IMPERIAL

Learning and Deliberation for Requisite Social Influence

Author

C. Patsalidis

CID: 01866599

Supervised by

Prof. J. Pitt

Second Marker

Dr M. Cattafi

A Thesis submitted in fulfillment of requirements for the degree of
Master of Engineering in Electrical and Electronic Engineering

Department of Electrical and Electronic Engineering
Imperial College London

Abstract

This thesis investigates the dynamics of self-regulated systems through a model comprising of a regulator and a regulated network of agents, interconnected by a feedback loop. The regulated system forms a social network, which allows each agent to be influenced by four informational sources based on the "*4voices*" algorithm. The regulator is designed to be adaptive, employing a reinforcement learning strategy to learn how to adhere to the expressions of the regulated units. This cybernetic system helps establish *pathways to requisite influence*. In this paper, we introduce a new model for a self regulated system, that uses second order cybernetics to create the role of the observer. This addition aims to fully institute the pathways for *requisite social influence* and allow for systemic stability by integrating feedback loops that incorporate the observer's influence on the system's dynamics. Additionally, we examine the impact of adopting a new reinforcement learning approach, specifically the Advantage Actor-Critic (A2C) method, on learning, and communication with the regulated agent network to optimise pathways of influence. Lastly, we incorporate a partially observable Markov decision process (POMDP), which places our agents under conditions where they learn to form opinions without full knowledge of the state space, through the use of a modified Kalman filter update algorithm, emphasising the effects of misinformation on agent beliefs.

Declaration of Originality

I hereby declare that the work presented in this thesis is my own unless otherwise stated. To the best of my knowledge the work is original and ideas developed in collaboration with others have been appropriately referenced. I affirm that I have submitted, or will submit, an electronic copy of my final year project report to the provided EEE link. I affirm that I have submitted, or will submit, an identical electronic copy of my final year project to the provided Blackboard module for Plagiarism checking. I affirm that I have provided explicit references for all the material in my Final Report that is not authored by me, but is represented as my own work. I have used ChatGPT, as an aid in the preparation of my report. I have used it to improve the quality of my English throughout, however all technical content and references comes from my original text

Copyright Declaration

Acknowledgments

I would like to express my deepest appreciation to my project supervisor, Prof. Jeremy Pitt, for his invaluable guidance and support throughout the duration of this project. His insightful feedback and expertise have been a major contributor in shaping the direction and quality of this thesis.

I also wish to extend my sincere thanks to Mrs. Asimina Mertzani for her seminal work in "Requisite Social Influence in Self-Regulated Systems," which served as a crucial basis for this thesis. Her knowledge and direction have been incredibly helpful, and her contributions to the field have provided a solid framework for my research.

Contents

1

Project Specification

Contents

1.1 Project Overview

The foundation of this thesis is built upon the principles and concepts presented in the work of 'Requisite Social Influence in Self-Regulated Systems' [1]. We explore autonomous, self-regulating systems in this study, with a particular emphasis on the feedback mechanism between a regulator and its regulated entities, where the regulator employs reinforcement learning techniques to learn from the expressions of the regulated agents. We highlight the role that ethical considerations have in the design of regulatory systems, defining fundamental characteristics of ethical regulation, based on Ashby's 2020 [2] principles. The idea of social influence is central to our discussion, demonstrating the interrelation of these systems by requiring efficient communication and influence exchange between the regulated and the regulator in a socially networked environment where agents can influence one another [3]. As seen further on, this idea can be represented in the form of the '*4voices*' algorithm, that explores where one can shift his attention to, based on the impact of individual or collective opinions.

Our objective is to expand upon the preexisting model introduced in 'Requisite Social Influence in Self-Regulated Systems' [1], and enhance the stability and adaptability of self-regulated cybernetic systems. By incorporating advanced reinforcement learning algorithms and the principles of second-order cybernetics, the project will develop a framework where regulators and agents interact more effectively within a complex, dynamic environment. In addition to leveraging second-order cybernetics, this thesis also considers the implications of third and fourth-order cybernetics in the design of these systems. We also aim to apply principles of partial observability, and belief systems, to gain insights on the behaviour of the regulated units when they participate in an imperfect system, and how this influences its overall stability.

1.2 Regulator

The existing regulator, which employs Deep Q-Networks (DQN), has demonstrated a satisfactory learning capacity from the expressions of the regulated units. A key objective of this project is to achieve a significant enhancement in the regulator's performance. Specifically, the goal is for the regulator to adapt more rapidly and efficiently to the given environment, thereby yielding consistently better results. This means that the actions proposed by the regulator should demonstrate a significant improvement in the expected level of 'good' collective expressions by the regulated units.

1.3 Higher Order Cybernetics

Second-order cybernetics emerged as a concept designed to offer diverse perspectives within a system, enhancing the capacity for self-reflection on actions and beliefs. Consequently, integrating this concept into our self-regulated system is intended to add more complexity. This integration is a flexible approach to social impact, whereby the expected results of this approach are not exclusively dependent on improving our model's performance. We also hope to uncover interesting insights about social influence and the further uses of second-order cybernetics.

Higher order cybernetics (Third and Fourth Order) extends the principles of cybernetics, which focus on feedback, control, and communication within systems, to more complex levels of observer involvement and self-reflection. In this study, we provide the theoretical tools necessary for applying these concepts to our model in future research.

1.4 Partial Observability and Belief

The current self-regulated system operates under the assumption of full transparency between agents and complete observability. In this model, each agent possesses perfect and truthful information for decision-making. However, this scenario does not accurately reflect the complexities of real-world social networks, where uncertainty and misinformation are common. This provides necessity for a belief system that allows the agents to more accurately map the environment that they are a part of.

Our goal is to observe how agents change their opinions when faced with uncertainty, and how these adaptations impact the overall system dynamics. Additionally, we aim to evaluate the effectiveness of the regulator's policy making, provided inaccurate information. We can gain insights into the robustness of self-regulation and the potential need for more sophisticated regulatory strategies.

1.5 Thesis Structure and Overview of Upcoming Sections

This thesis is organised into several key sections, leading up to a comprehensive understanding of the research undertaken and the results that were gathered. Below is an overview of what to expect in the subsequent sections:

1. *Background:* This section introduces and critically examines the paper 'Requisite Social Influence in Self-Regulated Systems' [1], and we explore the key concepts and methodologies that will facilitate the aforementioned modifications to the framework.

2. *Implementation:* Here, we outline the process of implementing the modifications to the existing model.

3. *Testing and Results:* We introduce the experiments and analyse our results, in order to gain insights on the proposed concepts.

4. *Conclusions and Future Work:* We reflect on the overall outcomes of the project, assessing the impact of our modifications. Additionally, we offer recommendations for future research based on the findings and limitations identified in this study.

5. *Ethics:* We outline some of the ethical implications of this project, in terms of AI compliance.

2

Background

Contents

2.1 Introduction

The background section delves into the world of cybernetics and reinforcement learning, offering the exploration of the foundational concepts that motivated this study. It discusses social influence as a requisite for ethical regulated systems [2], and the way we can establish pathways for *requisite social influence* [1]. We also review on-policy reinforcement learning methods, commenting on their relevance and application in enhancing the adaptability and efficiency of cybernetic systems. Additionally, we highlight the influence of second order cybernetics, on the development of the Milan systemic family therapy model [4], which would be a catalyst in the shaping of modern cybernetics.

Beyond second-order cybernetics, this study acknowledges the emerging concepts of third and fourth-order cybernetics. This level of cybernetics emphasises observer reflexivity and co-evolution of observers within a system [5], further shaping our understanding of influence and control in complex systems. Lastly, we introduce how a partially observable environment can be represented, and how belief systems can be used to provide our agents with a better understanding of their situation.

This introduction sets the stage for a detailed examination of the key principles and methodologies that inform the thesis.

2.2 Self-Regulated System

A self-regulated system is one that operates without externally imposed controls or regulations [1]. It is also an automatically functioning system, whereby an observer is expected to be idle and noninfluential. In the field of cybernetics, such a system comprises of a regulator that applies policies to the system it is trying to regulate, and a regulated system that internalises said policies and provides feedback on their effectiveness. Figure 2.1 provides an example of a self-regulated system in the context of ethical regulators [2].

In this subsection, we outline the foundational model that this thesis builds upon, which was initially introduced in 'Requisite Social Influence in Self-Regulated Systems' [1]. For further clarification on specific concepts or design choices, please refer to the original work.

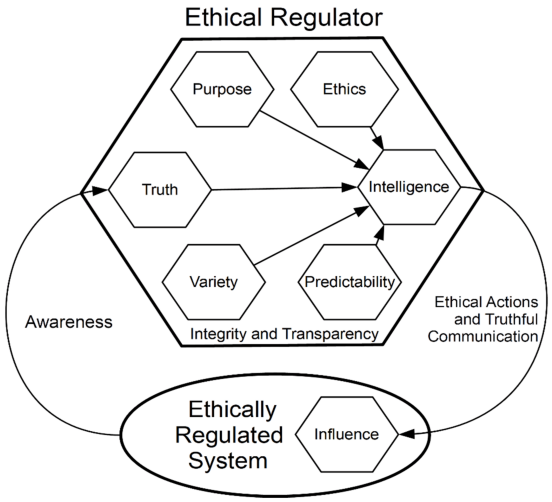

Figure 2.1: Ashby's 2020 Ethically Regulated System (ERT)

2.2.1 Requisite Influence

One of the six requirements that must exist for a cybernetic regulated system to function effectively and ethically is social influence [2]. The concept of influence emphasises the significance of pathways through which a regulator's actions impact the regulated system.

Figure 2.2 depicts the dynamics and structure of the self-regulated system, as related to social influence [1]. The system is composed of a regulator and a regulated system, with an emphasis on the influence and flow of information and actions. This model is used to underscore the integration of feedback loops that allow for the continuous flow of information between the regulator and the regulated system, helping establish *pathways to requisite social influence*.

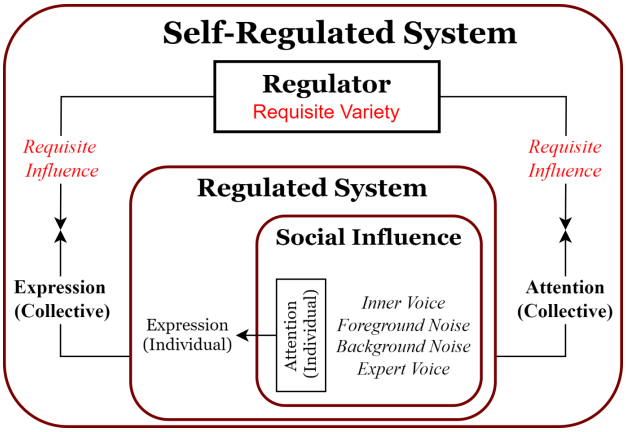

Figure 2.2: The Self-Regulated System

Regulated System

The regulated system consists of a number of agents within a social network. Within this network, agents are interlinked, facilitating influence between them, and its structure has a critical role in the overall behaviour of the system. It encompasses social influence, whereby each agent can both impact and be impacted by the opinions of other agents [3].

