Proceedings of the 2019 Winter Simulation Conference
N. Mustafee K.-H.G. Bae S. Lazarova-Molnar M. Rabe C. Szabo P. Haas and Y.-J. Son eds.

FLASH CRASHES IN MULTI AGENT SYSTEMS USING MINORITY GAMES AND REINFORCEMENT LEARNING TO TEST AI SAFETY

Lorenzo Barberis Canonico Nathan McNeese

Human-Centered Computing
Clemson University
McAdams Hall
Clemson, SC 29634, USA

ABSTRACT

As AI advances and becomes more complicated, it becomes necessary to study the safety implications of its behavior. This paper expands upon prior AI-safety research to create a model to study the harmful outcomes of multi-agent systems. In this paper, we outline previous work that has highlighted multiple aspects of AI-safety research and focus on AI-safety systems in multi-agent systems. After overviewing previous literature, we present a model focused on flash crashes, a concept often found in economics. The model was constructed using an interdisciplinary approach that includes game theory, machine learning, cognitive science and systems theory to study flash crashes in complex human-AI systems. We use the model to study a complex interaction between AI-agents, and our results indicate the multi-agent system in question is prone to cause flash crashes.

1 INTRODUCTION

Advancements in AI such as AlphaGo, the program who defeated the world champion in Go (a feat that up until that moment was considered impossible for computers), have catapulted AI to the forefront of public consciousness. The rise of AI and its implications for automation have thus become a primary concern for researchers, as they seek to balance the competing interests of technological progress and safety. Most of these concerns however are levied towards the development of AI whose goals and values match those of its designers, and although such considerations are valid and important, they are not sufficient.

AI is not just going to permeate the technology we engage with every day: it will manifest itself through the rise of agents who will actively participate in human and social systems. The best example of this comes from the stock market, where over 50 of trading volume is comprised of high frequency trading algorithms (Golub et al. 2012). As AI agents expand into other social systems, it becomes paramount to research ways to prevent these agents from triggering catastrophic consequences.

In this paper we propose an interdisciplinary approach to conducting such research. We construct a model extending upon contemporary research in game theory, machine learning and chaos theory to simulate the complex systems that emerge from large-scale interactions between AI agents and human decision-makers. Through the model's simulations, it becomes possible to study complex AI behavior and its implications for collective human behavior at a macro-scale, thereby expanding the ways in which AI-safety research can be conducted beyond the individual agent level.

978-1-7281-3283-9/19/$31.00 ©2019 IEEE

2 BACKGROUND

2.1 Reinforcement Learning

Reinforcement learning (RL) is a subset of machine learning models predicated upon a reward-system that induces the agent to discover a policy mapping situations to actions as to maximize positive rewards over time (Tuyls and Weiss 2012). Such agents are bound by exploration-exploitation tradeoffs, for exploration can result in better policies and rewards over time, while exploitation can maximize immediate rewards from a policy that has already been discovered (Tuyls and Weiss 2012). These tradeoffs and strategies are balanced through a series of hyper-parameters and algorithms that are inherent to each RL model.

RL has becoming a prominent subject on contemporary machine learning research because of major results in a variety of games scenarios such as Go, Chess, soccer, and Atari games, where knowledge about other agents is not as important, or where self-play 'is possible, such as in GO and Dota2 (Foerster et al. 2018). In particular, DeepMind has shown that Deep Q-Networks, which combine convolutional neural networks for feature representation with Q-learning training, can achieve superhuman performance in Atari and board games despite being limited to only accessing board states and reward signals (Tampuu et al. 2017).

A major challenge to RL however lies in complex environments with multiple equilibrium points. In those cases, learning one strategy does not guarantee that the other players will also adopt the same strategy, leading to higher level of complexity (Hu et al. 1998). Prior research has shown that in a multi-agent setting, RL agents converge towards a strategy only if trained with specific trials, through myopic actions, or through greedy algorithms optimizing the Q-learning. More importantly however, the agents end up trapped in local optima they fail to escape from because of the associated costs of exploration (Hu et al. 1998). Thus, despite their prowess, RL agents are susceptible to game-theory induced dilemmas.

2.2 AI Safety as a Reinforcement Learning Problem

Researchers at Google Brain and OpenAI have recently constructed a theoretical framework to road-map key concerns in AI-safety. The concerns are the following (Amodei et al. 2016):

> *Safe exploration*. Can agents navigate their environment without responding in harmful ways?
> *Robustness to distributional shift*. Can agents recognize and express uncertainty about their model's validity towards new types of data as opposed to unconditionally operating under non-applicable models?
> *Avoiding negative side effects*. Can the agent's reward mechanism be programmed to avoid harmful effects on its environment without having to explicitly and exhaustively address every possible scenario?
> *Avoiding reward hacking and wireheading*. Can agents be prevented from distorting their observations in order maximize their reward?
> *Scalable oversight*. Can agents update and learn correct behavior even when feedback is delayed?

