Proceedings of the 2023 Winter Simulation Conference
C. G. Corlu, S. R. Hunter, H. Lam, B. S. Onggo, J. Shortle, and B. Biller, eds.

TRANSPARENCY AS DELAYED OBSERVABILITY IN MULTI-AGENT SYSTEMS

Kshama Dwarakanath
Svitlana Vyetrenko
Tucker Balch

JP Morgan AI Research
383 Madison Avenue
New York, NY 10179, USA

Toks Oyebode

JP Morgan Regulatory Affairs
25 Bank Street, Canary Wharf
London, E14 5JP, UK

ABSTRACT

Is transparency always beneficial in complex systems such as traffic networks and stock markets? How is transparency defined in multi-agent systems, and what is its optimal degree at which social welfare is highest? We take an agent-based view to define transparency (or its lacking) as delay in agent observability of environment states, and utilize simulations to analyze the impact of delay on social welfare. To model the adaptation of agent strategies with varying delays, we model agents as learners maximizing the same objectives under different delays in a simulated environment. Focusing on two agent types - constrained and unconstrained, we use multi-agent reinforcement learning to evaluate the impact of delay on agent outcomes and social welfare. Empirical demonstration of our framework in simulated financial markets shows opposing trends in outcomes of the constrained and unconstrained agents with delay, with an optimal partial transparency regime at which social welfare is maximal.

1 INTRODUCTION

Multi-agent systems (MASs) are ubiquitous in applications such as robotics (Dudek et al. 1996), healthcare (Shakshuki and Reid 2015), finance (Byrd et al. 2020) and transportation (Burmeister et al. 1997) where the domain can be modeled as a system with interacting agents/components. Each agent is an entity that takes in sensory observations of the environment to make goal-oriented decisions (Dorri et al. 2018). There are numerous challenges associated with the design and development of MASs including communication and collaboration between agents, multi-agent learning and lastly, transparency or the degree of information dissemination. Since each agent in a MAS uses disseminated information to update its decisions, modifying transparency can lead to contrasting agent and system behaviors. Concurrently, simulations offer an effective framework to understand the impact of transparency on MASs to subsequently inform real world policy.

In this work, we consider stochastic MASs comprised of adaptive agents. The stochastic nature arises from various factors including intrinsic environmental randomness (uncontrollable by agents), randomness in agent behavior, along with the fact that the system evolves as a result of agent interactions. Each agent has access to a set of system observables that potentially informs their behavior. The more informative the set of observables are, the more knowledge the agent has about the system including the behavior of other agents. **We characterize transparency in such stochastic MASs by the degree of observability of agents.** A system comprised of agents with a higher degree of observability is said to be more transparent than one where agents have a lower degree of observability.

We seek to analyze the impact of varying observability in stochastic MASs on the strategies adopted by the agents, and subsequently on social welfare. We mathematically formulate a type of observability called delayed observability that is characterized by a single delay parameter controlling the degree of observability. We then let the agents adapt their strategies to this delay parameter. In order not to hard

979-8-3503-6966-3/23/$31.00 ©2023 IEEE

code (and bias/restrict) this adaptation, we equip agents with learning algorithms designed to maximize their objectives under varying delays. Therefore, we formulate this as a multi-agent reinforcement learning problem with partial observability. The key contributions of this work are as follows.

1. A mathematical formulation for transparency in multi-agent systems using the notion of delayed observability in stochastic games.
2. A framework for analyzing the impact of observability on strategies and outcomes of constrained and unconstrained agents, and on social welfare using multi-agent reinforcement learning. We adapt a social welfare metric that captures agents' average outcome as well as equality of outcomes.
3. Detailed empirical study of the proposed framework in simulated financial markets. We use a multi-agent market simulator to train our agents, and investigate the variation in their policies with observability. Our findings indicate the trade-offs between the degree of observability and agent outcomes, and suggest an intermediate degree of observability at which social welfare is maximized.

2 BACKGROUND AND RELATED WORK

2.1 Transparency

Transparency has been of interest in various disciplines including governance, medicine, financial markets and organisations (Ball 2009). It has diverse meanings ranging from being a tool to affect accountability and performance of governmental agencies (Kosack and Fung 2014) to one that builds patient trust in medical physicians (Chimonas et al. 2017), and stakeholder trust in organisations (Auger 2014). Regulators of financial markets are increasingly focused on improving information dissemination to traders to improve market efficiency (Gensler 2022; Treasury 2022). Recent work (Barsotti et al. 2022) looks at the interplay between transparent explanations shared by an institution and strategic adaptation by individuals subject to a classification model. They evaluate the impact of such feedback on faking behaviour by individuals.

We look at transparency as information availability to agents in a dynamic MAS that can be used to improve their objectives. There is work on studying the impact of information in networks of interacting agents in the game theory literature that involves classifying information in games into two types (Arefizadeh et al. 2022). The first type looks at the availability of the rules of the game to players in it, resulting in games with complete or incomplete information (Harsanyi 1967). The second type examines the visibility of agent actions to one another thus partitioning games into those with perfect or imperfect information.

Arnott et al. (1991) challenge the intuitive notion that increased information dissemination in traffic systems reduces congestion by investigating the equilibrium effects of such information on driver behaviors as well as mean travel times. Through mathematical models for commuter travel times and costs, the authors make the case that while proprietary information may benefit a single driver, informing all drivers can result in them being worse off than when uninformed. Das et al. (2017) look at designing information made available to agents in a congestion game in order to improve social welfare. The authors make the case for partial information being able to improve efficiency in such network routing games.

While transparency in systems is generally a positive notion as it is intuitively thought to improve information access and opportunity for participants, there exists literature in the financial domain to investigate this intuition deeper. They account for the fact that individual agent objectives in MASs may not be in line with each other or with overall social welfare, causing the MAS to be more sensitive to released information. Ehrmann and Fratzscher (2009) look into existing practice by central banks to refrain from information transparency right before policy meetings because transparency during these periods could lead to higher market volatility. The authors corroborate this rationale by analyzing past market data to empirically estimate effects of information dissemination. Walsh (2007) and Van der Cruijsen et al. (2010) explore the optimal degree of central bank transparency that can limit inflation. Here, they measure transparency by central bank announcements about its view on the economy or its own private information.