Under this regulated system, each agent has an individual expression that is determined by the attention they give to four different types of 'voices' [1]. These voices include:

- *Inner Voice:* This represents an agent's personal thoughts based on their own beliefs and experiences.
- *Foreground Noise:* This consists of the immediate sources of information and opinions that an agent encounters within their social circle (neighbours).
- *Background Noise:* It represents the collective opinions and viewpoints formed by the entire system in which the agent operates.
- *Expert Voice:* This is the influence of opinion by recognised experts within the network.

The collective expression that is given as input to the regulator, represents the summation of individual expressions of all the agents. Each agent individually goes through a process to determine which of the four voices they will be influenced by and therefore which voice will make up their individual expression. This is done by assigning a degree of trust that they have for each internal voice, and is updated for each iteration of the loop, based on a specific update method [1]. Following is a brief outline of the three methods:

- *Expert Deviation ('Exp'):* Agents reward the voice that deviates least from the expert voice by increasing its trust, and vice versa.
- *Collective Deviation ('Col'):* Agents increase trust to the voice that deviates the least from the average expression of the group, and reduce trust to the one that deviates the most.
- *Individual Immediate Effect ('Ind'):* Agents determine whether the new policy has benefited them or not, when compared to the previous policy (in terms of inner voice). They update the trust to the voice selected in the previous iteration, accordingly.

These voices represent the different sources of information and influence that a person might encounter and must consider when making decisions. The psychological inspiration behind this algorithm comes from the work of Charles Fernyhough in 'The Voices Within' [6], which calls attention to the concept of the '*polyphony of voices*'. It refers to the diverse nature of internal dialogues that occur within an individual and explores how our inner speech, comprising distinct voices with varying perspectives and contents, plays a critical role in self-reflection and decision-making.

Additionally, we encounter the '*cocktail party effect*' [7] as it details how humans can focus on a single conversation amid distracting sounds. It delves into the auditory system's capacity to distinguish and prioritise speech through various cues. Upon further research, it becomes evident that the engagement with the four distinct 'voices' can be linked to the impact of Covid-19 in 2020 on the education system, particularly with the shift to online learning. This transition introduced greater distractions for learners, effectively amplifying the background 'noise' typically experienced in physical classrooms [8]. Consequently, the discussed algorithm may exemplify how additional distractions can steer individuals away from their primary objectives, obscuring their focus on what is most beneficial for themselves and the broader collective.

Regulator

Requisite variety, as detailed in Ashby's work [2], is a principle in cybernetics that suggests a regulator must possess a sufficient diversity of actions to deal with the variety of the system it regulates. In the context of a self-regulated system, this means the regulator must be able to manage and respond to the complexity and unpredictability of the regulated system effectively. This model uses a Deep Q-Network reinforcement learning method [9], that enables the regulator to learn from the collective feedback from the regulated system, selecting policies that align with the optimal actions for the units.

2.2.2 Summary

As a basis for this study, we have a self-regulated system that operates autonomously, utilising a feedback loop between an internal regulator and a social network of agents. The regulator decides on a policy to be implemented by the regulated units. Each agent then internalises said policy, and according to its effectiveness, forms its own opinions from four distinct voices: the inner voice (personal opinion), the foreground noise (opinions of immediate neighbors), the background noise

(overall system opinion), and the expert voice (opinions of recognised authorities). Each agent updates its trust in these voices over time, based on their perceived reliability. The agent then selects which voice to attend to by comparing the trust given to each voice (details of this selection process can be found in the original paper [1]). The selected voice becomes the agent's expression that is communicated back to the regulator.

The regulator receives the collective expressions of all agents, which reflect the aggregated opinions of the system. This acts as a reward signal, on the effectiveness of its policies. The regulator then uses reinforcement learning (DQN) to adjust its policy selection, so that they better align with the system's needs and goals. This continuous loop of policy implementation, opinion formation, and feedback allows the system to dynamically adapt and maintain stability, whilst also helping to establish pathways to requisite social influence.

The remainder of this background section outlines the theoretical foundations for the modifications implemented (or proposed for future work), building upon the discussed, preexisting self-regulated system framework.

2.3 Reinforcement Learning and A2C

By receiving feedback on the results of its policies, the regulator in a self-regulating system can employ reinforcement learning (RL) to learn how to pay attention to the collective expression of the agents [1]. In this process, decisions are made by the regulator, who then uses the agents' collective response to influence future policy adjustments with the goal of improving the desired outcomes. This boosts the overall stability and performance of the system.

In this subsection we review the new advantage actor-critic algorithm (A2C), that is an on-policy, synchronous variant of A3C, that has shown to perform equally well. Amongst its other benefits that will be discussed further on, the A2C algorithm is designed to make more effective use of GPUs by waiting for each actor to finish its segment of experience before updating, thereby offering a cost-effective and efficient alternative to A3C, especially on single-GPU machines [10].

The motivation for further investigation into this algorithm, came due to a number of studies that compared the performance of A2C with Deep Q-Network (DQN), when attempting to solve specific problems. An example that stands out is a case study in Portugal, that tackled the issue of optimising the irrigation system for a tomato crop [11]. The study's key finding was that, in comparison to the agent taught with DQN, the agent trained with the A2C model was able to

reduce its water use by 20%. This was just one of several examples that suggested implementing A2C in the place of DQN.

2.3.1 A2C Algorithm

The Advantage Actor-Critic (A2C) algorithm follows a structured process to learn optimal policies and value functions within a reinforcement learning framework, and is classified as one of the better performing, state-of-the-art learning methods [12]. The following aims to explain in detail what each parameter of the model signifies and how they can be used to implement the A2C algorithm.

Markov Decision Process

Every parameter in the algorithm is important, since it influences how the agent behaves as it learns from interactions within its environment. The A2C approach is based on the Markov Decision Process (MDP) framework, which defines the structured environment in which an agent works to discover optimal policies [13]. States (s), actions (a), rewards (r), and transitions (s') define the MDP in the context of A2C and capture the decision-making situation in which the agent's goal is to maximise cumulative rewards over time.

Within this framework, the A2C algorithm employs both a policy model (actor) and a value function model (critic) [14]. The actor selects actions based on the current policy, parameterized by θ, which maps states to action probabilities, $P(s', r|s, a)$, aiming to maximize the expected return. The critic, on the other hand, evaluates the actions taken and determines their effectiveness. Figure 2.3 provides us with a high level overview of the A2C Markov Decision Process.

2.3.2 Actor and Critic

Actor-critic systems blend value-based and policy-based reinforcement learning, utilising two networks:

- The *Critic*: Estimates the value function, $Q_w(s, a)$. The aim of this component is to forecast the expected returns from a given condition, similar to value-based reinforcement learning techniques [14].

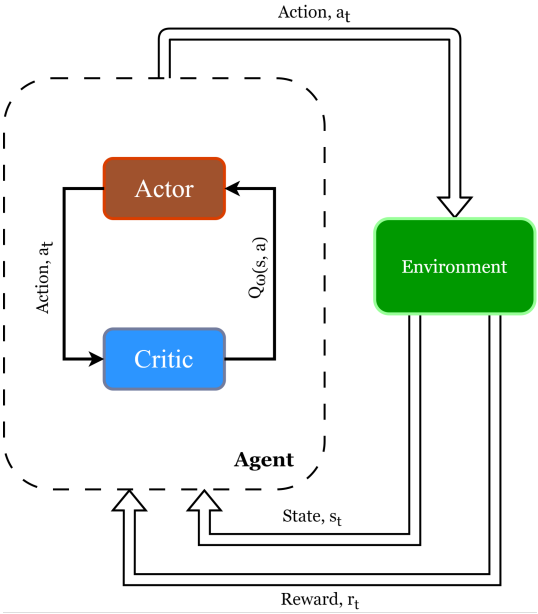

Figure 2.3: Simplified A2C Framework

- The *Actor*: Refines the policy based on the Critic's guidance. This is similar to policy-based reinforcement learning techniques, in which the goal is to directly learn a policy that links a state with the best course of action [14].

The Actor-critic algorithm employs an approximate policy gradient to navigate the learning process 2.1, where $J(\theta)$ is the loss function used to calculate the gradient with respect to θ. Expression 2.2 showcases the change in policy parameters (weights) of the actor model, where α denotes the learning rate, $\pi_\theta(s, a)$ is the policy function's output for action a in state s under parameters θ, and $Q_w(s, a)$ signifies the action-value function estimated by the Critic [15].

$$\nabla_\theta J(\theta) \approx \mathbf{E}_{\pi_\theta} \left[\nabla_\theta \log \pi_\theta(s, a) Q_w(s, a)\right] \tag{2.1}$$

$$\Delta\theta = \alpha \nabla_\theta \log \pi_\theta(s, a) Q_w(s, a) \tag{2.2}$$

One of the functions of the critic is to assess different possible outcomes of various actions that can be taken by the actor. This helps the system to avoid local maxima, and ensures that learning will continue until we have arrived to better strategies [16].

Advantage Function

The advantage function, denoted by Equation 2.3 is a a technique used by the critic, that reduces gradients by subtracting the cumulative reward from a baseline to reduce the variance of the policy gradient and result in better convergence than standard Q-values. We can also use TD error to arrive at Equation 2.4, which is a good estimator that eliminates the need to use two value functions [12].

$$A(s,a) = Q(s,a) - V(s) \tag{2.3}$$

$$A(s,a) = R + \gamma V(s') - V(s) \tag{2.4}$$

2.3.3 Conclusion

Through its actor-critic architecture, the A2C method combines the advantages of both policy-based and value-based approaches, marking a significant advancement in reinforcement learning. The aforementioned benefits of the Advantage Actor-Critic (A2C) algorithm, make it stand out as a more effective solution to the current regulator implementation.

Whilst A2C can in theory be a better fit to our model, we still need to experiment and gather insightful data, before we can declare with certainty that the current system has been improved. This is carried out in later sections, where the impact of this strategy on our outcomes is evident.

2.4 Cybernetics of Cybernetics

Second-order cybernetics, also known as "*cybernetics of cybernetics*", emphasises the observer's role within the system they are observing, which represents a significant shift in the study of systems [17]. Heinz von Foerster made a major contribution to this transition in 1974 when he distinguished between second-order cybernetics—the cybernetics of observing systems—and first-order cybernetics, which is centred on observed systems [18].

A simple way to understand this, is by considering the well-known example of the thermostat [19], which forms a thermostatic system comprised of a heater and a switch. In terms of

cybernetics, this setup does not distinguish between a controller and the controlled; instead, it functions as a circular, causative system.

The presence of an observer interacting with the thermostat, by setting it to a specific temperature, introduces another feedback loop into the system [20]. This interaction illustrates the essence of second-order cybernetics, where the observer's involvement directly influences the system's behavior, creating a more complex system.

2.4.1 Cybernetics in the context of family therapy

The field of family therapy has been profoundly impacted by the principles of second-order cybernetics, drastically altering the therapist's position in the therapeutic process. It's common knowledge in modern practice that therapists who monitor their clients are active participants in their treatment rather than passive observers. This involvement indicates that therapists are accountable for impacting the family system that is being observed, acknowledging their vital part in determining the dynamics and results of therapy.

