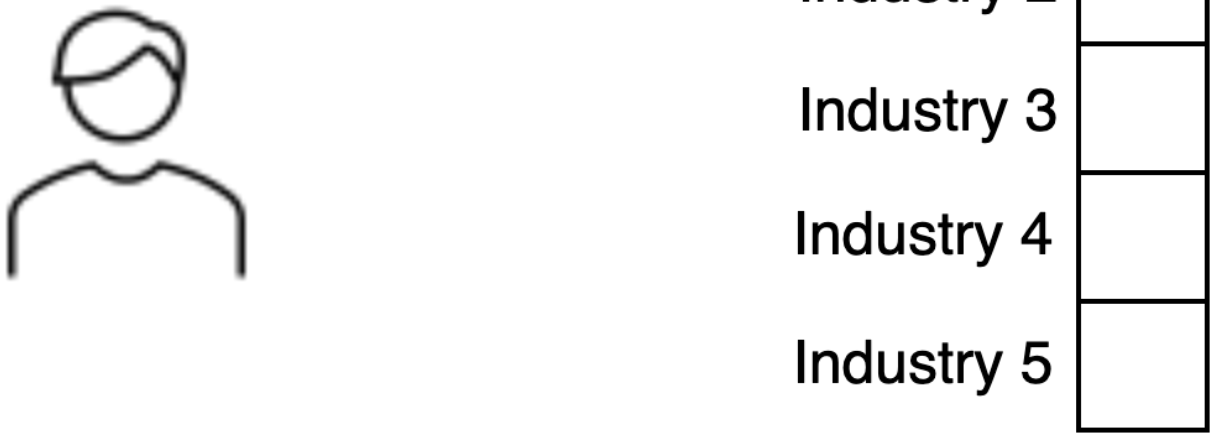


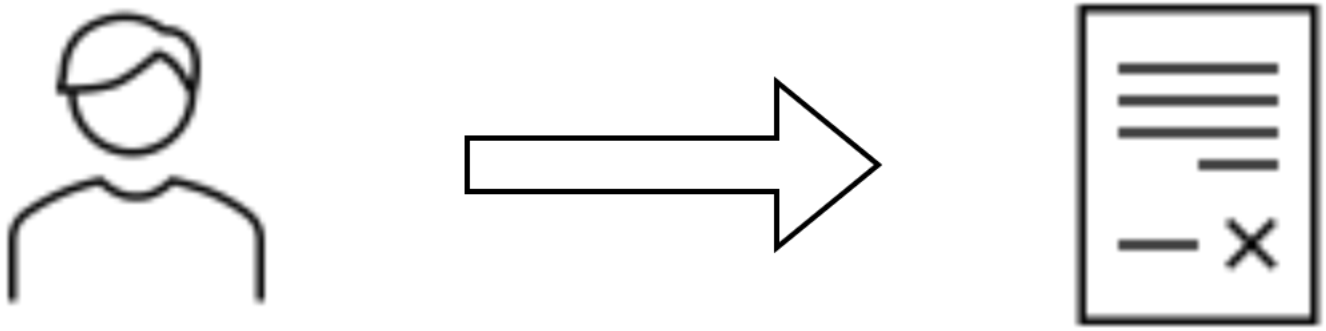


IO Matrix

	Industry 1	Industry 2	Industry 3	Industry 4	Industry 5
Industry 1					
Industry 2					
Industry 3					
Industry 4					
Industry 5					




Consum
Industry 1
Industry 2
Industry 3
Industry 4
Industry 5



Entity	Key Decisions
Business	Borrowing choices Trading activities Production capacity utilization
Individual	Employment B2C trade activities Saving Decision Investment and portfolio allocation
Central Bank	Determining interest rate

Alignment with Finance Community

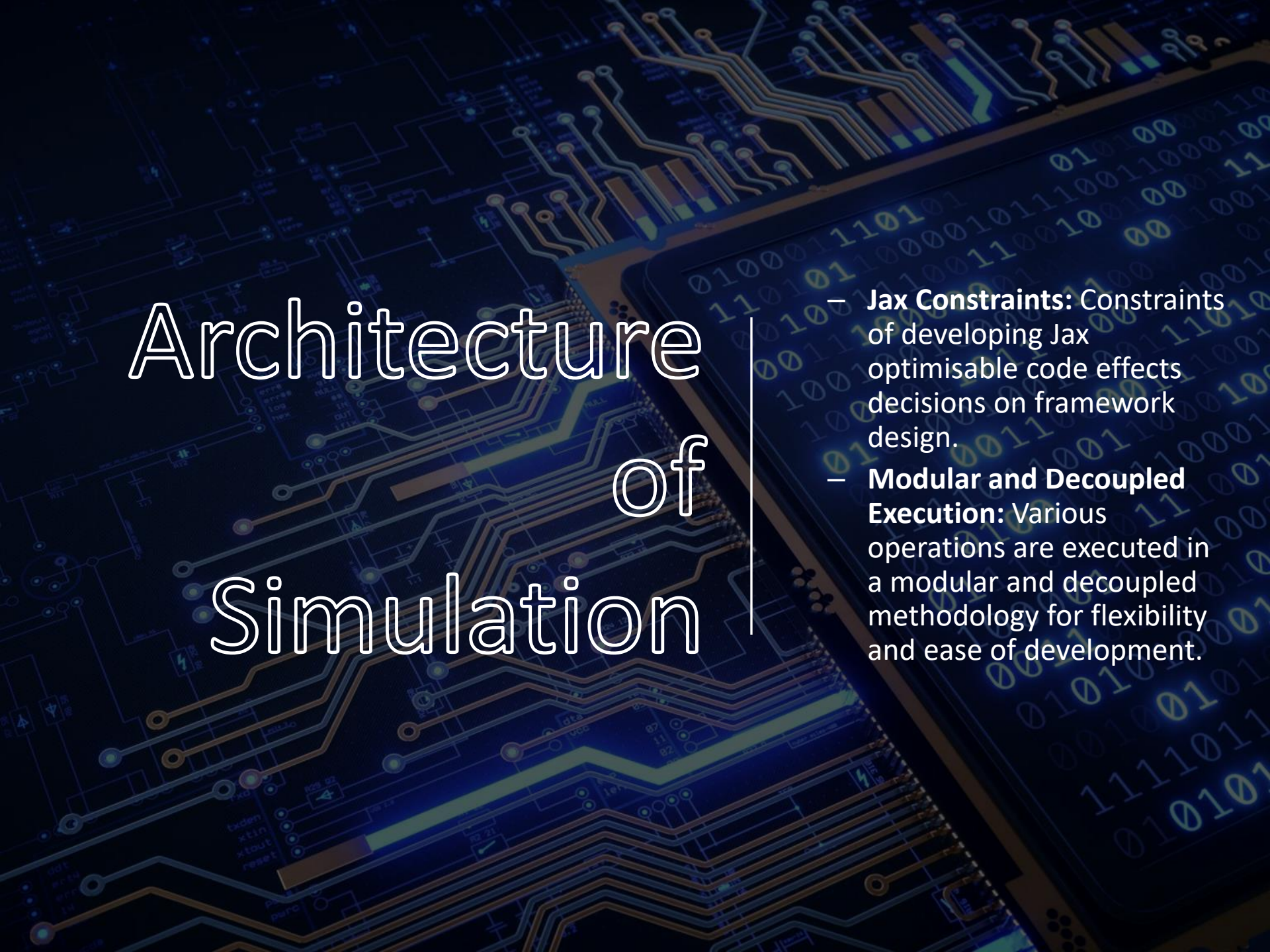
- Our model aligns with familiar financial terminology and concepts in order to facilitate understanding within the broader finance community.
- We utilize utility functions to express trade-offs between immediate and future returns, and calibrate the model based on observable phenomena.
- The asset endowment dynamics are explored in two configurations: a cash and risky vs riskless asset dynamic, and coupling asset returns to the revenues or valuation of simulated companies.