2.2 Partially Observable Stochastic Games and Multi-Agent Reinforcement Learning

We seek to evaluate how strategic agents in a MAS utilize information available under increased transparency regimes to adapt their behaviours. Given agent objectives, we allow these agents to use information to learn strategies that improve their objectives over time. A single agent that is learning to act in an uncertain environment is modeled by a Markov Decision Process (MDP) in the framework of reinforcement learning (Sutton and Barto 2018). In the MDP framework, multiple agents are modeled to be non-adaptive and accounted as part of the environment of the learning agent. Multiple adaptive agents with interacting or competing goals can be modeled by Markov Games in multi-agent reinforcement learning. Markov Games are an extension of MDPs involving the specification of a global state space, action spaces for each agent along with corresponding reward functions. Littman (1994) propose a Q-learning algorithm for 2-player zero-sum games where the objective of the first player is exactly opposite to that of the second.

An MDP is said to be partially observable (PO) if the environment state is not completely visible to the agent (Spaan 2012). Partially observable MDPs can be extended to PO Markov Games (or PO Stochastic Games) to precisely model stochastic MASs with varying degrees of observability (Shapley 1953; Hansen et al. 2004). A finite horizon Partially Observable Stochastic Game (POSG) is denoted by $\Gamma = \langle \mathscr{N}, \mathscr{S}, \{\mathscr{A}_i\}_{i=1}^n, \{\mathscr{O}_i\}_{i=1}^n, \mathbb{T}, \{\mathbb{O}_i\}_{i=1}^n, \{R_i\}_{i=1}^n, \gamma, H\rangle$ where

- $\mathscr{N} = \{1, 2, \cdots, n\}$ is the set of agents
- $\mathscr{S}$ is the state space
- $\mathscr{A}_i$ is the action space of agent i with $\mathscr{A} = \mathscr{A}_1 \times \mathscr{A}_2 \times \cdots \times \mathscr{A}_n$ denoting the joint action space
- $\mathscr{O}_i$ is the observation space of agent i
- $\mathbb{T} : \mathscr{S} \times \mathscr{A} \to \mathbb{P}(\mathscr{S})$ is the transition function that maps the current state and joint action to a probability distribution over the next state
- $\mathbb{O}_i : \mathscr{S} \to \mathbb{P}(\mathscr{O}_i)$ is the observation function that maps the current state to a probability distribution over observations of agent i
- $R_i : \mathscr{S} \times \mathscr{A} \to \mathbb{R}$ is the reward function of agent i
- $\gamma \in [0, 1)$ is a discount factor
- H is the horizon

The objective of each agent $i \in \mathscr{N}$ in a POSG is to find a sequence of their own actions that maximizes their expected sum of discounted rewards over the horizon

$$\max_{(a_i(0),\cdots,a_i(H-1))} \mathbb{E}\left[\sum_{t=0}^{H-1} \gamma^t R_i\left(s(t), a_1(t), \cdots, a_n(t)\right)\right]$$

where $s(t+1) \sim \mathbb{T}\left(s(t), a_1(t), \cdots, a_n(t)\right)\ \forall t$. A POSG is fully observable when $\mathscr{O}_i = \mathscr{S}$ with $\mathbb{O}_i$ being the Dirac delta function over $\mathscr{O}_i = \mathscr{S}$ for all $i \in \mathscr{N}$.

Arefizadeh et al. (2022) take a partial observability approach to optimal information transparency in network games where a principal needs to decide on agent neighborhoods for aggregate information sharing. They formulate this as an optimization problem seeking an optimal partition of agents to maximize social welfare. Gharesifard and Smith (2018) look at the effect of a lack of information about other agents' strategies on agent objectives. These two works differ from ours in that transparency refers to the knowledge of other agents' actions, rather than an observation of a global system state as in our case. We also deal with applications where the change in agent strategies in response to those of others is not easily defined due to the dynamic nature of their objectives, calling for the use of multi-agent simulations. Thus, we allow agents to adapt their strategies in order to maximize their objectives using reinforcement learning.

3 PROBLEM FORMULATION

3.1 Delayed Observability and Constrained Agents

In this work, we define the lack of transparency in MASs as the delay in observability in their POSG formulation. We introduce the notion of **delayed observability** as follows. Partition every state $s \in \mathscr{S}$ as $s = \begin{bmatrix} s_I & s_D \end{bmatrix}$ where s_I is a part of s that is immediately observable to all agents, while s_D is a part of s that is observable after a time delay $\delta \in [0, H]$. Thus, the observation of any agent $i \in \mathscr{N}$ at time t is

$$o_i(t) = \begin{bmatrix} s_I(t) & s_D(t-\delta) \end{bmatrix} \tag{1}$$

Note that we do not consider $\delta > H$ since that would correspond to looking into states from previous episodes. Equivalently, we formulate POSGs with delayed observability δ as those where $\mathscr{O}_i = \mathscr{S}$ and the observation function $\mathbb{O}_i$ is a indicator for observations satisfying (1), for all agents $i \in \mathscr{N}$. Note that a large delay δ would correspond to less transparency, and vice-versa. Additionally, we do not model any hidden and fully unobservable states in this work thereby eliminating the need for belief state updates.

For an MDP with partial observability (i.e. a POSG with a single agent), Åström (1965) showed that the agent's cumulative rewards increase with increase in observability. In this work, we look at the impact of the delay in observability δ on the strategies learnt by 2 players/agents in the POSG. These players are characterized by having the same reward function, but the actions of player 1 are *more constrained* than those of player 2. More rigorously, we have $\mathscr{A}_1 \subsetneq \mathscr{A}_2$ with $R_1(s, a_1, a_2) = R_2(s, a_1, a_2) \forall s \in \mathscr{S}, a_i \in \mathscr{A}_i$ (and $\mathscr{O}_1 = \mathscr{O}_2$ from before). We call player 1 the **constrained player** and player 2 the **unconstrained player**.