The idea of incorporating second-order cybernetics principles into family therapy was developed in Milan in the 1970s and 1980s by a group of therapists, influenced by the works of Gregory Bateson [21]. This method was called 'Milan Systemic Family Therapy' [4].

Circular questioning emerges as a key method, concentrating on the interactions among individuals within a system, facilitating a deeper understanding of how each member's behavior influences and is influenced by others. This methodology resonates with Gregory Bateson's principle of double description, emphasizing the notion that causality within a system is both reciprocal and circular [22]. By offering an alternative viewpoint to the system's participants, we are able to provide additional information and insights that can enhance their understanding of the system to which they belong.

2.4.2 Importance of Perspective

The importance of perspective in cybernetics, particularly when dealing with partial information, is highlighted through a discussion on how the principles of second-order cybernetics have transformed the understanding of complex systems, including socio-technical and environmental systems [23].

The idea of "generative defamiliarisation," which draws inspiration from Gertrude Stein's literary devices [24], is creating unfamiliarity out of the familiar by approaching it from fresh or unexpected angles. By challenging assumptions and adopting a new perspective, it helps observers gain new insights into how systems function and identify areas for possible action.

These insights on perception enable us to understand that new technologies and techniques can act as conduits to broader horizons. This concept forms the basis of the foundational principle of second-order cybernetics. It should also be emphasised that we are discussing a world/system characterised by its members possessing incomplete information, contingent upon their view point. This means that our observer will need to not only gain insight from their unique standpoint, but also use information that is not readily available to the rest of the regulated units, enhancing the realism of our model and aligning it with our earlier findings.

2.4.3 Second order Model

To integrate second-order cybernetics into our system, we must adapt the existing model that incorporates the "*4voices*" algorithm [1]. Figure 2.4 illustrates a framework where the system's observer is recognized as an active participant, fundamentally contributing to the system's dynamics.

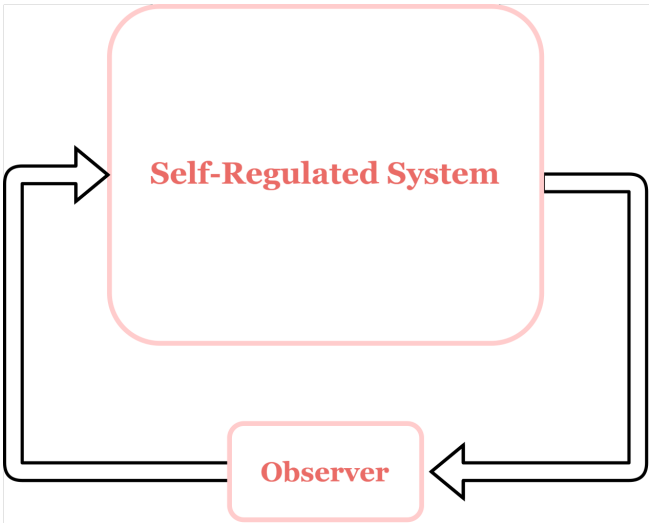

Figure 2.4: Proposed Complex System

The idea behind this restructure, is to implement the concepts of second order cybernetics, to allow an observer to provide insight to the regulated agents, through the information that they gain when overlooking the self regulated system. This method allows for the incorporation of previously

addressed ideas related to perception, providing the regulated units with a new perspective. A more detailed framework is provided in the Implementation section, where we introduce the modifications made to facilitate interactions between observer, regulator and regulated units.

2.4.4 Summary

Second-order cybernetics emphasises the observer's active role within the system, marking a significant shift from first-order cybernetics. The thermostat example [20] exemplifies this concept, for a system where observer interaction adds complexity through additional feedback loops. In family therapy, second-order cybernetics has transformed therapists from passive observers to active participants, using techniques like circular questioning [22] to understand reciprocal influences within systems.

In our model, we integrate second-order cybernetics by adapting the "4voices" algorithm to recognise the observer as an active participant. This allows the provision of valuable insights to the regulated agents, enhancing their perspectives and addressing the concept of generative defamiliarisation [24].

2.5 Third and Fourth Order Cybernetics

From the foundational principles of first-order cybernetics to the reflective nuances of second-order cybernetics, researchers have continuously expanded our understanding of how systems operate and adapt. This has given rise to third and fourth-order cybernetics, each offering deeper insights into the dynamics of observer interactions and the cognitive processes that underline system behaviour.

2.5.1 Third Order

The term and its theoretical framework were developed as an extension of the principles of second-order cybernetics, which itself evolved from first-order cybernetics. During its early formulation, a prominent figure named Ranulph Glanville emphasised the complexity of interactions among multiple observers and the reflexivity involved in such systems [5]. The idea was further explored and expanded upon by researchers such as Philip Boxer and Vincent Kenny. In the paper, "The Economy of Discourses: A Third Order Cybernetics?" [25], we are introduced to the necessity of third order cybernetics in understanding complex, adaptive systems where the roles of observer and

participant are interconnected and dynamic. This approach recognises that solving complex issues requires more than just different perspectives, as it requires an awareness of how these perspectives interact and influence each other. Third order cybernetics emphasises that each observer's viewpoint is also contingent on their interactions with other observers, which is something that needs to be acknowledged, in order to come to a viable solution [25]. This requires continuous reflection and communication among all observers to ensure that their collective insights contribute to the overall success of the system.

Implementing third order cybernetics in a self-regulated system, can enhance its effectiveness and robustness in several ways. Incorporating multiple observers within the system allows for a more diverse range of perspectives and solutions. Therefore, the regulated units will not have to depend on the methodology or viewpoint of a single observer; instead, they benefit from the collective efforts and perspectives of multiple observers working together for the collective good. This will ensure a more flexible and responsive system that can quickly adapt to changes and uncertainties.

Finally, When defining third-order cybernetics, it's crucial to acknowledge the different interpretations that may exist. For instance, in the context of ethical regulators and systems which this thesis is built upon, third-order cybernetics doesn't necessarily involve a series of additional feedback loops among various observers. Rather, it signifies an observer's ability to reevaluate and self-reflect on their actions [2]. This reflexivity is not only guided by the system's influence, but also by the observer's own morals and judgment. This highlights the importance of personal ethics and accountability when making decisions that affect a many other individuals.

2.5.2 Fourth Order

Fourth order cybernetics represents an advanced stage of cybernetics, making it an extension of first, second, and third order cybernetics. It emphasises self-awareness, self-observation, and holistic integration within complex systems [26]. This level of cybernetics is centered around understanding the functioning of our minds, including our use of rationality and language, and how these processes help us in staying coherent and balanced in our thinking.

Mancilla's view on fourth-order cybernetics emphasises high-level cognitive processes. He talks about rationality as a sort of cognitive machine that helps maintain coherence in our thoughts [27]. This perspective is rooted in constructivist epistemology, which means that knowledge is actively built by individuals rather than passively received. Mancilla's approach looks at how our individual rationality interacts with collective rationality in social contexts, including power structures, culture, and institutions [27].

On the other hand, other perspectives on fourth-order cybernetics, such as those integrating the Universal Dialectical Systems Theory (UDST) and the Cybernetics of Conceptual Systems, focus on a holistic approach [26]. These views aim to address complex socio-economic and environmental issues by achieving a balance between natural and man-made systems. This approach highlights the interrelation of all components within a system and the need for sustainable and balanced development.

Both views recognise the observer's active role within the system and the importance of different perspectives in understanding complex systems. However, Mancilla's framework dives deeper into the cognitive aspects and their interactions, while other views may concentrate more on the broader applications of these principles to achieve sustainable development in various systems.

2.5.3 Summary

Third and fourth-order cybernetics explore complex interactions and cognitive processes within systems. Third-order cybernetics focuses on the relationships of multiple observers, with continuous reflection and communication to enhance the robustness of a system [25]. It is also relevant in ethics, with personal accountability being a major driver in decision making [2].

Fourth-order cybernetics goes further with self-awareness and holistic integration. Mancilla's approach highlights cognitive processes [27], while other perspectives focus on achieving sustainable development [26]. Together, these advanced cybernetic frameworks help us understand and manage the complexity of modern systems more effectively.

2.6 Partial observability and Belief System

In many real-world systems, agents operate with limited information about their environment and the state of other agents. This concept, known as partial observability, is a critical aspect in fields like robotics, economics, and artificial intelligence [28]. Partial observability challenges agents to make decisions based on incomplete data, often requiring sophisticated mechanisms for belief updates and decision-making under uncertainty.

The misinformation effect shows how exposure to misleading information can alter an individual's memory and perception of an event [29]. The concept was further explored in various studies, such as those reviewed by Ayers and Reder (1998) [30], which delve into the theoretical explanations behind the misinformation effect, including the activation-based memory model. In our suggested environment where agents have only partial observability, misinformation can introduce noise into the system, leading to incorrect beliefs and subsequent actions. This is particularly relevant in our multi-agent system. where each agent's perception is influenced not only by their own observations but also by the information received from others (foreground and background noise), which may be flawed or biased.

2.6.1 Partially Observable Markov Decision Process (POMDP)

Here, we address scenarios in environments where agents are confronted with the challenge of incomplete and imperfect perception, a common reality that complicates learning processes traditionally based on Markov Decision Processes (MDPs) [31]. Given that MDP-based methods require complete observability, this limitation necessitates an exploration of alternatives. Specifically, we examine cases where an agent's observations about the environmental state are clouded by noise and incompleteness. As a result, we introduce an evolved formal model known as a Partially Observable Markov Decision Process (POMDP).

POMDPs extend MDPs by incorporating uncertainty in observations, allowing agents to maintain a belief state, which is a probability distribution over all possible states [32]. This belief state is updated as the agent receives new observations, enabling it to make informed decisions despite the lack of information. The belief update process involves Bayesian inference, which combines prior knowledge with observed evidence to refine the agent's understanding of the environment [33].

2.6.2 Belief System Update Mechanisms

To manage partial observability effectively, agents must update their beliefs about the state of the environment and other agents. In this thesis, we implement a belief system update mechanism using a Kalman filter-based approach [34]. This method provides a recursive solution to estimate the state of a dynamic system from noisy observations. The details on how our model is adapted to incorporate partial observability are provided in subsequent sections.

2.6.3 Kalman Filter Equations

The inspiration behind these equations is the need to estimate the state of a dynamic system from noisy measurements. The Kalman filter addresses this, by balancing the uncertainty in the prior estimate with the uncertainty in the new measurements. It is designed to update the estimate of the current state based on the latest measurements and the previous state estimate [34].

Kalman Gain Calculation

$$K_k = \frac{P_{k|k-1}}{P_{k|k-1} + R_k} \tag{2.5}$$

The Kalman gain K_k determines the weight given to the new measurement versus the prior estimate. It is computed based on the prior variance $P_{k|k-1}$ and the measured noise variance R_k [35].

Posterior Mean (State Estimate Update)

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + K_k(z_k - \hat{x}_{k|k-1}) \tag{2.6}$$

The posterior mean $\hat{x}_{k|k}$ updates the prior mean $\hat{x}_{k|k-1}$ by incorporating the new observation z_k, weighted by the Kalman gain K_k.