Reward Mixture / Reward as Calibration Goal

$$R_{i,t}^{Mix} \sim \phi(R_{i,immediate}, R_{i,t}, R_{global,immediate}, R_{global,t})$$

Architecture of Simulation

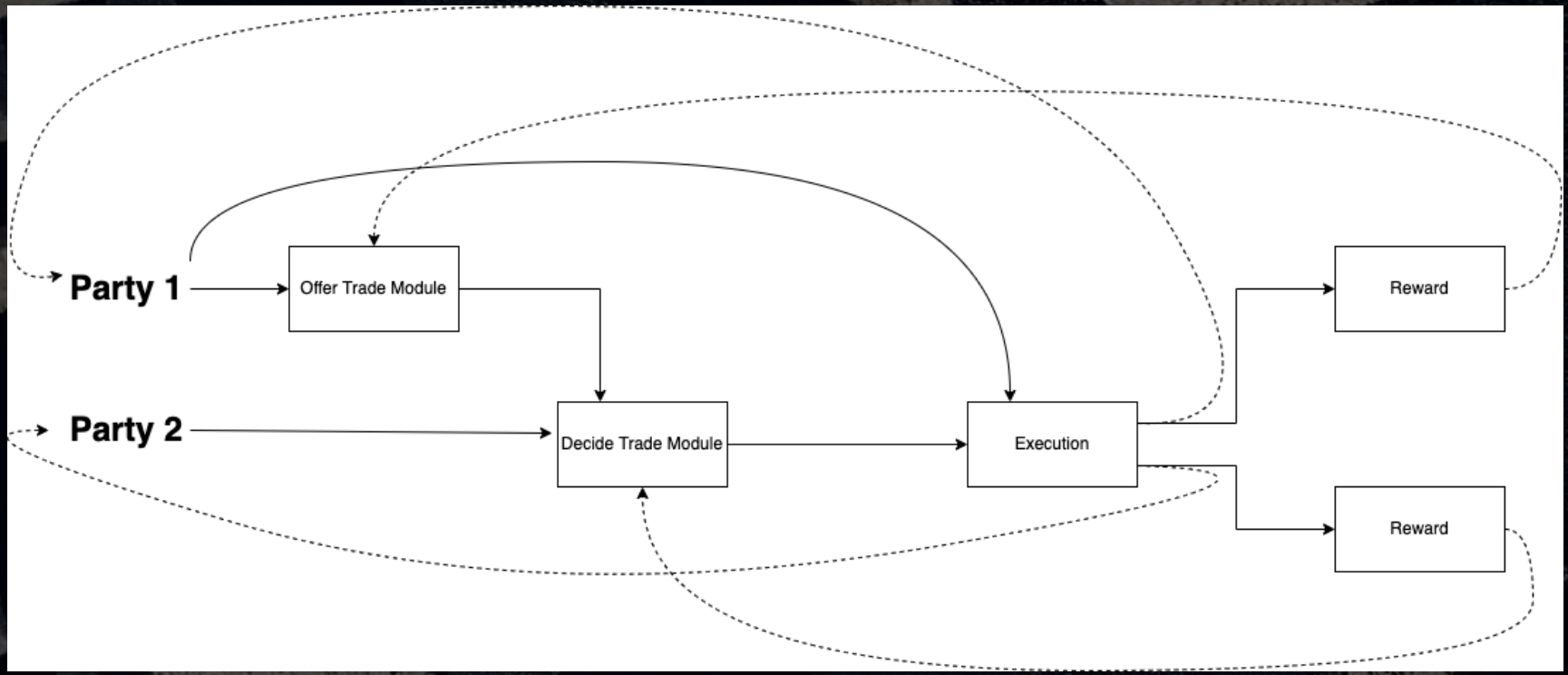
- **Micro Simulation:** Granular micro level interactions that can be trained.
- **Unified Framework:** JAX provides a unified framework for training and inference of deep neural networks as well as executing simulation operations.
- **Introduction of a Novel Framework:** This research introduces a framework to test financial agent-based models and simulations that can be optimized with deep reinforcement learning.



