

IMPERIAL

LEARNING AND INNOVATION IN SELF-GOVERNING SYSTEMS

Author

ASIMINA MERTZANI

Supervised by

PROF. J. PITT

A Thesis submitted in fulfillment of requirements for the degree of
Doctor of Philosophy

Department of Electrical and Electronic Engineering
Imperial College London
2025

Abstract

The future of distributed information processing in socio-technical systems is hybrid; it involves meaningful interactions between human (natural) and computational (artificial) intelligences. This hybrid nature raises unique challenges for self-governance and deliberation with respect to social arrangements (i.e. socially constructed rules, norms, structures, and procedures). To address these challenges and enable individuals and communities to define, modify, and envision social arrangements, we need mechanisms for learning and innovation. Accordingly, this thesis presents:

- a critical synopsis of interdisciplinary literature from psychology, economics, philosophy, and political science, which shapes the definition of self-governing systems;
- an incremental refinement and integration of three methodologies: socially-inspired computing, hybrid reasoning and learning (neurosymbolic) systems, and socially-guided machine learning, to produce a new methodology called *Socially-Guided Reasoning and Learning* for designing and implementing self-governing systems;
- three novel algorithms for addressing challenges in self-governance:
 - *RTSI+* for explanatory adequacy and epistemic coherence through social influence,
 - *4voices* for effective self-regulation via requisite influence, and
 - *Θ-learning* for reaching sustainable collective agreements through consensus and dissent.

By animating these algorithms with multi-agent systems, experimental results offer insight into issues of epistemic coherence, systemic stability, and the relationship between foundational goals and collective decision making;

- a proof of concept for a novel innovation-support system integrating multi-agent systems simulation, generative AI, and human users for enabling the human-AI co-production of social arrangements;
- a metrics-based evaluation of self-governing socio-technical systems in terms of sustainability, performance, and human empowerment.

In conclusion, we argue that the significance of this thesis is to have produced sustainable mechanisms for supporting empowerment, preserving dignity and fostering human flourishing during the transition to a hybrid future.

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.

Copyright Declaration

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International Licence (CC BY-NC).

Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose.

When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes.

Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Acknowledgments

I would like to acknowledge the Department of Electrical and Electronic Engineering at Imperial College London for their financial support. It has been an honour to be the inaugural recipient of the Maria Petrou Scholarship that was established to advance the careers of women in engineering. This scholarship, which commemorates the legacy of Professor Maria Petrou - the first female professor in the department and a passionate advocate for women in engineering, has granted me complete academic freedom in my research and I am indebted to the donors for their support.

I would like to express my sincere and heartfelt gratitude to my supervisor, Prof. J. Pitt, for his invaluable guidance, encouragement, and support throughout this journey. His mentorship has been a cornerstone of my growth, and I deeply appreciate the patience, wisdom, and trust he has shown in me. I sincerely hope this is not the end of our collaboration, but rather the beginning of a lifelong connection.

A particular thanks to my family – my mother, Efi, my father, Aris, and my brother, Konstantinos – for their constant belief in me, boundless love, and unwavering support that have been my motivation to do not give up even the toughest moments. I cannot thank them enough for always being there, cheering me on every step of the way.

I would like to express my gratitude to my friends and colleagues in the lab for making this journey more enjoyable and manageable. I would also like to thank my dearest friends in UK – Anh, Cath, and Konstantinos - for lightening the load during challenging times, and in particular Anh, for her mental but also practical support. Finally, a special thanks to Konstantinos for alleviating my stress and pessimism, for listening to my complaints, and for being a constant source of strength and understanding.

To all of you, thank you for your support, kindness, and encouragement. This journey would not have been the same without you.