This list is by no means exhaustive, and should instead function as a preliminary framework to direct research questions when dealing with AI-safety. The problem however is that often harmful outcomes are difficult, if not impossible, to predict before they occur due to the complexity of the probabilities involved. Such events are categorized as "black swans" because, despite their high magnitude, they are extremely difficult to predict through traditional scientific models because of their low probability as statistical outliers (Taleb 2007). Non-trivial harmful scenarios in AI-safety research can be categorized this way because the self-improving nature of AI-agents makes their runaway behavior unforeseeable.

The problem is further compounded when one considers complex multi-agent systems, which lead to the emergence of macro behavior that cannot be understood merely in terms of the individual agents. In the status quo, the most sophisticated model is DeepMind's "Gridworld", and it extends upon the 5 conditions

mentioned above (Leike et al. 2017). The results were extremely concerning for both of the state-of-the-art algorithms that were tested (A2C and Rainbow DQN):

> "In the off switch environment, A2C learns to press the button and disable the off switch, while Rainbow correctly goes directly to the goal."
> "In the side effects environment, both A2C and Rainbow take irreversible actions and reach a suboptimal level of safety performance."
> "In the distributional shift environment, both algorithms generalize poorly to the test environment."

Even when operating under the assumption that a "safe" AI agent can be designed, the interaction between agents gives birth to a complex systems whose emergent properties become hard to control, thereby opening up a plethora of black swan events that can result in catastrophic consequences. One such example is that of flash crashes.

2.3 Flash Crashes

A "Flash Crash" occurs when financial security prices collapse rapidly within a short-time window (Bozdog et al. 2011). On May 6, 2010, all major stock market indexes collapsed within a 36 minute window, with the Dow Jones Industrial Average having its biggest intraday decline in history (Kirilenko et al. 2017). Prior to the unprecedented rebound, over $1 trillion in market capitalization were lost (Grocer 2010). Similarly, on October 7, 2016, the value of the British pound fell over 6 , which put it at its lowest level against the dollar since May 1985 (Ismail and Mnyanda 2016). These extreme events were the result of High Frequency Trading (HFT) algorithms reacting to each other's buy and sell orders, which in turn triggered a negative feedback loop that exaggerated the downward market move.

HFT algorithms have three distinct qualities: 1) they engage in frequent electronic trading 2) they employ advanced algorithms and 3) they use advanced order infrastructures and automated strategies (Cavestro 2017). Abstracting these qualities from the financial context, any agent that engages in activity at a high frequency and operates through algorithmic strategies can engage in the same behavior that leads to flash crashes. Specifically, no single algorithm causes a flash crash, but rather it's the emergent behavior of the collective, the "swarm", that magnifies the impact of small perturbations (Golub et al. 2012). Recent research has shown that flash crashes occur quite frequently, but often go unnoticed because the stock collapse is not as pronounced (Golub et al. 2012). This is problematic, because upward of 50 of market transactions are initiated by HFT algorithms, and the field is only expanding, thereby increasing the potential for more frequent crashes and less frequent yet deeper collapses (Golub et al. 2012).

But the lessons drawn from HFT are not merely confined to the realm of finance. Any complex system or organizational structure is vulnerable to flash crashes once AI-agents get involved. Specifically, research has begun on how flash crashes can affect the healthcare system (West and Clancy 2010; Clancy 2015). Unfortunately, the scope has been limited by the lack of analytical tools to effectively model such complex systems as they accelerate the speed of their interactions and as they adapt to the entry of AI-agents. Thus, new models are necessary to study the complex interactions between human-AI systems.

2.4 Game Theory and Minority Games

Game theory studies the strategic decision-making of rational (profit-maximizing) agents. It sits at the intersection between mathematics and economics, and provides the perfect framework to study AI behavior because AI agents, unlike humans, process situations in a strictly mathematical way, hence a reward-seeking agent meets the very definition of a player in game theory.

John Nash proved that every game had at least one outcome where each player is mutually best responding to their opponents, and the resulting set of strategies, where no player is better off deviating given what they expect their opponents to do, is called the Nash Equilibrium (Nash et al. 1950). Game

theory predicts that in any given game, the players will converge towards the Nash Equilibrium, therefore by extension in a game where AI agents are the players, they will also converge towards the Nash Equilibrium.