Posterior Variance (Error Covariance Update)

$$P_{k|k} = (1 - K_k)P_{k|k-1} \tag{2.7}$$

The posterior variance $P_{k|k}$ represents the updated estimate of the error covariance, reduced by the factor $(1 - K_k)$, indicating the reduced uncertainty after incorporating the new measurement.

These equations form the core of the Kalman filter algorithm, enabling it to provide optimal state estimates in the presence of noise and uncertainty.

2.6.4 Belief Update Algorithm

Each agent follows the steps outlined in by Algorithm 1, to update its beliefs about the state of other agents. Each agent in the system has an initial set of priors for other agents and a set of observed noises. The condition $i \neq j$, ensures that an agent does not update beliefs about itself. This is necessary because an agent's beliefs about its own state are not relevant. This algorithm leverages the Kalman filter to continuously refine the beliefs of each agent in the system about the states of other agents

for *each agent* **i** *in the system* **do**
  for *each other agent* **j** *in the system* **do**
    if $i \neq j$ **then**
      Retrieve prior state:
      Prior Mean,
      Prior Variance
      Compute Kalman gain
      Compute state estimate (Posterior Mean)
      Compute Posterior Variance
      Update:
      Agent Belief (Posterior Mean)
      Kalman Filter States (Posterior mean, Posterior Variance)
    end
  end
end

Algorithm 1: Belief Update Algorithm Using Kalman Filter

2.6.5 Summary

This section addresses the challenges of partial observability in multi-agent systems, where agents operate with limited information about their environment and other agents [32]. Agents often make decisions based on incomplete data, which requires sophisticated belief update mechanisms to handle uncertainty, as exposure to misinformation can alter agents' beliefs and actions [30]. A Kalman filter-based approach is introduced for belief updates, providing a recursive solution to estimate the state of a dynamic system from noisy observations [35]. Key equations include the calculation of the Kalman gain, posterior mean, and posterior variance, which help balance uncertainties in prior estimates and new measurements.

3

Implementation

Contents

3.1 Technical Progression

This section outlines the key modifications implemented in our system to enhance its performance and stability, and gain insights on the various psychological concepts we introduced. We compare the previously used Deep Q-Network (DQN) with the Advantage Actor-Critic (A2C) algorithm, used in the regulator to aid with policy selection. This will provide us with enough evidence as to determine whether the performance of the system can be improved, by utilising a reinforcement learning algorithm that better suits our custom environment.

We also introduce an observer as part of our second-order cybernetics modification, whose role is to gather and process information on agent interactions and past voice selections, subsequently influencing both the regulated units and the regulator. This modification aims to promote coherent collective expressions and improve systemic stability by ensuring that agents' expressions are aligned with the system's goals.

Lastly, we implement partial observability by introducing noise to the individual voices observed by each agent. This modification simulates real-world scenarios where observations are imperfect and subject to various disturbances. We compare the effectiveness of the Kalman gain belief update algorithm against a simpler belief update method, assessing their performance under varying noise levels.

3.2 Experimentation

We follow a similar experimentation procedure, as in the 'Requisite Social Influence in Self-Regulated Systems' [1] paper. Below we outline the important characteristics of the problem that was defined regarding job scheduling in cloud computing, and the environment that was set up for learning to occur.

- The system operates in epochs where each agent delegates jobs to the cloud, which may or may not be included in the process queue for the next epoch.
- It involves a dynamic and non-deterministic environment with multiple heterogeneous, networked units.
- The preferable policy for job scheduling depends on context, opinion, and self-adaptation.
- Metrics to describe system operation are relatively straightforward to define, such as job size, urgency, processing order based on urgency and size, total and average costs, and delays.

This case illustrates the complexities and difficulties of continuous monitoring and management in a self-regulated system in the context of cloud computing job scheduling. For a more in-depth explanation of the problem we are trying to solve and its design, one can refer to the works of the original paper [1].

3.3 A2C Modification

In this section, we focus on the implementation details that underpin our experimental setup. We utilise Stable-Baselines3 [9], a comprehensive and user-friendly Python library, which facilitates the deployment and testing of deep reinforcement learning algorithms. Additionally, we outline the experimental framework and methodology used to evaluate the performance of our regulator, using two different reinforcement learning algorithms: A2C and DQN.

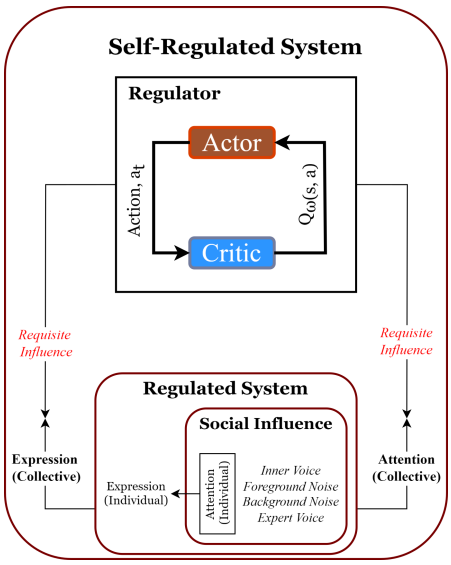

Figure 3.1: Modified Self-Regulated System to implement A2C reinforcement learning

Figure 3.1 illustrates the updated framework for implementing the A2C algorithm within the regulator, effectively replacing the previously utilised DQN approach.

3.3.1 Stable Baselines3

Stable-Baselines3 is a robust, reliable, and user-friendly Python deep reinforcement learning library. It sets itself apart with its rich documentation, simple, yet powerful API, and implementations of state-of-the-art algorithms. Following, we outline a few of its advantages over other reinforcement learning software [9]:

- **User Friendly:** Many libraries target experienced RL researchers requiring expert knowledge, Stable-Baselines3 is designed to be accessible for users with varying levels of experience. It offers axhaustive documentation, and tutorials.

- **Balanced Modularity:** Some libraries' modular nature enables them to quickly incorporate advancements from different papers, meaning that it is required by the users to fully comprehend the entire codebase in order to modify their algorithms. Stable Baselines3 simplifies this process, thus minimising the amount of code users need to understand to modify an algorithm.

3.3.2 Side-by-Side Comparison

Below we provide a comparative view of the results we obtain when running the experiment using DQN and A2C as our reinforcement learning algorithms. The y-axis indicates the policies that the regulator selects for the regulated system, which represent the number of agents whose jobs are chosen for processing. The x-axis represents the number of epochs passed since the start of the experiment. The colour of the points on the plot convey the level of noise that the regulated units feedback to the regulator (Collective Expression), with 'red' meaning high noise and bad policy selection, and vice versa.

The way the experiments are run, is by determining the attention of the agents with respect to their inner voices. In Figure 3.1 we can see that when the agents choose to listen to the voice that gives the best immediate effect ('Ind' Update), it becomes harder for the regulator to learn which policies to choose, as it is not being provided with constructive feedback, but rather individual opinions that have no formation or coherence. Therefore we cannot make any clear observations whether or not A2C is performing better than DQN.

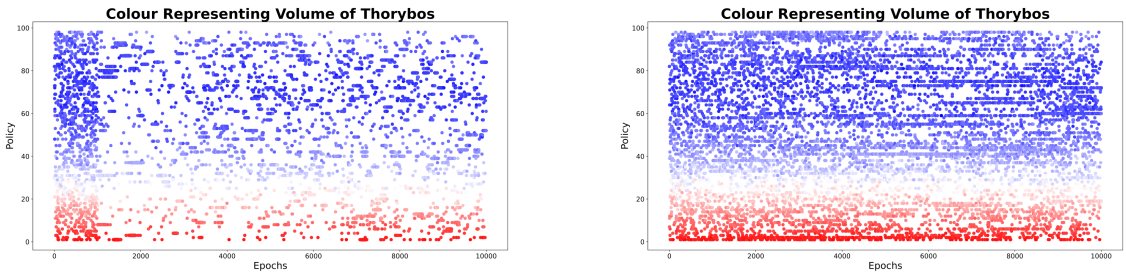

Figure 3.2: Policy selection of an RL regulator using DQN (left) vs A2C (right) and corresponding collective expression for a single run ('Ind' Update)

When we shift the attention of the agents to listen more to the expert opinions ('Exp' Update) or the collective average opinion ('Col' Update), we can more clearly pinpoint the differences between the two algorithms (Figures 3.2 and 3.3). For both cases, the Advantage Actor-Critic method seems to reduce the number of bad policies chosen by the regulator, since then number of 'red' points for later epochs is significantly less than that of DQN. We also understand that the regulator, when using A2C, is able to have less randomness when picking out the best policies, indicated by the smaller spread of 'blue' points.

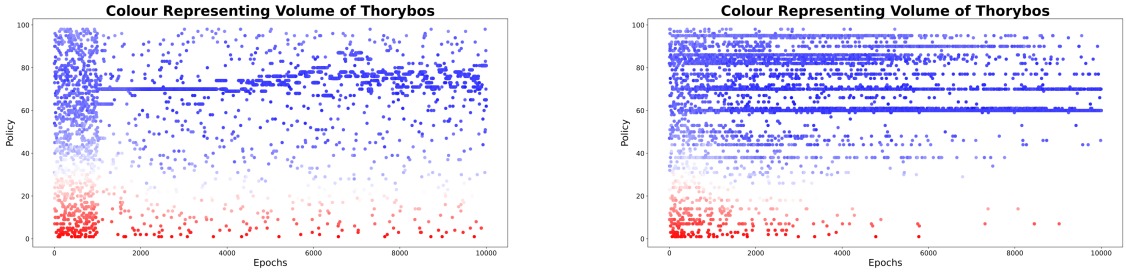

Figure 3.3: Policy selection of an RL regulator using DQN (left) vs A2C (right) and corresponding collective expression for a single run ('Col' Update)

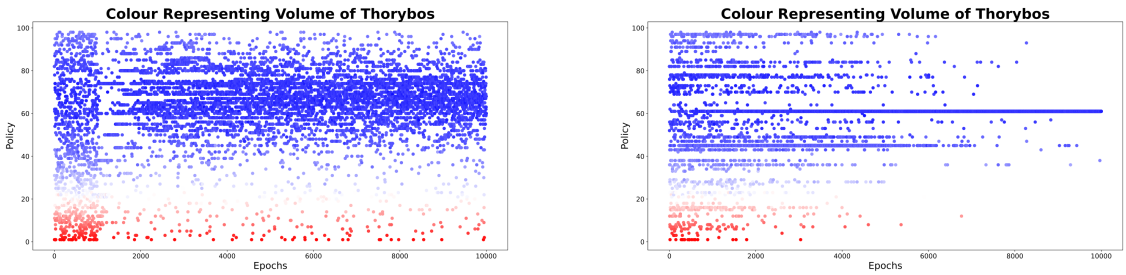

Figure 3.4: Policy selection of an RL regulator using DQN (left) vs A2C (right) and corresponding collective expression for a single run ('Exp' Update)

By averaging the performance of the regulator over 10 runs, as seen in Figure 3.4, we can see that the rate of learning for the A2C regulator is more gradual than that of the DQN. It is also apparent that after 4000 epochs, the range of good policies selected is smaller for the A2C, making it more precise and consistent with its policy selection (DQN: 55-75, A2C: 60-70). When looking at the number of 'light blue' points, compared to 'dark blue' points, we can see that for later epochs the ratio of great policies chosen ('dark blue') to good policies ('light blue') is higher for A2C than DQN.