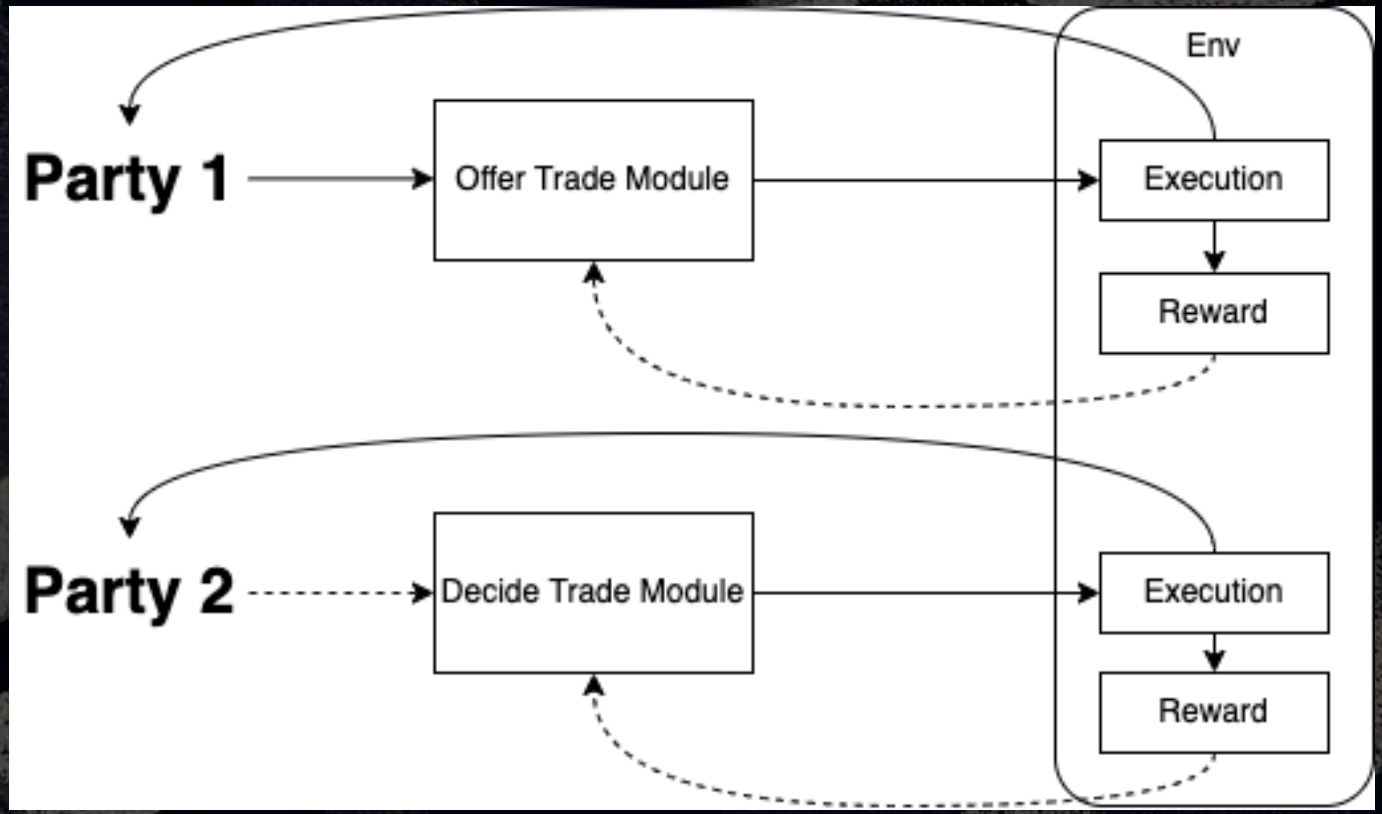
Architecture of Simulation

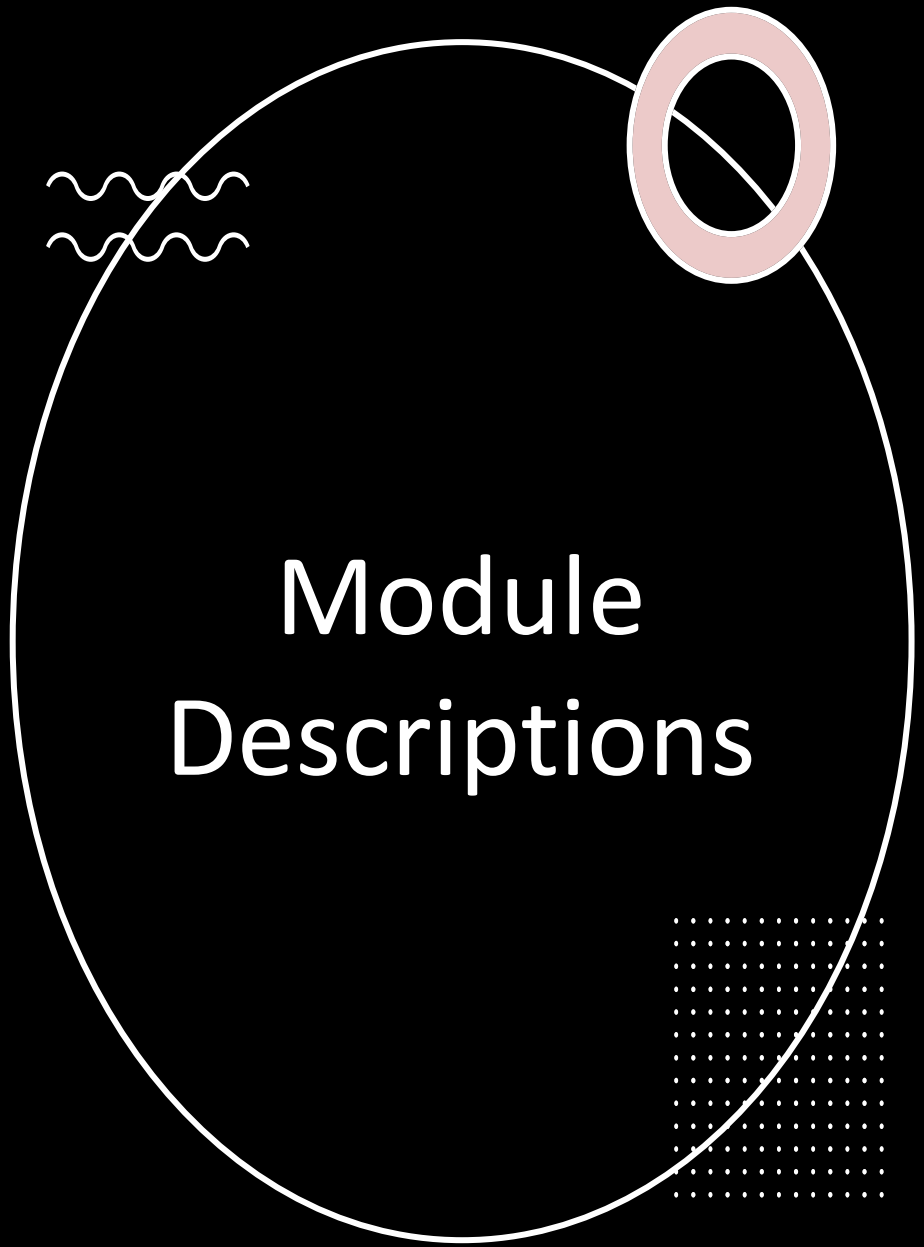
- **Jax Constraints:** Constraints of developing Jax optimisable code effects decisions on framework design.
- **Modular and Decoupled Execution:** Various operations are executed in a modular and decoupled methodology for flexibility and ease of development.

Architecture of Simulation (cont.)



Architecture of Simulation (cont.)