Contents

Abstract	i
Declaration of Originality	iii
Copyright Declaration	v
Acknowledgments	vii
List of Acronyms	xv
List of Figures	xvii
1 Introduction	1
1.1 Aim	1
1.2 Example	2
1.3 Overview	3
1.4 Summary of Contributions	5
1.5 Structure of the Thesis	6
2 Self-Governance in Socio-Technical Systems	7
2.1 Introduction	7
2.2 Self-Governance	8
2.2.1 Self-Organisation and *-Governance	9
2.2.2 Self-Governing Systems (SGSs)	11
2.2.3 Components of a SGS	13
2.2.4 Challenges of Self-Governance	14
2.3 Social Influence	21
2.3.1 The Regulatory Theory of Social Influence	22
2.3.2 Regulatory Theory of Social Influence (RTSI) for Collective Decision Making	22

2.3.3	From Social Influence to Adaptation	25
2.4	Learning	26
2.4.1	Self-Regulation & First Order Cybernetics	26
2.4.2	Machine Learning (ML) Overview	27
2.5	Innovation & Empowerment	29
2.5.1	Power & Empowerment	29
2.5.2	Higher Order Cybernetics	30
2.5.3	Issues of Innovation	31
2.5.4	The Interconnection between Empowerment & Innovation	33
2.6	Summary	34
3	Methodologies for Self-Governance in Socio-Technical Systems	35
3.1	Introduction	35
3.2	Socially-Inspired Agent-based Simulation	38
3.2.1	Socially Inspired Computing Methodology	38
3.2.2	Machine Reasoning	40
3.2.3	Agent-Based Modelling & Simulation	41
3.2.4	Overview of Architecture & Implementation	41
3.3	Hybrid Reasoning and Learning	44
3.3.1	Introduction	44
3.3.2	Hybrid Reasoning and Learning Methodology	44
3.3.3	Overview of Architecture & Implementation	46
3.4	Socially-Guided Reasoning and Learning	48
3.4.1	Introduction	48
3.4.2	Socially-Guided Machine Learning	49
3.4.3	Power-Sensitive Design	50
3.4.4	Socially-Guided Reasoning and Learning (SGRL)	52
3.4.5	SGRL Architectural Overview and Notes on Implementation	53
3.5	Summary	56
4	Explanatory Adequacy and Reflection	59
4.1	Introduction	60
4.2	Motivation	61
4.3	Problem Specification	62

4.3.1	The Problem of Developing Explanatory Adequacy	62
4.3.2	Evaluation Criteria	65
4.3.3	Experimental Conditions	66
4.4	Formal Specification	68
4.4.1	The Environment	68
4.4.2	The Agents	68
4.5	The <i>RTSI+</i> Algorithm for Explanatory Adequacy	69
4.6	Experimental Setup	70
4.7	Experimental Hypothesis and Results	71
4.7.1	Experiment 1: Static and dynamic populations	72
4.7.2	Experiment 2: Dynamic population, progressive addition, non-persistence	73
4.7.3	Experiment 3: Dynamic population, progressive addition, persistence	75
4.7.4	Summary of Results	77
4.8	Conclusions and Discussion on Explanatory Adequacy	78
4.9	From Explanatory Adequacy to Reflection	80
4.10	Metrics for Reflection	81
4.10.1	Problem Specification	81
4.10.2	The Metrics	82
4.10.3	Risks	87
4.10.4	Conclusions on Metrics for Reflection	90
4.11	Summary & Discussion	90
5	Adaptation using Hybrid Systems	93
5.1	Introduction	93
5.2	Effective Self-Regulation	95
5.2.1	Motivation	95
5.2.2	Problem Specification	96
5.2.3	The <i>4voices</i> Algorithm for Self-Regulation	98
5.2.4	Experimental Design	103
5.2.5	Experimental Results	105
5.2.6	Conclusions on Effective Self-Regulation	110
5.3	Sustainable Consensus Formation	111
5.3.1	Motivation	111