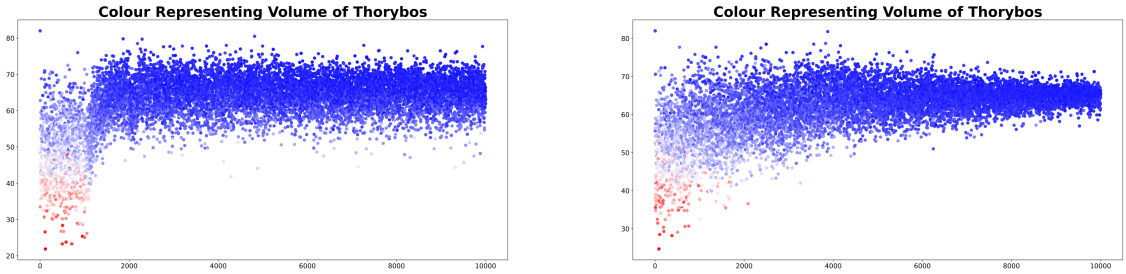

Figure 3.5: Policy selection of an RL regulator using DQN (left) vs A2C (right) and corresponding collective expression, averaging over 10 runs ('Exp' Update)

3.3.3 Result Interpretation

We have seen that there are differences in performance for both algorithms, and so we attempt to explain why this may be the case for our specific problem:

- The DQN technique may initially choose policies at random because it is trying to learn the Q-value function and study the environment. As it continues learning from the feedback, it begins to leverage its newly acquired understanding to choose better policies that generate less noise.

- The A2C system is predictable and synchronous; it updates by averaging over all actors once each actor completes their segment of experience. This might account for the A2C method's apparent improvement in outcomes, with essentially no evidence of the regulator selecting poor policies in subsequent epochs.

- Because of its dynamic and non-deterministic nature, the environment is ideal for A2C. As an on-policy approach, A2C discovers the best policy for the present, constantly shifting environment. This may result in learning that is more effective with improved performance on these types of tasks. Due to the aforementioned, it might be able to identify which policies generate less noise more quickly, which could stop the selection of bad policies in later epochs.

- Due to the computational demands associated with storing and retrieving past experiences through a replay buffer, DQN can be resource-intensive[36]. In contrast, A2C eliminates the need for a replay buffer, thus reducing computational overhead and potentially accelerating the learning process. This reduction in randomness and a more consistent selection of policies over numerous epochs could be attributed to the regulator utilising A2C.

3.4 Second Order Cybernetics Modification

The implementation of second-order cybernetics within our system involves introducing an 'observer' whose role is to gather, process, and influence the regulated units and the regulator. As illustrated in Figure 3.6, the observer collects data on agent interactions and the past voice selections of each agent. This collected information is then used to determine the rewards given to the regulator and to propose a voice selection for each agent. The past voice selection corresponds to one of the four voices that a specific agent attended to in the previous epochs, based on the designated update method ('Ind', 'Col', 'Exp'). Interactions are tracked to identify instances where one agent influences another, providing a comprehensive view of the system dynamics.

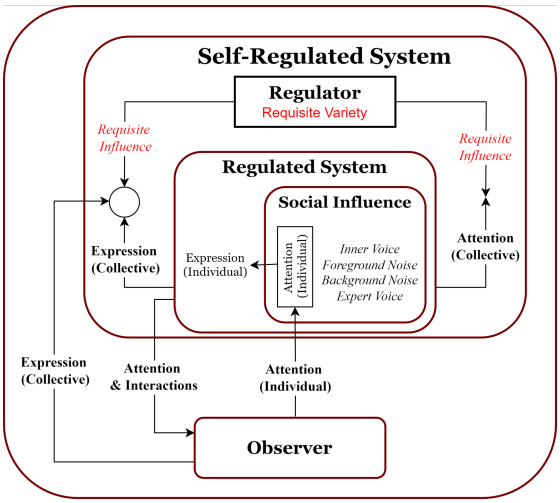

Figure 3.6: Modified Self-Regulated System to implement second order cybernetics

As we have seen previously, and as outlined in the 'Requisite Social Influence in Self-Regulated Systems' [1] paper, when the regulated units update their attention to a voice based on the best immediate effect ('Ind'), we do not have systemic stability. This instability occurs because the regulator cannot effectively learn and subsequently select optimal policies. To address this issue, our experiment focuses on how an observer can promote a more coherent collective expression that aligns more closely with the 'Exp' and 'Col' updates, thereby restoring systemic stability. By doing so, we aim to ensure that the observer facilitates better learning and decision-making processes, leading to improved overall performance and stability of the regulated system.

3.4.1 Observer Role

The observer has a level of trust associated with each agent, which fluctuates based on whether the agent's individual expression has improved or deteriorated compared to the previous epoch. This method of updating trust was selected to reflect how an agent might seek alternative sources of information or perspectives when the policies provided by the regulator consistently fail to meet its expectations. As a result, agents initially update their trust in a given voice based on the immediate effect ('Ind'). Over time, as the epochs progress, the trust in the observer increases. Eventually, the agents begin to follow the observer's voice attention propositions rather than relying solely on the 'Ind' update. Once that shift of attention occurs, then the observer will start to have an influence on the regulated units and regulator.

The observer in our system serves two primary purposes. Firstly, it proposes which of the four voices a specific agent should focus on. Secondly, it provides the regulator with a reward based on these voice propositions. The following outlines the process by which the observer operates:

Data Collection and Storage

- **Interaction Tracking:** For every agent in the system, we track the number of interactions they have had with every other agent. An interaction occurs when an agent's expression is influenced by another agent. This tracking helps us understand the frequency of interactions between agents.
- **Past Voice Selection:** We maintain a record of up to 10 past voice selections for each agent. This historical data is crucial for evaluating patterns and making informed decisions.

Selection of Influential Agents

- **Identifying Top Interactions:** For each agent, we identify the top 25^{th} percentile of other agents based on the number of interactions. This selection criterion ensures that we focus on agents with significant interaction histories, which are likely to provide valuable insights.
- **Capturing Past Voices:** From these top-interacted agents, we store one past voice selection from each. This step captures the most relevant historical data from influential peers.

- **Secondary Influences:** For each agent identified in the top 25th percentile, we determine the agent they have interacted with the most. We then store one past voice selection from these secondary influences as well. This process helps in capturing a broader perspective of influential voices.

Voice Selection Process

- **Finding Common Voice Selections:** We analyze the stored voice selections to find the most common voice selection among the influential agents and the most common voice selection from the current agent's 10 past selections.
  - If the most common voice selection from the influential agents matches the most common selection from the agent's own history, we propose that selection.
  - If the selections differ, we randomly choose between the two most common selections to ensure diversity and adaptability in decision-making.
- **Random Proposals:** There is a 0.2 probability that we propose a random voice selection from the four available options. This randomness is introduced to prevent stagnation and promote exploration, especially when there are insufficient past selections to analyse.

This process is repeated for every agent in the system, ensuring that each agent's voice selection is influenced by both their own historical data and the feedback of their interactions.

Regulator Feedback

- After determining the proposed voice selections for all agents, we identify the most common voice proposition among all agents. This most common voice proposition is then used to determine the reward that we feed to the regulator. The regulator uses this feedback to adjust policies and improve overall system performance.

3.4.2 Results and Interpretation

Firstly, let's examine Figure 3.7, which displays the collective expressions averaged over 10 runs for the 'Ind' update method (best immediate effect) without the observer implemented. It is evident that over the course of 10,000 epochs, the regulator fails to satisfy the regulated units, as it selects policies that do not align with the agents' expressions. This demonstrates systemic instability and indicates that the pathways of requisite social influence are not established, as the regulator appears unable to learn from the units' expressions.

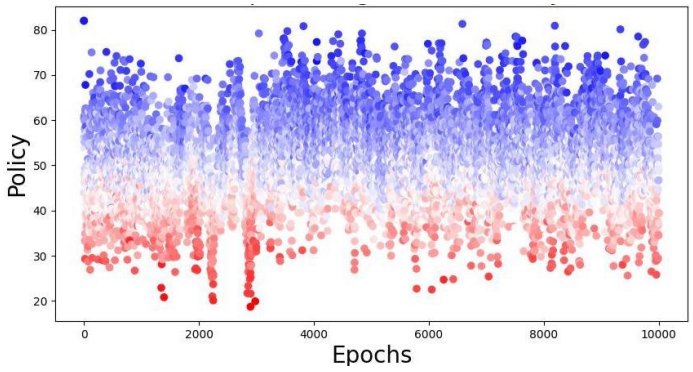

Figure 3.7: Policy selection of an RL regulator and corresponding expression, averaging over 10 runs ('Ind' Update)

In contrast, when the experiment is conducted with the observer implemented, as shown in Figure 3.4.2, we observe that after some iterations, the regulator begins to better align with the regulated units' expressions.

Initially, up until around epoch 3000, the system remains unstable with the regulator struggling to learn effectively. However, in the later epochs, we can clearly see a shift, as the regulator starts to select policies that yield greater satisfaction among the agents. This results in outcomes similar to those obtained with the 'Exp' and 'Col' update methods. This shift from instability to systemic stability marks the point where agents start to follow the observer's attention propositions rather than the 'Ind' method. This transition is further validated by the plots in Figure 3.4.2, which show the mean and standard deviation of the regulated units' individual expressions, where at around epoch 3000, coinciding with the shift in trust towards the observer, there is a noticeable change in the diversity of individual expressions.

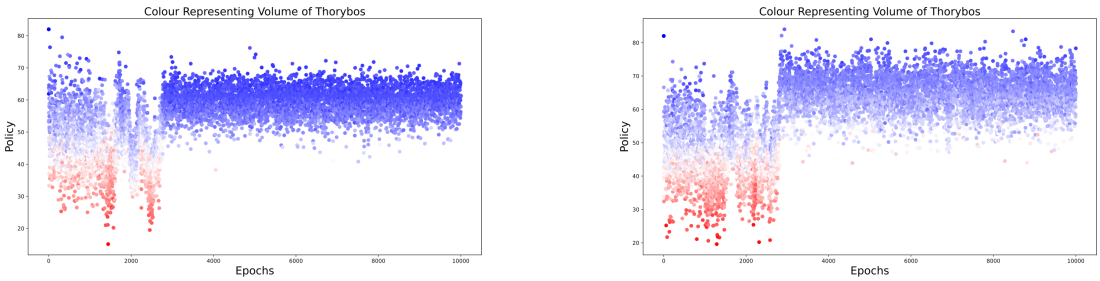

Figure 3.8: Policy selection of an RL regulator and corresponding expression for two separate attempts (left and right), averaging over 10 runs (Observer Implemented)

For further context, Figure 3.4.2 presents two separate scenarios to illustrate the variability in voice propositions that the observer might make from one run to the next. This variability arises because the observer proposes voice attention to the agents based on their interactions and historical values, making each run unique. Consequently, the observer influences the system differently depending on the system's characteristics.