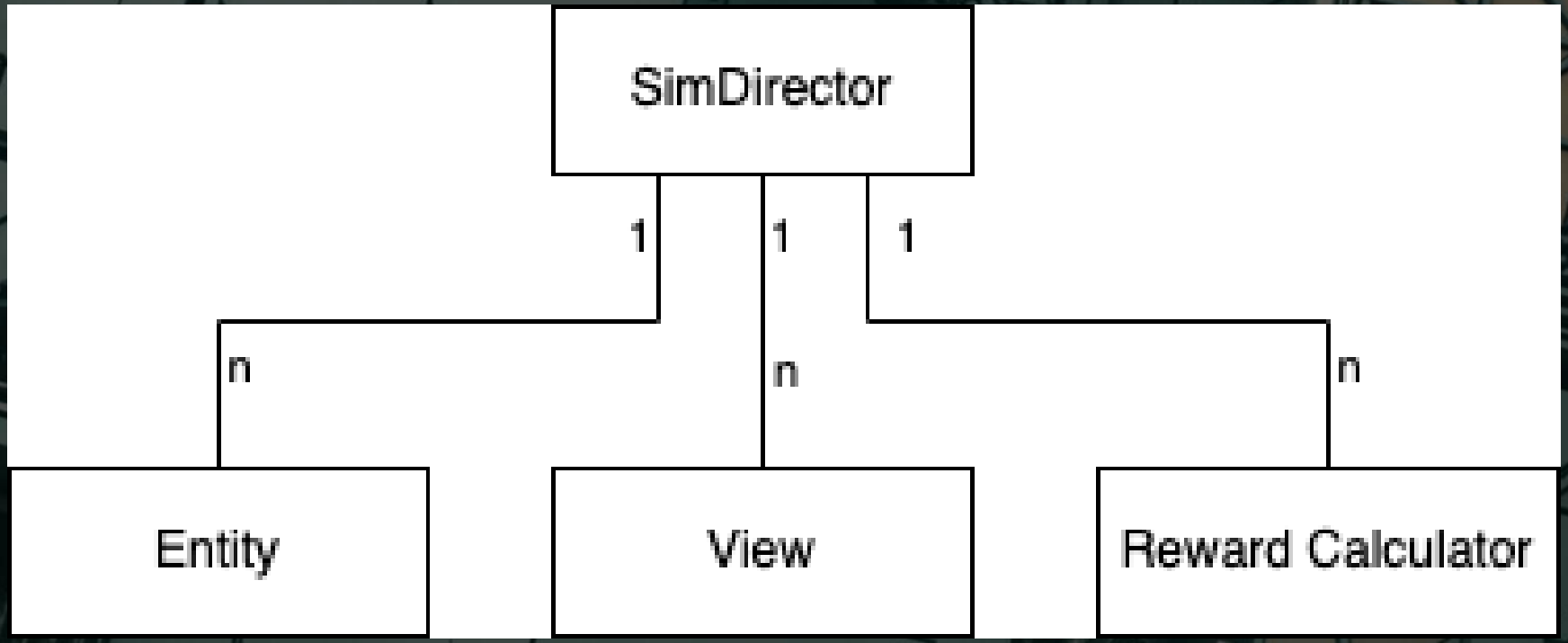




Module Descriptions

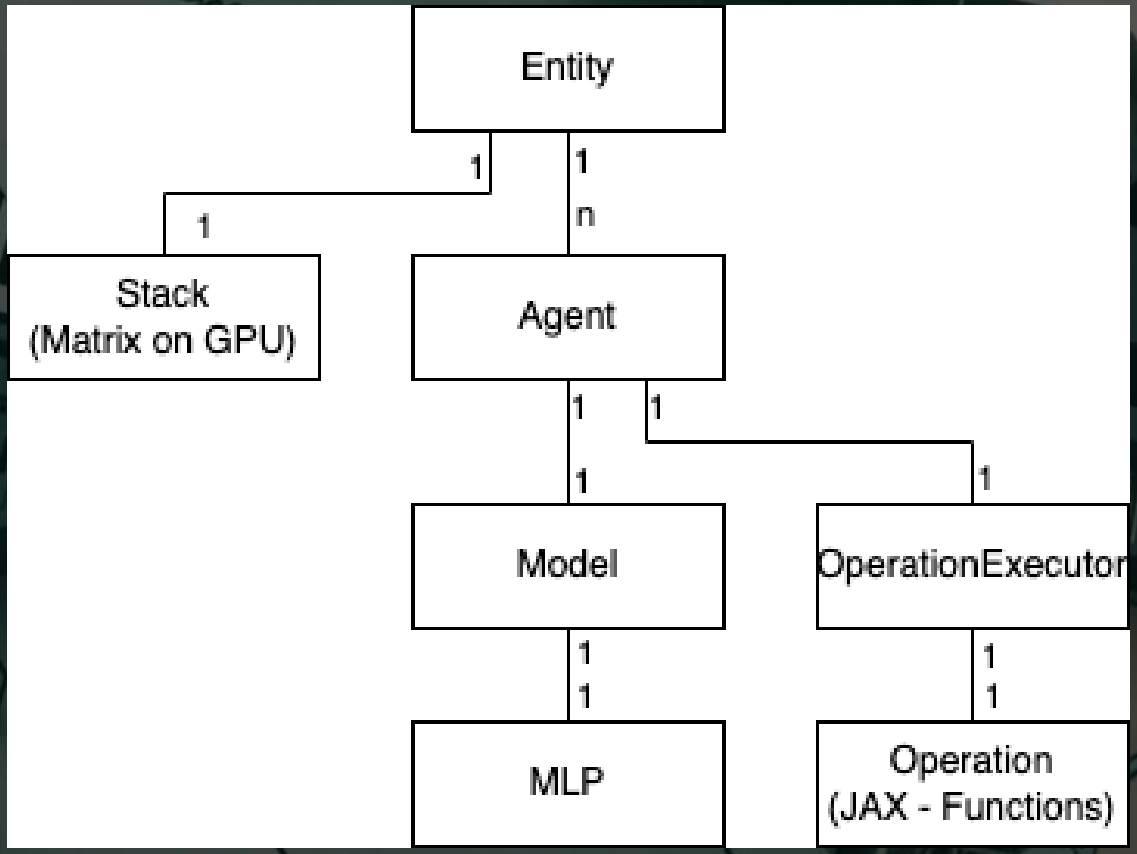
- SimDirector
- Reward Calculator
- Entity
- Stack
- Agent
- Model
- Operation Executor
- Operation
- View
- Logger

Module Visualizations



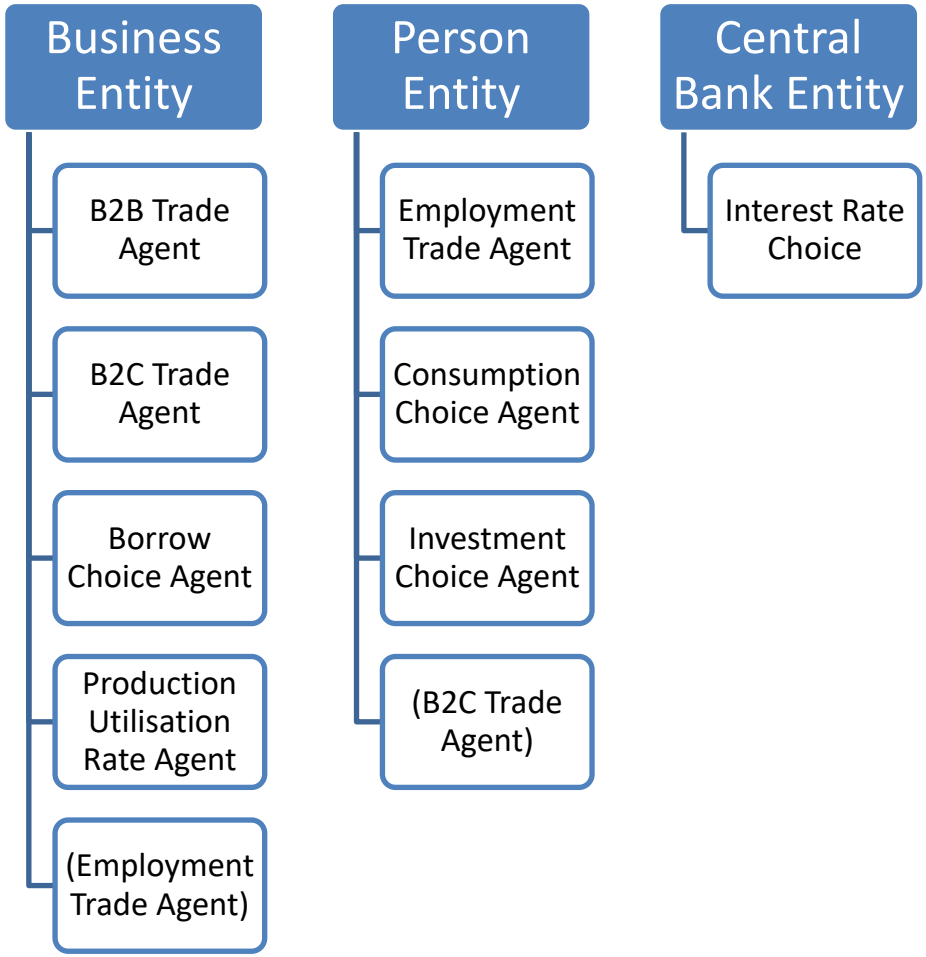
SimDirector orchestrates the simulation and underlying modules

Module Visualizations



Entity and underlying modules

Entity Map


Business Entity
B2B Trade Agent
B2C Trade Agent
Borrow Choice Agent
Production Utilisation Rate Agent
(Employment Trade Agent)
Person Entity
Employment Trade Agent
Consumption Choice Agent
Investment Choice Agent
(B2C Trade Agent)
Central Bank Entity
Interest Rate Choice

Mechanism of Interactions

- Trade operations consist of two parts: an Offer and a Decision. These operations utilize deep neural networks to take in an agent's embedding, market data, and output a vector specifying the goods offered and the cash requested.
- Choice operations involve an agent making a decision independently, based on the agent's embedding and market data.
- Simulation dynamics are executed without active decision making but are relevant to the operation of the simulation.