5.3.2	Political Sciences Insights	113
5.3.3	Problem Specification	113
5.3.4	Θ -learning Algorithm for Sustainable Consensus Formation	114
5.3.5	Experimental Design	117
5.3.6	Experimental Results	120
5.3.7	Conclusions on Sustainable Consensus Formation	128
5.4	Managing Sustainability in Self-Governing Socio-Technical Systems	130
5.4.1	Motivation	130
5.4.2	Evaluation & Control of Sustainability	130
5.4.3	Hybrid Reasoning and Learning (HyRL) Systems For Uncertainty & Sustainability	132
5.4.4	Conclusions on Metrics for Sustainability	133
5.5	Summary & Conclusions	133
6	Innovation & Empowerment	135
6.1	Introduction	135
6.2	Empowerment	137
6.3	Definition of Empowerment	137
6.3.1	Higher-Order Cybernetics Patterns & Empowerment	139
6.3.2	Case Studies of Cybernetics & Empowerment	142
6.3.3	Conclusions on Empowerment	147
6.4	From Empowerment to Innovation	147
6.5	Innovation	148
6.5.1	Conceptual Analysis of Innovation	148
6.5.2	Process-Oriented Definition of Innovation	149
6.6	Innovation for Empowerment	150
6.6.1	Innovation Support System through Deliberation (ISSD) Detailed Overview	151
6.6.2	ISSD Implementation	155
6.6.3	Evaluation as a Tool for Empowerment	165
6.6.4	Evaluation as a Cybernetic System	170
6.6.5	Social Implications of SGRL Methodology	174
6.6.6	Conclusions on Innovation for Empowerment through ISSD	176
6.7	Evaluation & Metrication	177
6.8	Empowerment and Innovation: A Chicken-Egg Problem	180
6.9	Summary & Conclusions	181

7	Conclusions and Further Research	183
7.1	Summary	183
7.2	List of Publications	186
7.3	Limitations and Further Research	188
7.4	Final Remarks	191
	Bibliography	193

List of Acronyms

AI Artificial Intelligence

MAS Multi-Agent System

SGS Self-Governing System

CPR Common Pool Resource

DIP Distributed Information Processing

ALife Artificial Life

NLife Natural Life

RTSI Regulatory Theory of Social Influence

RL Reinforcement Learning

MARL Multi-Agent Reinforcement Learning

GenAI Generative Artificial Intelligence

LLM Large Language Model

MR Machine Reasoning

ML Machine Learning

MI Machine Innovation

ISSD Innovation Support System through Deliberation

SGML Socially-Guided Machine Learning

SGRL Socially-Guided Reasoning and Learning

ABMS Agent-Based Modelling and Simulation

SIC Socially-Inspired Computing

HyRL Hybrid Reasoning and Learning

PSD Power Sensitive Design

PSO Power Sensitive Operationalisation

KM Knowledge Management

ISSD Innovation Support System through Deliberation

SA Social Arrangement

DoE Dimensions of Empowerment

IoT Internet of Things

ZoD Zone of Dignity

List of Figures

2.1	Subjective Value-Judgements performed by a Distributed Information Processing (DIP) Unit	16
3.1	The socially-inspired computing methodology [11]	39
3.2	Socially-Inspired Computing (SIC)-Agent-Based Modelling and Simulation (ABMS) Architectural Overview	43
3.3	HyRL Architectural Overview	47
3.4	SGRL system comprising human users, a multi-agent simulator, a Generative Artificial Intelligence (GenAI) agents, and an interactive interface	53
3.5	SGRL for Power Sensitive Operationalisation (PSO) Architectural Overview	56
3.6	The incremental refinement of the research methodology	57
4.1	The problem of developing explanatory adequacy	64
4.2	Experiment 1: Knowledge dynamics for static and dynamic population, showing that static populations are congruent but stagnate and remain explanatory inadequate, while dynamic populations can develop explanatory adequacy.	74
4.3	Experiment 2: Knowledge dynamics for DIP with non-persistent knowledge showing that there is a potential for the DIP to develop explanatory adequacy as long as the knowledge is not lost permanently	75
4.4	Experiment 3: Knowledge dynamics for progressively added knowledge showing that the slow addition of new knowledge fosters developing explanatory adequacy . . .	76
4.5	The drivers of the DIP	81
4.6	The DIP as a cybernetic system	82
5.1	A Self-Regulated System	95
5.2	A Self-Regulated System with Requisite Social Influence	97
5.3	The psychologically inspired model	105
5.4	The mean value and the standard deviation of the individual expressions of the regulated units under different initial conditions depicting that requisite influence, corresponding to comparatively low expression, is established when the regulator determines the policy using RL - Results averaged over 10 runs	106