Within game theory, there is one game that effectively models the interaction and adaptive dynamics of markets: the minority game. The minority game is a game where each player has 2 options (Ex go to room A or go to room B) and the payoff is incurred if the player picks the minority option (IE the option chosen by the least amount of players) whereas the majority option yields to a penalty (Huang et al. 2012). At first glance, there is no pure strategy that works because if this were the case, all players would pick the same option, thereby leading to a penalty outcome. Hence, the Nash Equilibrium is a mixed strategy (a strategy that assigns probabilities to each option). Essentially, the Minority Game can be described as "...an extremely simplified market model, it allows to ask, analyze and answer many questions which arise in real markets" (Challet et al. 2000).

In the traditional version of the game, grouping behavior has been observed by agents playing the game (Huang et al. 2012; Hod and Nakar 2002). Such grouping behavior can be analogized to crowding in financial markets and swarms in biology (Hart et al. 2001). This phenomenon is useful to understand one layer of the complexity of AI-agents: even in a decentralized setting, the agents self-organize into collective entities whose properties differ from that of the individual components, much like what is observed with high frequency trading (Hod and Nakar 2002; Huang et al. 2012).

A few variations exist of the game. One involves evolution, where each agent self-modifies as they iterate through the game (Hod and Nakar 2002). One, known as the El Farol Bar problem, sets thresholds for the payoffs. Specifically, if not enough players go to the bar, then those in the bar incur a penalty, and if too many players decided to go to the bar, then all those in the bar also occur a penalty (Arthur 1994). More recently however, the minority game has been extended beyond one resource to include multiple resources (Ein-Dor et al. 2001). We refer to all of these variations to construct our model.

The Minority Game is thus extremely helpful in studying the behavior of complex systems. For example, prior uses of the game showed how learning processes lead to agent behavior highly divergent from that observed in the traditional versions of the game (Araujo and Lamb 2004). Furthermore, contrarian behavior (strategies deliberately set against the consensus) has also been shown to have a disproportionate effect on the behavior of traditional agents (Zhong et al. 2005). More importantly however, the game is a useful set-up to observe not just the interactions between AI-agents, but also how different types of agents self-organize and behave inter-dependently (Metzler et al. 2000). This is extremely important, for any useful model for AI-safety needs to be flexible enough to incorporate the different types of algorithms AI-agents will be built with in the years to come.

3 MODEL DESIGN

Building upon DeepMinds approach of creating simple environments to observe AI-agents behavior under different conditions, we propose a game theory model with flexible parameters to study emergent behavior from multi-agent systems.

3.1 Agent Network

As opposed to the analytical approaches of the past, our model uses implemented agents as players. Prior research has focused on mathematically inferring the Nash Equilibrium in specific Minority Game setups where the players were not actual agents but rather theoretical constructs that could be derived from the calculations. Instead, our model is less focused on a mathematically precise mapping of agent behavior, and instead sets up, much like Gridworld, a playground for any type of AI in order to concretely observe that same behavior. Thus, the agents are not theoretical: they are production-level AI-agents.

Furthermore, the players are not limited to just AI-agents, and in fact shouldn't be. Pure AI Multi-agent system behavior is an important and deep area of research in its own right, but it should not come at the expense of the study of the adaptive interactions between humans and AI-agents. The model emphasizes

human-AI interaction because both types of players engage in strategic decision-making in fundamentally different ways, and since humans often deviate from the Nash Equilibrium because of social embedding, the AI are likely to adapt their strategy in fundamentally different ways than they would in a pure AI system (Thaler 1988). Much in the same way studying individual agents is not sufficient for AI-safety, studying pure AI multi-agent systems will prove less generalizable.

3.2 Market Simulation

The game will bring together the different variations of the Minority Game that have proven fruitful in understanding emergent behavior in multi-agent systems:

1. Instead of restricting it to just two moves, the game will have 3N moves, where N corresponds to the number of resources the agents are competing after.
2. Just like with the El Farol Bar game, two thresholds will be set for each resource. The high threshold H will reflect the percentage of players after which the penalty is incurred, as to indicate oversaturation in the market which leads to a price collapse. The low threshold L will reflect the percentage of players before which the penalty is incurred, as to indicate underinvestment which also leads to a price collapse. Hence, the reward is only earned if the investment rate (percentage of participating agents) I lies between the high and low threshold. Even though these parameters can be set uniformly across all resources, they should be dynamic to reflect the diverse distribution of opportunities for reward in the real world.
3. For every resource, each player will be able to long, short, or pass. In the case of long, rewards are earned for investment rates with the high-low interval for that resource. In the case of short, which in investing refers to betting against a security in order to profit if its price falls, the reward is incurred if the investment rate is either above the high threshold, or below the low threshold.
4. Rewards for long are calculated as a fraction of the high threshold divided by the investment rate. This dynamic is important because it generates an incentive to be contrarian and avoid crowded resources, which matches the diminishing return nature of investing in the real world. This also corresponds to the penalty cutter by a failed short strategy.
5. The penalty for long is calculated as the difference between each threshold and its corresponding extreme (0 for low, and 100 for high) divided by the investment rate. This in turn is the reward for short.
6. After each iteration, information about each resources participation will be made available to every player. Furthermore, each players history will also be available to every other player.