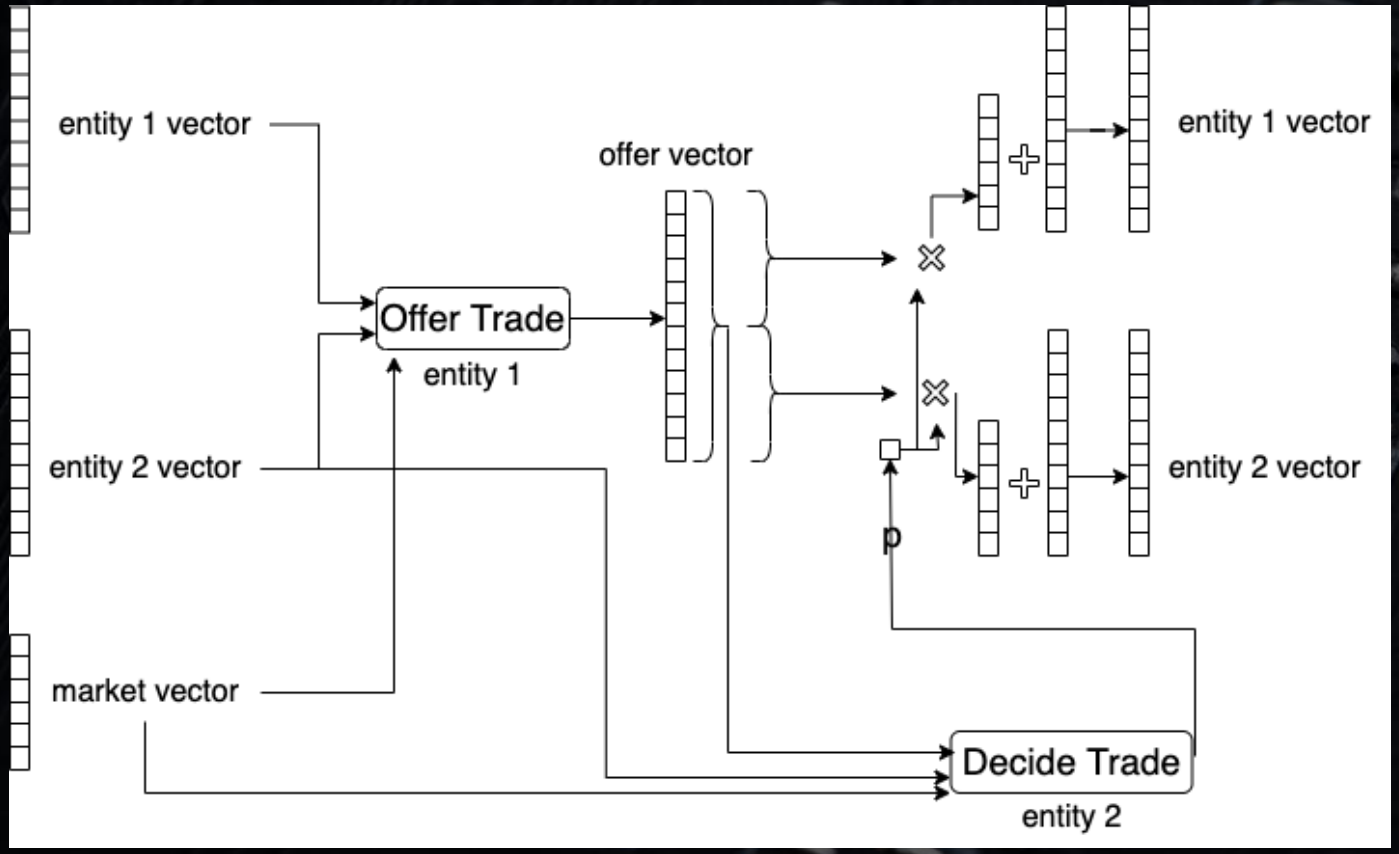
Mechanism of Interactions (continued)

$$a_{i,t}^{offer} \sim \pi(o_{i,t}^{agent}, o_{i,t}^{market}, h_{i,t}; \theta)$$

$$a_{i,t}^{decision} \sim \pi(o_{i,t}^{agent}, o_{i,t}^{offer}, o_{i,t}^{market}, h_{i,t}; \theta)$$

$$a_{i,t}^{choice} \sim \pi(o_{i,t}^{agent}, o_{i,t}^{market}, h_{i,t}; \theta)$$

Trade Module



The Trade Module: Offer and Decision Models

Training and the Reinforcement Learning Problem

- **MAPPO** is a multi-agent adaptation of the Proximal Policy Optimization (PPO) algorithm.
- **PPO** is an on-policy, model-free reinforcement learning method designed to improve the stability and sample efficiency of policy gradient methods.
- MAPPO extends PPO to handle multi-agent environments by employing a centralized training approach with decentralized execution and shown to be effective (Yu et al., 2022).

The State and Action Space in Multi-agent Environments

- In a multi-agent environment, the **state space** is given by the concatenation of individual agent states, $S = \{s_1, s_2, \dots, s_n\}$.
- The joint **action space** is the combination of individual agent actions, $A = \{a_1, a_2, \dots, a_n\}$.
- The goal of MAPPO is to learn a set of **policies** for all agents, $\pi = \{\pi_1, \pi_2, \dots, \pi_n\}$,that maximize the expected **cumulative reward** for each agent.

Objective Function for Agent i in MAPPO

Objective Function for Agent i in MAPPO

The objective function for agent i in MAPPO is given by:

$$L^{MAPPO}(\theta_i) = \mathbb{E}_{s,a,r,s'}\Big[\min\Big(\frac{\pi_i(a_i|s_i)}{\pi_{i,old}(a_i|s_i)}A_i^{GAE}(s_i,a_i),\ \mathrm{clip}\Big(\frac{\pi_i(a_i|s_i)}{\pi_{i,old}(a_i|s_i)},1-\epsilon,1+\epsilon\Big)A_i^{GAE}(s_i,a_i)\Big)\Big]$$

Advantage Estimations and Value Function for Agent i

Advantage estimations $\hat{A}_{t,i}$ for agent i are calculated with a truncated version of Generalized Advantage Estimation (GAE) for T timesteps:

$$\hat{A}_{t,i} = \delta_{t,i} + (\gamma\lambda)\delta_{t+1,i} + ... + (\gamma\lambda)^{T-t+1}\delta_{T-1,i}$$
$$\delta_{t,i} = r_{t,i} + \gamma V_{\theta_{t,i}}(s_{t+1,i}) - V_{\theta_{t,i}}(s_{t,i})$$
$$V_{\theta_{t,i}}(s_{t=0,i}) = 0$$

The value function of the critic for agent i is clipped with the same ϵ hyperparameter of the actor:

$$L_i^V(\theta) = \max\Bigg[(V_{\theta_{t,i}} - V_{target_{t,i}})^2,$$
$$\text{clip}(V_{\theta_{t,i}}, V_{\theta_{t-1,i}} - \epsilon, V_{\theta_{t-1,i}} + \epsilon) - V_{target_{t,i}}\Bigg]^2$$