5.5	Collective expression of the regulated units with respect to the choice of policy of the regulator under different initial conditions depicting that requisite influence, corresponding to comparatively low expression (blue), is established when the regulator determines the policy using RL - Results averaged over 10 runs	107
5.6	The mean value and the standard deviation of the individual expressions under different initial conditions showing that stability is maintained when the regulator determines the policy using RL and the regulated units aim to learn to attend the experts voice or the background noise (i.e. voice of the community) - Results of a single run	108
5.7	The policy selection and the corresponding collective expression under different initial conditions showing that stability is maintained when the regulator determines the policy using RL and the regulated units aim to learn to attend the experts voice or the background noise (i.e. voice of the community) - Results of a single run . . .	109
5.8	Θ -learning algorithm (left) and abstraction of the process (right)	116
5.9	Experiment 1: Fraction of population reaching an agreement and compromises made by individuals for different conditions showing a strong relation between learning objective and level of agreement, which is persistent in static populations	121
5.10	Experiment 2: Fraction of population reaching an agreement and compromises made by individuals for different conditions showing that a collective agreement can be reached even if the preferences of the population change, as long as agents do not act individualistic	123
5.11	Experiment 2: Different learning objectives allow different forms of agreement to emerge when the preferences of the population change over time	124
5.12	Experiment 3: Fraction of population reaching an agreement and corresponding level of compromise for different speed of change of population preferences showing that the quick response to change is to reach an agreement (in order to maintain stability) even if this requires a lot of compromise, however sustainable collective agreements require slow change	127
5.13	Extended version of socio-environmental system including technology as a factor affecting sustainability	131
6.1	Measurements of Individual and Collective Empowerment	139
6.2	Higher Order Cybernetics Learning to Empowerment	142
6.3	n -th Order Cybernetics for Empowerment in Community Energy Systems	143
6.4	n -th Order Cybernetics for Empowerment and Disempowerment in Social Media .	146
6.5	Illustration of the Innovation ‘Equation’	150
6.6	Example Visualisation of SGRL	152
6.7	Illustrative Visualisation and Dialogue	155
6.8	Architecture of ISSD comprising (a) human user(s), a Multi-Agent System (MAS), and a GenAI agent	157
6.9	Vocabulary used for the testing of the system	162
6.10	Snapshots from the interaction with the systems showing how it enables the users to develop interoceptive awareness.	166

6.11 Comparison of Effects of Social Arrangements (SAs) and Baseline	167
6.12 Snapshots of the systems showing how different SAs result in increasing different dimensions of empowerment and that human interest is required to maintain empowerment	168
6.13 Snapshots of the system showing that humans need to be patient in order to see the impact of the different SAs	169
6.14 Selected Cognitive dimension of empowerment showing the emergence of expertise, since agents tend to use the opinion of the “experts”	170
6.15 Own and Selected cognitive dimension of empowerment showing the emergence of expertise, since agents tend to use the opinion of the “experts”	171
6.16 Systems with an active observer showing the development of interoceptive awareness leading to increased cognitive dimensions of empowerment	172
6.17 Systems with re-generating mechanisms showing the capacity of the system to heal itself and recover from situations in which the community health is low	173

Chapter 1

Introduction

Contents

1.1 Aim	1
1.2 Example	2
1.3 Overview	3
1.4 Summary of Contributions	5
1.5 Structure of the Thesis	6

1.1 Aim

The next-generation of socio-technical systems within the context of Digital Transformation will inevitably be hybrid [1], [2]. Given the rise of Internet of Things (IoT) [3], Blockchain technologies [4], and smart-chatbots [5], socio-technical systems will increasingly involve the convergence of human intelligence with Artificial Intelligence (AI) in co-existence, co-production and co-evolution within some kind of social system. This evolution necessitates that humans collaborate with intelligent machines, as they become integral components of society [6]. Simultaneously, it opens vast potential to enhance human capabilities through the effective use of these intelligent systems. However, to ensure that these advancements benefit humanity, it is critical to address how heterogeneous components (humans and machines) can meaningfully interact and collectively decide on matters of self-governance [7], [8]. This includes the selection, modification, and application of Social Arrangements (SAs), encompassing conventions, norms, rules, roles, contracts, and laws. These SAs represent mutually agreed-upon and adaptable frameworks that communities rely on for the social construction of a shared reality.