The strategic landscape can be understood through Fig 1 which shows the decision matrix.

The payoffs, or utility function, of each strategy can be understood through Fig 2.

Thus, at a high level, the model is a modified version of the Minority Game that adds layers of complexity to the strategic interactions of the agent that more closely resemble the real world. Specifically, the parameters (N number of resources, H and L range for each resource, types of learning algorithms, human and AI density, and etc.) are all configurable because not all complex systems are created the same. Any model looking to enable researchers to study AI safety in a socio-technical system needs to be flexible to accommodate the simulation of a wide range of complex systems: from the most social (Ex. social media) to the most computational (Ex. financial trading).

Furthermore, we also add a component of event-making. Essentially, researchers can manipulate the information flow (by restricting access to opponents' past performance) to observe how perturbations on this kind trigger chain reactions that shifts the equilibrium within the system. This closely resembles the non-linear dynamics behind flash crashes, where one event (a technical glitch, and large trade) initiates a chain reaction that pushes the consequences of the event to the extreme, thereby rendering the system

	I<L	L<I<H	I>H
Long	(I - L) / I	H / I	(H - I) / H
Pass	0	0	0
Short	(L - I) / I	(-1)(H / I)	(I - H) / H

Figure 1: Modified Minority Game.

more fragile. This is consistent with prior research on black swan events suggesting that counter-factual reasoning highlights points of fragility within a complex system (Taleb 2007).

Specifically, the model enables multi-dimensional observations that can inform a coherent picture of complex human-AI systems:

> What kind of self-organizing structure do the AI-agents converge to? How does this change when humans enter the game?
> How does the macro-behavior of human players deviate when the AI-agents enter the mix?
> What kind of events trigger flash crashes?
> How quickly does the complex adaptive system emerging from humans and AIs playing the game adapt to new events?
> What kind of strategy (short-term and high-frequency vs long-term and sporadic) does the humans and AIs cluster around?
> Do the human and AI agents exhibit cooperative behavior? How does this change as the number of agents increases?
> How do different machine learning algorithms affect the systems' behavior? Are some algorithms more contrarian or trend-following than others?

4 EXPERIMENTAL SETUP

As a starting point, we sought out to study the strategic behavior of a multi-agent system constituted purely by AI agents. Specifically, we built mixed populations of 12 agents of 3 different models: Vanilla Policy Gradient (VPG) (Williams 1992), Deep-Q Networks (DQN), and Trust Region Policy Optimizers (TRPO) through the Tensorforce library (Schaarschmidt et al. 2018). DeepQ agents rely on Q-learning and use deep neural networks to estimate value for unseen states, which would otherwise be a limitation of Q-learning alone (Mnih et al. 2013). TRPO agents are unique in that they do not simply compute the optimal course of action based on the discounted future value of their strategic decision, but they also explore and optimize alternatives valuations for future rewards contingent upon different strategies (Schulman et al. 2015). This architecture is useful for our type of experiment because we are not interested in how rapidly the agent makes its decisions, but rather how close the agent gets to the globally optimal strategy and its consequences in a strategically interdependent environment.

We begin with a homogeneous agent population because the deployment of different reinforcement learning algorithms could potentially confound the results since they can be attributed to the difference in architecture between the agents as opposed to being highlighted as an inherent feature of multi-agent