As shown in Figure 3.4.2, the agents' attention can shift to any of the voices, each affecting the system's stability in distinct ways. The y-axis indicates the specific agent and each dot signifies the voice that the agent attended to during the epoch (Blue - Individual Voice, Black - Foreground Noise, Green - Expert Voice, Red - Background Noise). In Figure 3.4.2, the plot on the left demonstrates an outcome where, over 10 runs, the observer frequently suggests shifting attention to the expert or background noise, leading to a particular pattern of system stability. Conversely, the graph on the right shows a scenario where the observer more often proposes a shift to individual or foreground noise, which might result in lower overall satisfaction among the agents, as indicated by the fewer dark blue dots.

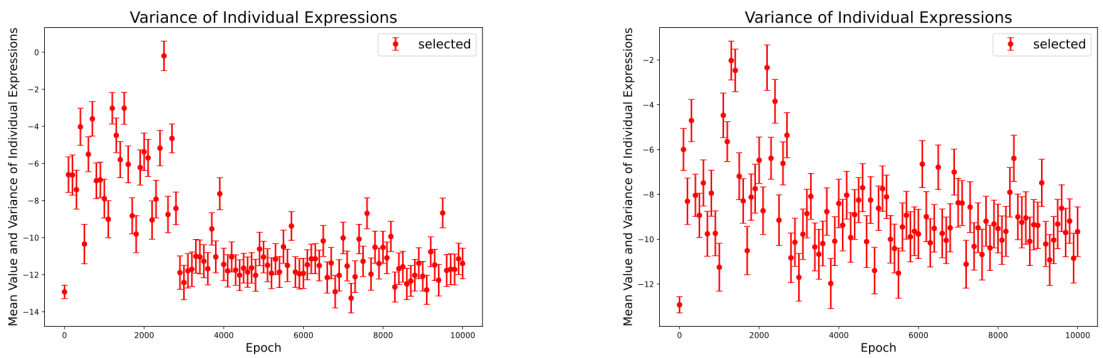

Figure 3.9: Mean and standard deviation of individual expressions (Observer Implemented)

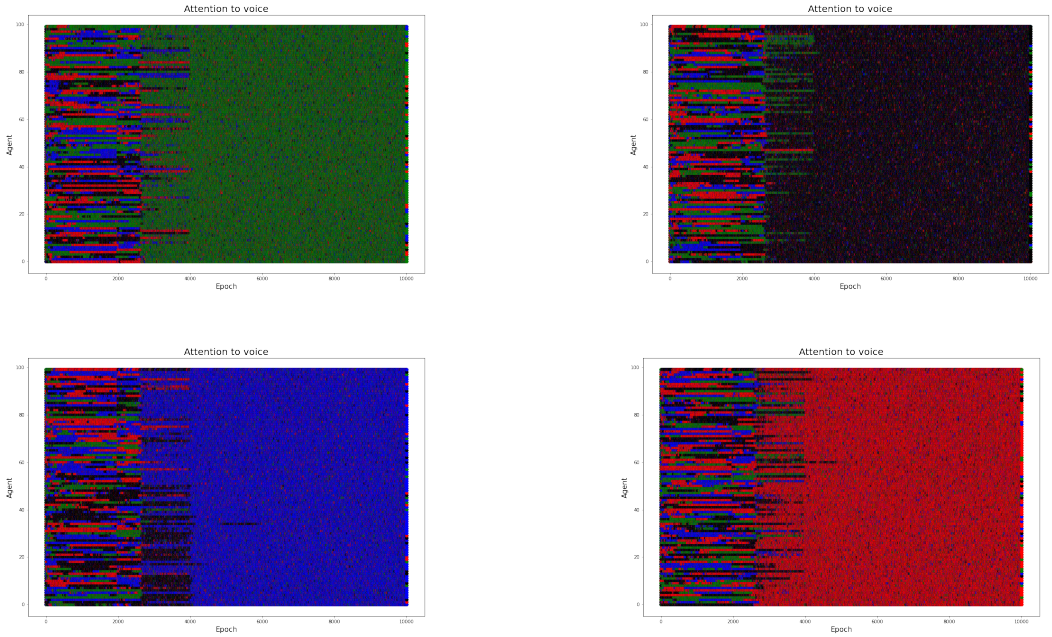

Figure 3.10: Individual Voice selection in four different scenarios

3.5 Partial Observability Modification

In the base system, each agent has access to the exact, accurate individual voice that each other agent expresses at any given point in time. To make the system more realistic, we modify this setup, as seen in Figure 3.11, by introducing noise to the individual voices observed by each agent. This means that the perceived voices differ between agents, introducing variability and uncertainty.

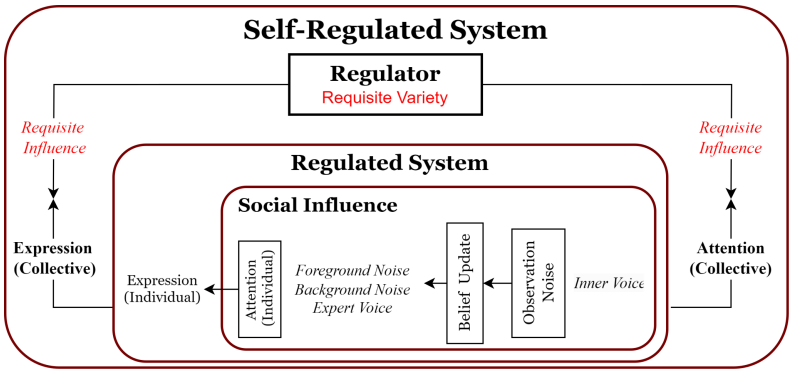

Figure 3.11: Modified Self-Regulated System to implement partial observability and belief update

As we can see, a belief update system is also incorporated, where each agent, based on its observations of other agents' voices, is responsible for updating its beliefs about what those voices might be, due to their noisy nature.

3.5.1 Noisy Observations

We introduce noise to the individual voices by adding random noise to the observed state of other agents (individual noise). This noise is generated using a Gaussian distribution with a mean of 0 and a specified variance. This approach simulates the real-world scenario where observations are imperfect and subject to various disturbances. It is important to note that this modification will have an effect on the foreground and background noises, meaning that we expect to see how misinformation coming from other agents can impact the individual expression, according to the 4 voices algorithm.

In addition to introducing noisy observations, we implement the Kalman gain belief update algorithm for each agent. This algorithm is designed to optimally update beliefs by balancing the uncertainty in the prior estimate with the uncertainty in the new measurements. The performance of the belief system using the Kalman filter is compared to a simpler belief update system to assess its effectiveness.

3.5.2 Simple Belief Update System

The simple belief update system uses a straightforward approach to update the new belief based on the observed noise. The update formula is as follows:

$$New\ Belief\ =\ g\ *\ (Observed\ Noise\ -\ Prior\ Belief) \tag{3.1}$$

,where g is a constant between 0 and 1. For our experiment, we use $g\ =\ 0.5$, meaning that each new observation contributes equally to the prior belief, balancing between the old belief and the new observed data.

This method is simpler, as it does not account for the varying uncertainty in the measurements and the prior estimates. Instead, it applies a fixed gain to adjust the beliefs. The simple belief update system provides a baseline for comparison. It highlights the impact of a more sophisticated algorithm like the Kalman filter. The reasoning behind this choice is outlined below:

- **Baseline Comparison:** The simple update method serves as a control to demonstrate the effectiveness of the Kalman filter. By comparing results from both methods, we can quantify the improvements brought by the Kalman gain algorithm.

- **Easy to Implement:** Its simplicity makes it easy to implement and understand. It involves basic arithmetic operations without the need for advanced statistical computations.
- **Limited Computational Resources:** The simple update method is computationally less intensive compared to the Kalman filter, making it suitable for quick, approximate updates.

3.5.3 Experimentation

The experiment is conducted by comparing how the noise levels of the collective change as we increase the variance of added noise from 1 to 10 for each of the aforementioned belief update methods. Additionally, we plot the mean and standard deviation of the root mean squared errors (RMSE) for each agent at every epoch. The RMSE is calculated as follows:

1. For a specific agent, calculate the difference between its belief and the actual voice of each other agent, then square this difference.
2. Sum these squared errors for all other agents, divide by the number of agents, and take the square root of the result.
3. Perform this calculation for the beliefs of every agent.
4. Compute the mean and standard deviation of the RMSEs for the current epoch.

The mathematical formulation of the RMSE for a given agent and epoch is expressed as:

$$RMSE_{i,k} = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(\hat{x}_{j,k} - x_{j,k})^2} \tag{3.2}$$

where:

- N is the number of agents.
- $\hat{x}_{j,k}$ is the belief of agent i about the state of another agent, j, at epoch k.
- $x_{j,k}$ is the actual voice (true state) of the other agent, j, at epoch k.

This procedure allows us to quantify the accuracy of the belief update methods under varying levels of noise, providing insights into their performance and robustness.

3.5.4 Results and Interpretation

When conducting our experiment with added noise at a variance of 1, we obtain the results found in Figures 3.12 and 3.13. These results showcase that even with noisy observations, the regulator can still learn and select policies effectively. However, closer inspection of our plots reveals instances where the collective expression of the agents diverges from what it should ideally be, despite the regulator choosing a good policy. For example, in Figure 3.12, first column, first row, which shows the collective expressions for one run using the Kalman update, we observe the impact of noise on the agents. There are instances where red dots ('bad' policy) appear instead of blue dots ('good' policy), or light blue/red dots ('average' policy) instead of dark blue dots. This indicates that agents tend to estimate the opinions of other agents to be worse than they actually are, thereby negatively influencing their own perception of the policy.

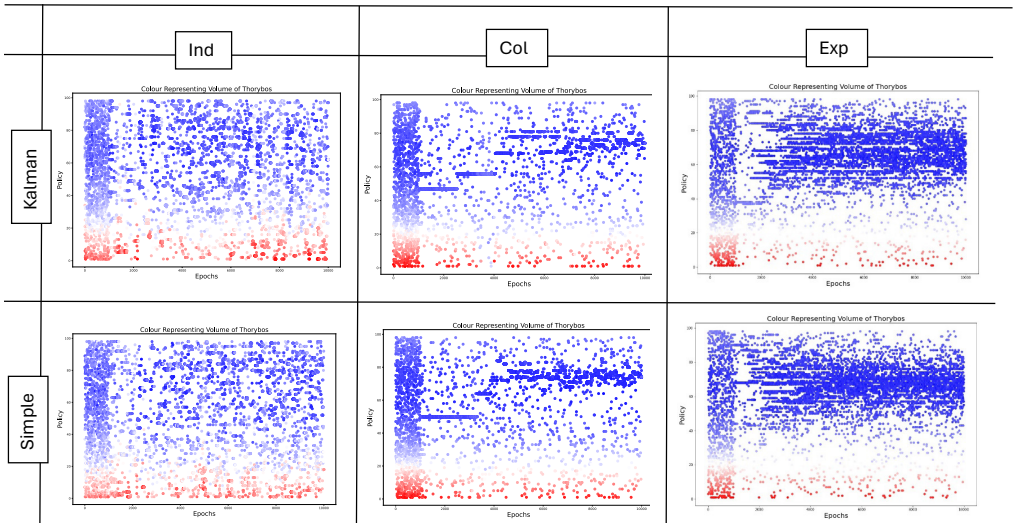

Figure 3.12: Policy selection of an RL regulator and corresponding expression for different types of attention, for a single run (Kalman vs Simple, Noise Variance = 1)

This reflects the misinformation effect, where agents are influenced by incorrect information despite their own experiences with the given policy. The misinformation leads to skewed perceptions and lower overall satisfaction among the agents, highlighting the importance of accurate information.