Target Value and Composite Objective Function for Agent i

The target value for agent i is calculated as:

$$V_{target_{t,i}} = \hat{A}_{t,i} + V_{\theta_{t,i}}(S_{t,i})$$

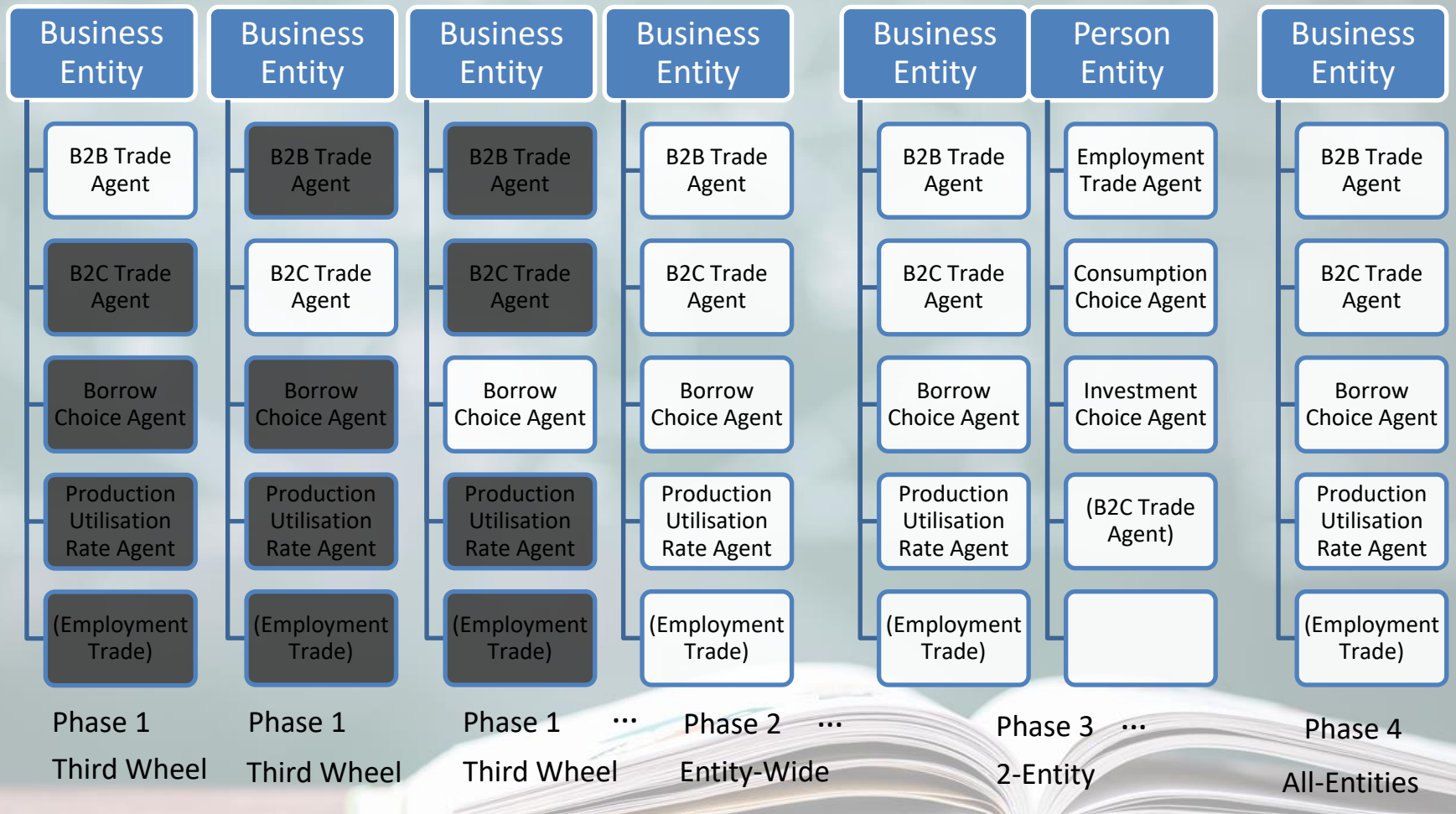
The composite objective function consists of the actor's clipped surrogate objective function, the clipped squared error loss of the critic's value function, and S, an entropy bonus:

$$L_i^{CLIP+V+S}(\theta) = \hat{\boldsymbol{E}}_t \left[L_{t,i}^{CLIP}(\theta) - c_1 L_{t,i}^{V}(\theta) + c_2 S\left[\pi_\theta\right](s_{t,i}) \right]$$

Training Process

- During training, the policy and value networks for each agent are updated using stochastic gradient descent with multiple epochs of mini-batch updates.
- The use of GAE and the PPO clipping objective help to stabilize the learning process and improve sample efficiency.
- MAPPO has proven to be effective in a variety of multi-agent tasks, including cooperative, competitive, and mixed scenarios.
- It is particularly suitable for settings where the interactions between agents are complex and the environment is partially observable, and when stable learning and efficient use of samples are crucial.

Curriculum Learning

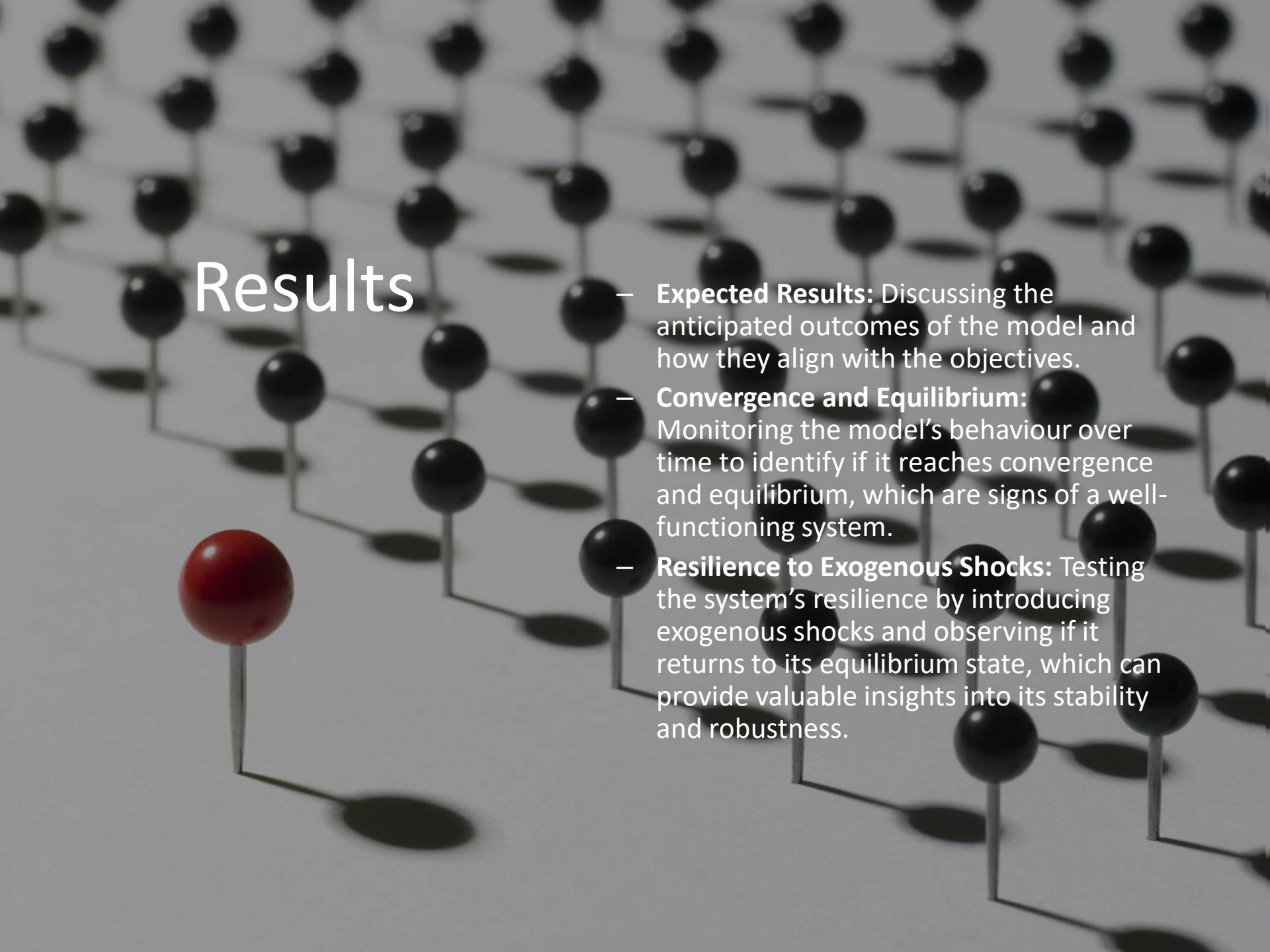


Validation (Methods)