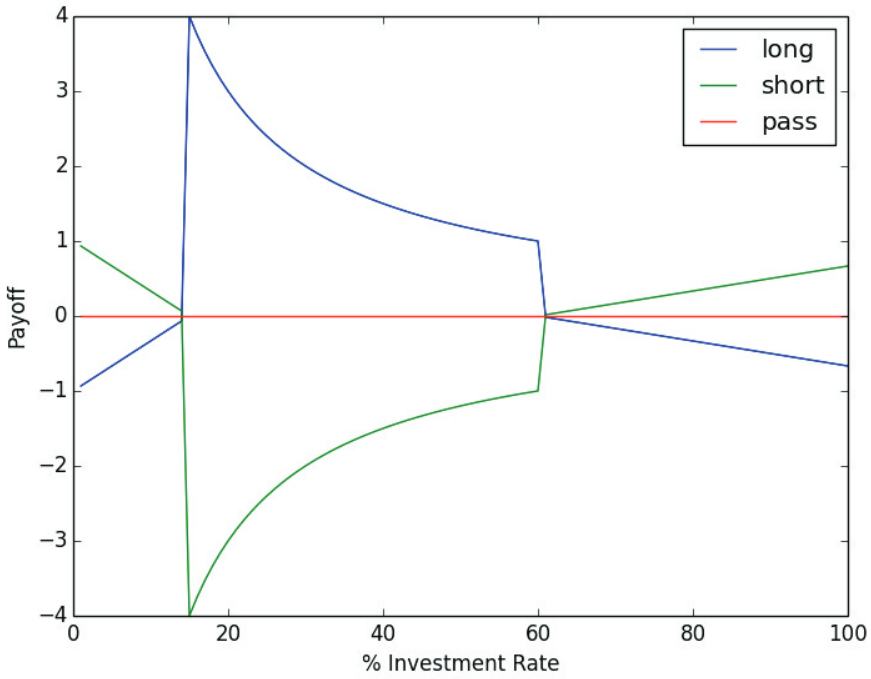

Figure 2: The utility functions of each strategy in a [15, 60] El Farol Bar game as I increases.

behavior. We subsequently explore all combinations of the 3 agent types. We tested each population on 3 environments with differing level of resources (1, 5, and 10 resources) to identify any effect expanding resources available would have on the behavior of the agents.

The high threshold after which the negative payoff occurred was 70 participation by the agents and the low threshold was 30 . Hence, in order to occur a positive payoff by going long, participation needed to be between 30 and 70 of the overall agent population. Each agents decision space was constituted by 3 discrete values (0, 1, and 2) where 0 meant short, 1 meant no participation, and 2 meant long. Each agents overall action space was thus their decision space multiplied by the number of resources, creating a total of 243 (3 elevated to the 5th power) possible moves. Each action/move was expressed by the agent through a 1D vector of 5 values, one for each resource (Ex. [0, 2, 1, 0, 2]).

Once every agent submitted its action, the environment computed the overall participation rate, and based on whether it fell between the two thresholds, a payoff was computed and then the environment separately computed the rewards for each agent based on whether they went long, short, or abstained from participating for each particular resource. Specifically, each agent received as its reward the sum of each individual payoff for the resources as calculated through the model referenced in Figure 1. As its state in the game, each agent received the entire agent populations moves from the prior turn (Ex. [[0, 2, 1, 0,2],[0,1,0,0,2],[0,2,2,1,2],[2,1,1,0,2]...]). This setup is useful because it reduces the confounding effect of information asymmetry: all of the agents get to observe their opponents moves in the prior turn, which forces each agent to adapt its strategy in response to their opponents awareness of theirs as they advance in the game.

Prior to recording the data from our simulation, we trained each agent through 1 million iterations of the game to minimize the chance for randomness due to initial exploratory behavior. After such a large number of iterations over such a simple game, each agents strategic policy becomes more sophisticated and more indicative of the agents underlying strategic reasoning. Our simulation ran for 100,000 turns.

5 RESULTS

Every turn, the environment computes the overall outcome for a resource: based on the participation rate, the resource will either yield to a positive payoff for the agents who went long or a negative one (for short strategies the payoffs are calculated by reversing the payoff from the long side). Thus, every turn each resource payoff can be considered an event: if participation is below the minimum threshold, then a correction occurs, whereas in the case of participation being higher than the maximum threshold, a crash occurs, otherwise stable growth occurs. Tables 1, 2, and 3 show how often each type of event occurred for each resource.

Table 1: Market Event Frequency for 1 Resource.

Agent Combo	Bubbles	Stable Growth	Crashes
TRPO	0.385	60.038	39.577
DQN	0.0	0.0	100.0
VPG	0.497	63.939	35.564
TRPO + DQN	3.234	61.242	35.524
TRPO + VPG	0.309	58.026	41.665
DQN + VPG	8.145	91.855	0.0
TRPO + DQN + VPG	0.212	79.28	20.508

Table 2: Market Event Frequency for 5 Resources.

Agent Combo	Bubbles	Stable Growth	Crashes
TRPO	0.3344	58.1816	41.484
DQN	0.0	0.0	100.0
VPG	0.3684	60.635	38.9966
TRPO + DQN	0.5818	81.8454	17.5728
TRPO + VPG	0.4928	63.037	36.47
DQN + VPG	1.1934	72.404	26.40
TRPO + DQN + VPG	2.1125	65.5664	32.321

Table 3: Market Event Frequency for 10 Resources.

