RAFAEL MOLINARI CHEANG

A CENTRALIZED NORM SYNTHESIS AND NORM ENFORCEMENT FRAMEWORK FOR GOVERNING MIXED-MOTIVE MULTIAGENT REINFORCEMENT LEARNING ENVIRONMENTS

São Paulo
2023

RAFAEL MOLINARI CHEANG

A CENTRALIZED NORM SYNTHESIS AND NORM ENFORCEMENT FRAMEWORK FOR GOVERNING MIXED-MOTIVE MULTIAGENT REINFORCEMENT LEARNING ENVIRONMENTS

Dissertation presented to Escola Politécnica of Universidade de São Paulo to obtain Master of Sciences degree in Electrical Engineering.

São Paulo
2023

RAFAEL MOLINARI CHEANG

A CENTRALIZED NORM SYNTHESIS AND NORM ENFORCEMENT FRAMEWORK FOR GOVERNING MIXED-MOTIVE MULTIAGENT REINFORCEMENT LEARNING ENVIRONMENTS

Final Version

Dissertation presented to Escola Politécnica of Universidade de São Paulo to obtain Master of Sciences degree in Electrical Engineering.

Concentration Area:

Computer Engineering

Advisor:

Prof. Dr. Jaime Simão Sichman

São Paulo
2023

Este exemplar foi revisado e corrigido em relação à versão original, sob responsabilidade única do autor e com a anuência de seu orientador.

São Paulo, 30 de janeiro de 2024

Assinatura do autor: _____

Assinatura do orientador: _____

Catalogação-na-publicação

Cheang, Rafael Molinari
A centralized norm synthesis and norm enforcement framework for governing mixed-motive multiagent reinforcement learning environments / R. M. Cheang -- versão corr. -- São Paulo, 2024.
70 p.

Dissertação (Mestrado) - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Computação e Sistemas Digitais.

1.Jogos de incentivo misto 2.Aprendizado por reforço 3.Agentes normativos I.Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia de Computação e Sistemas Digitais II.t.

ACKNOWLEDGMENTS

I am grateful to my family for their support, love, and sacrifices that made this academic journey possible. To my fiancée Ana Flavia, your patience, encouragement, and belief in me have been my anchor during challenging times. I would also like to thank my friends Marcos Menon and Flavio Nakasato, that helped me with throughout this process with many insightful conversations, and my advisor, Jaime, for his invaluable guidance and mentorship, which played a pivotal role in shaping this research. I am deeply thankful to each of you for the support that has carried me through this transformative experience.

This research has been carried out with the support of *Itaú Unibanco S.A.*, through the scholarship program of *Programa de Bolsas Itaú* (PBI), and it is also financed in part by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Finance Code 001, Brazil.

Any opinions, findings, and conclusions expressed in this manuscript are those of the authors and do not necessarily reflect the views, official policy or position of Itaú-Unibanco and CAPES.

RESUMO

Jogos de incentivos mistos compreendem um subconjunto de jogos em que os incentivos individuais e coletivos não estão totalmente alinhados. Esses jogos são relevantes porque ocorrem com frequência no mundo real, bem como em sistemas multiagentes, e seus resultados poderiam ser melhores para as partes envolvidas caso aspectos coletivos fossem considerados. Instituições e normas oferecem boas soluções para governar sistemas com incentivos mistos, mas na literatura, são usualmente estudadas e incorporadas de forma distribuída. Neste trabalho, propomos um *framework* para melhorar os resultados coletivos obtidos em ambientes de aprendizado por reforço multiagente de incentivos mistos. O *framework* propõe aprimorar o ambiente com um sistema normativo controlado por um agente externo de aprendizado por reforço. Ao empregá-lo, mostramos que é possível alcançar bem-estar social usando apenas arquiteturas tradicionais de agentes de aprendizado por reforço, mesmo em um sistema formado apenas por agentes egoístas.

Palavras-Chave – Jogos de incentivo misto, Aprendizado por reforço, Agentes normativos.

ABSTRACT

Mixed-motive games comprise a subset of games in which individual and collective incentives are not entirely aligned. These games are relevant because they can be matched to frequently occurring events in the real-world, as well as in multiagent systems, and their outcomes could be better for the involved parties if collective aspects were considered. Institutions and norms offer good solutions for governing mixed-motive systems, but in the literature, they are usually studied and incorporated into the system in a distributed fashion. In this work, we propose a framework for reaching socially good outcomes in mixed-motive multiagent reinforcement learning environments by enhancing the environment with a normative system controlled by an external reinforcement learning agent. By employing this framework, we show that it is possible to reach social welfare using only traditional reinforcement learning agent architectures, even in a system of self-interested agents.

Keywords – Mixed-motive games, Reinforcement learning, Normative agents.

LIST OF FIGURES

LIST OF TABLES

LIST OF ALGORITHMS

CONTENTS

1 INTRODUCTION

Mixed-motive games comprise a subset of games in which individual and collective incentives are not entirely aligned. These games are defined by two basic properties (DAWES, 1980): *a)* every individual is incentivized to socially defect and *b)* all individuals are better off if all cooperate than if all defect.

Opposing the view that groups will find ways to act so as to serve their own interests — as individuals often do —, when their incentives point to a different direction than that of their members, a collective action problem may emerge (OLSON, 1965) and drive the whole system to a state socially unwished-for.

Olson (1965) develops the notion of a collective action problem starting from the *raison d'etre* of *organizations*. These, as Olson describes, are groups that serve to further the interests of their members. The problem emerges when the individuals of such groups also have antagonistic incentives to those common to the collective. In this case, individuals are left to choose between harming the organization as a whole in favor of their own benefit or passing on the opportunity for bigger gains in favor of the group. A collective action problem happens when the former is systematically preferred over the latter.

Consider a market competition scenario: every competitor has, at the same time, the incentive to keep prices high, which is a function of the market's supply — the lower the supply, the higher the price —, and the incentive to increase supply in order to increase revenue. In case most competitors opt to increase their supplies, prices will be driven down, which is a bad outcome for the collective[1].

Global warming is another case of the collective action problem emerging from mixed incentives. In this scenario, every player — be it an individual, institution, or government — have an incentive to emit as much greenhouse gases as desired — for matters of comfort, financial gains, or popularity —, regardless of how much others are emitting. If to these ends the collective emissions surpass some threshold, the system increasingly dips into an

[1] It is a bad outcome among the group of firms within such market, not necessarily among the society as a whole.

undesirable state that is bad for everyone.

The collective action problem is not particular to communities of people in the real world, it may also happen in systems of artificial agents known as multiagent systems (MAS). In the context of MAS, the collective action problem is all so relatable to the issue of maintaining the system's social order (CONTE, 2001; CASTELFRANCHI, 2000). Castelfranchi (2000) defines social order as the problem of reaching desirable, predictable, and stable emergent outcomes from local actions in a system composed of agents with different beliefs and goals. Preserving social order is not only an issue in mixed-motive games as difficulties in coordinating actions can also hinder a system's stability when agents agree to cooperate with each other (WOOLDRIDGE, 2009, p. 200), but it is especially pervasive in these types of games as their misaligned incentives become another obstacle.

As the agents' learning capabilities improve with the advent of technologies such as reinforcement learning (RL), so increases their ability to optimize for their own benefit. This is reminiscent of the motto "*people respond to incentives*" (MANKIW, 2018, p. 7), which is the root cause of the collective action problem in the real world. Thus, as the learning capabilities of artificial agents increase, the chances of falling into the collective action problem in mixed-motive MAS also increase.

Social norms and norm enforcement mechanisms are tools of an institutional machinery that can be used for governing mixed-motive systems in order to sustain its social order (VERHAGEN, 2000). These can be implemented either in centralized — when a central governing body is tasked with running the institutional apparatus by itself — or decentralized fashion — when the normative system is conducted by the agents themselves.

1.1 Motivation

Decentralized approaches have been used in the past to prevent the collective action problem in MAS (HUGHES et al., 2018; ECCLES et al., 2019). However, decentralized solutions either imply *a)* pro-social behavior from the the agents or *b)* some form of direct or indirect retaliatory capacity — i.e. having the choice not to cooperate in future interactions. We acknowledge the effectiveness of these mechanisms in some cases but also recognize they are no *panacea*.

For instance, how can one — agent or group of agents — successfully drive a complex

MAS towards social order from within without assuming anything about others' beliefs, intentions, or goals, given that punishing uncompliant behavior is not desirable or allowed? This problem is akin to many situations in modern society; thus far it is impossible to know the beliefs and intentions of every person we might interact with, and not every problem we face is ideally solvable by a "taking matters into own hands" approach.

Consider as an example the problem with burglary. As a society, we don't expect social norms and good moral values to solve the problem completely, although they certainly change the rate at which it happens. Similarly, we don't expect the victim of a burglary to return the favor with a response of similar intensity — like stealing from the aggressor's house.

A similar issue may also occur in MAS. Consider a system of self-driving autonomous vehicles. Every vehicle has an incentive to get to its destination as fast as possible. If, to this end, a vehicle engages in careless maneuvers and risky overtakes to gain a few extra seconds, how could another vehicle sharing the road respond to this non-compliant behavior?

We could assume that all agents in this system are pro-social to some degree, and thus, such an event would never happen. Preventing socially bad outcomes by having agents acting empathically is an option that has been explored (HUGHES et al., 2018; CHEN; WANG, 2019). However, this might not always be a good premise. In the above example, the system itself is embedded in a competitive environment of firms fiercely fighting for market share. Performance, in the form of getting to the destination faster, might represent getting a bigger slice of the pie. Does the designer behind the agent have the right incentives to design altruistic agents? Social defection for the sake of financial gains is not unthinkable by any means in the automobile industry[2].

Alternatively, we could endow agents with the ability to punish defection, thus changing the expected payoff of such recklessness (AXELROD, 1984). But could any form of punishment be accomplished without compromising the safety of passengers? Furthermore, even if we agree upon the safety of reciprocating, there are many situations where direct retaliation might be undesirable. For instance, how do we address fairness in these systems? If highly interconnected, even a small violation could be met with a huge wave of public bashing, similar to the problem of internet cancel culture[3].

Thus, a clear need exists for addressing the collective action problem in mixed-motive

[2]https://www.bbc.com/news/business-34324772

[3]https://nypost.com/article/what-is-cancel-culture-breaking-down-the-toxic-online-trend/

MAS when two conditions are true: *a)* we have no prior knowledge about the agents' architectures, thus it is not possible to safely assume agents will always behave pro-socially, and *b)* agents' are not allowed to punish each other.

1.2 Objective

Our aim with this work is to propose a general purpose framework for steering mixed-motive MAS out of socially bad outcomes when conditions *a* and *b*, cited in Section 1.1 hold.

In particular, we propose to use a combination of *reinforcement learning* and *centralized norm enforcement techniques*. Besides the RL agent players, we propose to enhance such environments with a normative system controlled by a RL regulator agent. This regulator will be able to learn to adjust norms and sanctions of the system it is regulating according to its social outcomes.

Such general purpose framework could be useful in a couple of scenarios: *a)* to train an RL agent to assume the role of regulator and possibly prevent the collective action problem in mixed-motive MAS, and *b)* to learn how different norms and norm-enforcement mechanisms may influence mixed-motive human interactions (agent-based social simulation research).

Furthermore, it is also our goal to test this framework on a version of a famous mixed-motive game and, without loss of generality, to assess some of the main drivers for its successful implementation. Our goals can be summarized in the following research questions:

1. Can we successfully train a regulator agent to prevent the collective action problem in mixed-motive MAS environments?

2. What effect does the frequency in which the norm is changed have on the system's social outcome?

3. What effect does the harshness of the penalty applied to those who violate the norm have on the system's social outcome?

1.3 Methodology

In case it is not safe to assume other agents are pro-social and it is not desirable for agents to directly punish each other, we need to resort to centralized governance of some kind. Jones and Sergot (1994) propose two complementary models of centralized norm enforcement:

1. *Regimentation*: Provides an imposing normative framework to which agents have to abide to, therefore non-compliant behavior does not occur;

2. *Regulation*: Assumes agents can violate norms, and violations may be sanctioned when detected.

A drawback of the former is that it constraints agents' autonomy (NARDIN, 2015). Furthermore, implementing a regimentation system is not necessarily trivial; edge cases may arise such that violations may still occur (JONES; SERGOT, 1994). On the other hand, the latter preserves — to some degree — agents' autonomy by allowing their actions to violate the norms.

As mentioned in Section 1.2, we propose to use a combination of *reinforcement learning* and *centralized norm enforcement* techniques for regulating mixed-motive environments. We focus our attention to the multiagent reinforcement learning (MARL) case, which is a subset of MAS, since it presents a challenging case for preventing the collective action problem, as the agents are continuously learning and improving their individual rewards regardless of how well the collective is doing.

1.4 Expected results

If successful, this work's expected results, that also characterize its main contributions to the literature will be:

1. A framework for dealing with the collective action problem in mixed-motive MARL environments that is compliant with conditions a and b cited on the last paragraph of Section 1.1.

2. A framework for solving the collective action problem in mixed-motive MARL environments that can be implemented using only off-the-shelf RL architectures such as DQN (MNIH et al., 2013) and A2C (MNIH et al., 2016).

1.5 Organization of the manuscript

This manuscript is organized in 6 chapters. Chapter 2 presents the relevant background for the project divided into three key topics, games and mixed-motive games, multiagent systems and normative multiagent systems, and reinforcement learning. Chapter 3 presents some relevant work, that have tackled the broad issue of social order in multiagent systems through regulation, the collective action problem in mixed-motive MARL environments, and a brief discussion of how our work relates to this corpus. Chapter 4 presents an overview of the proposed framework, its formal model, and an algorithm that shows how it could be implemented. Chapter 5 presents the experiments, how they relate to the research questions, the environment used to run them, and their results. Finally, Chapter 6 presents some final considerations and further work.

2 BACKGROUND

2.1 Games and mixed-motive games

2.1.1 Mixed-motive games

Mixed-motive games is the name given to denote a subset of games from game theory; a branch of mathematics developed and applied mostly by economists (PETERS, 2015). Shoham and Leyton-Brown (2008) simply define game theory as "the mathematical study of interaction among independent, self-interested agents"[1]. In other words, game theory is the study of how independent agents plan and achieve their goals given their expectations on how other agents in the system plan to achieve their own goals, and the repercussions of this complex layering of expectations for the system itself.

Game theory is mostly interested with the study of games and their solutions. A game can be defined by the set of rules and information structure that constrain the agents' *moves* on a *play* — an instance of a game[2]. A solution to a game is a systematic description of the emergent outcomes found in a collection of its plays assuming some degree of rationality from the agents playing it (OSBORNE; RUBINSTEIN, 1994).

Based on the above, a mixed-motive game, or social dilemma (DAWES, 1980), can be described as one in which players' preferences over outcomes are partially aligned and partially opposed. This duality commonly means individual rationality does not build up to group rationality (KOLLOCK, 1998), i.e., the aggregate effects of selfishness yield a socially bad outcome.

[1] The word "agents" is used generically to refer to any autonomous, pro-active entity with its own set of *beliefs* and *goals*

[2] Refer to (NEUMANN; MORGENSTERN, 1944) for a formal and thorough definition.

2.1.2 Two-players mixed-motive games

Consider a simple two-players symmetric game in which players can choose to cooperate (C) or defect (D) with the other. Such a game can be defined by its set of players $P = \{1,\ 2\}$, the set of actions each player can choose from $A_1 = A_2 = \{C,\ D\}$, and a mapping from a combination of players' actions to two real value rewards — one for each player — $R : A_1 \times A_2 = \{C_1C_2,\ C_1D_2,\ D_1C_2,\ D_1D_2\} \rightarrow \mathbb{R}^2$. Here, we denote C_1C_2 for the case when both players cooperate, C_1D_2 for the case when player 1 cooperates and player 2 defects, and so forth. Without loss of generality, let reward (R), sucker (S), temptation (T) and punishment (P), denote, in order, the four corresponding rewards earned by player 1 for each of combination of actions in $\{C_1C_2,\ C_1D_2,\ D_1C_2,\ D_1D_2\}$, e.g., if player 1 cooperates and player 2 defects, player 1 will receive a reward of S and player 2 will receive a reward of T. In these settings, it is possible to define a handful of mixed-motive games by modifying the relative values of the rewards earned by each player. This generic two-players, two-actions game is summarized in Table 1.

	C_1	D_1
C_2	R, R	T, S
D_2	S, T	P, P

Table 1: Rewards earned by each player for every combination of actions in a two-players, two-actions game. The first value in each cell is earned by player 1 while the second value is earned by player 2.

We can build the prisoner's dilemma (PD) game — the most well-known mixed-motive game — on top of this generic framework to serve as an example. PD is a two-players game, formally defined by the relative reward values such that the inequalities $T > R > P > S$ and $2R > S + T$ are satisfied. The background story behind the formal definition is a tale of two outlaws caught by the police, and sent to different rooms for interrogation. During interrogation, each prisoner has the option to snitch the partner (defect), for which he would receive a lesser punishment — at the cost of a more severe punishment to his accomplice —, or keep quiet (cooperate). Since the punishment for being snitched is greater than the penalty reduction for snitching, mutual cooperation is preferred over mutual defection.