- **How to validate the model:** Essential strategies for model validation.
- **Cyclic Properties:** Understanding and examining the cyclic properties of the model.
- **Observed Principles:** Identification of key principles observed.

Investigating Different Countries

- Simulations are bootstrapped with populations reflecting the sectoral distributions of respective countries.
- The model uses parallelism between discounted rewards in reinforcement learning and time-discounting in multi-step portfolio optimization to communicate the financial interpretation of the agent-based model.
- We propose incorporating heterogeneous employee profiles to further enhance the model's realism and relevance.



Results

- **Expected Results:** Discussing the anticipated outcomes of the model and how they align with the objectives.
- **Convergence and Equilibrium:** Monitoring the model's behaviour over time to identify if it reaches convergence and equilibrium, which are signs of a well-functioning system.
- **Resilience to Exogenous Shocks:** Testing the system's resilience by introducing exogenous shocks and observing if it returns to its equilibrium state, which can provide valuable insights into its stability and robustness.

Challenges: Addressing Stationarity and Training Challenges

- Multi-agent systems face stationarity challenges compared to single-agent reinforcement learning.
- The dynamics agents learn actively change during training.
- Various methodologies address this issue.
- PPO, which shows strong empirical performance, is chosen for our multi-agent pension model.

Challenges: Stateful agents

- The proposed model can be augmented with sequential neural networks.
- Such networks, like LSTMs or transformers, have proven successful in capturing memory and attention to information.
- Learning Meta-Strategies

Challenges: Rewards

- Rewards for the trained agents should account for both **intertemporal** reward attribution and reward attribution in the presence of multiple **agents** determining actions.

Challenges: Input Representations

- Schemes to represent input states for agents are necessary for capturing systemic risks.
- They enable system-level policy training.
- For instance, graph neural networks can be used to allow agents to understand the actions and states of other agents.

Presented Model and Future Work

- This paper introduces a **multi-agent reinforcement learning model** for the **pension ecosystem**, aiming to **optimize contributor saving** and **investment strategies**.
- The utilization of multiple agents facilitates the investigation of the **impact** of **endogenous and exogenous shocks, business cycles, and policy decisions** on contributor behaviour.
- The model produces **synthetic and diverse income trajectories**, which paves the way for the development of more **inclusive** savings strategies.
- This research surpasses traditional econometric models **by deriving meta-strategies** for contributor agents that are robust to evolving environmental dynamics.
- Future work will focus on the **calibration** of the model and **interpretation** of the simulations within the context of the **pension** ecosystem.

References

Asano, Y. M., Kolb, J. J., Heitzig, J., & Farmer, J. D. (2019). *Emergent inequality and endogenous dynamics in a simple behavioral macroeconomic model*. arXiv. http://arxiv.org/abs/1907.02155

Campanale, C., Fugazza, C., & Gomes, F. (2015). Life-cycle portfolio choice with liquid and illiquid financial assets. *Journal of Monetary Economics*, *71*, 67–83. https://doi.org/10.1016/j.jmoneco.2014.11.008

Campbell, J. Y., & Viceira, L. M. (2002). *Strategic asset allocation: Portfolio choice for long-term investors*. Oxford University Press.

Cocco, J. F., Gomes, F. J., & Maenhout, P. J. (2005). Consumption and Portfolio Choice over the Life Cycle. *Review of Financial Studies*, *18*(2), 491–533. https://doi.org/10.1093/rfs/hhi017

Cont, R., & Wagalath, L. (2016). FIRE SALES FORENSICS: MEASURING ENDOGENOUS RISK: FIRE SALES FORENSICS: MEASURING ENDOGENOUS RISK. *Mathematical Finance*, *26*(4), 835–866. https://doi.org/10.1111/mafi.12071

England, B. of. (n.d.). *Announcement of additional measures to support market functioning*. https://www.bankofengland.co.uk/news/2022/october/bank-of-england-announces-additional-measures-to-support-market-functioning

Frostig, R., Johnson, M. J., & Leary, C. (2018). Compiling machine learning programs via high-level tracing. *Systems for Machine Learning*, *4*(9).

Guvenen, F., Ozkan, S., & Song, J. (2012). *The Nature of Countercyclical Income Risk* [Working Paper]. National Bureau of Economic Research. https://doi.org/10.3386/w18035

Harris, C. R., Millman, K. J., Walt, S. J. van der, Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., Kerkwijk, M. H. van, Brett, M., Haldane, A., Río, J. F. del, Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Impavido, G., & Tower, I. (2009). How the financial crisis affects pensions and insurance and why the impacts matter. In *IMF Working Paper* (Vol. wp09151). IMF.

Merton, R. C. (1971). Optimum consumption and portfolio rules in a continuous-time model. *Journal of Economic Theory*, *3*(4), 373–413. https://doi.org/10.1016/0022-0531(71)90038-X