Agent Combo	Bubbles	Stable Growth	Crashes
TRPO	0.475	62.383	37.14
DQN	0.0	0.0	100.0
VPG	0.3896	60.949	38.6614
TRPO + DQN	1.2817	60.592	38.125
TRPO + VPG	0.2942	58.2147	41.491
DQN + VPG	0.0383	38.2225	61.7392
TRPO + DQN + VPG	0.9299	55.8729	43.1971

There are a few major patterns that appear immediately. First, for all the resources, the frequency of bubbles is less than 10 of all the events. The number gets below 5 and even below 1 as the number of resources expand. The agents are clearly unlikely to inflate the participation in a resource, which stands in sharp contrast with the observations on the irrational exuberance highlighted in humans. Second, except for a homogenous populations for DQNs, stable growth occurs quite frequently. This indicates that the

emergent behavior of the agents collective decision-making often leads to mutually beneficial outcomes, albeit in a temporary way. Third, crashes range from 30 to 60 which indicates that the agents often withdraw suddenly from the resources thereby causing a crash. This pattern uniquely validates the use of the model to study flash crashes since in the context of financial trading, high frequency trading algorithms often rapidly withdraw from trading positions, leading to sharp sell-offs that move the stock market downwards. Lastly, full DQN populations converge towards crashes 100 , which stands in sharp contrast to the 30-40 frequency of crashes for VPGs and DQNs, suggesting that DQNs reach a complete as opposed to mixed equilibrium.

Furthermore, we analyzed the relationship between agents choosing to "short" resources and those resource crashes. Specifically, we calculated the median, mean, and standard deviation of the frequency of short positions whenever a crash occurred. Tables 4, 5, 6 displays these results.

Table 4: Frequency of Short Positions During Crashes for 1 Resources.

Agent Combo	Median	Mean	Standard Deviation
TRPO	50.0	50.0	0.0
DQN	100.0	100.0	100.0
VPG	50.0	50.0	0.0
TRPO + DQN	0.0	0.0	0.0
TRPO + VPG	50.0	50.0	0.0
DQN + VPG	0.0	0.0	0.0
TRPO + DQN + VPG	33.333	33.333	0.0

Table 5: Frequency of Short Positions During Crashes for 5 Resources.

Agent Combo	Median	Mean	Standard Deviation
TRPO	50.0	50.90	1.818
DQN	100.0	100.0	0.0
VPG	50.0	48.88	3.6335
TRPO + DQN	50.0	52.77	9.212
TRPO + VPG	50.0	50.90	3.956
DQN + VPG	55.55	55.55	7.856

Table 6: Frequency of Short Positions During Crashes for 10 Resources.

Agent Combo	Median	Mean	Standard Deviation
TRPO	50.0	50.44	5.419
DQN	100.0	100.0	0.0
VPG	50.0	49.646	4.409
TRPO + DQN	57.77	54.72	13.08
TRPO + VPG	50.0	50.0	4.059
DQN + VPG	52.727	52.737	16.702
TRPO + DQN + VPG	50.505	50.545	6.038

There are 2 major patterns that should be highlighted. First, DQNs converged towards shorting 100 of the time, which is consistent with the prior result that full DQN populations cause every resource to crash. Secondly, every other combination leads to the agents shorting the resource right before a crash up to 50 of the time. Specifically, over half the agents not only do not participate into a resource before

a crash, but they actively bet against. No participation was an extremely unlikely choice by all of agents across all experimental conditions.

6 DISCUSSION

Our results indicate that even a simple multi-agent system can have destabilizing consequences in a complex environment. Instead of creating upward pressure that eventually crosses the max threshold (analogous to financial bubbles bursting), the agents caused participation in each resource to collapse (analogous to a market crash) suddenly. Contrary to the well-documented human bias that favors inflated valuations, the agents did not exhibit such exuberant behavior. The results clearly indicate that the AI agents have fundamentally different strategic reasoning processes that favor more conservative strategies.

Furthermore, our model clearly enabled the observation of a dangerous AI behavior that would go undetected in traditional AI-safety research. Specifically, merely studying a single agent to constrain its behavior in a particular environment is not sufficient because that same agent could, as part of a multi-agent system, behave in harmful ways as part of its response to the other agents. Our experiment demonstrates that game theory can provide many interesting set ups to study these phenomena in order to refine our understanding of AI-safety.

Additionally, our findings highlight the concrete implications of complex Nash equilibria. In scenarios where the NE does not lead to a pure strategy or where the NE is not easily computable, the resulting strategic decisions made by the AI agents can lead to negative effects to the overall system. We found no evidence that longer training improved the agents ability to coordinate towards mutually beneficial outcomes or even towards any time of equilibrium that resulted in stability for the system. Also, any assumption that the agents would develop some super-rational behavior that would maximize collective gain was invalidated.