Considering the values $R = 3$, $T = 4$, $S = 0$ and $P = 1$, it is straightforward to see how the dilemma plays out. Both players have the incentive to defect regardless of what the other does, since the value they get for defecting is greater, either in case the other chooses to cooperate — $T = 4 > R = 3$ — or defect — $P = 1 > S = 0$. We call defect in the Prisoner's dilemma game a *dominating strategy*, i.e., an action that yields a greater

reward than any other, regardless of what the other player does (AXELROD, 1984). In case both players choose their dominating strategy (defect), a state of equilibrium is reached, where no player has the incentive to switch actions. This state is known as a *Nash equilibrium*. In our example, if both players chose their dominating strategies, they both would get a reward of 1, even though a more advantageous state for the pair could be reached; one where they would have cooperated, and for that, would have gotten a reward of 3. Table 2 sums up the dynamics of the PD game.

	C_1	D_1
C_2	3, 3	4, 0
D_2	0, 4	1, 1

Table 2: The matrix form of a PD game. Players in a PD are driven by the system's incentives towards mutual defection since it is more advantageous for both to defect than to cooperate, regardless of what the other player does.

It is possible to create other two famous mixed-motive games by changing the relative payoffs of R, T, S, and P (KOLLOCK, 1998). If they satisfy the inequality $R > T > P > S$ we have the game of *assurance* or *stag hunt*, while if they satisfy the inequality $T > R > S > P$ we have the game of *chicken*.

2.1.3 N-players mixed-motive games

Mixed incentives are by no means particular to two-players games. A mixed-motive game exists in every situation where individuals in a group have to choose between a greater group outcome and lesser own payoff, or greater own payoff and lesser group outcome. Table 3 illustrates the point from the perspective of an agent participating in a hypothetical three-players, two-actions mixed-motive game.

	Ag_1	Ag_2	Ag_3
Ag_1 a_1	10	0	0
Ag_1 a_2	4	4	4

Table 3: Agent's one (Ag_1) possible choices of action and their respective payoffs to each agent in a hypothetical three-players, two-actions mixed-motive game. If rational, agent one will choose action one, that yields a higher payoff to himself ($10 > 4$), rather than action two, that yields a higher payoff to the group ($10 < 12$).

One example of an n-players mixed-motive game in the real world is the public goods game (PGG) (OLSON, 1965). The PGG describes a set of problems wherein a group of people needs to pay an upfront cost to maintain a shared good that is accessible to all, but it's of every member's interest to "free-ride" instead of paying.

A common example of PGG is our municipal tax system. The local government employs the money from our taxes — among other expenditures — to improve public spaces that every resident has the right to use. The lack of payment from one resident will most likely not affect the maintenance of our roads and parks, but if tax evasion becomes a norm, the community — especially those who paid their taxes — will be punished by having public spaces that are not well maintained. [3]

A similar yet different n-players mixed-motive game named the tragedy of the commons is also commonly seen in real-world scenarios. The tragedy of the commons is a term introduced by Hardin (1968) to describe a set of social problems that cannot be solved solely by technological advances; instead, a behavior change is needed. Like the PGG, it is a group dilemma. Still, unlike the former, it is associated with individuals being incentivized to increase their short-term payoff at the cost of inflicting a long-term punishment to everyone in the game.

Hardin himself describes a didactic example in his article; a group of herders, having access to a common piece of land, may allow as many of their cows to graze on it. Every herder has the individual incentive to let as many of his cows in, but if all herders behave accordingly, the grass will soon be depleted[4], and the cows will have nothing to eat. The tragedy of the commons is a mixed-motive game most often associated with environmental issues, such as the global warming cited in Chapter 1.

One way to understand the origins and the differences between these games is to classify them under two criteria: excludability and rivalry. Excludability is a variable that controls whether or not the resource is accessible to the general public. Rivalry on the other hand, refers to whether or not the consumption of the good by one entity prevents others from consuming it. Table 4 gives an example of how some goods can be classified under both criteria.

Similar to the two-players case, when all players play by their dominating strategies in n-players mixed-motive games the system as a whole is driven to a sub-optimal equilibrium. As such, the theoretical findings regarding solving the collective action problem in group dilemmas point to a non-endogenous resolution — commonly privatizing the resource or regulating its consumption (HARDIN, 1968; OSTROM, 2000). These are commonly based on three assumptions *a)* Resource users are norm-free utility maximizers with no bounded rationality; *b)* Designing rules to change incentives is an easy task; *c)*

[3]The problem is being simplified to make a point. It is not accounting for the fact that if you don't pay your taxes, you are likely to be punished, which in the real world changes the payoff of tax evasion.

[4]Assuming there are enough cows to eat all the grass in a somewhat short period of time.

	Excludable	Nonexcludable
Rival	**Private goods** housing, food, car	**Tragedy of the commons** timber, fish
Nonrival	**Club goods** cable TV, internet, cinema	**Public goods** public parks, national defense

Table 4: Classification of goods according to the excludability and rivalry criteria. Both PGG and the tragedy of the commons emerge due to the nonexcludability property that opens up the possibility to free ride (MANKIW, 2018, p. 213).

The resolution to these problems demands intervention from central authority (OSTROM, 1999).

Nonetheless, empirical work has provided evidence that these assumptions do not always conform with reality. In practice, experimental studies have shown instances of n-player mixed-motive games being solved by local communities, without the need for a regulatory central authority, and that social norms play a substantial role in solving them (OSTROM, 1999; OSTROM, 2000). Still, as communities grow in size and human interactions grow in complexity, we commonly resort to some form of norms or rules — formal or not — to dictate expected behavior.

The issue of disaligned group and member incentives is not particular to human groups and societies. They can also be present in systems of artificial agents in MAS. Such systems, may also present macro-patterns that are harmful to those within it, as a symptom of an unfavourable structure of incentives.

2.2 Multiagent systems and normative multiagent systems

2.2.1 Multiagent systems

A multiagent system (MAS) is one in which *autonomous* agents, with some degree of *rationality*, coexist in and interact with an environment in order to accomplish an individual or collective *goal* (WOOLDRIDGE, 2009). Despite not existing a definition upon which the notion of an agent is universally agreed, autonomy, rationality, and goal-orientation are properties commonly found in most agents.

These three properties are complementary and interrelated. Autonomy relates to an aspect agents have of acting independently, without the need for being explicitly told what

to do (WOOLDRIDGE, 2009, p. 3), which grants agents the ability of being proactive and purposeful. Purposefulness often implies the existence of preferences over outcomes or an end goal; an end result or state the agent desires to reach. The property of acting towards accomplishing a goal is referred as goal-orientation, and is also key to the notion of an agent (HOEK; WOOLDRIDGE, 2003). Finally, in order to successfully achieve its goals, an agent is better off picking actions that bring it closer to such desired states rather than pushing it away from it. The measure of how good an agent is at action-picking given its context is referred to as rationality (BOWLING; VELOSO, 2001; HOEK; WOOLDRIDGE, 2003).

Agents in MAS interact with an exogenous entity known as the environment. From the agent's perspective, the agent-environment interaction is a continuous cycle of choosing an action based on the current environment state, and reaching a new state — as a function of the agent's action and the environment transition function —, in which the agent will act again. Such interactions occurs through two of the agents' sets of components: the sensors and the actuators.

An agent's set of sensors is its entry-point to the agent-environment interaction. The sensors are responsible for perceiving an agent's surrounding, similar to how we, as humans, sense the environment around us through our senses. The output at this perception stage can be referred as *percepts*, which are abstract representations of the current state of the environment (RUSSELL; NORVIG, 2010, p. 34). After the perception step, the percepts or a sequence of them is passed to the agent function, that parses such information and decides what action to execute next.

The agent's actuators on the other hand are responsible for executing its actions, that may change the state of the environment, thus changing the agent's perception of it. Such cycle of perceiving and acting may continue indefinitely until a stop condition is met. An abstract representation of the agent-environment interaction is depicted in Figure 1

2.2.2 Normative multiagent systems (NMAS)

As briefly discussed at the end of Section 2.1.3, MAS hold many similarities with human societies in that, like us humans, agents may have heterogeneous preferences and may differ in how they assess their surroundings and act toward their goal — mathematically speaking, different agents may have different agent functions. As such, MAS may also be subject to the harmful symptoms commonly found in mixed-motive human systems such as miscoordination (MANKIW, 2018, p. 261), collusion (MANKIW, 2018, p. 338), and

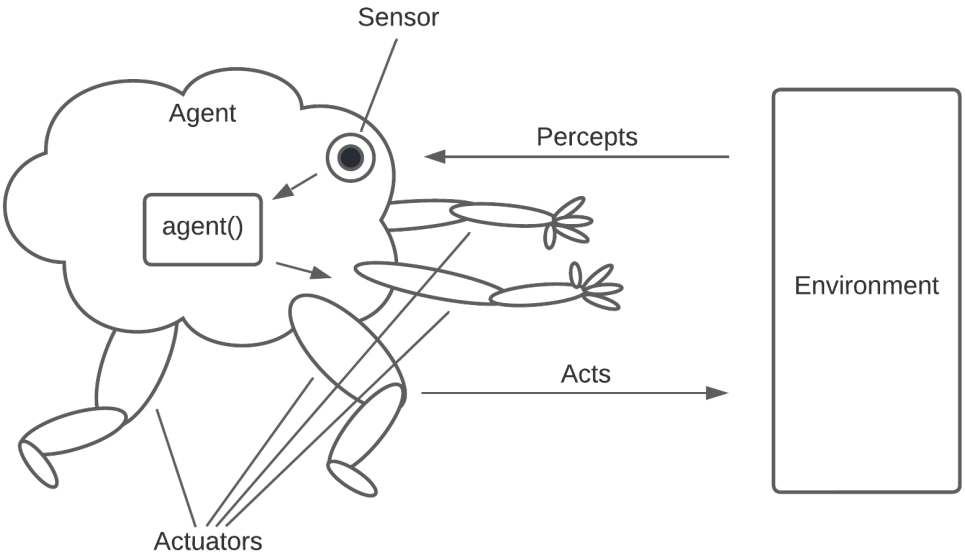

Figure 1: An agent-environment basic interaction in multiagent systems. An agent senses the state of the environment through its set of sensors, decides what to do based on its percepts, and acts thorough its actuators (KURDI; STANNETT; ROMANO, 2015; RUSSELL; NORVIG, 2010, p. 35).

negative externalities (MANKIW, 2018, p. 189).

One way of preventing these issues both in the real-world and in MAS is through the use of regulation and oversight. Such apparatus involve the creation of norms that dictate the socially desired behavior of agents, as well as the establishment of oversight bodies that ensure that these norms are being followed.

A norm enhanced MAS can be regarded as a normative multiagent system (NMAS), i.e. a MAS in which these normative concepts may influence the overall rewards earned by its agents, and therefore the outcome of the system (NARDIN, 2015). In these settings, despite not having an unified definition, a norm is typically understood to be a standard or guideline that is widely accepted and expected to be followed within a particular group or society (ULLMANN-MARGALIT, 1977).

In NMAS, norms that are not complied with might be subject to being sanctioned. Sanctions can be generally classified into *direct material sanctions*, that have an immediate negative effect on a resource the agent cherish, such as a fine, or *indirect social sanctions*, such as a lowering effect on the agent's reputation, that can influence its future within the system (CARDOSO; OLIVEIRA, 2009). Nardin (2015) also describes a third type of sanction; *psychological sactions* are those inflicted by an agent to himself as a function of the agent's internal emotional state.

Norm enforcement in such systems can be arranged either in a centralized or dis-

tributed manner (LIMA; NARDIN; SICHMAN, 2018). They differ in whether the normative machinery is sustained and enforced by a single entity — be it an agent or an organization — (centralized), or not (distributed).

Crawford and Ostrom (1995) describes how norms, alongside with shared strategies and rules, often originate many day-to-day behavioral patterns that we call *institutions*. For instance, the expectation two people will shake hands when greeting each other gives birth to this institution (shaking hands when greeting)[5]. Not complying with it might be perceived as an act of rudeness, thus, more often than not, people do shake hands when greeting.

Conversely, the simple occurrence of a behavioral pattern without the connotation of it being proper or improper can also be viewed as an institution, or as Crawford and Ostrom (1995) calls it, institution-as-equilibria (shared strategies). Many people in Brazil have the habit of having feijoada for lunch on Saturdays. Having something else for lunch on a Saturday in Brazil is not considered uncompliant; if anything, some Brazilians might find it unusual at most.

Finally, Crawford and Ostrom (1995) also considers the patterns of interaction that emerge from rules as institutions (institutions-as-rules). These rest on the assumption that actions inconsistent with those proscribed by some rule can be sanctioned, and that the mere existence of a formal sanction is capable of creating such patterns. Most of the literature considers rules, as per defined by Crawford and Ostrom (1995), as a type of norm, and so will we in this work.

In short, we highlight the fact that norms are key to describe and can explain and drive many institutions present in all spheres of our society. They can mold behavior and generate patterns of interaction within groups of people (OSTROM, 2000). They can also be incorporated into MAS as soft constraints on agents' actions (BOELLA; TORRE; VERHAGEN, 2006) to improve their macro-properties (VERHAGEN, 2000; NARDIN, 2015) without entirely compromising a fundamental aspect of MAS, i.e., agents' autonomy.

In order to formalize the conception of norms, Crawford and Ostrom (1995) proposes the ADICO grammar of institutions. The grammar is defined within the five dimensions:

- ***A**ttributes:* is the set of variables that defines to whom the norm is applied.
- ***D**eontic:* is a holder for the three modal operations from deontic logic: *may* (per-

[5]We are not implying that shaking hands is the only method of greeting other people around the world, nor that it is the common method used everywhere.

mitted), *must* (obliged), and *must not* (forbidden). These are used to distinguish prescriptive from nonprescriptive statements.

- *A**i**m:* describes a particular action or set of actions to which the deontic operator is assigned.
- ***C**onditions:* defines the context — when, where, how, etc. — an action is obliged, permitted or forbidden.
- ***O**r else:* defines the sanctions imposed for not following the norm

This grammar can be useful to turn the somewhat abstract concept of a norm into something tangible, and to operationalize the norm creation and norm revision processes. For instance, the norm *All Brazilian citizens, 18 years of age or older, must vote in a presidential candidate every four years, or else he/she will be unable to renew his/her passport* as per defined in the ADICO grammar, can be broken down into: ***A:*** Brazilian citizens, 18 years of age or older, ***D:*** must, ***I:*** vote in a presidential candidate, ***C:*** every four years, ***O:*** will be unable to renew his/her passport.

The use of a norms and oversight to regulate a mixed-motive environment becomes even more justifiable when we add learning capabilities to the agents in the system. This is the case, because learning agents — especially reinforcement learning (RL) agents — like us humans, are proficient in optimizing for their own rewards which, given the mixed-motive nature of the system, means falling prey to the collective action problem.

2.3 Reinforcement learning

2.3.1 Single-agent reinforcement learning

The reinforcement learning task mathematically formalizes the path of an agent interacting with an environment, receiving feedback — positive or negative — for its actions, and learning from them. Figure 2 illustrates the general idea of the agent-environment interaction. This formalization is accomplished through the Markov decision process (MDP), defined in the following.

Definition 1. *A Markov Decision Process (MDP) is defined by the tuple* $\langle \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P}, \gamma \rangle$ *where*

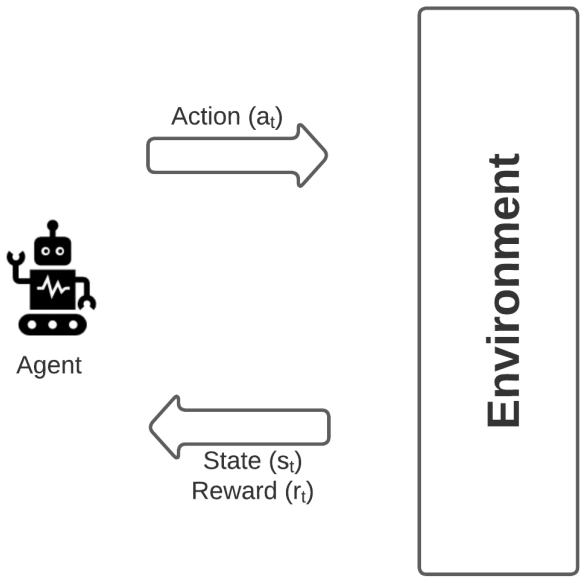

Figure 2: A single step of an RL run. The agent senses the current state of the environment, acts on it, changes its state, and receives a reward. The current state of the environment, the action, the reward, and the new state of the environment can be used to assess the value of taking such action in such current state — and extrapolate for similar actions in similar states for that matter.