In financial markets, constrained players include market makers whose actions are subject to regulatory constraints of liquidity provision (Chakraborty and Kearns 2011). And, unconstrained players would be firms trading similar volumes as market makers, without being constrained to provide liquidity. In traffic systems, constrained players would be (public) government transit systems that are restricted to a smaller set of routes as opposed to other (unconstrained) commuters or private transit agencies. It is then natural to expect that increased observability affects unconstrained players in a different fashion than constrained players. The interaction between the two leading to potentially interesting implications to social welfare.

3.2 Social Welfare Function

Social welfare refers to the notion of goodness of current state of affairs with respect to the society as a whole (Sen 2018). It can be quantified through social welfare functions (SWFs) that can rank social states as being less or more desirable for social welfare. Given utilities/outcomes for individuals in a population, there are many SWFs studied in the literature on optimal taxation. Let $Y_i\left(s(0), \pi_1(\cdot), \cdots, \pi_n(\cdot)\right)$ denote the outcome for player i over horizon H when players use respective policies $\pi_j : \mathscr{S} \to \mathscr{A}_j$, $\forall j \in \mathscr{N}$ starting from game state $s(0)$. For player outcomes $Y = \begin{bmatrix} Y_1 & \cdots & Y_n \end{bmatrix}$, the utilitarian SWF is $\text{SWF}(Y) = \sum_{i \in \mathscr{N}} Y_i$.

Given an average population outcome $\bar{Y} = \frac{1}{n} \sum_{i \in \mathscr{N}} Y_i$, different outcome distributions between players correspond to different levels of equality. We draw from Zheng et al. (2022) in capturing the trade-off between profitability and equality by using a product of both as a SWF as

$$\text{SWF}(Y) = \text{Equality}(Y) \times \text{Profitability}(Y)$$

with profitability measured by $\bar{Y}$. Popular (in)equality metrics in the social choice literature include the Gini index (Sen 2018) and the Theil-L index (Sen et al. 1997). Since we have two types of players (likely with multiple players of each type), we use generalized entropy indices to measure (in)equality as in Dwarakanath et al. (2022), Speicher et al. (2018). They are attractive due to their property of subgroup decomposability while containing several inequality indices as special cases.

We consider the following SWFs (although our framework is flexible to the use of any other):

1. Using the generalized entropy index with parameter $\kappa \notin \{0, 1\}$ to measure equality as

$$\text{SWF}(Y) = \exp\left(-\text{GE}_\kappa(Y)\right) \times \bar{Y} \tag{2}$$

where $\mathrm{GE}_\kappa(Y) = \frac{1}{n\kappa(\kappa-1)} \sum_{i\in\mathscr{N}} \left[\left(\frac{Y_i}{\bar{Y}}\right)^\kappa - 1 \right]$.

2. Using the Theil-L index given by $\mathrm{Theil}_L(Y) = -\frac{1}{n}\sum_{i\in\mathscr{N}} \ln\left(\frac{Y_i}{\bar{Y}}\right)$ to measure equality as

$$\mathrm{SWF}(Y) = \exp\left(-\mathrm{Theil}_L(Y)\right) \times \bar{Y}. \tag{3}$$

The negative sign in the exponential in (2)-(3) is because these indices are metrics for inequality, with the generalized entropy index containing the Theil-L index as a special case when $\kappa = 0$.

The goal of this work is to evaluate the impact of the delay in observability δ on

- strategies learnt by the (constrained) player 1 and (unconstrained) player 2
- outcomes of both players resulting from playing out learnt strategies
- social welfare measured by (2) and (3)

We emphasize that we equip both players with learning algorithms to learn strategies that utilize available observations to maximize their objectives without any constraints on the way strategies change with delay in observability. Thus, we formulate this problem as a 2-agent reinforcement learning problem for every value of δ. We specifically experiment with the example of simulated financial markets.

4 APPLICATION TO MARKETS

4.1 Multi-Agent Market Simulator

Consider exchange-based markets (e.g. US equities market) comprising a variety of traders that send their order requests to a centralized exchange that matches buy and sell orders. Traders are allowed to specify both the price and the direction (buy or sell) of their orders sent to the exchange. In order to simulate trades in such a market, we employ a multi-agent market simulator called ABIDES (Byrd et al. 2020; Amrouni et al. 2021). ABIDES provides a selection of trading agents with different trading incentives and behaviors. The simulation engine handles all communication between trading agents and the exchange. Figure 1 shows an example snapshot of orders collected at the exchange wherein each purple/green rectangle represents a block of sell/buy orders respectively. The price of the cheapest sell order is called the best sell price while that of the most expensive buy order is called the best buy price. The current stock price (also called mid-price) is the average between the best sell and best buy prices, with the difference between the best sell and best buy prices called the spread. Figure 1 shows an example with a mid-price of 10000.5 cents and spread of 1 cent. We populate the simulated market with the following traders:

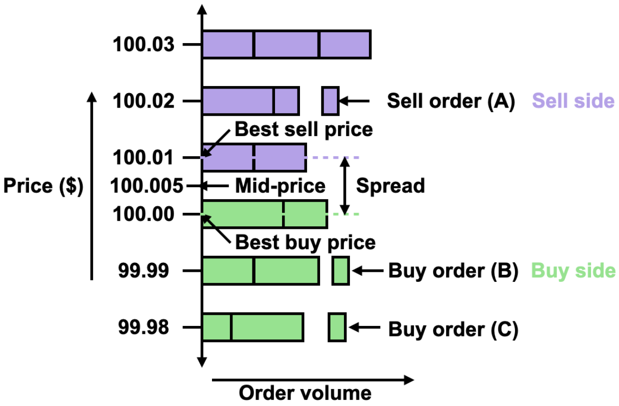

Figure 1: Snapshot of buy and sell orders at an exchange. Mid-price moves when traders submit orders that cross the spread i.e., a buy order with price greater than or equal to the best sell price or vice-versa.