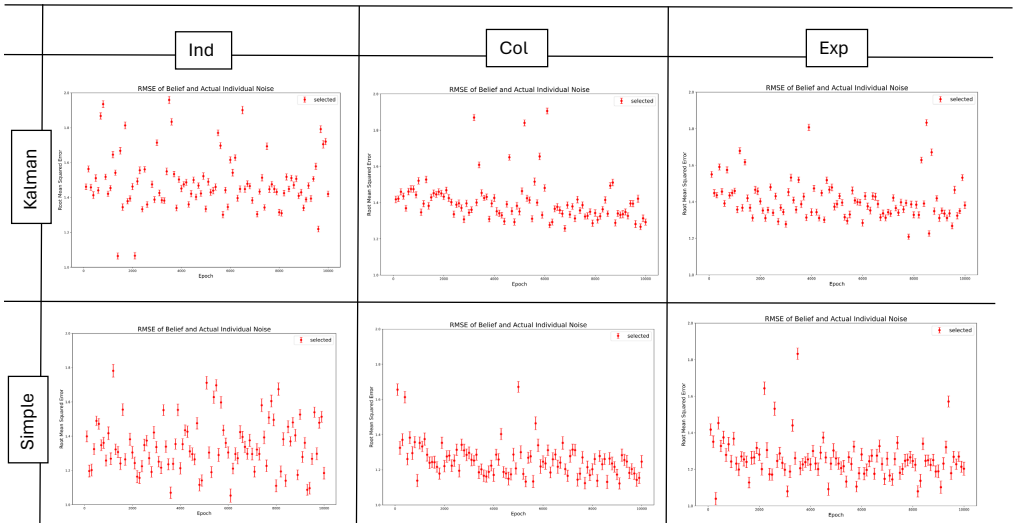

Figure 3.13: RMSE for different types of attention, for a single run (Kalman vs Simple, Noise Variance = 1)

This effect becomes even more pronounced when we increase the variance of added noise to 10. As we can see in Figure 3.14, second row, where we display the outcome of a single run using the simple update method, there is a noticeable overall reduction in agent satisfaction. As the noise variance increases, the agents' ability to accurately estimate the opinions and states of other agents worsens. This level of noise amplifies the misinformation effect, causing agents to form beliefs that are further from reality. Consequently, even when the regulator selects policies that would otherwise be effective, there is a fall in the collective satisfaction levels.

In contrast, the Kalman update run (first row) demonstrates performance that is similar to the scenario with lower noise variance. This consistency is evidenced by the fact that the Root Mean Squared Error (RMSE), seen in Figures 3.13 and 3.15, remains at similar levels between the low and high noise variance cases. This stability is different from the simple update method, where the RMSE increases significantly from approximately 1.3 to 5.75. In addition, as we reset the beliefs of each agent every 2000 epochs, it is clear from the Kalman gain RMSE graphs (Figure 3.15), that it is able to stabilise after every reset.

This scenario signifies the critical importance of robust belief update mechanisms, such as the Kalman filter, in mitigating the adverse effects of high noise levels. The simple update method, which is not satisfactory when adjusting dynamically to varying noise levels, shows a clear drop in performance and satisfaction, highlighting its limitations under more extreme conditions.

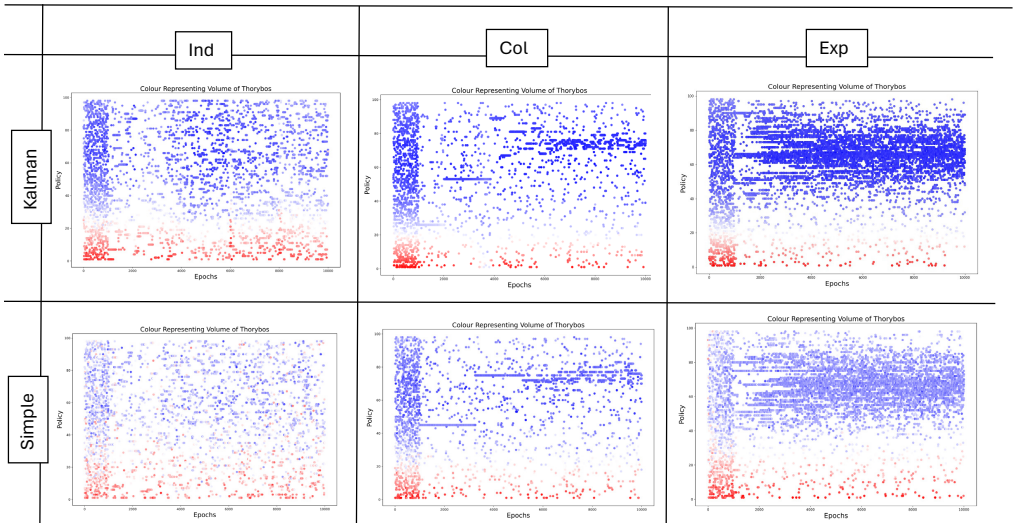

Figure 3.14: Policy selection of an RL regulator and corresponding expression for different types of attention, for a single run (Kalman vs Simple, Noise Variance = 10)

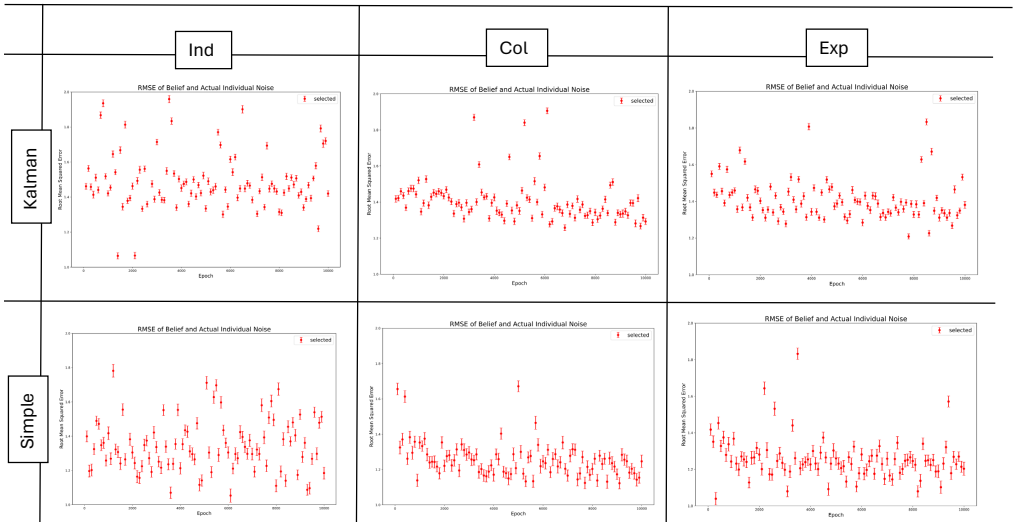

Figure 3.15: RMSE for different types of attention, for a single run (Kalman vs Simple, Noise Variance = 10)

Figure 3.16 showcases the outcomes when we average 10 runs at noise variance of 10. The plots validate our previous findings, as they illustrate that the Kalman filter-based belief update method outperforms the simple belief update method across all types of attention types. Additionally, the plots highlight the detrimental impact of misinformation on the satisfaction levels of the regulated units.

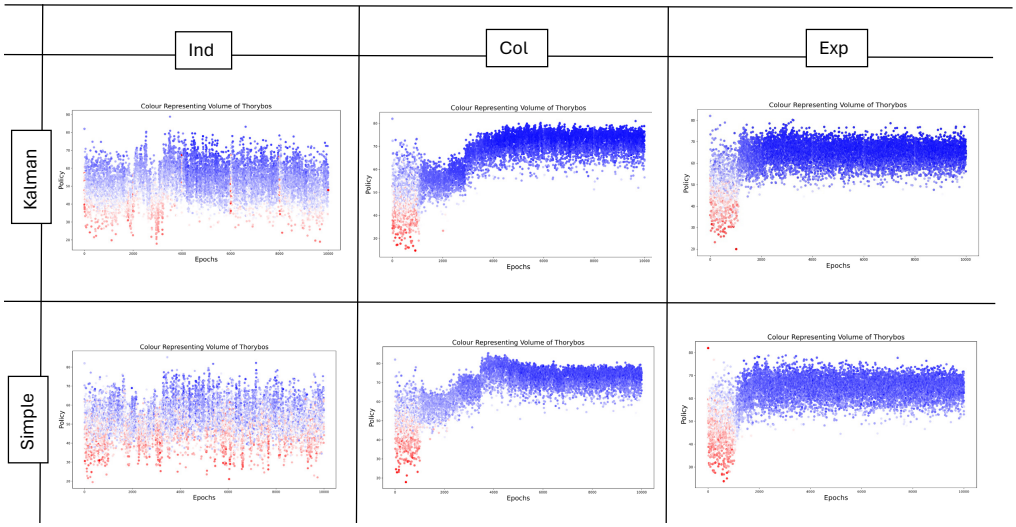

Figure 3.16: Policy selection of an RL regulator and corresponding expression for different types of attention, averaging over 10 runs (Kalman vs Simple, Noise Variance = 10)

4

Conclusions and Further Work

Contents

4.1 Conclusions

In this thesis, we have explored the dynamics of self-regulated systems through the implementation of various modifications to enhance systemic stability and agent satisfaction. The primary contributions of this research include the integration of the Advantage Actor-Critic (A2C) algorithm, the introduction of second-order cybernetics through the implementation of an observer, and the incorporation of partial observability by adding noise to agents' observations.

The implementation of the A2C algorithm as a replacement for the previously used Deep Q-Network (DQN) demonstrated notable improvements in the regulator's ability to select effective policies. The A2C algorithm provided a more stable and efficient learning process, reducing the overall noise in the system and improving agent satisfaction.

Introducing an observer into the system allowed for a more coherent collective expression among agents. The observer's role in gathering and processing information on agent interactions and past voice selections contributed to more informed and adaptive decision-making by the regulator. This modification led to a noticeable shift from systemic instability to stability.

By introducing noise to the individual voices observed by each agent, we simulated a more realistic environment where observations are imperfect. The comparison between the Kalman gain belief update algorithm and a simpler belief update method revealed the superiority of the Kalman filter in maintaining accuracy and stability under varying noise levels. Despite the added noise, the Kalman filter was able to provide optimal state estimates, thereby mitigating the adverse effects of misinformation.