- $\mathcal{S}$ *represents a finite set of environment states;*
- $\mathcal{A}$*, a finite set of agent actions;*
- $\mathcal{R}$*, a reward function* $\mathcal{R} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow \mathbb{R}$ *that defines the immediate — possibly stochastic — reward an agent gets for taking action* $a \in \mathcal{A}$ *in state* $s \in \mathcal{S}$*, and transition to state* $s' \in \mathcal{S}$ *thereafter;*
- $\mathcal{P}$*, a transition function* $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$ *that defines the probability of transitioning to state* $s' \in \mathcal{S}$ *after taking action* $a \in \mathcal{A}$ *in state* $s \in \mathcal{S}$*; and*
- $\gamma \in [0, 1]$*, a discount factor of future rewards (SUTTON; BARTO, 2018, p. 47).*

In these settings, the agent's goal is to maximize its long-term expected reward G_t, given by the infinite sum $(R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + ... + \gamma^n R_{t+n+1})$. Solving an MDP ideally means finding an optimal *policy* $\pi_* : \mathcal{S} \rightarrow \mathcal{A}$, i.e., a mapping that yields the best action to be taken at each state — the action a corresponding to the highest long-term expected reward G_t subject to the discount factor γ at a given state s.

The relative goodness of a policy π can be determined by a value function $v_\pi : \mathcal{S} \rightarrow \mathbb{R}$. In practice, the value function informs the expected reward — or an estimate of it — to

be earned by the agent for being in state s and following policy π thereafter. A policy π is said to be better than or equal to a policy π' if and only if $v_\pi(s) \geq v_{\pi'}(s)$ for all $s \in S$ (SUTTON; BARTO, 2018, p. 62).

RL algorithms can be divided into two broad categories: *model-free RL* and *model-based RL*. The difference between them relies on whether a model of the environment — reward and transition functions — is needed for learning (model-based) or not (model-free). Here, we focus on model-free RL, since it's most commonly covered and the category used throughout this project.

Model-free RL can be further split into two classes of algorithms: *value-based* and *policy gradient*. The policy search task in value-based algorithms is dependent on the computation of a monotonically improving value function. This is generally accomplished by executing back and forth two intermediary steps, namely policy evaluation — updating the value function based on past action(s) and reward(s) — (SUTTON; BARTO, 2018, p. 74) and policy improvement (SUTTON; BARTO, 2018, p. 76) — updating the policy based on the recently updated value function. Algorithms such as Sarsa (SUTTON; BARTO, 2018, p. 129), Q-learning (SUTTON; BARTO, 2018, p. 131), and DQN (MNIH et al., 2013) fall into this category.

Conversely, policy gradient algorithms learn a *parameterized policy* so actions can be chosen without the need of a value function. The goal in this case is to learn parameters $\theta \in \mathbb{R}^d$ such that the policy $\pi(a|s, \theta)$, which yields the probability of action a being taken given the environment is in state s with parameters θ, maximizes some scalar performance measure $J(\theta)$. This goal is achieved by successively updating the values of θ through a gradient ascent algorithm, that is, $\theta_{t+1} = \theta_t + \alpha \nabla J(\theta)$ (SUTTON; BARTO, 2018, p. 321). Intuitively, policy gradient algorithms try to maximize the frequency in which actions that yield good outcomes given the state of the environment are picked by the agent. Algorithms such as REINFORCE (WILLIAMS, 1992), DDPG (LILLICRAP et al., 2016), A2C (MNIH et al., 2016), and PPO (SCHULMAN et al., 2017) fall into this category[6].

[6]DDPG, A2C, and PPO are frequently categorized into the actor-critic family of RL algorithms that use both a value function and a parameterized policy. Actor-critic algorithms are commonly placed under the policy gradients umbrella since they use the value function to estimate the goodness of an action/state pair, which in turn, can be used to compute $J(\theta)$ and update the policy parameters θ.

2.3.2 Multiagent reinforcement learning

Multiagent reinforcement learning (MARL) refers to the set of RL tasks where multiple agents — two or more — coexist and interact with an environment and with each other. The MDP counterpart in MARL is the Stochastic Game or Markov Game (LITTMAN, 1994), defined in the following.

Definition 2. *A Markov Game (MG) can be formally defined by the 6-tuple* $\langle \mathcal{N}, \mathcal{S}, \{\mathcal{A}^i\}_{i\in\mathcal{N}}, \{\mathcal{R}^i\}_{i\in\mathcal{N}}, \mathcal{P}, \gamma \rangle$*, where*

- $\mathcal{N} = \{1, ..., N\}$ *denotes the set of* $N > 1$ *agents;*
- $\mathcal{S}$*, a finite set of environment states;*
- $\mathcal{A}^i$*, agent's* i *set of possible actions.*

Let $\mathcal{A} = \mathcal{A}^1 \times ... \times \mathcal{A}^N$ *be the set of agents' possible joint actions. Then*

- $\mathcal{R}^i$ *denotes agent's* i *reward function* $\mathcal{R}^i : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ *that defines the immediate reward earned by agent* i *given a transition from state* $s \in \mathcal{S}$ *to state* $s' \in \mathcal{S}$ *after a combination of actions* $a \in \mathcal{A}$*;*
- $\mathcal{P}$*, a transition function* $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ *that defines the probability of transitioning from state* $s \in \mathcal{S}$ *to state* $s' \in \mathcal{S}$ *after a combination of actions* $a \in \mathcal{A}$*; and*
- $\gamma \in [0, 1]$*, a discount factor on agents future rewards (ZHANG; YANG; BAŞAR, 2021).*

Such formalism holds some similarities with the single-agent case. For once, the *goal* from a single agent's perspective is the same — to learn an optimal policy so as to maximize long-term expected reward. Another shared commonality is that the learning strategies introduced in Section 2.3.1 — value-based and policy gradient methods — can still be used in the multiagent paradigm.

Still, one key difference between RL and MARL lies on the fact that the environment transitions to a new state as a function of the combined actions of all agents on the

latter, as opposed to the former, where it transitions solely as a function of one agent's action. The fact that all the agents' actions may influence the state transition has some bad repercussions for the multiagent case. First, it brings about engineering challenges such as the sequence in which agents and environment should act and change respectively (TERRY et al., 2021), how tie-breaking should be implemented (TERRY et al., 2021) and the increase in dimensionality. Also, the Markov property — the premise that all the information needed for an action to be picked is encompassed within the current environment state, i.e., $p(a|s_t) = p(a|s_t, s_{t-1}, s_{t-2}, s_{t-3}, ...)$ — is violated and the environment ceases to be stationary, which hampers agents learning.

Finally, a game theoretic aspect which is central to multiagent systems is added to the mix. Since the environment transitions as a function of the joint actions of all agents, an agent i has to optimize its policy not only with respect to the state of the environment, but also, relative to the joint policy of all other agents in the system (π^{-i}). Thus, once again, arises the notions of *best responses* and *Nash equilibria* (GRONAUER; DIEPOLD, 2022).

Definition 3. *An agent's i best response is the policy π^i_* such that its value function $v^i_{\pi^i_*,\pi^{-i}}(s) \geq v^i_{\pi^i,\pi^{-i}}(s)$, with π^{-i} being the joint policy of all other agents in the system, for all states $s \in \mathcal{S}$ and all policies $\pi^i \in \Pi^i$, with Π^i being the set of all possible policies for agent i.*

Definition 4. *A Nash equilibrium is a joint policy equilibrium in which all the agents' policies are best responses to all the other agents policies, i.e. $v^i_{\pi^i_*,\pi^{-i}}(s) \geq v^i_{\pi^i,\pi^{-i}_*}(s)$, for all agents in the system, states $s \in \mathcal{S}$, and policies $\pi^i \in \Pi^i$.*

MGs are a common place for mixed-motive games and therefore, the occurrence of the collective action problem. In the next chapter we present how the MARL community, as well the MAS community have dealt not only with this problem, but also with the broader issue of social order, starting with the latter.

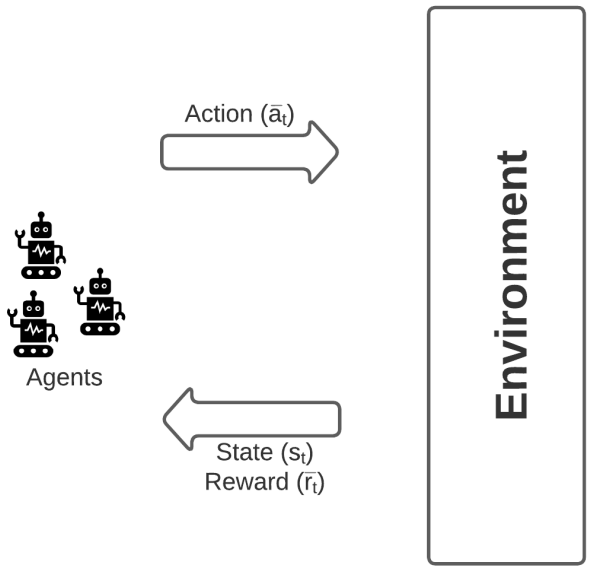

Figure 3: A single step of a MARL run. Multiple agents act on the environment, the combination of actions change its state, and each agent receives a reward.

3 RELATED WORK

3.1 Regulating MAS

The idea of regulating systems of heterogeneous agents through a formal institution is about as old as the problem of attaining social order from local actions and interactions (CASTELFRANCHI, 2000). A key step towards the development of a social control framework was the proposal of electronic institutions (EI) (NORIEGA, 1997; ESTEVA et al., 2001; ESTEVA et al., 2004), which specifies among other definitions, a set of rules that determines what the agents in the system ought to do or not under predefined circumstances. These institutions are inspired, and play a similar role to the one traditional norm-setting institutions play in real-world societies (BOU et al., 2009).

Though an important step, EIs had some limitations when compared to real-world institutions. For once, EIs were conceived at design time and were not capable of evolving over time (BOU; LÓPEZ-SÁNCHEZ; RODRÍGUEZ-AGUILAR, 2007). This issue presented some challenges for their adoption since *a)* regulating complex systems is a hard task, especially when the rules of the game are set *a priori*, and *b)* because conceiving fully functional EIs at design time is hard, a desirable property of software may be lost, i.e., the deployed system may not be self-managed.

This latter issue gave birth to the proposal of an autonomic electronic electronic institution (AEI) (BOU; LÓPEZ-SÁNCHEZ; RODRÍGUEZ-AGUILAR, 2007; BOU et al., 2009), that as the name suggests, is an electronic institution with autonomic capabilities (norm-evolving at run-time). The main objective of an AEI is for the institution to accomplish its goal by measuring some of the system's metrics, assessing whether or not the goal is being accomplished, and adapting the system's norms in case it is not through the use of an evolutionary algorithm.

3.2 Agent-centric approaches in MARL

The RL community has also seen its fair share of proposals for solving the issue of social order, and more specifically, the collective action problem in mixed-motive MARL environments. That being said, its take on the problem differs from that of the MAS community previously presented in that most of its proposals have tackled the problem from an agent-centric perspective; their solutions involve tailoring agents' architectures to the specific needs of mixed-motive scenarios.

As discussed before, these solutions can work just fine in closed systems, where one has control over the agents being deployed, or even in systems where agents are allowed to punish each other, but not as much in open systems where firm retaliation[1] is not allowed. They can be generally grouped in two: strategies that leverage reciprocity mechanisms, where agents learn to punish defective behaviors, and pro-social intrinsic motivation strategies, that reward agents for pro-social behavior.

3.2.1 Reciprocity

Reciprocity has been a notorious strategy for agents in mixed-motive games since the days of the Axelrod's tournaments (AXELROD, 1980a; AXELROD, 1980b), when researcher Robert Axelrod promoted two tournaments and invited game theorists to submit strategies to play one another in repeated plays of the prisoner's dilemma game.

This strategy is as simple as it is effective, an agent playing a reciprocity strategy defects when it recognizes antisocial behavior and cooperates when it recognizes prosocial behavior. Note that reciprocity is inherently related to the notion of a normative system; it is a mechanism that incentivizes actions that are compliant and discourages actions that are not.

These strategies have been implemented in RL agents by simply adding the capability of firmly punishing others to the agents' set of actions. By doing this, agents were capable of learning to reciprocate through self-play. Among the works that have leveraged reciprocity mechanisms to combat the collective action problem in mixed-motive MARL, we highlight those of Pérolat et al. (2017), that implemented agents with the ability of tagging other agents out of the game for a period of time, Lerer and Peysakhovich (2018), that implemented agents with two switchable policies, one fully cooperative and one fully defective, and Eccles et al. (2019), that implemented reciprocity through imitation.

[1] By firm retaliation we mean that the punishment inflicted by one agent to another is not negligible.

3.2.2 Prosocial intrinsic motivation

Another active avenue of research is to deviate from rational rational egoist model and endow RL agents with prosocial *intrinsic motivation*. As seen in Section 2.3, traditional RL agents are rewarded by the environment after choosing an action a in state s and transitioning to state s'. This reward can be regarded as *extrinsic*, i.e. the reinforcement is given to the agent as a signal of how well it is solving a problem of clear practical value (SINGH; BARTO; CHENTANEZ, 2004). Conversely, *intrinsic motivation* can be modeled as a term that composes the agents' rewards together with the extrinsic; this can be understood as a reward that is not related to the specific task in hand, but is rather earned because it is inherently enjoyable (SINGH; BARTO; CHENTANEZ, 2004).

Intrinsic motivation can be used as a way to model complex abstract patterns such as morality, empathy, or influence. Among the works that leverage prosocial intrinsic motivation to deal with the collective action problem in mixed-motive environments, we highlight those of Hughes et al. (2018), that incorporated inequity aversion preferences in RL agents, Peysakhovich and Lerer (2018), that modeled prosociality by including other agents' rewards as agent's intrinsic motivation, Jaques et al. (2019), that used the intrinsic motivation term of the reward to implement a model of social influence, and McKee et al. (2020), that tested the effects of population heterogeneity in mixed-motive scenarios.

3.3 Discussion

The proposed work is similar to the AEI framework in that it addresses most of the same problems (social order in MAS) by leveraging the use of norms, but different in that it is reduced in scope (the collective action problem in mixed-motive MARL), and it uses RL for norm adaptation instead of an evolutionary algorithm.

It also deviates significantly from those solutions put forward by the RL community; it does not assume anything about the agents' in the system, be it their intentions or their internal architecture. Instead, we opt to add another agent to the system and delegate to it the role of an overseeing regulator, that is capable of sensing the macro-state of the environment and changing the norms of the system in order to improve the social outcome.

This description somewhat fits the AI Economist framework proposed by Zheng et al. (2020), that closely portrays the general idea of a central authority learning system-level rules to guide the behavior of adaptable agents. The framework allows the training of

RL *social planners*, that learn optimal tax policies in a MARL environment of *economic actors* by observing macro-properties of the system (productivity and equality). Economic actors in this framework are free to roam around a 2d grid-world and harvest resources to build houses in order to earn money. Resources can also be traded between them. Similar to what happens in real-world societies, a percentage of all earnings — conditioned on the amount earned by each actor — is collected by the social planner and redistributed equally between all actors.

Although the work does not deal explicitly with the collective action problem in mixed-motive MARL, nor it explicitly deals with normative concepts, it inspired many of the ideas used here.

4 A NORM-ENHANCED MARKOV GAME

4.1 Overview

We propose a norm-enhanced Markov Game (neMG) for governing mixed-motive MGs by making use of an RL regulator agent and some added normative concepts. A neMG comprises two types of RL agents: $N > 1$ *players* and one *regulator*. Players are simple RL agents, analogous to the ones that interact with regular versions of MG environments, with the difference that they are aware of the norm of the game, which is available to them as it is embedded in the environment's state. The regulator, on the other hand, is able to act exclusively on the environment's norm at a predefined frequency measured in terms of players' steps, which we refer as a period. This agent senses the state of the environment through a social metric — i.e. a system-level diagnostic — and the efficacy of its actions is signaled back by the environment as a reward based on the system's social outcome.