1. **Market Maker (MM)**: Player 1 that is constrained to place orders on both the buy and sell sides of the market, following from its regulatory definition (Wah et al. 2017; Federal Reserve Bank of New York. 2023). MM can choose price levels for the buy, sell orders relative to the current stock price, and earns the difference between the sell and buy prices if orders are matched on both sides.
2. **Principal Trader (PT)**: Player 2 that can choose to place either a buy order, a sell order, buy and sell orders, or hold (no order), along with the price at which to do so relative to the current stock price. PT is hence unconstrained, with the volumes of each buy/sell order being equal for both MM and PT.
3. **Background traders**: Non-adaptive traders that utilize available market signals to trade in a rule based fashion. They include consumer traders employing a uniformly random trading strategy with random arrivals to the market, as well as intelligent traders that use momentum strategies or exogenous stock value to trade (Byrd et al. 2020).

4.2 POSG Formulation for Markets

The **state** of the environment at time t includes

- *Quotes*: Prices and volumes of quoted orders (pre-trading) on buy, sell sides over period $t-L,\cdots,t$
- *Spread*: Market spread defined as the difference between the best sell and best buy prices at t
- *Depth*: Half of the price difference between worst sell (most expensive sell) and worst buy (cheapest buy) orders at t
- *Inventory*: Volume of stocks held by players 1 and 2 over $t-1,t$
- *Cash*: Cash held by players 1 and 2 over $t-1,t$
- *Momentum*: Momentum signals for stock price over 1, 10 and 30 time steps defined as the ratio of current mid-price to that 1/10/30 time steps before
- *Trades*: Price and volumes of traded buy and sell orders over period $t-M,\cdots,t$

Note that the length of history for quotes L and that for trades M are hyper-parameters.

The **observation** for player $i \in \mathscr{N}$ at time t includes the immediately observable states of *Quotes*, *Spread*, *Depth*, *Inventory* of player i, *Cash* of player i and *Momentum*. It also includes delayed *Trades* information with delay parameter δ. Hence at t, player i has access to traded volumes and prices over the period $t-M-\delta,\cdots,t-\delta$. Intuitively, low δ implies knowledge of more recent trades in the market and hence, access to more relevant information than for high δ.

The **action** for (constrained) player 1 represented by the MM includes

- *Half-spread*: Distance from current stock price at which MM symmetrically places orders on the buy and sell sides. Figure 1 shows an example snapshot of MM orders (A) and (B) placed at the exchange with half-spread 1.5 cents when the stock price is 10000.5 cents.

The **action** for (unconstrained) player 2 represented by the PT includes

- *Half-spread*: Distance from current stock price at which player 2 places orders
- *Order side*: Player 2 can choose to place either a single order on the buy or sell side or place on both sides like Player 1 to provide liquidity or hold (and do nothing).

Clearly, the action space of player 2 contains that of player 1. This means that player 2 has more capability in choosing its actions than player 1 by the design of their action spaces. Figure 1 shows an example snapshot of a PT buy order (C) with half-spread 2.5 cents when the stock price is 10000.5 cents.

The **reward** for player $i \in \mathcal{N}$ at time t is the change in value of its portfolio from $t-1$ to t, given by

$$\begin{aligned} R_i(t) = {} & Cash(t) + Inventory(t) \times Mid\text{-}price(t) \\ & - Cash(t-1) - Inventory(t-1) \times Mid\text{-}price(t-1) \end{aligned}$$

Given the formulation above, we learn policies that maximize discounted cumulative rewards for both players as they interact with each other and the environment for different delays δ. Player outcomes are represented by their profits over the horizon, defined as the difference in portfolio value at $t = H$ to that at $t = 0$. Notice that the undiscounted sum of above reward over H precisely gives agent profits. We therefore measure player outcomes by the undiscounted cumulative rewards realized from using learnt policies over H. These outcomes are subsequently used to calculate the SWFs in (2) and (3).

5 EXPERIMENTAL RESULTS

5.1 Training

Our simulated market contains 24 background traders (20 consumer, 4 intelligent), 1 learning MM and 1 learning PT. For every δ, the horizon is a single trading day starting at 9:30am and ending at 4pm. The MM and PT place orders every minute, giving $H = 390$ steps per episode. Hence, we vary δ between $\delta = 0$ and $\delta = 390$, with $\delta = 0$ being the most observable scenario where all states are immediately observable. And, $\delta = 390$ being the least observable scenario where a strict sub-part of the states s_I are observable. As a caveat, we allow the PT to only place a buy or sell order or hold in the experiments in order to further differentiate between the actions of the MM and PT, although allowing the PT to place on both sides of the market fits the exact formulation. The background traders come in at random times in the trading day. We use $\gamma = 0.9999$ in our experiments since the value of money at the end of the day is nearly the same as that in the beginning.

Given the formulation in section 4.2, we have continuous states with discrete levels of *Half-spread* and categorical *Order side* options. We use the policy gradient method called Proximal Policy Optimization (PPO) from the RLlib package (Schulman, J., F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. 2017; Liang et al. 2018) to learn policies for the MM and PT. The states and rewards are normalized to enable efficient exploration in a continuous state, discrete action setup. Figure 2 is a plot of (moving averages of) discounted cumulative rewards of the MM and PT as a function of training episodes. We observe convergence in rewards with training episodes for the values of δ considered. The difference in scale of rewards between players is because the MM places orders on both sides while the PT typically places on one side or holds, with each order being of equal size.