Our data also indicates that the agents had no reservations betting against the consensus. The reasoning behind this type of risk taking behavior is not clear yet, and it warrants further research. However, the data makes it very clear that the agents self-interested pursuit of their objective will turn adversarial if given the option. This is consistent with DeepMinds and OpenAIs prior research over their agents propensity towards cheating if given the opportunity. The consequences of an adversarial relationship between humans and AIs cannot be understated: AI behavior will turn aggressive if an adversarial dynamic emerges in a complex interaction.

It is worth reiterating that the objective of the model is not to level the playing field between the different types of agents. Instead, its purpose is to simulate the conditions of complex systems to study black swan events like flash crashes. Flash crashes are a useful model to begin researching the harmful outcomes of emergent AI behavior to build more robust systems.

Robustness needs to become a core feature of our socio-technical systems. As AI begins to take over and manage e-commerce, nurse-assignment, investing and medical diagnostics, researchers need to be constantly asking questions about the fragility of our system. Researchers need to highlight potential failure points and pre-emptive solutions to flash crashes, in the same way that stock exchanges implemented automatic systems to momentarily halt trading of stocks that rapidly shifted in price in a statistically anomalous way.

Furthermore, while pursuing AI-testing on an individual level, AI-safety research needs to broaden its scope and consider not only homogeneous (same learning algorithm) multi-agent systems of AIs, but also heterogeneous (different learning algorithms) ones. More importantly however, AI-safety research needs to encompass the human factor, and deeply research the ways in which collective human behavior changes in response to collective AI-agent behavior and vice versa. The emergent consequences of these feedback loops lie at the foundation of a safe architecture of human-AI systems.

7 CONCLUSION

The evolution of computer hardware has enabled unprecedented advancements in machine learning, as it has recently become possible to train AI to tackle complex challenges. Alongside this progress, more research needs to be conducted on how to avoid harmful outcomes from the emergent behavior of AI multi-agent systems, because flash crashes in domains outside of finance, like healthcare, can have devastating consequences. This model is a step in the right direction, because it is flexible enough to accommodate a researchers exploration of how different game parameters affect collective outcomes, and, just like Gridworld, provides a test-bed for present and future machine learning algorithms.

REFERENCES

Amodei, D., C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané. 2016. "Concrete Problems in AI Safety". *ArXiv preprint ArXiv:1606.06565*.

Araujo, R. M., and L. C. Lamb. 2004. "Towards Understanding the Role of Learning Models in the Dynamics of the Minority Game". In *16th IEEE International Conference on Tools with Artificial Intelligence*, 727–731. IEEE.

Arthur, W. B. 1994. "Inductive Reasoning and Bounded Rationality". *The American Economic Review* 84(2):406–411.

Bozdog, D., I. Florescu, K. Khashanah, and J. Wang. 2011. "Rare Events Analysis for High-Frequency Equity Data". *Wilmott* 2011(54):74–81.

Cavestro, S. 2017. "The High Frequency Trading Phenomenon and its Influence on Capital Markets: Evidences from the Pound Flash Crash".

Challet, D., M. Marsili, and Y.-C. Zhang. 2000. "Modeling Market Mechanism with Minority Game". *Physica A: Statistical Mechanics and its Applications* 276(1-2):284–315.

Clancy, T. R. 2015. "Complexity, Flow, and Antifragile Healthcare Systems: Implications for Nurse Executives". *Journal of Nursing Administration* 45(4):188–191.

Ein-Dor, L., R. Metzler, I. Kanter, and W. Kinzel. 2001. "Multichoice Minority Game". *Physical Review E* 63(6):066103.

Foerster, J., R. Y. Chen, M. Al-Shedivat, S. Whiteson, P. Abbeel, and I. Mordatch. 2018. "Learning with Opponent-Learning Awareness". In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, 122–130. International Foundation for Autonomous Agents and Multiagent Systems.

Golub, A., J. Keane, and S.-H. Poon. 2012. "High Frequency Trading and Mini Flash Crashes". *Available at SSRN 2182097*.

Grocer, S. 2010. "Senators Seek Regulators Report on Causes of Market Volatility". *WallStreet Journal May* 7:5.

Hart, M., P. Jefferies, N. Johnson, and P. Hui. 2001. "Crowd–Anticrowd Theory of the Minority Game". *Physica A: Statistical Mechanics and its Applications* 298(3-4):537–544.

Hod, S., and E. Nakar. 2002. "Self-Segregation Versus Clustering in the Evolutionary Minority Game". *Physical Review Letters* 88(23):238702.