4.2 Formal model

The formal model below describes how players and regulators interact with the environment, the framework builds upon a regular version of a Markov Game:

Definition 5. *Let $\langle \mathcal{N}, \mathcal{S}, \{\mathcal{A}^i\}_{i\in\mathcal{N}}, \{\mathcal{R}^i\}_{i\in\mathcal{N}}, \mathcal{P}, \gamma\rangle$ be the regular version of the Markov Game to be enhanced. Then, a norm-enhanced Markov Game (neMG) can be formally defined by a 13-tuple $\langle \phi, \mathcal{N}_p, \mathcal{S}_p, \{\mathcal{A}_p^i\}_{i\in\mathcal{N}_p}, \{\mathcal{R}_p^i\}_{i\in\mathcal{N}_p}, \mathcal{P}_p, \gamma_p, m, \mathcal{S}_r, \mathcal{A}_r, \mathcal{R}_r, \mathcal{P}_r, \gamma_r\rangle$, where*

- *ϕ denotes the neMG's set of norms;*
- *$\mathcal{N}_p = \mathcal{N}$ denotes the set of $N > 1$ players;*

- $\mathcal{S}_p = \mathcal{S} \times \phi$, *the players' finite set of environment states;*
- $\mathcal{A}_p^i = \mathcal{A}^i$ *player's i set of possible actions.*

Let $\mathcal{A}_p = \mathcal{A}_p^1 \times ... \times \mathcal{A}_p^N$ *be the set of players' possible joint actions. Then*

- $\mathcal{R}_p^i$ *denotes player's i reward function* $\mathcal{R}_p^i : \mathcal{S}_p \times \mathcal{A}_p \times \mathcal{S}_p \rightarrow \mathbb{R}$ *that defines the immediate reward earned by player i given a transition from state* $s_p \in \mathcal{S}_p$ *to state* $s'_p \in \mathcal{S}_p$ *after a combination of actions* $a_p \in \mathcal{A}_p$*;*
- $\mathcal{P}_p$*, a transition function* $\mathcal{P}_p : \mathcal{S}_p \times \mathcal{A}_p \times \mathcal{S}_p \rightarrow [0, 1]$ *that defines the probability of the players' environment transitioning from state* $s_p \in \mathcal{S}_p$ *to state* $s'_p \in \mathcal{S}_p$ *after a combination of actions* $a_p \in \mathcal{A}_p$*;*
- $\gamma_p \in [0, 1]$*, a discount factor on players future rewards;*
- $m \in \mathbb{N}$*, the amount of players' steps per period;*
- $\mathcal{S}_r$*, the regulator's set of states;*
- $\mathcal{A}_r$*, the regulator's set of actions;*

Let r_j^i *denote the reward earned by player i at a relative time step j of a given period*[1]*, and n the number of players in a neMG. Then*

- $\mathcal{R}_r$ *denotes the regulator's reward function* $\mathcal{R}_r = \sum_{i=1}^{n} \sum r_j^{i\,2}$*, that determines the immediate reward earned by the regulator at the end of a period given by the sum of all players' rewards over the period;*
- $\mathcal{P}_r$*, the normative transition function* $\mathcal{P}_r : \phi \times \mathcal{A}_r \rightarrow \phi$ *that defines norm update following a regulator's action; and*
- $\gamma_r \in [0, 1]$*, the regulator's discount factor.*

[1] e.g. r_3^2 refers to the third reward earned by player 2 within the period.
[2] $\sum r_j^i$ refers to the sum of rewards earned by player i in the given period.

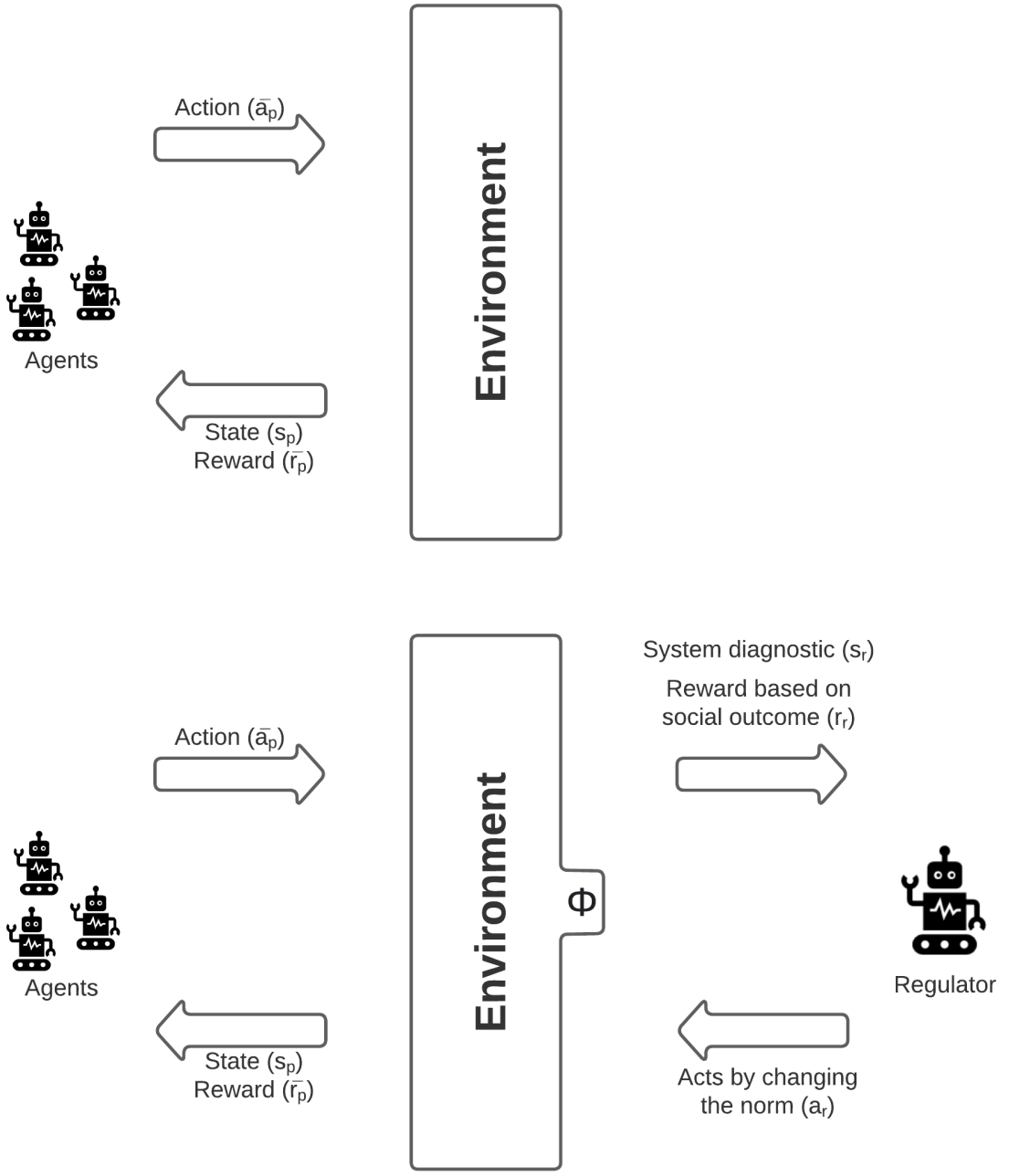

Figure 4: A comparison between an MG and an neMG, the upper part depicting an MG and the lower part, an neMG. An neMG builds on an existing MG by adding normative concepts and a regulator to it. The added regulator is also an RL agent, and as such, can also choose an action a_r in state s_r changing it to s'_r and receiving a reward r_r in the process. The regulator senses the environment through a macro-level diagnostic (the state of the environment from its perspective), acts on its norm (ϕ), and receives a reward based on the system's social outcome.

4.3 Algorithm

In these settings, a neMG could be run following two RL loops; an outer one relative to the regulator, and an inner one relative to the players. Algorithm 1 exemplifies how these could be implemented.

Algorithm 1: neMG Pseudocode

```
1  algorithm parameters: number of players (n), steps per period (m);
2  initialize policy and/or value function parameters;
3  foreach episode do
4      initialize environment (set initial states s_{r0} and s_{p0});
5      foreach period do
6          regulator adjusts norm (φ) by consulting its policy π_r in state s_r;
7          for m steps do
8              set current player i;
9              current player acts based on its policy π_p^i in state s_p, state transitions
                 to s'_p, player observes its reward r_p^i, and updates its policy π_p^i;
10         end for
11         regulator observes next state s'_r, its reward r_r and updates its policy π_r;
12     end foreach
13 end foreach
```

Training on an neMG happens across multiple episodes, a term often used in the RL literature to denote instances of games that end on a terminal state (SUTTON; BARTO, 2018, p. 54). An episode begins with the initialization of the environment's states (line 4). At every period, the regulator acts by adjusting the environment's norm based on its percept, players in the game act for m steps, and the regulator receives an immediate reward, update its policy, and the environment transitions to the next state (lines 5-10). In this case, period size (m) is the variable used to control the frequency in which the regulator acts and is measured in terms of players' steps. At every step, players act based on their percepts, the state transitions, players receive an immediate reward from the environment, and players update their policy (lines 8-9). Note that the norm does not appear anywhere in the players' loop because it is embedded within the environment state.

5 EXPERIMENTS

5.1 Recalling the research questions

Before getting into the details of the experiments, we recall the research questions introduced in Section 1.2 in order to relate them.

1. Can we successfully train a regulator agent to prevent the collective action problem in mixed-motive MAS environments?
2. What effect does the frequency in which the norm is changed have on the system's social outcome?
3. What effect does the harshness of the penalty applied to those who violate the norm have on the system's social outcome?

These questions are addressed in three experiments, each one serving the purpose of answering one of the questions. We start describing the experiments by introducing the environment used for running them.

5.2 The tragedy of the commons environment

The neMG framework was tested on a mixed-motive MARL environment that emulates the tragedy of the commons game described in Section 2.1.3 and that closely resembles the environment used by Ghorbani et al. (2021) (GHORBANI; HO; BRAVO, 2021). In this environment, agents consume units of a common resource that replenish as a function of the amount of resources left in a previous step — i.e. if the resource level fall to zero, the replenishment will also be zero. It also allows for the existence of a regulator agent by including all elements introduced in Section 4.2.

The environment is composed of two different but related parts: the players' environment and the regulator's environment, which are both described in the following.

Players' environment: A period (p) begins with an initial quantity (R_0) of the common resource. At every simulation step, players in the environment can observe the current resource level (R_j), the consumption limit (l_j), and the fine multiplier (λ_j); and can choose how much of the resource to consume (c_j), with $0 \leq c_j \leq c_{max}$, where c_{max} is a consumption limit that represents a physical limit in an analogous real-world scenario. Upon such decision, the environment's resource level is updated following the simple rule $R_j := R_j - c_j$.

Every n simulation steps — n being the number of players — the resource grows by a quantity given by the logistic function $\Delta R = rR_j(1 - \frac{R_j}{K})$ — akin to how some natural resources grow in the real world (GHORBANI; HO; BRAVO, 2021) —, with ΔR being the amount to increase; r, the growth rate; R_j, the current resource quantity; and K, the environment's carrying capacity — an upper bound for resources.

The environment also encodes the ADICO variables as described in Section 2.2, which is the normative framework used to operationalize norm synthesis and norm adaptation in this environment — ϕ in the formal model. The A, D, and C dimensions remain fixed for this experiment since *a)* the norm applies to all players, *b)* the norm always defines a forbidden action, and *c)* the norm is valid throughout the episode, no matter the conditions. The I and O dimensions, on the other hand, may be changed by the regulator agent; i.e., every m steps, that denotes the frequency in which the regulator adjusts the norm, the regulator may change how much of the resource a player is allowed to consume (l) and the fine applied to those who violate this condition ($f(c, l, \lambda)$) — by setting the value of λ. Thus the ADICO information that enhances this environment is made up of:

- ***A:*** all players;
- ***D:*** forbidden;
- ***I:*** consume resources above the consumption limit (l_j);
- ***C:*** always;
- ***O:*** pay a fine of $fine = (c_j - l_j) \times (\lambda_j + 1)$.

A violation only turns into a fine at a predefined percentage of occurrences. Such percentage is denoted here as $P(punish)$, and defines the probability a player will be caught once it exceeds the consumption limit — e.g. if $P(punish) = 0.3$, violations will

be punished 30% of the time. The fine value is subtracted from the violator's consumption in the same step the norm is violated following the simple update rule $c_j := c_j - fine$. This is the case because players' rewards in this environment are directly proportional to their consumption.

At every players' environment invocation, players act, one at a time, for a total of m steps. This means that in case the regulator changes the norm at every 100 steps, and there are 5 players in the game, then each player will act for a total of 20 times before the norm is once again changed. The amount of times a player acts on a single norm iteration is denoted *n_cycles*. Algorithm 2 summarizes the players' environment execution, which relates to lines 7-10 in Algorithm 1.

Algorithm 2: Pseudocode for the players' environment

```
1  for n_cycles do
2      foreach player do
3          player i consumes some quantity of resources (c_j) based on its policy π_p^i
             and the state of the environment (R_j, l_j, λ_j);
4          resources update (R_j := R_j − c_j);
5          if consumption (c_j) > norm-set consumption limit (l_j) then
6              random := random number between 0 and 1;
7              caught := (random < P(punish));
8              if caught then
9                  player i is penalized (c_j := c_j − fine);
10             player updates its policy π_p^i based on its net consumption (c_j) and the
                 new state of the environment (R_j, l_j, λ_j);
11     end foreach
12     resources replenish (R_j := R_j + ΔR);
13 end for
```

Regulator's environment: Before a new norm is set, the regulator can evaluate the system-level state of the environment by observing how much of the resource is left (R_p), and a short-term and long-term sustainability measurement (S_{sp} and S_{lp} respectively), given by $S = \sum_{k=p-t}^{p} \frac{rp_k}{c_k}$ defined for $c_k > 0$ and $t \geq 0$, with

- t being the number of periods considered as short-term and long-term (respectively one and four for all simulations in this work);
- rp_k, the total amount of resources replenished in period k;
- c_k, the total consumption in period k;
- p, the current period.

The initial values at the beginning of the simulation for these observable variables are drawn from uniform distributions, i.e. $R_0 \sim \mathcal{U}(10000, 30000)$, $S_{s0} \sim \mathcal{U}(0.4, 0.6)$, and $S_{l0} \sim \mathcal{U}(0.4, 0.6)$.

In this environment, the regulator acts by increasing or decreasing the values relative to the **I** and **O** ADICO variables, i.e. the consumption limit (l) — with changes limited to a value of 400 (Δl_{max}) and up until a maximum value of c_{max} ($l_{max} = c_{max}$) — and the fine multiplier (λ) — with changes limited to a value of 0.5 ($\Delta\lambda_{max}$) and up until a maximum value of three (λ_{max}). The initial values of both the consumption limit and the fine multiplier are drawn from normal distributions in the first period of the simulation, i.e. $l_0 \sim \mathcal{N}(375, 93.75)$ and $\lambda_0 \sim \mathcal{N}(1, 0.2)$. The values of l and λ set by the regulator are used throughout the period by the players in each their steps (lines 3 to 10 in algorithm 2).

At the end of the period, the success of past norms is feed-backed to the regulator by the environment as a reward value directly proportional to the period's — all players' — total consumption.

A run — or episode, as it is commonly regarded in the RL literature — has two stop conditions; it finishes at the end of a period in case resources are completely depleted or after a thousand steps. Thus, the regulator will take an action for a maximum of $m/1000$ times, which is referred here by the max period (p_{max}) variable — e.g. if $m = 100$, then $p_{max} = 10$. Algorithm 3 summarizes the execution process of the regulator's environment, which relates to lines 5-12 in Algorithm 1.

Algorithm 3: Pseudocode for the regulator's environment

```
1 while resources(R_p) > 0 or p < p_max do
2 |   regulator acts by increasing/decreasing the consumption limit (l) based on its
  |     policy (π_r) and the state of the environment (R_p, S_sp, S_lp);
3 |   regulator acts by increasing/decreasing the fine multiplier (λ) based on its
  |     policy (π_r) and the state of the environment (R_p, S_sp, S_lp);
4 |   players' environment executes;
5 |   regulator updates its policy (π_r) based on the total period's consumption (c_p)
  |     and the new state of the environment (R_p, S_sp, S_lp);
6 |   p := p + 1;
7 end while
```

A summary with all environment related variables used in this experiment and their descriptions are presented in Table 5.

Variable name	Description
n	number of players
m	number of steps in a period
R_0	initial quantity of common resource
R_i	quantity of common resource at step i
K	environment's carrying capacity (resources upper bound)
r	resources growth rate
ΔR	replenishment amount at a single step
c_i	single player consumption at step i
c_{max}	players max consumption (hard limit)
$P(punish)$	probability an agent will receive a fine if it exceeds the consumption limit
l	norm-set consumption limit
λ	norm-set fine multiplier
$fine$	fine paid by an agent
p	current period
t	number of periods considered for calculating S_{sp} and S_{lp}
S_{sp}	short-term sustainability metric at period p
S_{lp}	long-term sustainability metric at period p
rp_p	period's total replenishment
c_p	period's total consumption
l_{max}	max norm-set consumption limit
Δl_{max}	max change in norm-set consumption limit
λ_{max}	max norm-set fine multiplier
$\Delta\lambda_{max}$	max change in norm-set fine multiplier

Table 5: Summary of the environment's variables and their abbreviations.

5.3 Experimental settings

Besides the fixed-valued variables introduced in the last section, we propose testing the model with changes along three major axes; with or without an active regulator, with fixed fine multipliers of different sizes or with the regulator setting it, and with different period sizes. The cases are distributed in three experiments, each one serving the purpose of testing how this implementation of the framework behaves given variations on each axis.