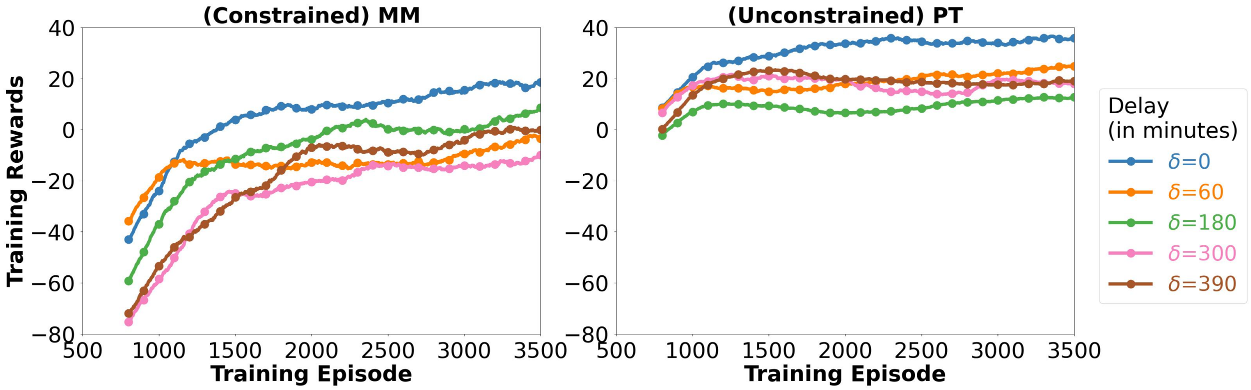

Figure 2: Discounted cumulative rewards while training demonstrating convergence in learning.

5.2 Impact of Delay on Player Outcomes

The learnt policies of both players are played out in 500 test episodes to collect their realized rewards. Figure 3 is a plot of the cumulative rewards of the MM and PT in test episodes (i.e. outcomes) as a function of the delay in observability. The solid line with circular markers represents the mean values, and the shaded region shows the 95% confidence interval. Interestingly, we observe that the outcomes for the constrained MM improve as delay increases, while those of the PT degrade. Recall that observability increases as the delay δ decreases (going right to left on the x-axis). This is to say that increased observability empowers the unconstrained player with more information allowing it to improve its outcomes. On the other hand, the constrained player loses out due to inability to act on the new information due to constraints on its actions. Note that the kink at $\delta = 300$ arises due to the results at $\delta = 390$ varying from the general trend. This is because of the large change in the environment at $\delta = 390$ when no delayed states are observable.

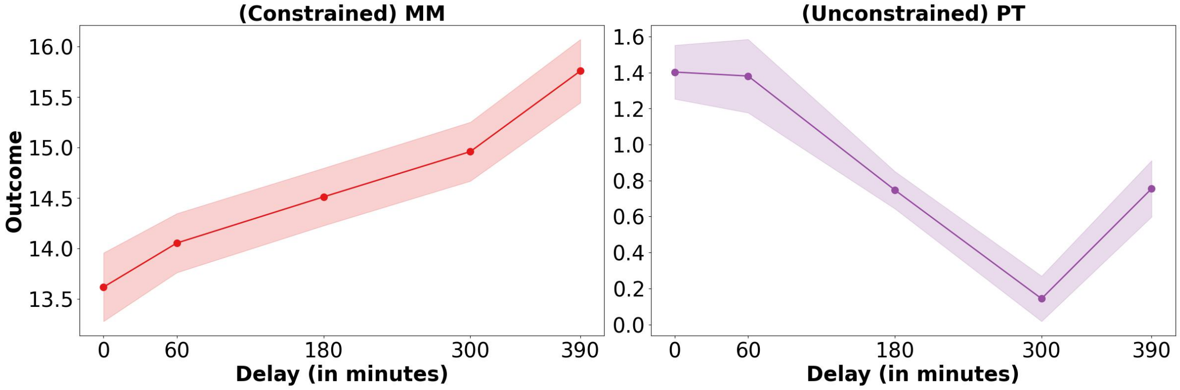

Figure 3: Cumulative rewards measuring player outcomes. With increase in observability (decrease in delay), constrained player outcomes worsen while unconstrained player outcomes improve.

5.3 Impact of Delay on Learnt Strategies

We first plot the average *Half-spread* (average computed across time steps and test episodes per delay) for the MM and PT in Figure 4. We observe a (near) monotonically increasing trend in both strategies as δ increases, with the PT choosing lower *Half-spread* at every δ than the MM. Orders with lower *Half-spread* have prices closer to the current stock price meaning that they are more competitively priced and have higher chances of being matched. Therefore, we make the case that increased observability about recent trades allows the PT to place more competitively priced orders than the MM. The second PT action of *Order side* is categorical and corresponds to multiple directions of placing an order or holding. We plot the average % of hold decisions in a trading day as a function of δ in Figure 5. We see that the % of hold decisions decreases as observability increases. This means the PT places buy/sell orders more frequently at low δ, and prefers to hold more at high δ. Thus, observability enables the PT to trade more frequently.

To investigate which observation features are most impactful towards PT decisions, we use the explainability tool called SHAP (for SHapley Additive exPlanations). We examine the PPO policy network that takes in observations to give out PT actions. SHAP uses cooperative game theory to decompose the network output locally into a sum of effects attributed to each input feature (Lundberg and Lee 2017). SHAP values are scores measuring the importance of observation features towards the action prediction. For illustration, we use SHAP for the PT action of *Order side* to calculate global feature importances for market observations (section 4.2) in Figure 6. Note that delayed features of traded volume and traded price are colored in olive. These global importances are calculated as the average of absolute SHAP values across a dataset comprising (observation, action) pairs.