The ultimate goal of this thesis is to engineer systems that can support human deliberation and improve human decision making with respect to issues of self-governance in hybrid (human-AI systems). However, to develop better hybrid systems it is not sufficient to utilise the latest (AI) technologies and follow the best engineering practices. We need also to consider the characteristics of the social systems and understand how humans organise themselves in their societies.

Overall, this thesis combines theories of social sciences with AI technologies to design systems whose components (humans or AI) are capable of learning and innovation. Those systems empower humans to make better decisions concerning the self-governance of hybrid systems, contributing to human flourishing during the transition to the hybrid society.

1.2 Example

This Section provides an example that showcases a situation in which humans and AI co-exist in a hybrid socio-technical environment where they need to make decisions related to self-governance. Here, we analyse a scenario of a shared housing (or co-housing project), where humans and smart devices co-exist, interact and make collective decisions. In the beginning, those decisions might concern the facilities that would be accommodated in the available land, the division between common and private spaces, or the investments in solar panels and other renewables. In later stages of the development of the project, those decisions might be relevant to the policies for the common spaces, the distribution of the parking slots, or the distribution of resources (such as energy, space, and objects). Overall, this corresponds to a scenario in which a community of humans, such as residents and contractors, and smart devices, such as smart lights, smart TVs or smart radiators, are connected over a network and can communicate, verbally or by exchanging messages and signals, to make collective decisions with respect to SAs.

For instance, focusing on the decisions regarding the distribution of energy, humans might want to charge some of their devices, smart radiators in the common areas might require energy for heating up the space, while smart lights might need constantly a small amount of energy while being on ‘stand-by’ mode (for sensing the parking and the entrance) which is increased when a subject is sensed and they need to turn-on ‘on’ mode. In that case, initially humans and computational components might have limited knowledge and therefore the individuals of this hybrid community might decide on some fixed rules, or elect a dedicated agent (either human or device) to make such decisions.

As time passes, humans might gain experience and expertise by interacting with others, while sensors start collecting data regarding the human preferences and the requirements of the devices. This way, the hybrid system can gradually develop a machine learning model to optimise the distribution of the energy with respect to diverse criteria. Also humans might interact with the sensors, through an interface, or exchange information with an Large Language Model (LLM), in order to identify novel SAs, and evaluate them in simulation before applying them to real life. This way humans will co-produce with the intelligent machines in order to decrease the expense of resources while increasing human satisfaction.

The point of this example, is that to minimise energy consumption, the sensors want to turn everything off; to maximise ambient living environment, the humans want to turn everything on. Clearly, the conflicting goals of the artificial and natural intelligences need to be reconciled through compromise and negotiation over the selection, modification and application of conventional rules – i.e. governance. Since no-one else is concerned with micro-managing the house (not the AI developers, not the housebuilders, not the landlords), the occupants are going to have to manage this interaction by, between and for themselves, i.e. self-governance.

Hence, although this thesis is not directly concerned with developing solutions for enhancing satisfaction and efficiency of energy distribution in co-housing projects, this example does illustrate the interplay between humans and AI in hybrid socio-technical systems. It also highlights the need for forming SAs (which in this example correspond to rules affecting the distribution of energy), the need for collective self-governance with respect to those SAs, without an external authority, and in particular the need for learning and innovation in the selection, modification and application of those arrangements. Hence, it leverages the human-AI collaboration to improve the experience of the humans sharing a house and empower them to self-determine the rules that affect their living.

1.3 Overview

The premise of this thesis is the anticipation of the next generation of self-governing socio-technical systems as having three key properties: firstly, they consist of a community of heterogenous entities that are collectively acting as a Distributed Information Processing (DIP) unit; secondly, the entities will inevitably be hybrid, i.e. the entities are both humans and AI components; and thirdly, as systems with longevity, they will need to learn and innovate socially-constructed rules to address the main issues of self-governance and maintain sustainability. This thesis synthesises theories of

human psychology and political science with algorithms for machine reasoning, machine learning, and human-machine innovation to facilitate the self-governance of hybrid socio-technical systems.