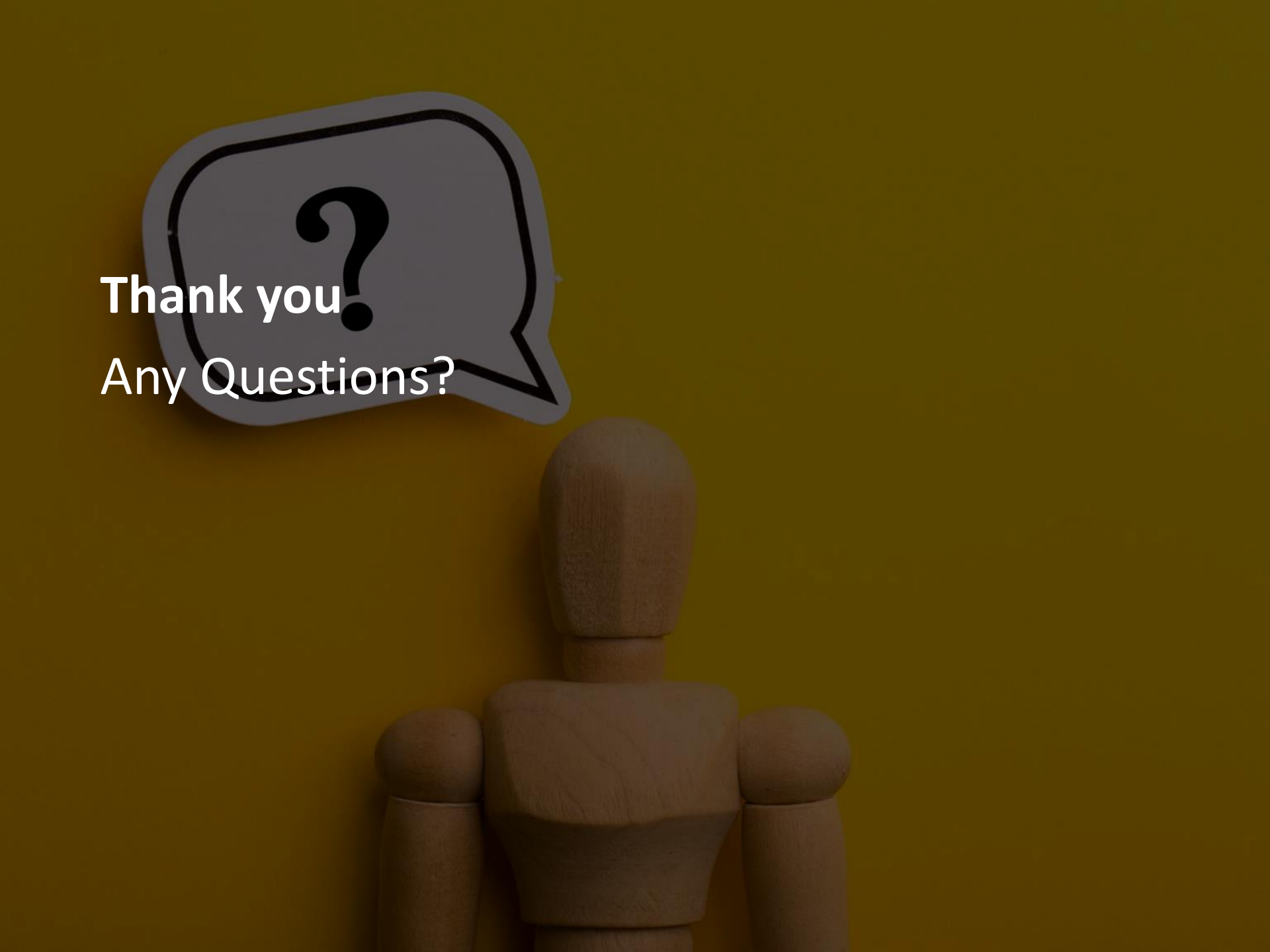
Papaioannou, M. G., & Rentsendorj, B. (2015). Sovereign Wealth Fund Asset Allocations—Some Stylized Facts on the Norway Pension Fund Global. *Procedia Economics and Finance*, *29*, 195–199. https://doi.org/10.1016/S2212-5671(15)01122-3

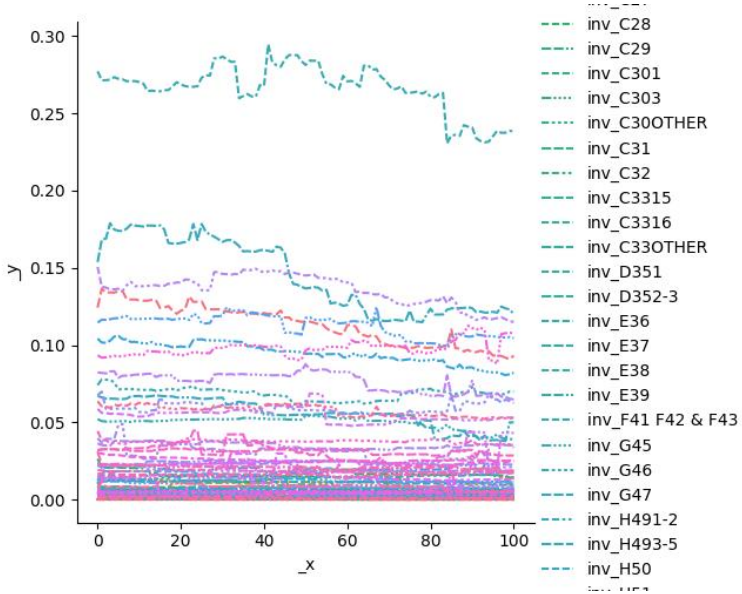
Pichler, A., & Farmer, J. D. (2022). Simultaneous supply and demand constraints in input–output networks: The case of Covid-19 in Germany, Italy, and Spain. *Economic Systems Research*, *34*(3), 273–293. https://doi.org/10.1080/09535314.2021.1926934

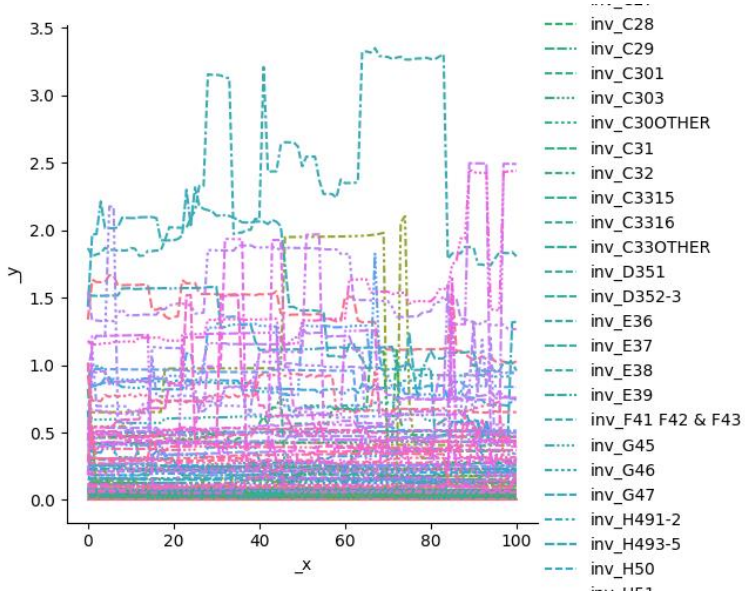
Samvelyan, M., Rashid, T., Witt, C. S. de, Farquhar, G., Nardelli, N., Rudner, T. G. J., Hung, C.-M., Torr, P. H. S., Foerster, J., & Whiteson, S. (2019). *The StarCraft Multi-Agent Challenge*. arXiv. http://arxiv.org/abs/1902.04043

Yang, Y., & Wang, J. (2021). *An Overview of Multi-Agent Reinforcement Learning from Game Theoretical Perspective*. arXiv. http://arxiv.org/abs/2011.00583

Yu, C., Velu, A., Vinitsky, E., Wang, Y., Bayen, A., & Wu, Y. (2022). *The Surprising Effectiveness of PPO in Cooperative, Multi-Agent Games*. arXiv. http://arxiv.org/abs/2103.01955


Thank you
Any Questions?


0.30
0.25
0.20
0.15
0.10
0.05
0.00
y
0
20
40
60
80
100
_x
inv_C28
inv_C29
inv_C301
inv_C303
inv_C30OTHER
inv_C31
inv_C32
inv_C3315
inv_C3316
inv_C33OTHER
inv_D351
inv_D352-3
inv_E36
inv_E37
inv_E38
inv_E39
inv_F41 F42 & F43
inv_G45
inv_G46
inv_G47
inv_H491-2
inv_H493-5
inv_H50


3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
y
0
20
40
60
80
100
_x
inv_C28
inv_C29
inv_C301
inv_C303
inv_C30OTHER
inv_C31
inv_C32
inv_C3315
inv_C3316
inv_C33OTHER
inv_D351
inv_D352-3
inv_E36
inv_E37
inv_E38
inv_E39
inv_F41 F42 & F43
inv_G45
inv_G46
inv_G47
inv_H491-2
inv_H493-5
inv_H50