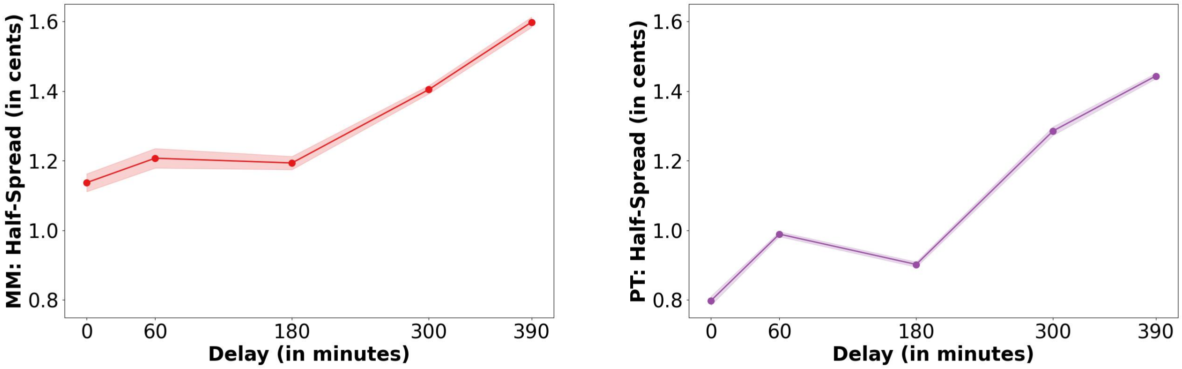

Figure 4: Learnt strategies of MM and PT: *Half-spread* of orders. See the (near) monotonic trend in strategies with delay.

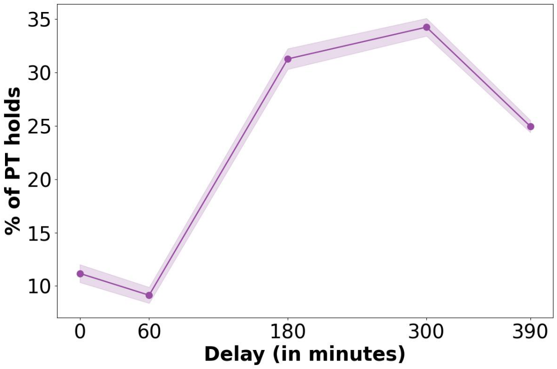

Figure 5: Learnt strategy of PT: % of hold decisions. PT trades more frequently (holds less) at low delays.

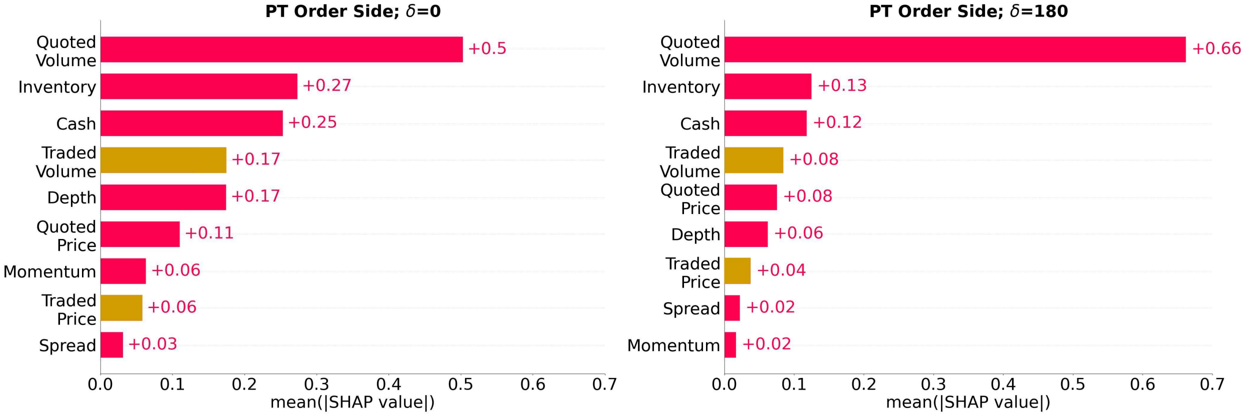

Figure 6: Impact of features on PT action of *Order side* for $\delta = 0$ (left) and $\delta = 180$ (right). See the reduction in importance for delayed features of ‘Traded Volume’ and ‘Traded Price’ with increase in δ.

The left plot in Figure 6 shows feature importances when $\delta = 0$, where quoted volume is the most impactful feature with traded volume coming in fourth. The right plot in Figure 6 shows feature importances when $\delta = 180$, where the importance scores for traded volume and traded price decrease from their values in the left plot. Thus, the delayed features lose their importance in defining actions as the delay is increased.

We also see that the importance of quoted volume increases as the delay in traded volume is increased. This goes to say that the PT uses quoted volume as a substitute for traded volume when the latter is delayed.

5.4 Impact of Delay on Social Welfare

Figure 7 is a plot of SWFs computed using agent outcomes in test episodes as a function of δ. The left plot uses (2) with $\kappa = 6$ while the right plot uses (3). We observe both SWFs increase with δ up to $\delta = 300$ after which they fall down. Thus, there lies an intermediate delay at which social welfare is maximized.

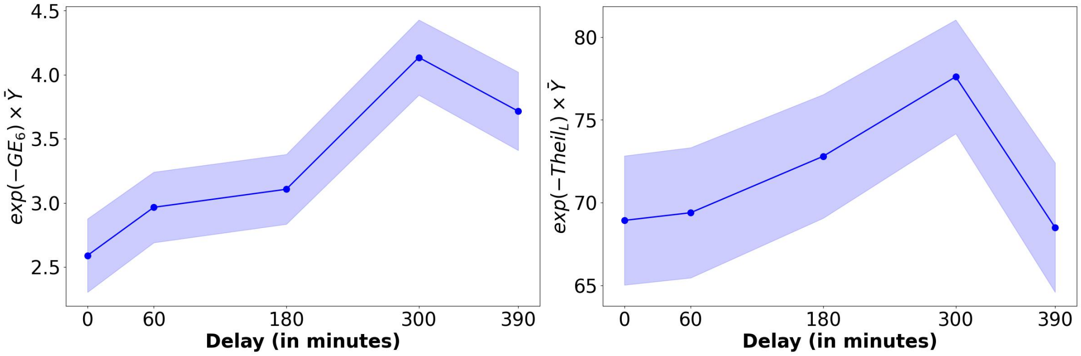

Figure 7: Social Welfare as measured by (2) in the left plot and (3) in the right plot. See that social welfare is maximized at intermediate values of delay in both cases.