The first experiment compares a default version of the tragedy of the commons experiment as an MG versus it as an neMG. The second experiment tests the effect the frequency in which the regulator acts has on the outcome of the system. Finally, the third experiment tests how the system behaves when the fine multiplier is fixed versus when it varies, and also the effect harsher punishment has on the environment versus blander punishment. Table 6 presents how the 9 proposed test cases vary along said axes and their respective experiments.

We investigate the effects these variations have on the system through some metrics: the total and net comsumption metrics are used as proxy for social outcome, the average relative and absolute difference between consumption and consumption limit per episode tracks how rationally players are behaving, and the average fines paid per 1000 episodes can also be used as a measure of players effectiveness.

The tragedy of the commons environment was built using both the OpenAI gym (BROCKMAN et al., 2016) and pettingzoo (TERRY et al., 2021) frameworks. Agents in this simulation were built using traditional RL architectures — SAC (HAARNOJA et al., 2018) for the regulator and A2C (MNIH et al., 2016) for the players — using the Stable Baselines 3 framework (RAFFIN et al., 2021), and players were trained on a shared policy. The learning rates for all agents were set to 0.00039.

5.3.1 Experiment 1: Regulator effect

This first experiment allows us to take a first look into the workings of the neMG framework by showing how the added normative concepts proposed by it affect the environment. This is accomplished by testing the same version of the tragedy of the commons game in two different scenarios: with normative concepts (neMG) and without them (MG).

The MG version of the game — without regulator nor any normative concept — is

Experiment	Name	Active regulator	Value of fine multiplier	Period size
Experiment 1	*noRegulator*	no	-	100
	default100	yes	var	100
Experiment 2	*default50*	yes	var	50
	default100	yes	var	100
	default200	yes	var	200
	default500	yes	var	500
Experiment 3	*default100*	yes	var	100
	fixedMultiplier0.5	yes	0.5	100
	fixedMultiplier1	yes	1	100
	fixedMultiplier2	yes	2	100
	fixedMultiplier3	yes	3	100

Table 6: Summary of implementation test cases. The *default100* case is used as a base case in all three experiments, and thus, it's present in all of them.

presented by the *noRegulator* test case. This test case is used as a benchmark to test the effects of adopting the neMG framework to regulate mixed-motive MARL environments. The *noRegulator* case is tested against the *default100* case, that consists of the same version of the tragedy of the commons game as the one in the *noRegulator* case, with the same values for variables — as far as possible —, but with the added normative concepts proposed by the neMG framework, i.e., an RL regulator agent acting on the system's norms. The values used for variables in both of these test cases are shown in Table 7. Each case was run 10 times.

In this experiment, we expect the neMG to outperform the MG version of the game since, in the former, agents will almost always have the incentive to consume below the consumption limit[1], which, if we assume to be their expected behavior, will grant the regulator the ability to control resources level and prevent its depletion.

Results

Figure 5 shows the average total consumption per episode for the *noRegulator* and *default100* cases, as well as the average net consumption (consumption - fines) per episode for the *default100* case. The results are also summarized in Table 8. As predicted by the Nash equilibrium, we notice there isn't much hope for generalized cooperation in case selfish agents are left playing the game by themselves — i.e. resources quickly deplete in the beginning of each episode.

[1]The exception would be when $\lambda = 0$. Then overconsuming pays off equally as much as consuming at the set consumption limit since the fine paid for a norm violation will be equal to the difference between consumption and limit.

Variable name	Values
n	5
m	100
R_0	$\sim \mathcal{U}(10000, 30000)$
R_i	var
K	50000
r	0.3
ΔR	var
c_i	var
c_{max}	1500
$P(punish)$	-; 100%
l	-; var
λ	-; var
$fine$	-; var
p	var
t	-; 1 (short-term), 4 (long-term)
S_{sp}	-; var
S_{lp}	-; var
rp_p	var
c_p	var
l_{max}	-; 1500
Δl_{max}	-; 400
λ_{max}	-; 3
$\Delta\lambda_{max}$	-; 0.4

Table 7: Values for variables used in the *noRegulator* and *default100* cases. The first entry in each row is the value used in the *noRegulator* case whereas the second is the value used in the *default100* case. Rows with only one value means it is used in both cases.

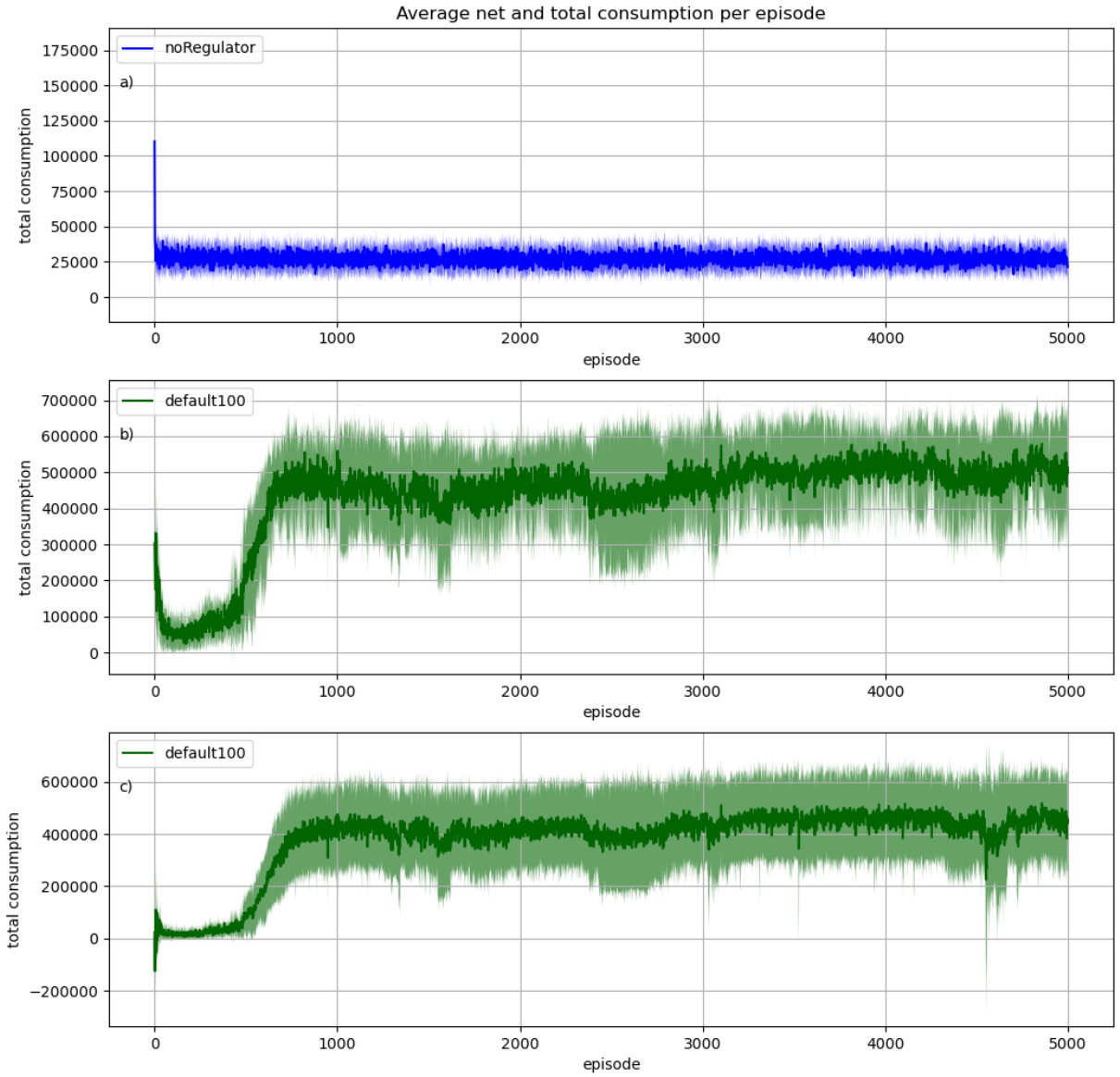

Figure 5: The average net and total consumption per episode over a 10 simulation run for the *noRegulator* and *default100* cases. The blue line in *(a)* shows the average total consumption for the *noRegulator* case while the green line in *(b)* shows the average total consumption for the *default100* case. The green line in *(c)* shows the average net consumption (consumption - fines) for the *default100* case. The shaded areas in all graphs cover the region one standard deviation above and below the average.

Total or net consumption	Case	Episode of reference	Result
Total consumption	*noRegulator*	5000	21,252 ± 6,696
	default100	5000	**512,459 ± 152,263**
Net consumption	*default100*	5000	**453,683 ± 181,435**

Table 8: Total and net consumption for all cases in experiment 1. Episodes of reference were chosen based on where the graphs seem to converge.

Conversely, this is not the case when the regulator is put in place. After a short period of randomness at the beginning of simulation, players learn not to consume from the resource since they frequently get punished when doing so. Around episode 500, players progressively learn to consume around as much of the resource as the set limit and the regulator increasingly learns to adjust such limit so as to keep resources at a sustainable level. A comparison between an episode at the beginning of a simulation and one at the end, after most of learning has taken place, is shown in Figure 6.

Every once in a while, resources deplete before the thousandth time step either due to innacuracies from the player's actions (i.e. players consume over the norm-set consumption limit until there is no resources left), from the regulator (i.e. the regulator sets the consumption limit higher than it should) or a combination of both, which explains in parts the total consumption variation depicted by the green shaded area in Figure 5. One instance of such combined inaccuracies is shown in Figure 7. We hypothesize that part of these innacuracies might be intrinsic to the nature of our problem; neither the agents nor the regulator have any contextual knowledge about norms or the environment they are in, they learn and act based purely on past states and rewards. Therefore it is not trivial for players — as it might be for a human in a similar real-life situation — to consume below the consumption limit as much as it is for them to consume around it.

The behavior described above — agents consuming around the limit rather than just under the limit — can be inferred from Figure 8. This figure presents the average absolute and relative distances between the consumption limit and consumption per episode. The absolute distance (the red curve) shows, on average, how far the agents are consuming from the limit, regardless if consumption is above or below it. The relative distance (blue curve) on the other hand, shows the average distance between consumption and limit including positive and negative values, i.e., an average consumption of 100 above limit in one run counters an average consumption of 100 below limit in another given these values are relative to the same episode within their runs. Since the average distance between consumption and limit remains close to 100 from episode 1000 onward, and the average relative distance for this same range hovers close to 0, one can conclude that

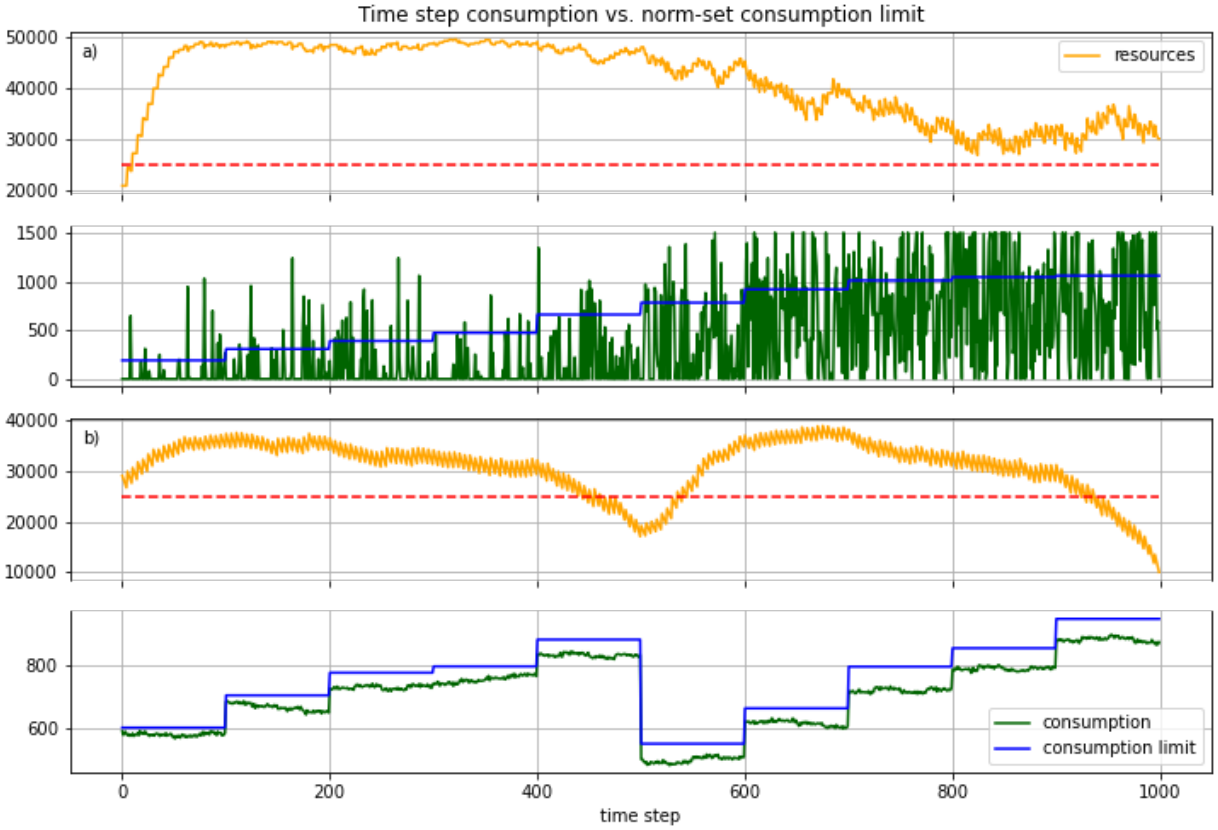

Figure 6: Time step consumption vs. consumption limit set by the regulator at an earlier episode *(a)* and at a later episode *(b)*. The orange line shows the resource level at all time steps and the dotted red line shows the resource level in which the replenishment rate is greatest (25000). In *(a)* players and the regulator act somewhat randomly and, for this reason, resources are kept at a sustainable range but consumption is sub-optimal. Players in *(b)* learn to approximate their consumption to the norm-set consumption limit and the regulator learns to decrease such limit at times when resources are lower and increase it when resources are higher. Resources in this episode are still kept at a sustainable range and consumption sharply increases in comparison to *(a)*.

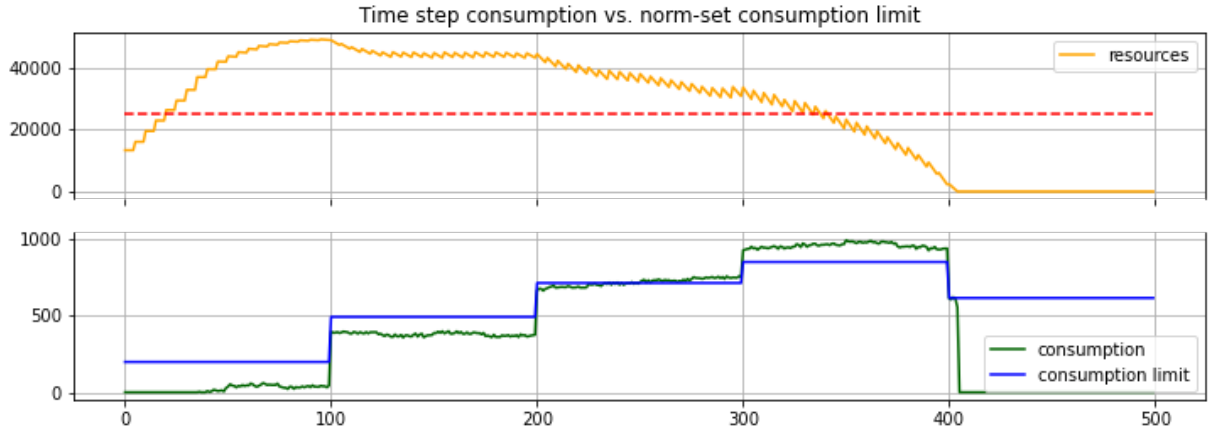

Figure 7: An instance of resources depleting before the thousandth time step at a later episode. Resources are kept at a healthy level in the first 300 steps of the episode and the regulator increasingly raises the consumption limit. Starting from step 300, agents overconsume on an already high consumption limit, which dips the system into an unrecoverable state. The regulator has no time to decrease such limit to prevent the depletion of resources.

agents are sometimes consuming above the set limit, sometimes consuming below it, but these quantities cancel out when averaging.

Finally, Figure 9 presents the average fines paid per 1000 episodes in the *default100* test case. The graph shows a sharp drop on fines paid when comparing the two first bins. This behavior is in accordance with expectation; the first thousand episodes encompass the first phase of learning when agents aren't trained yet, and thus, are more susceptible to violating the norm and being punished. Counterintuitively though, the average amount of fines paid increases in each successive bin thereafter, which is another hint that players are not learning to consume below the set limit, but rather around it. This pattern could also be explained by the agents' innacuracies or training instabilities cited above.