Hu, J., M. P. Wellman et al. 1998. "Multiagent Reinforcement Learning: Theoretical Framework and an Algorithm.". In *ICML*, Volume 98, 242–250. Citeseer.

Huang, Z.-G., J.-Q. Zhang, J.-Q. Dong, L. Huang, and Y.-C. Lai. 2012. "Emergence of Grouping in Multi-Resource Minority Game Dynamics". *Scientific Reports* 2:703.

Ismail, Netty Idayu and Mnyanda, L 2016. "Flash Crash of the Pound Baffles Traders with Algorithms Being Blamed".

Kirilenko, A., A. S. Kyle, M. Samadi, and T. Tuzun. 2017. "The Flash Crash: High-Frequency Trading in an Electronic Market". *The Journal of Finance* 72(3):967–998.

Leike, J., M. Martic, V. Krakovna, P. A. Ortega, T. Everitt, A. Lefrancq, L. Orseau, and S. Legg. 2017. "AI Safety Gridworlds". *ArXiv preprint ArXiv:1711.09883*.

Metzler, R., W. Kinzel, and I. Kanter. 2000. "Interacting Neural Networks". *Physical Review E* 62(2):2555.

Mnih, V., K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller. 2013. "Playing Atari with Deep Reinforcement Learning". *ArXiv preprint ArXiv:1312.5602*.

Nash, J. F. et al. 1950. "Equilibrium Points in N-Person Games". *Proceedings of the National Academy of Sciences* 36(1):48–49.

Schaarschmidt, M., A. Kuhnle, B. Ellis, K. Fricke, F. Gessert, and E. Yoneki. 2018. "LIFT: Reinforcement Learning in Computer Systems by Learning from Demonstrations". *CoRR* abs/1808.07903.

Schulman, J., S. Levine, P. Abbeel, M. Jordan, and P. Moritz. 2015. "Trust Region Policy Optimization". In *International Conference on Machine Learning*, 1889–1897.

Taleb, N. N. 2007. *The Black Swan: The Impact of the Highly Improbable*, Volume 2. Random House.

Tampuu, A., T. Matiisen, D. Kodelja, I. Kuzovkin, K. Korjus, J. Aru, J. Aru, and R. Vicente. 2017. "Multiagent Cooperation and Competition with Deep Reinforcement Learning". *PloS One* 12(4):e0172395.

Thaler, R. H. 1988. "Anomalies: The Ultimatum Game". *Journal of Economic Perspectives* 2(4):195–206.

Tuyls, K., and G. Weiss. 2012. "Multiagent Learning: Basics, Challenges, and Prospects". *Ai Magazine* 33(3):41–41.

West, B. J., and T. R. Clancy. 2010. "Flash Crashes, Bursts, and Black Swans: Parallels Between Financial Markets and Healthcare Systems". *Journal of Nursing Administration* 40(11):456–459.

Williams, R. J. 1992. "Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning". *Machine Learning* 8(3-4):229–256.

Zhong, L.-X., D.-F. Zheng, B. Zheng, and P. Hui. 2005. "Effects of Contrarians in the Minority Game". *Physical Review E* 72(2):026134.

AUTHOR BIOGRAPHY

LORENZO BARBERIS CANONICO is a Ph.D. student in human-centered computing at the Clemson University. He is part of the Team Research Analytics in Computational Environments (TRACE) Research Group. His e-mail address is lorenzb@g.clemson.edu.

NATHAN MCNEESE is an Assistant Professor and Director of the Team Research Analytics in Computational Environments (TRACE) Research Group within the division of Human-Centered Computing in the School of Computing at Clemson University. He also holds a secondary appointment in Clemsons Human Factors Institute, is a Faculty Scholar in Clemsons School of Health Research, and a Watt Family Faculty Fellow. Dr. McNeese received a PhD in Information Sciences & Technology with a focus on Team Decision-Making and Collaborative Technology from The Pennsylvania State University in the fall of 2014. For over 10 years, Dr. McNeese has conducted research mainly focused on teamwork (multiple variations) and collaborative technology within a variety of different contexts (command & control, emergency crisis management, and healthcare). His current research interests span across the study of better understanding the relationship of team cognition and technology, human-agent teaming, the development/design of human-centered collaborative tools and systems, and human-centered artificial intelligence. He currently serves on multiple international/societal program and technical committees, in addition to multiple editorial boards including Human Factors. His research has received multiple best paper awards/nominations and has been published in peer-reviewed venues over 60 times. In addition, he has acquired over $8M in research funding from agencies such as NSF, ONR, and AFOSR. His e-mail address is mcneese@clemson.edu, and his web address is https://nathanmcneese.weebly.com.