5.5 Discussion

In our setup, $\delta = 390$ is a fully opaque regime when no delayed states are observable, with $\delta = 0$ being a fully transparent regime when all states are immediately observable. With increase in transparency, both agents were seen to place more competitively priced orders with lower *Half-spread* that in turn reduces the market spread. Social welfare was found to be highest at intermediate δ that correspond to a partial transparency regime. This can provide an indication to policymakers on optimal transparency regimes that may not be intuitive by looking at outcomes of agents alone. Although our experiments were performed in a generic exchange market simulator that is not calibrated to a specific financial market, similar findings have been observed with the introduction of transparency in real markets as shown in Table 1.

Table 1: Trends that have been observed in real markets with increase in transparency. Our findings in simulated markets are in line with those observed in real markets.

Metric	Definition	Market	Trend in real data	Trend in our work
Price dispersion	Volume weighted difference between mid-price, traded price	Interest Rate Swaps	Fell by 12-19% (Benos et al. 2020)	Market spread reduced
Bid-ask spread	Difference between best buy, best sell prices	US Corporate Bonds	Reduced (Mizrach, B. 2015)	Reduced

6 CONCLUSION

We consider the problem of defining and evaluating the impact of transparency in multi-agent systems comprising adaptive agents using simulations. By defining transparency (or lack thereof) as delay in observability for agents, we propose a multi-agent reinforcement learning framework to evaluate the effects of varying observability on agent strategies and social welfare. We specifically look at the interplay between constrained and unconstrained agents. The framework is illustrated with experiments in simulated financial markets comprising constrained and unconstrained traders. We observe that increasing observability improves outcomes for unconstrained agents albeit with degrading outcomes for the constrained agents. We also experimentally demonstrate that social welfare is maximized at intermediate values of delay.

ACKNOWLEDGMENTS

This paper was prepared for informational purposes by the Artificial Intelligence Research group of JPMorgan Chase & Co and its affiliates ("J.P. Morgan") and is not a product of the Research Department of J.P. Morgan. J.P. Morgan makes no representation and warranty whatsoever and disclaims all liability, for the completeness, accuracy or reliability of the information contained herein. This document is not intended as investment research or investment advice, or a recommendation, offer or solicitation for the purchase or sale of any security, financial instrument, financial product or service, or to be used in any way for evaluating the merits of participating in any transaction, and shall not constitute a solicitation under any jurisdiction or to any person, if such solicitation under such jurisdiction or to such person would be unlawful.

REFERENCES

Amrouni, S., A. Moulin, J. Vann, S. Vyetrenko, T. Balch, and M. Veloso. 2021. "ABIDES-Gym: Gym Environments for Multi-Agent Discrete Event Simulation and Application to Financial Markets". In *Proceedings of the Second ACM International Conference on AI in Finance*. November 3rd-5th, Virtual, 1-9.

Arefizadeh, S., S. Ozgoli, S. Bolouki, and T. Başar. 2022. "Compartmental Observability Approach for the Optimal Transparency Problem in Multi-Agent Systems". *Automatica* 143:110398.

Arnott, R., A. de Palma, and R. Lindsey. 1991. "Does Providing Information to Drivers reduce Traffic Congestion?". *Transportation Research Part A: General* 25(5):309–318.

Åström, K. J. 1965. "Optimal Control of Markov Processes with Incomplete State Information". *Journal of Mathematical Analysis and Applications* 10:174–205.

Auger, G. A. 2014. "Trust me, Trust me Not: An Experimental Analysis of the Effect of Transparency on Organizations". *Journal of Public Relations Research* 26(4):325–343.

Ball, C. 2009. "What is Transparency?". *Public Integrity* 11(4):293–308.

Barsotti, F., R. G. Koçer, and F. P. Santos. 2022. "Transparency, Detection and Imitation in Strategic Classification". In *Proceedings of the 31st International Joint Conference on Artificial Intelligence*. July 23rd-29th, Vienna, Austria.

Benos, E., R. Payne, and M. Vasios. 2020. "Centralized Trading, Transparency, and Interest Rate Swap Market Liquidity: Evidence from the Implementation of the Dodd–Frank Act". *Journal of Financial and Quantitative Analysis* 55(1):159–192.

Burmeister, B., A. Haddadi, and G. Matylis. 1997. "Application of Multi-Agent Systems in Traffic and Transportation". *IEE Proceedings-Software* 144(1):51–60.

Byrd, D., M. Hybinette, and T. H. Balch. 2020. "ABIDES: Towards High-Fidelity Multi-Agent Market Simulation". In *Proceedings of the 2020 ACM SIGSIM Conference on Principles of Advanced Discrete Simulation*. June 15th-17th, Miami, Florida, 11-22.

Chakraborty, T., and M. Kearns. 2011. "Market Making and Mean Reversion". In *Proceedings of the 12th ACM Conference on Electronic Commerce*. June 5th-9th, San Jose, California, 307-314.

Chimonas, S., N. J. DeVito, and D. J. Rothman. 2017. "Bringing Transparency to Medicine: Exploring Physicians' Views and Experiences of the Sunshine Act". *The American Journal of Bioethics* 17(6):4–18.

Das, S., E. Kamenica, and R. Mirka. 2017. "Reducing Congestion through Information Design". In *2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. October 3rd-6th, Monticello, Illinois, 1279-1284.

Dorri, A., S. S. Kanhere, and R. Jurdak. 2018. "Multi-Agent Systems: A Survey". *IEEE Access* 6:28573–28593.

Dudek, G., M. R. Jenkin, E. Milios, and D. Wilkes. 1996. "A Taxonomy for Multi-Agent Robotics". *Autonomous Robots* 3:375–397.

Dwarakanath, K., S. Vyetrenko, and T. Balch. 2022. "Equitable Marketplace Mechanism Design". In *Proceedings of the Third ACM International Conference on AI in Finance*. November 2nd-4th, New York City, New York, 232-239.