5.3.2 Experiment 2: Period effect

This second experiment provides us with a way of testing the effect that different period sizes — the frequency in which the regulator acts — have on the overall performance of the system. To this end, we resort again to the *default100* case and use it as benchmark to test against versions of the game with the same variables, with the exception of the period size (m). These were set to 50 (*default50* case), 200 (*default200* case), and 500 (*default500* case). Table 9 brings the values for variables used in all cases. For each case the simulation was once again run 10 times.

Here we hypothesize that the smallest period size case (*default50*) will have a better

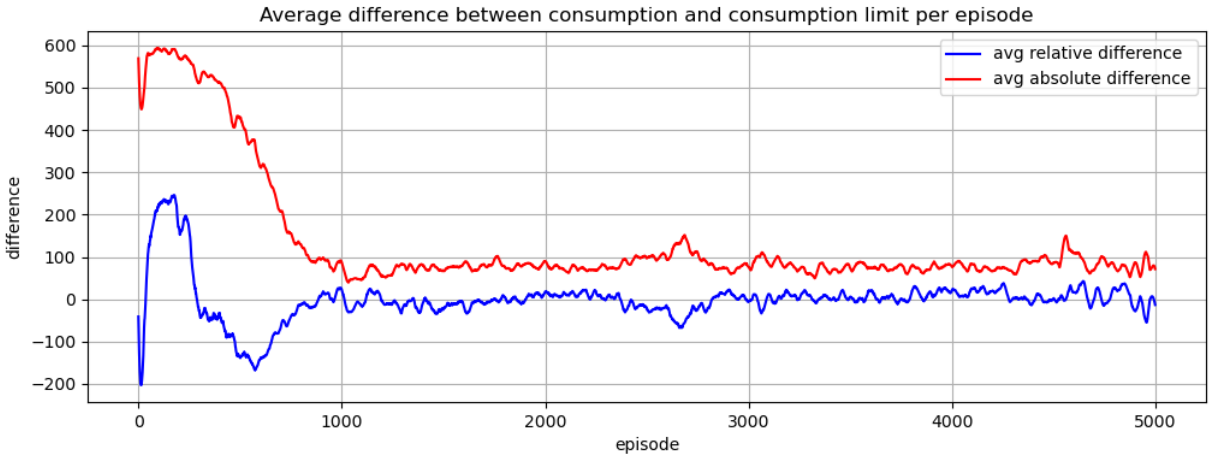

Figure 8: Average absolute and relative distances between consumption and limit. The red curve shows the average absolute difference (not considering positive and negative values) between consumption and consumption limit per episode, while the blue curve shows the average relative distance (considering positive and negative values) between consumption and consumption limit per episode.

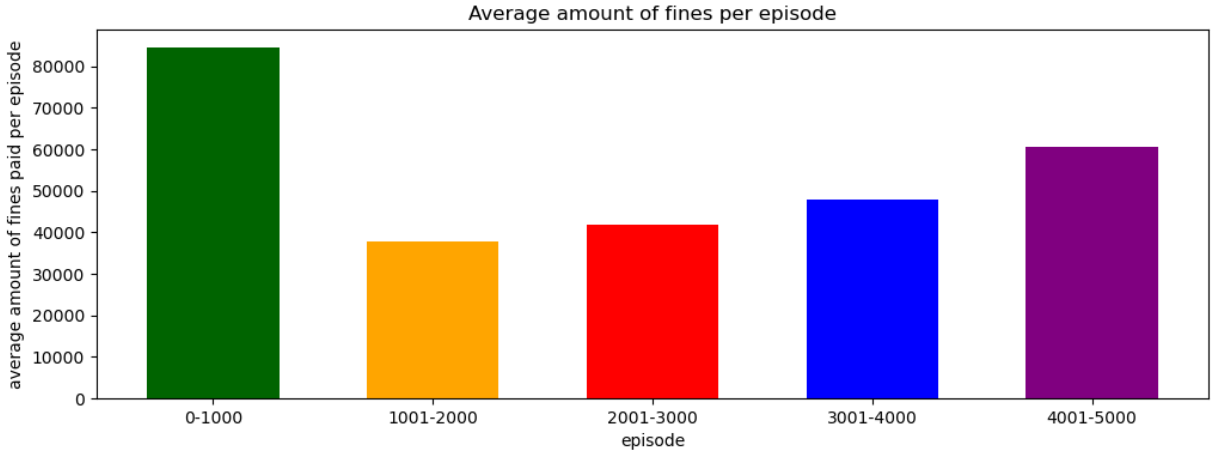

Figure 9: Average fines paid per 1000 episodes in the *default100* test case.

Variable name	Values
n	5
m	50; 100; 200; 500
R_0	$\sim \mathcal{U}(10000, 30000)$
R_i	var
K	50000
r	0.3
ΔR	var
c_i	var
c_{max}	1500
$P(punish)$	100%
l	var
λ	var
$fine$	var
p	var
t	1 (short-term), 4 (long-term)
S_{sp}	var
S_{lp}	var
rp_p	var
c_p	var
l_{max}	1500
Δl_{max}	400
λ_{max}	3
$\Delta \lambda_{max}$	0.4

Table 9: Values for variables used in experiment 2. The value of m was set to 50 for the *default50* case, to 100 for the *default100* case, to 200 for the *default200* case, and to 500 to the *default500* case. All the other values remained the same throughout all cases.

Total or net consumption	Case	Episode of reference	Result
Total consumption	*default50*	4000	**609,520 ± 175,206**
	default100	5000	512,459 ± 152,263
	default200	4500	378,015 ± 214,528
	default500	5000	271,586 ± 281,915
Net consumption	*default50*	4000	**537,525 ± 204,455**
	default100	5000	453,683 ± 181,435
	default200	4500	311,340 ± 188,166
	default500	5000	259,227 ± 281,599

Table 10: Total and net consumption for all cases in experiment 2. Episodes of reference were chosen based on where the graphs seem to converge.

performance when compared to the other cases for three main reasons: *a)* the regulator will act more often and thus should learn faster; *b)* since the regulator acts more often, it should have greater control over the system and more chances to correct its path before it derails; and *c)* players should also learn faster since they spend less time acting on noisy percepts that occur from the moment resources deplete to the moment the period ends and the environment is reset.

Results

Figure 10 presents the average net and total consumption per episode for each case in experiment 2 and Table 10 summarizes the results. The results corroborate our hypothesis; the *default50* case seems to reach — on average — a higher consumption (around 600,000) than the three other cases, before the four-thousandth episode, when it drops about 33%. We conjecture this drop occurs due to some training instability common to RL such as off-policy divergence (SUTTON; BARTO, 2018, p. 260).

For the test cases in which the regulator's actions are more infrequent, total consumption did not stabilize at — in the *default200* case — or even reach — in the case of *default500* — the same levels as the test cases in which the regulator act more frequently (*default50* and *default100*). This behavior is expected since this system level metric is highly dependent on the regulator's ability to set the right consumption limit, and its learning is dependent on the frequency in which it acts. Also, player's learning could have been harmed in these cases, since players spend more time acting on states with depleted resources, where their actions have no effect on their rewards.

The players' behavior throughout the simulation can be analyzed for each case in experiment 2 in Figure 12. It presents the average relative and absolute differences between consumption and consumption limit for the cases in experiment 2. They show an inter-

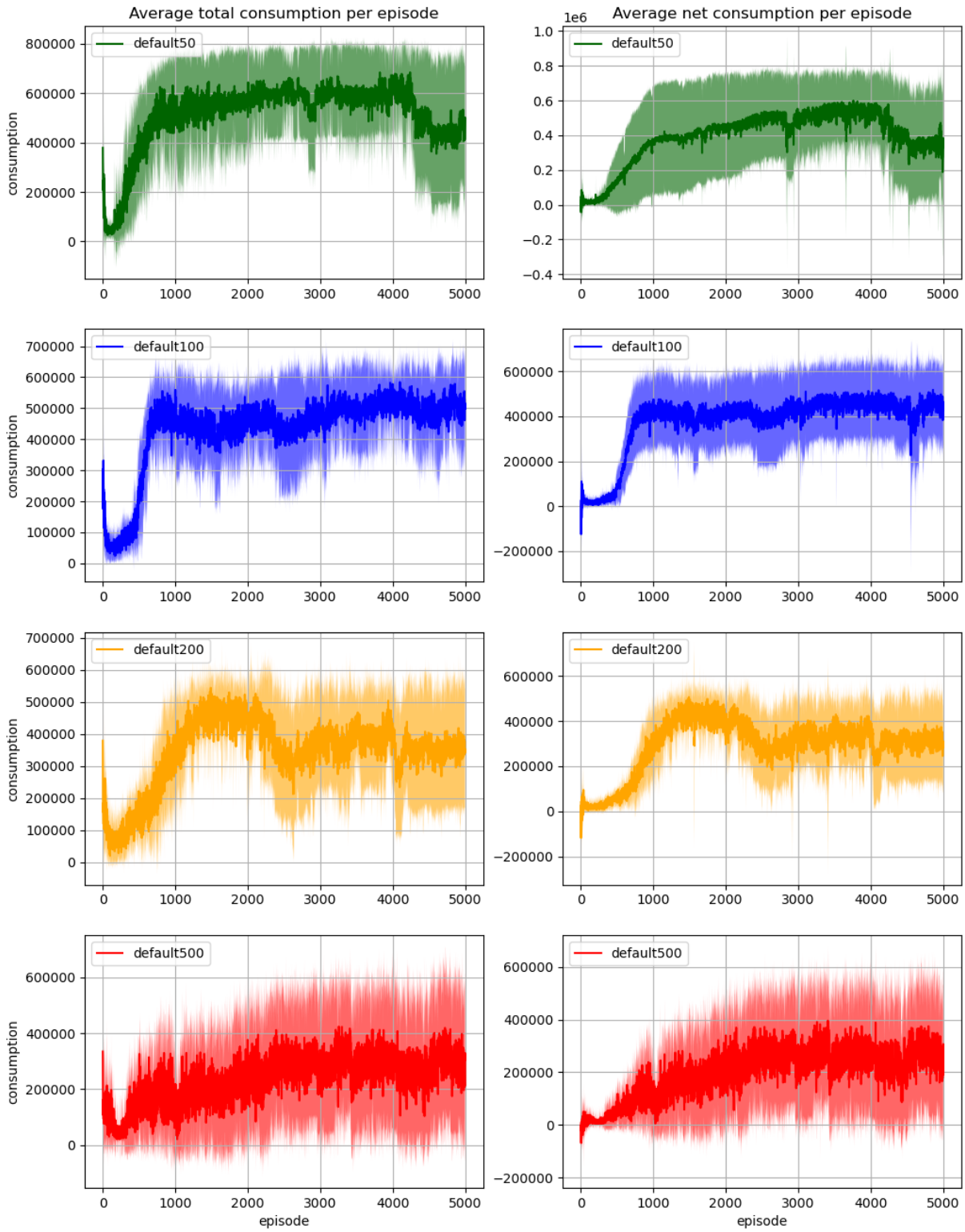

Figure 10: The average total and net consumption per episode for all cases in experiment 2 (*default50*, *default100*, *default200*, and *default500*). The shaded area in each graph covers the area of one standard deviation above and one standard deviation below the mean for each episode.

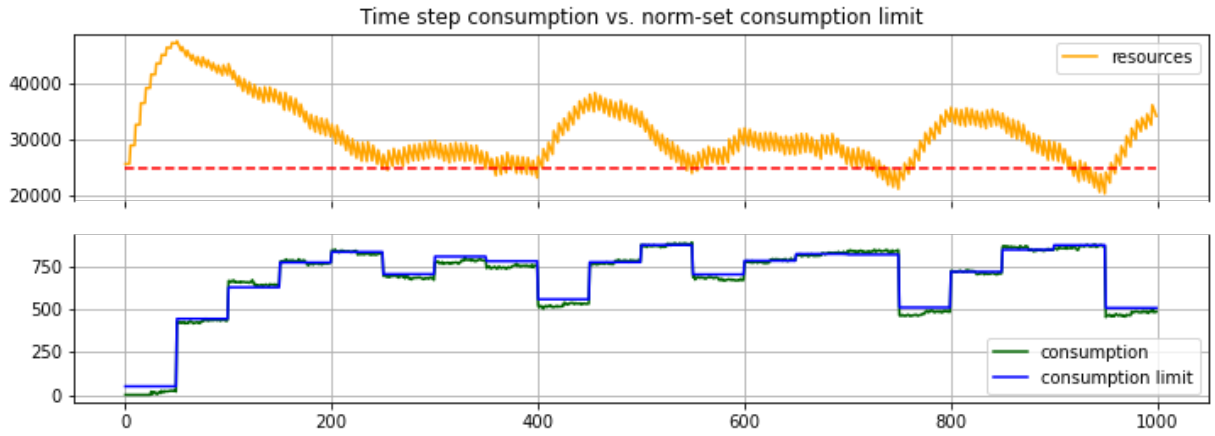

Figure 11: Resources level at a later episode when $m = 50$. The regulator manages to keep resources near the optimal level (25000), represented by the dotted red line.

esting pattern for the *default100*, and *default200*. At the beginning of simulation, while trying to maximize for consumption, players consume above the set consumption limit for a short period before they realize that overconsumption does not pay off. Then, after a small period of consuming below the limit — that could be explained by all the negative reinforcements players received for overconsuming —, agents learn to consume around it for the remaining episodes. The same pattern does not show up in the *default50* case. In it, players learn to overconsume and remain doing so for a while, slowly decreasing their consumption difference to the limit, before reaching a relative average of zero, and later increasing their consumption relative to the limit once again. In the *default50* case, two patterns deviate from the expectation: *a)* rational players should learn not to consistently overconsume, like they did in the other cases; and *b)* rational agents should not go back to overconsuming once they learned to consume around the set limit. This last pattern is directly linked to the decrease in the system's performance after episode 4000, and once again, could be caused by learning instabilities mentioned above.

As for the average amount of fines paid per episode, Figure 13 presents a big discrepancy between the *default50* case and the three other cases. This noticeable difference may be surprising at first, since one could expect the regulator's learning to be somewhat decoupled from the agents' learning, and if that was the case, the agents fines should not vary too much across the four test cases in this experiment. This seems to be a reasonable premise, since the agents are almost always incentivized to consume just below the set consumption limit, regardless of how much the limit is. In this case, the main difference across cases is the amount of time agents spend acting on environments with no resources. Since the environment only resets at the end of a period, players in the *default500* case spend more time acting on a depleted environment than players in the *default50* case, i.e.

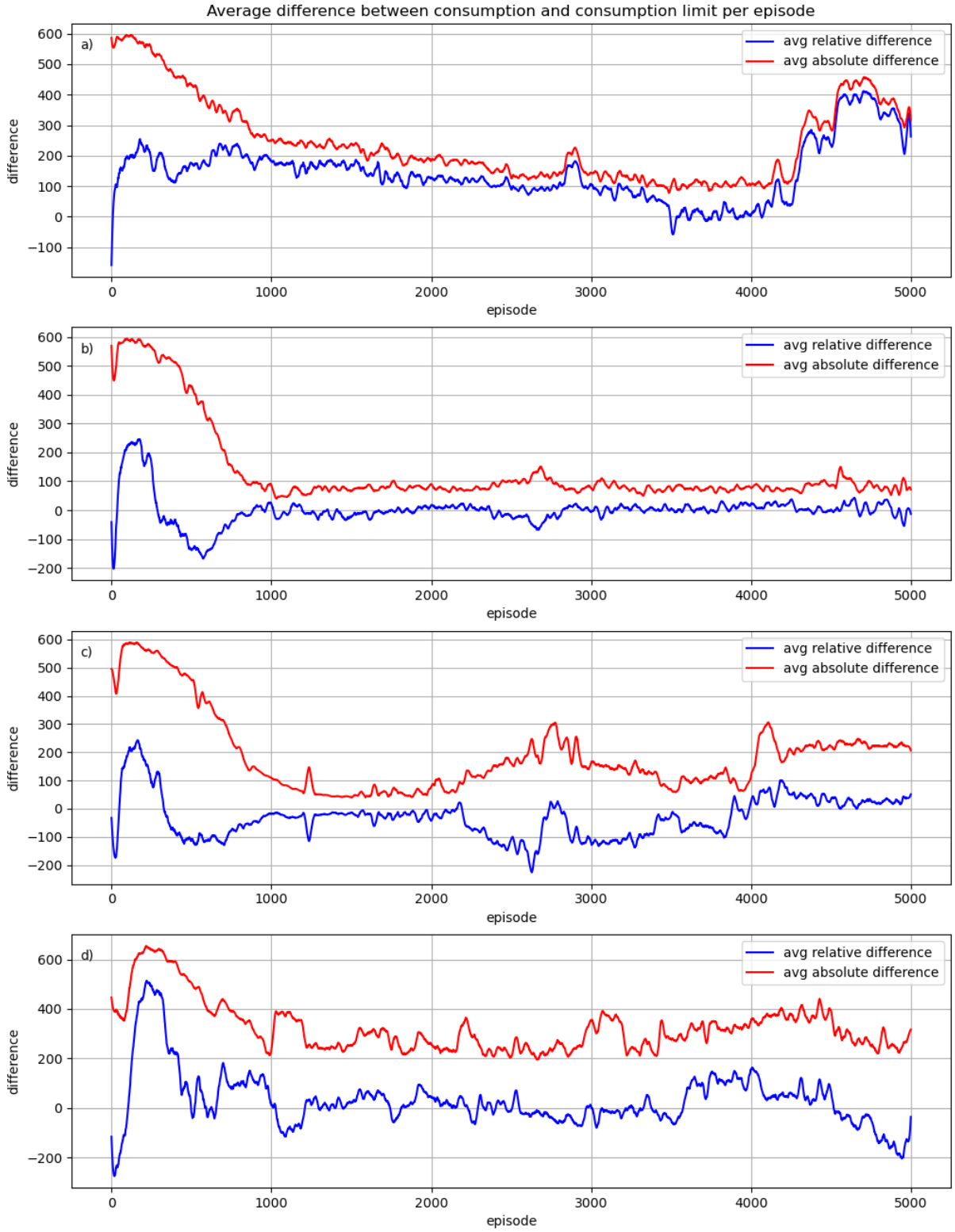

Figure 12: Average absolute and relative distances between consumption and limit for all cases in experiment 2. The red curve shows the average absolute difference (not considering positive and negative values) between consumption and consumption limit per episode, while the blue curve shows the average relative distance (considering positive and negative values) between consumption and consumption limit per episode for the test cases *a) default50*, *b) default100*, *c) default200*, and *d) default500*.