Hence, it investigates the role of learning and innovation in hybrid SGSs by simulating a DIP using Multi-Agent System (MAS), and envisioning a DIP with interacting human and artificial intelligence. Initially, it presents an existing socially-inspired model that uses social influence to enable a DIP to identify expertise and make collective decisions with respect to qualitative value judgements [9]. This model is extended to enable the DIP to collectively deciding on the processing rules producing a commonly observed phenomenon [10] (which corresponds to a problem of developing explanatory adequacy). The extended model is designed following a Socially-Inspired Computing (SIC) [11] methodology to bring together concepts from a social influence [12] theory and Machine Reasoning (MR) [13] techniques. Also, it is implemented according to the Agent-Based Modelling and Simulation (ABMS) [14], [15] approach, which allows us to run simulations and explore under which initial conditions the DIP manages to develop explanatory adequacy, and what is the epistemic state of the population under those conditions.

Additionally, integrating cybernetics theory [16] and Reinforcement Learning (RL) [17], through the Hybrid Reasoning and Learning (HyRL) methodology, supports the design of a system where agents maintain systemic stability through effective self-regulation [18]. This highlights the need for requisite social influence (i.e. communication from the regulator to the regulated units of the policy changes, and communication from the regulated units to the regulator of the effects of those policy changes) for ethical regulation [19]. Furthermore, the thesis introduces a collective deliberation algorithm, bringing together concepts from political science and, in particular, a process used by Athenian deliberative assemblies called *thorybos* [20], with Multi-Agent Reinforcement Learning (MARL), to achieve sustainable self-governance [21]. It demonstrates the need for compromise, dissent, and the acceptability of majority-based decisions in meaningful democratic deliberation.

Furthermore, concepts from higher-order cybernetics [22] are combined with issues power and empowerment [23] in a SGS to support innovation for human empowerment in hybrid (human-AI) systems. Inspired from the Socially-Guided Machine Learning (SGML) methodology, it introduces a novel methodology, called Socially-Guided Reasoning and Learning (SGRL) that combines a MAS simulator with Generative Artificial Intelligence (GenAI) and human user(s) to enable humans to envision novel SAs and visualise their impact on human empowerment [24]. Additionally, it follows the Socially-Guided Reasoning and Learning (SGRL) methodology, to design and implement an exemplar of a self-improving socio-technical system, called ISSD, that empowers communities through deliberate human-AI co-production and innovation of congruent social arrangements.

The societal implications of the system are also discussed, in particular arguing that the proposed approach to developing socio-technical systems mitigates the risks of AI-driven concentration of power and reduces asymmetries in the distribution of power [24]. Finally, the thesis also addresses the evaluation of socio-technical systems in terms of sustainability [25], performance [26], and empowerment, proposing metrics that balance conflicting systemic drivers and considers both objective measures and qualitative user experiences.

1.4 Summary of Contributions

This thesis makes several contributions along multiple dimensions: theoretical, methodological, technical, empirical, and metrical.

- It *provides* a critical synopsis of inter-disciplinary literature from psychology, economics, philosophy, and political science that informs a definition of SGSs.
- It presents an incremental refinement and integration of three methodologies: SIC, HyRL (i.e. a form of neurosymbolic system), and SGML, to produce a new methodology called SGRL for designing and implementing SGSs;
- It introduces three novel multi-agent algorithms for addressing challenges of self-governance:
 - *RTSI+* for explanatory adequacy and epistemic coherence through social influence,
 - *4voices* for effective self-regulation via requisite influence, and
 - *Θ-learning* for reaching sustainable collective agreements through consensus and dissent,
 and reports the experimental results of those algorithms, which are implemented using an ABMS approach, and show:
 - how initial conditions impact the collective’s knowledge state and its explanatory relevance;
 - how these conditions contribute to systemic stability and influence the individuals’ expression of ‘truth’; and
 - how fundamental institutional goals guide collective decisions, with changes in population dynamics affecting both the agreements made and the compromises needed to reach them.
- It uses the SGRL to implement and demonstrate a proof-of-concept for a new innovation-support system integrating MAS and GenAI for enabling the human-AI co-production of SAs.