Ehrmann, M., and M. Fratzscher. 2009. "Purdah: On the Rationale for Central Bank Silence around Policy Meetings". *Journal of Money, Credit and Banking* 41(2/3):517–528.

Federal Reserve Bank of New York. 2023. "Primary Dealers". http://www.newyorkfed.org/markets/primarydealers, accessed 5th July.

Gensler, G. 2022. "'The Name's Bond:' Remarks at City Week". http://www.sec.gov/news/speech/gensler-names-bond-042622, accessed 9th August 2023.

Gharesifard, B., and S. L. Smith. 2018. "Distributed Submodular Maximization With Limited Information". *IEEE Transactions on Control of Network Systems* 5(4):1635–1645.

Hansen, E. A., D. S. Bernstein, and S. Zilberstein. 2004. “Dynamic Programming for Partially Observable Stochastic Games”. In *Nineteenth National Conference on Artificial Intelligence*. July 25th-29th, San Jose, California, 709-715.

Harsanyi, J. C. 1967. “Games with Incomplete Information played by ‘Bayesian’ Players, I–III Part I. The Basic Model”. *Management Science* 14(3):159–182.

Kosack, S., and A. Fung. 2014. “Does Transparency Improve Governance?”. *Annual Review of Political Science* 17:65–87.

Liang, E., R. Liaw, R. Nishihara, P. Moritz, R. Fox, K. Goldberg, J. Gonzalez, M. Jordan, and I. Stoica. 2018. “RLlib: Abstractions for Distributed Reinforcement Learning”. In *International Conference on Machine Learning*. July 10th-15th, Stockholm, Sweden, 3053-3062.

Littman, M. L. 1994. “Markov Games as a Framework for Multi-Agent Reinforcement Learning”. In *Machine Learning Proceedings*, edited by W. W. Cohen and H. Hirsh, 157–163. New Jersey: Elsevier.

Lundberg, S. M., and S. Lee. 2017. “A Unified Approach to Interpreting Model Predictions”. In *Advances in Neural Information Processing Systems 30*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, 4765–4774. Long Beach: Curran Associates, Inc.

Mizrach, B. 2015. “Analysis of Corporate Bond Liquidity”. http://www.finra.org/sites/default/files/OCE_researchnote_liquidity_2015_12.pdf, accessed 8th August 2023.

Schulman, J., F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. 2017. “Proximal Policy Optimization Algorithms”. http://arxiv.org/abs/1707.06347, accessed 9th August 2023.

Sen, A. 2018. *Collective Choice and Social Welfare*. 1st ed. Cambridge, Massachusetts: Harvard University Press.

Sen, A., M. A. Sen, J. E. Foster, S. Amartya, J. E. Foster et al. 1997. *On Economic Inequality*. 1st ed. Oxford, England: Oxford University Press.

Shakshuki, E., and M. Reid. 2015. “Multi-Agent System Applications in Healthcare: Current Technology and Future Roadmap”. *Procedia Computer Science* 52:252–261.

Shapley, L. S. 1953. “Stochastic Games”. *Proceedings of the National Academy of Sciences* 39(10):1095–1100.

Spaan, M. T. 2012. “Partially Observable Markov Decision Processes”. In *Reinforcement Learning: State-of-the-Art*, edited by M. Wiering and M. Otterlo, 387–414. Berlin, Germany: Springer.

Speicher, T., H. Heidari, N. Grgic-Hlaca, K. P. Gummadi, A. Singla, A. Weller, and M. B. Zafar. 2018. “A Unified Approach to Quantifying Algorithmic Unfairness: Measuring Individual & Group Unfairness via Inequality Indices”. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. August 19th-23rd, London, UK, 2239-2248.

Sutton, R. S., and A. G. Barto. 2018. *Reinforcement Learning: An Introduction*. 2nd ed. Cambridge, Massachusetts: MIT Press.

U.S. Treasury 2022. “Treasury Launches New Effort to Improve Resilience of its Market”. http://home.treasury.gov/news/press-releases/jy0831, accessed 8th August 2023.

Van der Cruijsen, C. A., S. C. Eijffinger, and L. H. Hoogduin. 2010. “Optimal Central Bank Transparency”. *Journal of International Money and Finance* 29(8):1482–1507.

Wah, E., M. Wright, and M. P. Wellman. 2017. “Welfare Effects of Market Making in Continuous Double Auctions”. *Journal of Artificial Intelligence Research* 59:613–650.

Walsh, C. 2007. “Optimal Economic Transparency”. *International Journal of Central Banking* 3(1):5–36.

Zheng, S., A. Trott, S. Srinivasa, D. C. Parkes, and R. Socher. 2022. “The AI Economist: Taxation Policy Design via Two-level Deep Multiagent Reinforcement Learning”. *Science advances* 8(18):eabk2607.

AUTHOR BIOGRAPHIES

KSHAMA DWARAKANATH is a Research Scientist at J.P. Morgan AI Research working on using reinforcement learning to design and learn trading agents with diverse objectives in simulated multi-agent markets. Her interests lie in the fields of reinforcement learning, multi-agent simulations and mechanism design. Her email address is kshama.dwarakanath@jpmorgan.com.

SVITLANA VYETRENKO is an Executive Director at J.P. Morgan AI Research leading a team focusing on generative time series models, multi-agent simulations and reinforcement learning. Her email address is svitlana.s.vyetrenko@jpmchase.com.

TOKS OYEBODE is an Executive Director in Regulatory Affairs for J.P. Morgan’s Corporate & Investment Bank (CIB). He leads CIB’s policy engagement on regulatory developments impacting derivatives and fixed income market structure. Toks was a Technical Specialist at the UK’s Prudential Regulation Authority. His email address is toks.oyebode@jpmorgan.com.

TUCKER BALCH is a Research Director at J.P. Morgan AI Research and a professor of Interactive Computing at Georgia Tech (on leave). He is interested in problems concerning multi-agent social behavior in domains ranging from financial markets to tracking and modeling the behavior of ants, honeybees and monkeys. His email address is tucker.balch@jpmchase.com.