4.2 Limitations and Future Work

While the observer played a crucial role in improving system stability, there is potential for further enhancement. Implementing a more sophisticated observer module that performs additional functions, such as predictive analytics or more complex decision-making processes, could further improve the system's adaptive capabilities. Future implementations could explore multi-level observation, where the observer not only gathers data but also analyzes patterns over time to provide more nuanced feedback and suggestions to the regulator and agents. Additionally, for future work, we could use the theoretical frameworks from third and fourth-order cybernetics to further enhance our system. Third order cybernetics, which emphasises the complexity of interactions among multiple observers and the reflexivity involved in such systems, could see us develop a framework that incorporates multiple observers, each with distinct roles and perspectives. Such system could involve self-aware agents capable of self-observation. This could involve the creation of more complex environments for our regulated units to navigate and the introduction of more challenging problems for them to solve.

Our experiments with partial observability have been limited to specific noise levels and types. Expanding these experiments to include a wider range of noise variances and different types of observational errors could provide a better understanding of the system's robustness, and gain insight on the behaviour of the agents. Implementing more sophisticated belief update methods, beyond the Kalman filter, could also potentially reduce the root mean squared error (RMSE) of observations, leading to greater accuracy in belief updates and thus, better methods of counteracting the effects of misinformation and uncertainty.

Future research could explore the technical and ethical implications of incorporating human observers into self-regulated systems. This includes developing methodologies to manage the variability introduced by human decision-making, which brings into light the challenge of designing systems that are both robust and flexible enough to accommodate human unpredictability without compromising on ethical standards. Research could involve taking a look at books that delve into the irrationality of human decision-making, including acts of dishonesty or selfishness [37]

5 Ethics

Contents

5.1 Ethical and Safety Implications

This section explores the ethical and safety considerations within AI development, guided by the AI Act's classification and compliance requirements [38] for varying risk levels. It highlights the project's alignment with lower-risk criteria, and the responsibility of keeping alignment with preventing negative social biases.

5.1.1 Risk Level and Compliance Requirements

The AI Act categorizes AI systems based on their risk to safety, fundamental rights, and freedoms. Since the project operates in a controlled, theoretical environment without direct human interaction or impact on real-world scenarios, it would likely be classified as a lower-risk AI system.

For lower-risk AI systems, the compliance requirements are less strict. The focus would be more on transparency, data governance, and ensuring that the system does not have inherent biases or

flaws that could be problematic if applied in real-world scenarios in the future. Keeping thorough documentation of the system's capabilities, limitations, and operational mechanisms is crucial, for transparency.

5.1.2 Exemptions for Research and Development

The AI Act often provides exemptions or more lenient requirements for AI systems used strictly for research and development purposes. As the project is part of academic research and experimental development, it should not be subject to the same regulations as AI systems intended for real-world deployment.

5.2 Future Real-World Plans

Should this project progress to involve real human interactions in real-world applications, its foundation on ethical regulatory systems suggests it would remain within lower-risk categories under AI regulations. Adhering to the project's ethical guidelines in future developments, should safeguard against transitioning into high-risk categorization, ensuring compliance with AI laws and regulations. That is, this endeavour should maintain an awareness of its social implications, avoiding reinforcing negative biases or inequalities.

5.3 Environmental Safety

Environmental safety regulations, are commonly associated with physical projects that have a direct impact on the environment. These do not apply to this project due to its theoretical and digital nature. There is no engagement that would pose a risk to environmental safety. Therefore, this aspect of safety regulation is not relevant to the project's scope, which is centered on computational simulations and theoretical frameworks within cybernetics and artificial intelligence.

6

User Guide

The codebase for this project is available on GitHub at the following repository: cpatsalidis/Imperial-FYP-2024. Detailed information regarding the functionality of the code and the specific contributions of this project can be found in the README section of the repository.

Bibliography

[1] A. Mertzani and J. Pitt, "Requisite social influence in self-regulated systems," in *16th International Conference on Agents and Artificial Intelligence (ICAART)*, 2024, pp. 133–140.

[2] M. Ashby, "Ethical regulators and super-ethical systems," *Systems*, vol. 8, no. 4, 2020, ISSN: 2079-8954. DOI: 10.3390/systems8040053. [Online]. Available: https://www.mdpi.com/2079-8954/8/4/53.

[3] A. K. Nowak, R. R. Vallacher, A. Rychwalska, *et al.*, *Target in control: Social influence as distributed information processing*. Springer, 2019.

[4] L. Boscolo, G. Cecchin, L. Hoffman, and P. Penn, *Milan systemic family therapy: Conversations in theory and practice.* Basic Books, 1987.

[5] R. Glanville, "The question of cybernetics," *Cybernetics and Systems: An International Journal*, vol. 18, no. 2, pp. 99–112, 1987.

[6] C. Fernyhough, *The Voices Within: The History and Science of How We Talk to Ourselves.* Welcome Collection, 2017.

[7] B. Arons, "A review of the cocktail party effect," *Journal of the American Voice I/O society*, vol. 12, no. 7, pp. 35–50, 1992.

[8] C. Wang, "Comprehensively summarizing what distracts students from online learning: A literature review," *Human Behavior and Emerging Technologies*, vol. 2022, Article ID 1483531, 2022. DOI: 10.1155/2022/1483531.

[9] A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, and N. Dormann, "Stable-baselines3: Reliable reinforcement learning implementations," *Journal of Machine Learning Research*, vol. 22, no. 268, pp. 1–8, 2021.

[10] V. Mnih, A. P. Badia, M. Mirza, *et al.*, *Asynchronous methods for deep reinforcement learning*, 2016. arXiv: 1602.01783 [cs.LG].

[11] K. Alibabaei, P. D. Gaspar, E. Assunção, *et al.*, "Comparison of on-policy deep reinforcement learning a2c with off-policy dqn in irrigation optimization: A case study at a site in portugal," *Computers*, vol. 11, no. 7, 2022, ISSN: 2073-431X. DOI: 10.3390/computers11070104. [Online]. Available: https://www.mdpi.com/2073-431X/11/7/104.

[12] D. Mehta, "State-of-the-art reinforcement learning algorithms," *International Journal of Engineering Research and Technology*, vol. 8, pp. 717–722, 2020.

[13] F. Garcia and E. Rachelson, "Markov decision processes," *Markov Decision Processes in Artificial Intelligence*, pp. 1–38, 2013.

[14] G. Paczolay and I. Harmati, "A new advantage actor-critic algorithm for multi-agent environments," in *2020 23rd International Symposium on Measurement and Control in Robotics (ISMCR)*, 2020, pp. 1–6. DOI: 10.1109/ISMCR51255.2020.9263738.

[15] F. Ding, G. Ma, Z. Chen, J. Gao, and P. Li, "Averaged soft actor-critic for deep reinforcement learning," *Complexity*, vol. 2021, pp. 1–16, 2021. DOI: 10.1155/2021/6658724. [Online]. Available: https://doi.org/10.1155/2021/6658724.

[16] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Advances in Neural Information Processing Systems*, S. Solla, T. Leen, and K. Müller, Eds., vol. 12, MIT Press, 1999. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/1999/file/464d828b85b0bed98e80ade0a5c43b0f-Paper.pdf.

[17] H. Von Foerster and H. von Foerster, "Cybernetics of cybernetics," *Understanding understanding: Essays on cybernetics and cognition*, pp. 283–286, 2003.

[18] B. Scott, *Second-order cybernetics: an historical introduction*. Kybernetes, 2004, vol. 33, pp. 1365–1378. DOI: https://doi.org/10.1108/03684920410556007.

[19] R. K. Pitman, "A cybernetic model of obsessive-compulsive psychopathology," *Comprehensive psychiatry*, vol. 28, no. 4, pp. 334–343, 1987.

[20] F. Parra-Luna, *Systems Science and Cybernetics - Volume III*. EOLSS Publications, ISBN: 9781848262041. [Online]. Available: https://books.google.co.uk/books?id=2-VRCwAAQBAJ.

[21] G. Bateson, *Steps to an ecology of mind*. Balantine, 1972.

[22] M. Diorinou and E. Tseliou, "Studying circular questioning "in situ": Discourse analysis of a first systemic family therapy session," *Journal of Marital and Family Therapy*, vol. 40, no. 1, pp. 106–121, 2014. DOI: https://doi.org/10.1111/jmft.12005. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jmft.12005. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/jmft.12005.

[23] H. A. Love, "On aerial perspective, socio-technical systems, and interdisciplinarity: Reading modernism alongside cybernetics," *IEEE Technology and Society Magazine*, vol. 42, no. 4, pp. 35–41, 2023. DOI: 10.1109/MTS.2023.3340245.

[24] G. Stein, *Everybody's autobiography*. Vintage, 2013.

[25] P. Boxer and V. Kenny, "The economy of discourses: A third order cybernetics?" *Human Systems Management*, vol. 9, no. 4, pp. 205–224, 1990.

[26] S. Božičnik and M. Mulej, "A new–4th order cybernetics and sustainable future," *Kybernetes*, vol. 40, no. 5/6, pp. 670–684, 2011.

[27] R. G. Mancilla *et al.*, "Introduction to sociocybernetics (part 3): Fourth order cybernetics," *Journal of Sociocybernetics*, vol. 11, no. 1/2, 2013.

[28] Y. Rizk, M. Awad, and E. W. Tunstel, "Decision making in multiagent systems: A survey," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 10, no. 3, pp. 514–529, 2018. DOI: 10.1109/TCDS.2018.2840971.

[29] E. F. Loftus and J. C. Palmer, "Reconstruction of automobile destruction: An example of the interaction between language and memory," *Journal of verbal learning and verbal behavior*, vol. 13, no. 5, pp. 585–589, 1974.

[30] M. S. Ayers and L. M. Reder, "A theoretical review of the misinformation effect: Predictions from an activation-based memory model," *Psychonomic Bulletin & Review*, vol. 5, no. 1, pp. 1–21, 1998.

[31] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: A survey," *Journal of Artificial Intelligence Research*, vol. 4, pp. 237–285, 1996. DOI: https://doi.org/10.1613/jair.301.

[32] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra, "Planning and acting in partially observable stochastic domains," *Artificial Intelligence*, vol. 101, no. 1, pp. 99–134, 1998, ISSN: 0004-3702. DOI: https://doi.org/10.1016/S0004-3702(98)00023-X. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S000437029800023X.

[33] H. E. Kyburg Jr, "Bayesian and non-bayesian evidential updating," *Artificial intelligence*, vol. 31, no. 3, pp. 271–293, 1987.

[34] R. E. Kalman, "A new approach to linear filtering and prediction problems," 1960.

[35] L. D. Stone, R. L. Streit, and S. L. Anderson, "Bayesian single target tracking," in *Introduction to Bayesian Tracking and Particle Filters*, Springer, 2023, pp. 5–44.

[36] G. Dao and M. Lee, "Relevant experiences in replay buffer," in *2019 IEEE symposium series on computational intelligence (SSCI)*, IEEE, 2019, pp. 94–101.

[37] D. Ariely, *Predictably Irrational: The Hidden Forces That Shape Our Decisions*. HarperCollins, 2008.

[38] T. Madiega and S. Chahri, “Artificial intelligence act,” Members’ Research Service PE 698.792, Jun. 2023.