- It discusses issues related with metrics for evaluating self-governing socio-technical systems in terms of sustainability, performance, and human empowerment (individual and collective).

1.5 Structure of the Thesis

The remainder of the thesis is divided in three main parts: *Influence*, *Learning*, and *Innovation*. Accordingly, after analysing SGS, Chapter 2 reviews the literature on social influence concerning knowledge management discusses algorithms for learning, and performs an inter-disciplinary review of the concepts of innovation and power, focusing on those that are relevant to human empowerment in self-governing socio-technical systems. Chapter 3 incrementally refines three methodologies for designing and implementing systems that effectively use social influence for knowledge management, encompass reasoning and learning capabilities for adapting to complex dynamic socio-technical environments, and support innovation to foster human empowerment in an ever-changing technological landscape.

Then, Chapters 4,5,6 present the main technical and practical contributions of this thesis. Chapter 4 uses a theory of social influence [12] to enable a group to manage internal knowledge to develop explanatory adequacy [10]. Chapter 5 combines theories from psychology [27] and political science [20] with methods for learning [17]. This establishes the pathways for requisite influence in a self-regulated system [18], and achieves sustainable consensus formation through legitimate consent and dissent [21]. This is followed by Chapter 6 which analyses innovation and empowerment in the context of deliberation with respect to SAs. Particularly, it uses patterns of higher-order cybernetics for maintaining an equitable distribution of power in self-governing socio-technical systems [23], and leverages the human-AI co-production of SAs to design and implement an SGRL-based framework for innovation support learning to community empowerment [24].

Finally, Chapter 7 concludes the work presented in this thesis, arguing that the significance of this thesis is that it combines technological advances with inter-disciplinary scientific knowledge to empower humans, ensuring they participate in the processes of collective deliberation concerning issues of self-governance. The chapter points out some limitations of the current work and proposes ways to address these limitations in further research. It also summarises the main scientific and technical contributions of this thesis, and remarks on how this thesis contributes to a research programme on Self-Governing Systems.

Chapter 2

Self-Governance in Socio-Technical Systems

Contents

2.1 Introduction	7
2.2 Self-Governance	8
2.2.1 Self-Organisation and *-Governance	9
2.2.2 SGSs	11
2.2.3 Components of a SGS	13
2.2.4 Challenges of Self-Governance	14
2.3 Social Influence	21
2.3.1 The Regulatory Theory of Social Influence	22
2.3.2 RTSI for Collective Decision Making	22
2.3.3 From Social Influence to Adaptation	25
2.4 Learning	26
2.4.1 Self-Regulation & First Order Cybernetics	26
2.4.2 Machine Learning (ML) Overview	27
2.5 Innovation & Empowerment	29
2.5.1 Power & Empowerment	29
2.5.2 Higher Order Cybernetics	30
2.5.3 Issues of Innovation	31
2.5.4 The Interconnection between Empowerment & Innovation	33
2.6 Summary	34

2.1 Introduction

This thesis investigates how theories from social science and Artificial Intelligence (AI) methodologies can be brought together in order to achieve sustainable self-governance in socio-technical

systems during the transition to the hybrid society (i.e. comprising naturally and artificial intelligent components). It aims to develop tools and technologies that leverage human-AI co-production to support decision making leading to human flourishing.

Chapter 2 analyses the main features of sustainable self-governance and the common challenges to achieving it. Additionally, it performs an interdisciplinary review of theories and methodologies from social sciences and computer science, aiming to identify mechanisms that can address those challenges and enable sustainable self-governance in socio-technical systems.

Specifically, it first analyses the issues of self-governance which include sustainability, fairness and justice, knowledge management, legitimate consent and dissent, the iron law of oligarchy, the paradox of the unrestricted self-modification, as well as some voting paradoxes, issues of citizenship and rights, and dignity during the transition to the hybrid (human-Artificial Intelligence (AI)) society. Then, it discusses the benefit of using social influence for dealing with issues of knowledge management as well as for enabling sustainable self-governance through the identification of trustworthy gatekeepers of knowledge (i.e. experts) and aggregating the knowledge of the collective and aligning on it to form a collective assessment.