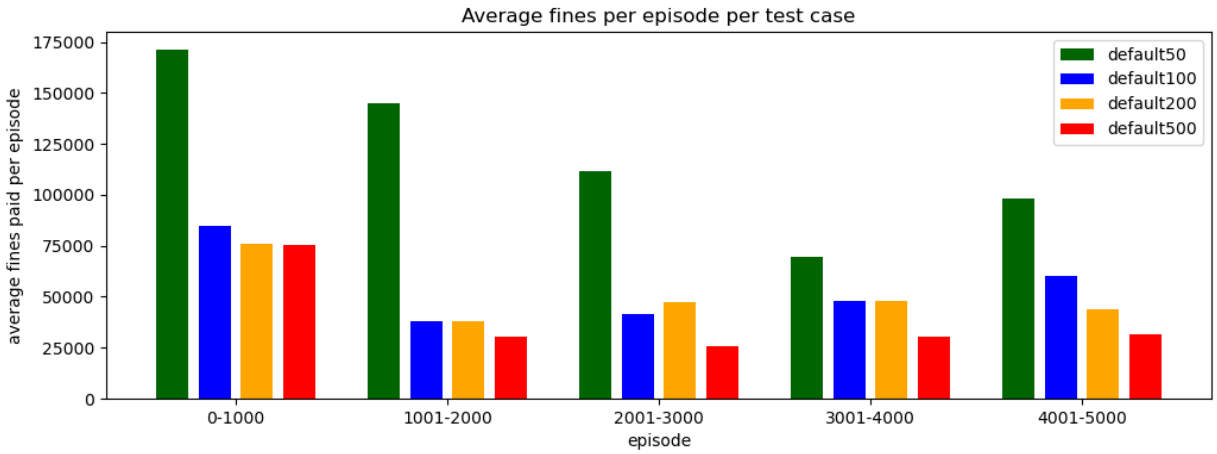

Figure 13: Average fines paid per 1000 episodes in experiment 2.

resources in the *default500* and *default200* cases are not always available, which mean these agents are not able to violate the norms even if they want to.

5.3.3 Experiment 3: Fine multiplier effect

In this last experiment we test the effect harsher punishment has on the system's performance versus blander punishment. This is accomplished by fixing the fine multiplier at different levels across four different test cases ($\lambda = 0.5$, $\lambda = 1$, $\lambda = 2$, $\lambda = 3$) and leaving only the task of setting the consumption limit to the regulator. Since fines are just a proxy metric for negative rewards in our environment, this experiment has the intent of testing how these mixed-motive systems behave for different scales of punishment and how these changes may affect the agents' learning path. We also compare these cases against the *default100* case, to check if there are any noticeable advantages in allowing this extra flexibility to the regulator. Table 11 shows the values of all variables used in the test case.

Here, we expect that a higher fine multiplier should discourage agents to consume over the limit. This could have a positive effect in that agents will be less likely to explore consuming above the limit, but it may also hinder players' learning in case negative reinforcements are too severe.

Results: Figure 14 presents the average total and net consumption per episode for each of the five test cases in experiment 3 (*default100*, *fixedMultiplier0.5*, *fixedMultiplier1*, *fixedMultiplier2*, and *fixedMultiplier3*) and the results are summarized in Table 12. We notice a tendency for convergence at a higher consumption level for the two cases with greatest fine multipliers (*fixedMultiplier2* and *fixedMultiplier3*) when compared to the two cases with the smallest fine multipliers (*fixedMultiplier0.5* and *fixedMultiplier1*). This

Variable name	Values
n	5
m	100
R_0	$\sim \mathcal{U}(10000, 30000)$
R_i	var
K	50000
r	0.3
ΔR	var
c_i	var
c_{max}	1500
$P(punish)$	100%
l	var
λ	0.5, 1, 2, 3, var
$fine$	var
p	var
t	1 (short-term), 4 (long-term)
S_{sp}	var
S_{lp}	var
rp_p	var
c_p	var
l_{max}	1500
Δl_{max}	400
λ_{max}	3
$\Delta\lambda_{max}$	0.4

Table 11: Values for variables used in experiment 3. The value of λ was set to 0.5 for the *fixedMultiplier0.5* case, to 1 for the *fixedMultiplier1* case, to 2 for the *fixedMultiplier2* case, to 3 to the *fixedMultiplier3* case, and it was allowed to change for the *default100* case. All the other values remained the same throughout all cases.

Total or net consumption	Case	Episode of reference	Result
Total consumption	*default100*	5000	512,459 ± 152,263
	fixedMultiplier0.5	5000	324,032 ± 247,439
	fixedMultiplier1	5000	414,934 ± 264,354
	fixedMultiplier2	5000	507,438 ± 74,121
	fixedMultiplier3	4000	**531,076 ± 95,277**
Net consumption	*default100*	5000	453,683 ± 181,435
	fixedMultiplier0.5	5000	229,781 ± 209,780
	fixedMultiplier1	5000	370,120 ± 271,507
	fixedMultiplier2	5000	463,129 ± 61,083
	fixedMultiplier0.5	4000	**524,033 ± 100,554**

Table 12: Total and net consumption for all cases in experiment 3. Episodes of reference were chosen based on where the graphs seem to converge.

effect likely happens due to the strength of the signal being sent to the agents in the form of fines. The smaller the fine multiplier, the lesser is the punishment received for violating the norm and weaker is the players' learning signal. The stronger signal does a better job at encouraging players to consume below the limit, which is good for them in the long run.

At a first glance, the greatest advantage of having a high fixed multiplier versus allowing the regulator to change it is the reduced variability in training. However, when analyzing the players' behavior in both scenarios, a small but important difference emerges. Figure 15 shows the average difference between consumption and consumption limit for all cases in experiment 3. When comparing the cases *default100*, *fixedMultiplier2*, and *fixedMultiplier3*, it is possible to notice that players in the latter two cases do a better job at consuming below the limit than in the former. This pattern can also be observed in the graph in Figure 16. In it, we notice that players in the *default100* case get consistently more punishments in all learning stages than those in the *fixedMultiplier2* and *fixedMultiplier3* cases.

We also notice that a higher fine multiplier is not associated with more fines paid by agents. The harsher punishment had the effect of quickly limiting overconsumption in the first stages of learning, as opposed to a more prolonged period of overconsumption observed in the cases with lower fine multipliers. As training progressed, players in the cases with harsher punishement were more consistent in adhering to the norm.

Once again, we observe some instabilities in training, and although it is hard to pinpoint exactly what are the root causes, we conjecture that harsher punishment might be associated with more stable training in these mixed-motive systems.

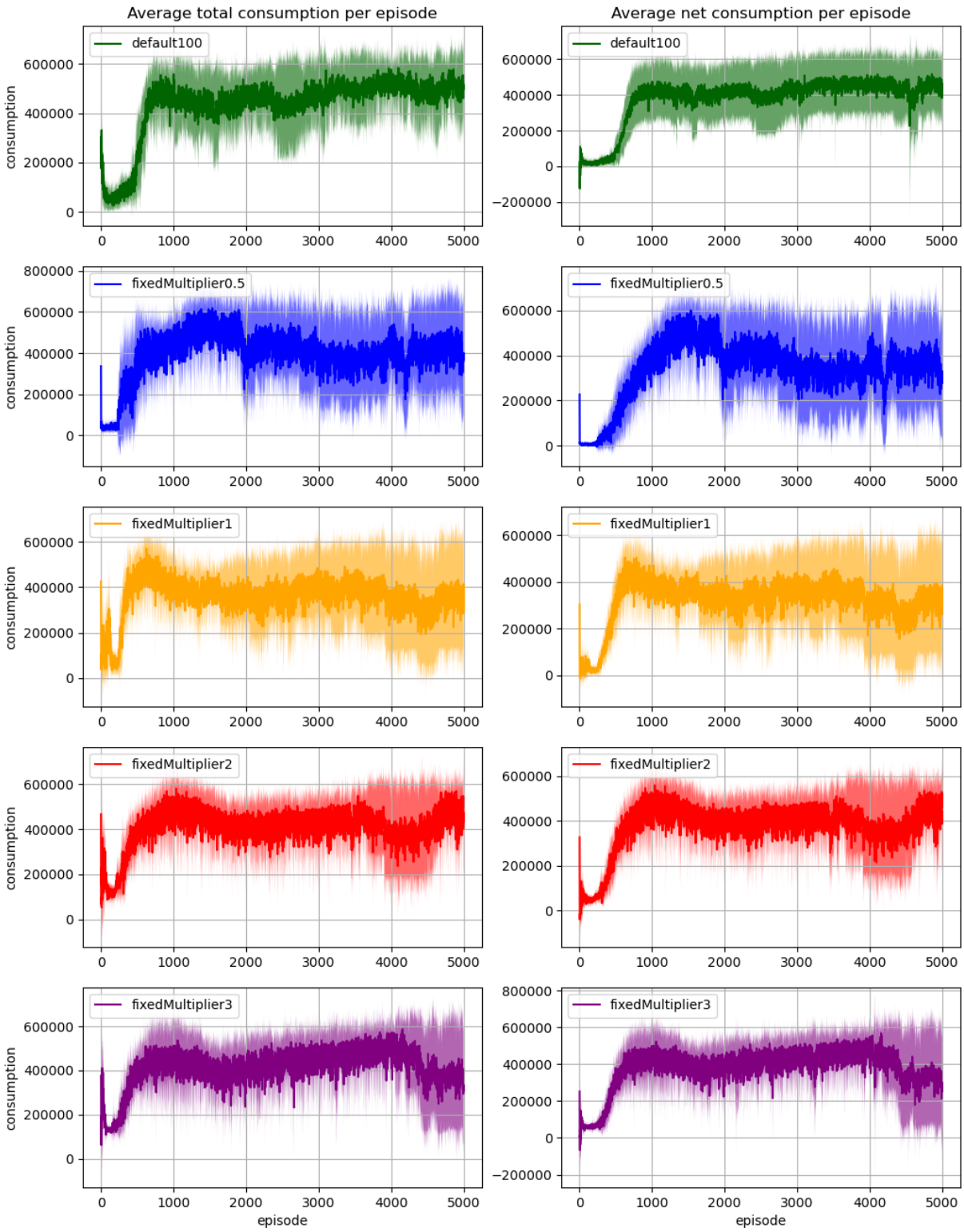

Figure 14: The average total and net consumption per episode for all cases in experiment 3 (*default100*, *fixedMultiplier0.5*, *fixedMultiplier1*, *fixedMultiplier2*, and *fixedMultiplier3*). The shaded area in each graph covers the area of one standard deviation above and one standard deviation below the mean for each episode.

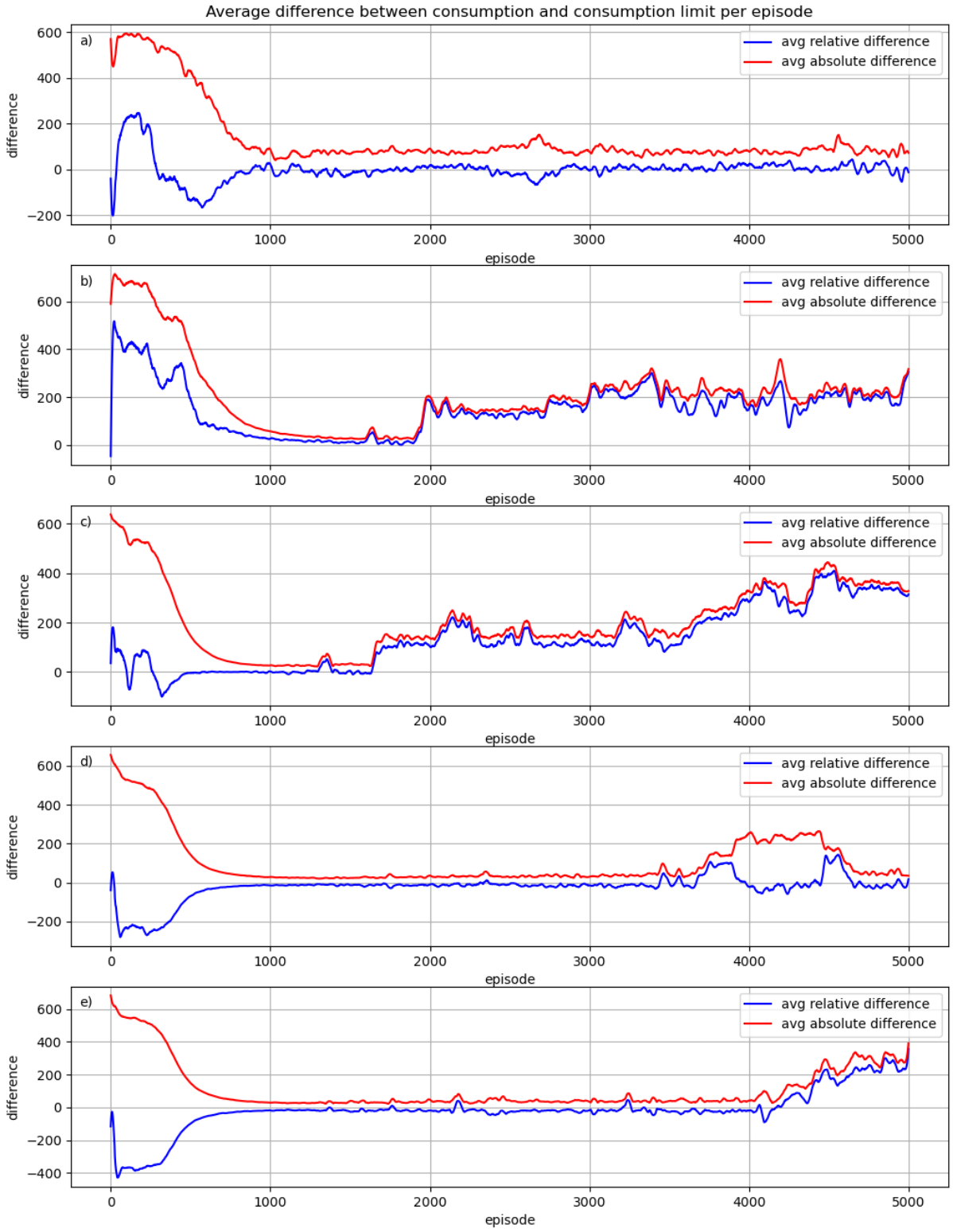

Figure 15: Average absolute and relative distances between consumption and limit for all cases in experiment 3. The red curve shows the average absolute difference (not considering positive and negative values) between consumption and consumption limit per episode, while the blue curve shows the average relative distance (considering positive and negative values) between consumption and consumption limit per episode for the test cases *a) default100*, *b) fixedMultiplier0.5*, *c) fixedMultiplier1*, *d) fixedMultiplier2*, and *e) fixedMultiplier3*.

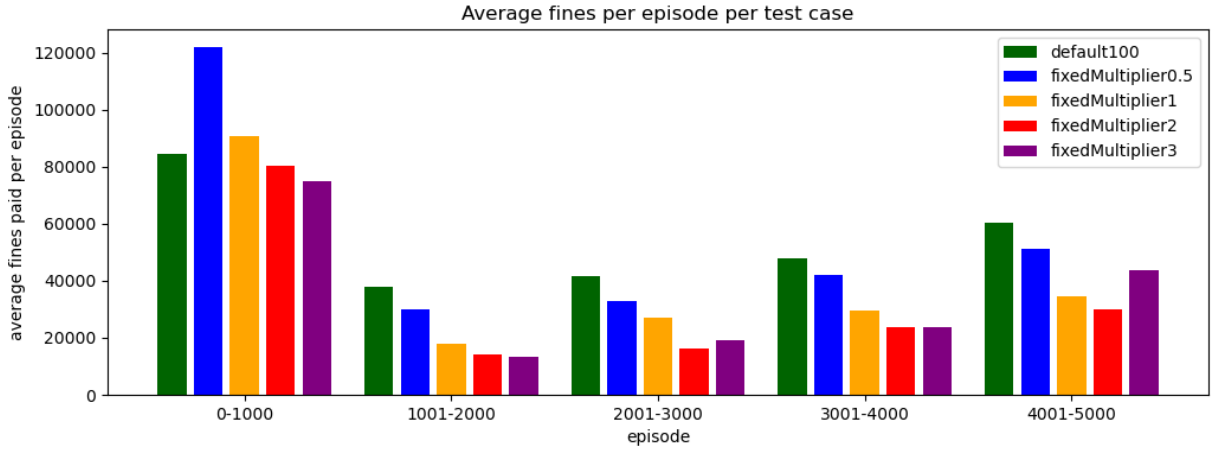

Figure 16: Average fines per 1000 episodes in experiment 3.

5.3.4 Discussion

The tragedy of the commons experiment serves to validate the idea of a norm-enhanced Markov Game (neMG). That is to say that it could be an option to be considered for regulating mixed-motive multiagent systems when the general case *a* — that no prior knowledge about the agents' beliefs, goals, incentives in the system is known — and constraint *b* — that agents in the system are not allowed to punish each other — are both valid.

Experiment 1 shows that it is possible to interfere with a system's outcome by purposefully employing selfish learning agents with different objectives to interact on a shared environment. With the advent of ever so powerful learning mechanisms — such as those found in recent RL architectures —, thinking about the agents' objectives, and how they interact with the environment, becomes a viable option for synthesizing norms or regulating systems at run time as opposed to pre-defining them at design time.

Experiment 2 shows that the frequency in which the regulator acts in the system proved to be a sensible variable. Increasing such frequency grants the regulator greater control by allowing it a bigger margin for it to correct the system's path once the system starts to behave undesirably. This could be especially useful in dynamic systems, where negative outcomes might scale exponentially.

Finally, experiment 3 gives us a hint to how the punishment variable — the **Or else** variable from the *ADICO* framework such as λ — impact learning in and the overall performance of a mixed-motive neMG. Greater punishment seems to grant more stability during training and also positively impact system's performance. That being said, we do not know the extent to which this pattern is valid, more experiments should be conducted

to test if it holds for even greater values of λ.

6 FINAL CONSIDERATIONS

Multiagent systems are part of a trend towards greater and widespread computational power (WOOLDRIDGE, 2009, p. 3) that harnesses the potential of autonomous, goal-oriented agents to solve ever so complex problems. This is reminiscent to how humans solve problems in societies. We coordinate, cooperate, and negotiate with one another in order to settle disputes, reach agreements, and move forward as collective.

Still we have come to agree that letting everyone freely pursuit their goals through any means deemed necessary in our modern-day societies may take us quickly down a dangerous road. In a system where incentives can point to many different directions, all sorts of emergent exploits may lead to negative externalities. For instance, two people may agree on a deal beneficial to them both but that goes against the interests of one or more third parties.

In many of these cases we resort to central regulation of some shape or form. If many parallels can be drawn between multiagent systems and real-world communities, why shouldn't we exploit this apparatus that has been employed for centuries in the real-world, and is very present in our everyday lives, to solve problems in communities of artificial agents?

Delegating norm enforcement to an external central authority might seem counter-intuitive at first, as we tend to associate distributed solutions with robustness. It also might seem to go against the findings of Elinor Ostrom (OSTROM, 1999; OSTROM, 2000), who showed that the collective action problem could be solved without the need of a regulatory central authority and for that, won the nobel prize in economics in 2009[1].

That being said, central regulation is still an important mechanism to govern complex systems. Many of the world's modern social and political systems use it in some form or shape. With this work, we try to show that central regulation is also a tool that could be useful in governing MAS and MARL, especially when it is not desirable for actors in the

[1] https://www.nobelprize.org/prizes/economic-sciences/2009/ostrom/facts/

system to punish each other.

As further work, we plan to *a)* test how other variables such as the probability of punishment influence the performance of the system, *b)* further investigate the learning instabilities observed in experiments 2 and 3, and *c)* to test this very same method in other mixed-motive MARL environments.

REFERENCES

AXELROD, R. Effective choice in the prisoner‘s dilemma. *Journal of Conflict Resolution*, v. 24, n. 1, p. 3–25, 1980.

AXELROD, R. More effective choice in the prisoner‘s dilemma. *Journal of Conflict Resolution*, v. 24, n. 3, p. 379–403, 1980.

AXELROD, R. *The evolution of cooperation*. New York: Basic, 1984.

BOELLA, G.; TORRE, L.; VERHAGEN, H. Introduction to normative multiagent systems. *Computational and Mathematical Organization Theory*, Kluwer Academic Publishers, USA, v. 12, n. 2–3, p. 71–79, oct 2006.

BOU, E.; LÓPEZ-SÁNCHEZ, M.; RODRÍGUEZ-AGUILAR, J. A. Towards self-configuration in autonomic electronic institutions. In: *Coordination, Organizations, Institutions, and Norms in Agent Systems II*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007. p. 229–244.

BOU, E.; LÓPEZ-SÁNCHEZ, M.; RODRÍGUEZ-AGUILAR, J. A.; SICHMAN, J. S. Adapting autonomic electronic institutions to heterogeneous agent societies. In: *Organized Adaption in Multi-Agent Systems*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009. p. 18–35.

BOWLING, M.; VELOSO, M. Rational and convergent learning in stochastic games. In: *International Joint Conference on Artificial Intelligence*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001. (IJCAI'01, v. 2), p. 1021–1026.

BROCKMAN, G.; CHEUNG, V.; PETTERSSON, L.; SCHNEIDER, J.; SCHULMAN, J.; TANG, J.; ZAREMBA, W. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

CARDOSO, H. L.; OLIVEIRA, E. Adaptive deterrence sanctions in a normative framework. In: *International Joint Conference on Web Intelligence and Intelligent Agent Technology*. [S.l.]: IEEE Computer Society, 2009. p. 36–43.

CASTELFRANCHI, C. Engineering social order. In: *Engineering Societies in the Agents World*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2000. p. 1–18.

CHEN, J.; WANG, C. Reaching cooperation using emerging empathy and counter-empathy. In: *International Conference on Autonomous Agents and MultiAgent Systems*. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, 2019. (AAMAS '19), p. 746–753.

CONTE, R. Emergent (info)institutions. *Cognitive Systems Research*, Elsevier Science Publishers B. V., NLD, v. 2, n. 2, p. 97–110, may 2001.

CRAWFORD, S. E. S.; OSTROM, E. A grammar of institutions. *American Political Science Review*, Cambridge University Press, v. 89, n. 3, p. 582–600, 1995.

DAWES, R. M. Social Dilemmas. *Annual Review of Psychology*, v. 31, n. 1, p. 169–193, 1980.

ECCLES, T.; HUGHES, E.; KRAMÁR, J.; WHEELWRIGHT, S.; LEIBO, J. Z. Learning reciprocity in complex sequential social dilemmas. 2019.

ESTEVA, M.; CRUZ, D. de la; ROSELL, B.; ARCOS, J. L.; RODRÍGUEZ-AGUILAR, J.; CUNÍ, G. Engineering open multi-agent systems as electronic institutions. In: *National Conference on Artifical Intelligence*. [S.l.]: AAAI Press, 2004. (AAAI'04), p. 1010–1011.

ESTEVA, M.; RODRÍGUEZ-AGUILAR, J. A.; SIERRA, C.; GARCIA, P.; ARCOS, J. L. On the formal specifications of electronic institutions. In: *Agent Mediated Electronic Commerce, The European AgentLink Perspective.* Berlin, Heidelberg: Springer-Verlag, 2001. p. 126–147.

GHORBANI, A.; HO, P.; BRAVO, G. Institutional form versus function in a common property context: The credibility thesis tested through an agent-based model. *Land Use Policy*, v. 102, p. 105237, 2021.

GRONAUER, S.; DIEPOLD, K. Multi-agent deep reinforcement learning: A survey. *Artificial Intelligence Review*, Kluwer Academic Publishers, USA, v. 55, n. 2, p. 895–943, Feb 2022.

HAARNOJA, T.; ZHOU, A.; ABBEEL, P.; LEVINE, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In: DY, J.; KRAUSE, A. (Ed.). *International Conference on Machine Learning*. [S.l.]: PMLR, 2018. (Proceedings of Machine Learning Research, v. 80), p. 1861–1870.

HARDIN, G. The tragedy of the commons. *Science*, American Association for the Advancement of Science, v. 162, n. 3859, p. 1243–1248, 1968.

HOEK, W.; WOOLDRIDGE, M. Towards a logic of rational agency. *Logic Journal of IGPL*, v. 11, p. 135–159, 03 2003.

HUGHES, E.; LEIBO, J. Z.; PHILLIPS, M.; TUYLS, K.; DUEÑEZ-GUZMAN, E.; CASTAÑEDA, A. G.; DUNNING, I.; ZHU, T.; MCKEE, K.; KOSTER, R.; ROFF, H.; GRAEPEL, T. Inequity aversion improves cooperation in intertemporal social dilemmas. In: *Advances in Neural Information Processing Systems*. [S.l.]: Curran Associates, Inc., 2018. v. 31.

JAQUES, N.; LAZARIDOU, A.; HUGHES, E.; GULCEHRE, C.; ORTEGA, P. A.; STROUSE, D. J.; LEIBO, J. Z.; FREITAS, N. de. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In: *International Conference on Machine Learning*. [S.l.]: PMLR, 2019. v. 97.

JONES, A. J. I.; SERGOT, M. On the characterization of law and computer systems: The normative systems perspective. In: _____. *Deontic Logic in Computer Science: Normative System Specification*. USA: John Wiley and Sons Ltd., 1994. p. 275–307.

KOLLOCK, P. Social dilemmas: The Anatomy of Cooperation. *Annual Review of Sociology*, v. 24, n. 1, p. 183–214, 1998.

KURDI, O.; STANNETT, M.; ROMANO, D. M. Modeling and simulation of tawaf and sa'yee: A survey of recent work in the field. In: *European Simulation and Modelling Conference*. Leicester, UK: [s.n.], 2015. p. 441–447.

LERER, A.; PEYSAKHOVICH, A. *Maintaining cooperation in complex social dilemmas using deep reinforcement learning*. 2018.

LILLICRAP, T. P.; HUNT, J. J.; PRITZEL, A.; HEESS, N.; EREZ, T.; TASSA, Y.; SILVER, D.; WIERSTRA, D. *Continuous control with deep reinforcement learning*. [S.l.]: arXiv, 2016.

LIMA, I. C. A. de; NARDIN, L. G.; SICHMAN, J. S. Gavel: A sanctioning enforcement framework. In: *EMAS@AAMAS*. [S.l.: s.n.], 2018.

LITTMAN, M. L. Markov games as a framework for multi-agent reinforcement learning. In: *International Conference on Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1994. (ICML'94), p. 157–163.

MANKIW, N. G. *Principles of Economics, 8th edition*. Cambridge, Massachusetts: Cengage Learning, 2018.

MCKEE, K. R.; GEMP, I.; MCWILLIAMS, B.; DUÈÑEZ-GUZMÁN, E. A.; HUGHES, E.; LEIBO, J. Z. Social diversity and social preferences in mixed-motive reinforcement learning. In: *International Conference on Autonomous Agents and MultiAgent Systems*. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, 2020. (AAMAS '20), p. 869–877.

MNIH, V.; BADIA, A. P.; MIRZA, M.; GRAVES, A.; LILLICRAP, T.; HARLEY, T.; SILVER, D.; KAVUKCUOGLU, K. Asynchronous methods for deep reinforcement learning. In: *International Conference on Machine Learning*. New York, New York, USA: PMLR, 2016. (Proceedings of Machine Learning Research, v. 48), p. 1928–1937.

MNIH, V.; KAVUKCUOGLU, K.; SILVER, D.; GRAVES, A.; ANTONOGLOU, I.; WIERSTRA, D.; RIEDMILLER, M. *Playing Atari with Deep Reinforcement Learning*. 2013.

NARDIN, L. G. *An Adaptive Sanctioning Enforcement Model for Normative Multiagent Systems*. Tese (Doutorado) — Universidade de São Paulo, 2015.

NEUMANN, J. von; MORGENSTERN, O. *Theory of Games and Economic Behavior*. [S.l.]: Princeton University Press, 1944.

NORIEGA, P. *Agent-mediated auctions: the fishmarket metaphor*. Tese (Doutorado) — Universitat Autònoma de Barcelona, 1997.

OLSON, M. *The Logic of Collective Action: Public Goods and the Theory of Groups*. Cambridge, Massachusetts: Harvard University Press, 1965.

OSBORNE, M. J.; RUBINSTEIN, A. *A course in game theory*. Cambridge, USA: The MIT Press, 1994.

OSTROM, E. Coping with tragedies of the commons. *Annual Review of Political Science*, v. 2, n. 1, p. 493–535, 1999.

OSTROM, E. Collective action and the evolution of social norms. *Journal of Economic Perspectives*, v. 14, n. 3, p. 137–158, 09 2000.

PÉROLAT, J.; LEIBO, J. Z.; ZAMBALDI, V.; BEATTIE, C.; TUYLS, K.; GRAEPEL, T. A multi-agent reinforcement learning model of common-pool resource appropriation. In: *Advances in Neural Information Processing Systems*. [S.l.]: Curran Associates, Inc., 2017. v. 30.

PETERS, H. *Game Theory: A Multi-Leveled Approach*. [S.l.]: Springer, 2015.

PEYSAKHOVICH, A.; LERER, A. Prosocial learning agents solve generalized stag hunts better than selfish ones. In: *International Conference on Autonomous Agents and MultiAgent Systems*. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, 2018. (AAMAS '18), p. 2043–2044.

RAFFIN, A.; HILL, A.; GLEAVE, A.; KANERVISTO, A.; ERNESTUS, M.; DORMANN, N. Stable-baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, v. 22, n. 268, p. 1–8, 2021.

RUSSELL, S.; NORVIG, P. *Artificial Intelligence: A Modern Approach*. 3rd. ed. [S.l.]: Prentice Hall Press, 2010.

SCHULMAN, J.; WOLSKI, F.; DHARIWAL, P.; RADFORD, A.; KLIMOV, O. *Proximal Policy Optimization Algorithms*. 2017.

SHOHAM, Y.; LEYTON-BROWN, K. *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*. USA: Cambridge University Press, 2008.

SINGH, S.; BARTO, A. G.; CHENTANEZ, N. Intrinsically motivated reinforcement learning. In: *International Conference on Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 2004. (NIPS'04), p. 1281–1288.

SUTTON, R. S.; BARTO, A. G. *Reinforcement Learning: An Introduction*. Second edition. Cambridge, MA, USA: The MIT Press, 2018.

TERRY, J. K.; BLACK, B.; GRAMMEL, N.; JAYAKUMAR, M.; HARI, A.; SULLIVAN, R.; SANTOS, L.; PEREZ, R.; HORSCH, C.; DIEFFENDAHL, C.; WILLIAMS, N. L.; LOKESH, Y. Pettingzoo: A standard api for multi-agent reinforcement learning. In: *Advances in Neural Information Processing Systems*. [S.l.: s.n.], 2021.

ULLMANN-MARGALIT, E. *The Emergence of Norms*. [S.l.]: Oxford University Press, 1977.

VERHAGEN, H. *Norm Autonomous Agents*. Tese (Doutorado) — Stockholm University, 07 2000.

WILLIAMS, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, Kluwer Academic Publishers, USA, v. 8, n. 3–4, p. 229–256, May 1992.

WOOLDRIDGE, M. *An Introduction to MultiAgent Systems*. 2nd. ed. [S.l.]: Wiley Publishing, 2009.

ZHANG, K.; YANG, Z.; BAŞAR, T. Multi-agent reinforcement learning: A selective overview of theories and algorithms. In: _____. *Handbook of Reinforcement Learning and Control*. [S.l.]: Springer International Publishing, 2021. p. 321–384.

ZHENG, S.; TROTT, A.; SRINIVASA, S.; NAIK, N.; GRUESBECK, M.; PARKES, D. C.; SOCHER, R. *The AI Economist: Improving Equality and Productivity with AI-Driven Tax Policies*. 2020.