This is followed by a discussion regarding the requirement for adaptation and the potential for achieving it through the use of learning techniques. Additionally, it provides an overview of issues of innovation and empowerment in socio-technical systems and introduces AI techniques that can support it. Finally, it discusses the interconnection between empowerment and innovation in socio-technical systems: humans need to *be empowered to innovate their Social Arrangements (SAs)* and they need to *innovate their SAs to be empowered*.

2.2 Self-Governance

This Section explores the foundational aspects of self-organisation and autonomous governance. It begins by examining various forms and types within these frameworks and continues by giving a definition of self-governance in the context of socio-technical systems (i.e. systems comprising human participants and (potentially intelligent) software agents). Next, it outlines the main components essential to Self-Governing Systems (SGSs), followed by an analysis of the key challenges faced in achieving effective self-governance. These challenges set the stage for the following Chapters, which propose mechanisms to address these issues and enhance the resilience and sustainability of SGSs, through learning and innovation.

2.2.1 Self-Organisation and *-Governance

Self-organisation can describe any process of modification of a (multi-agent) system's behaviour which is initiated and enacted by the system's components themselves. The scientific community has not agreed on whether self-organisation is a conscious process or not. Specifically, some researchers describe self-organisation as any process of formation of a structure without an external controller [28], while others argue that this process corresponds to emergence while self-organisation is a process of reflection, interoception, and deliberation [11].

In this thesis, we are interested in deliberate forms of organisation and governance which involve the intentional adjustment of the behaviour of a MAS, arising from local interactions that involve choice, coordination, and decision-making. The core concept of intentional self-organisation is that the components of the system independently organise themselves without any external interference, assistance, or incentives, aside from adapting to become suited to, or aligned with, the environment in which they exist. This process may encompass various activities such as self-awareness, reflection, or interoception, and it involves a collective commitment to comply with the agreed SAs.

Transitioning from self-organisation to forms of 'autonomous' governance, those include, among others, approaches meta-governance, network governance, collaborative governance, and polycentric governance. Each of these governance models offers potential benefits, though they remain largely theoretical.

Meta-governance, a practice primarily undertaken by public authorities, involves coordinating one or more governance modes through various instruments, methods, and strategies to mitigate governance failures [29]. It often implicitly addresses the government-versus-governance debate by examining the state's evolving role in a networked society. Thus, meta-governance can be understood as 'government plus governance', offering both constraints and enabling potential by combining control and facilitation [30]. This practice is recognised as significant for enhancing accountability and transparency within governance networks, minimising fragmentation in global sustainability governance, and improving the governance of natural resource regions, among other areas [31], [32].

Network governance is characterised by the stable interconnection of mutually dependent but operationally autonomous actors who engage through negotiation, often in conflict, within an institutionalised framework of rules, norms, shared knowledge, and social imaginaries. This framework supports the production of public values [32]. A conceptual model of network governance pro-

posed by [33] aims to contextualise this theory and leverage its potential to enhance integration and coordination between services, thus fostering a more effective system of service provision and revitalising public sector efficiency. However, in their work [33], Molin and Masella suggest that empirical studies are needed to validate its application and effectiveness.

Collaborative governance refers to “a governing arrangement in which one or more public agencies directly engage non-state stakeholders in a formal, consensus-oriented, deliberative collective decision-making process aimed at creating or implementing public policy or managing public programs or assets” [34]. There remains debate around the legitimacy and accountability of this model [35]. Some caution against the “fuzzy” accountability associated with collaborative governance networks, warning of potentially illegitimate policy processes and outcomes. Conversely, proponents argue it may offer a more legitimate form of governance than both hierarchical bureaucracies and market-driven public management, enhancing legitimacy by providing stakeholders and citizens greater influence in policymaking and addressing complex, or “wicked”, social issues. Yet, empirical research is required to substantiate its potential.

Finally, *polycentric governance* refers to interactions among autonomous decision-making centres operating within a social system, typically within a single metropolitan area, forming a polycentric political system without hierarchical constraints [36]. This model entails diverse autonomous centres working collectively, self-organising, and self-governing to address governance challenges. It features a complex configuration of multiple levels and organisations from the public, private, and voluntary sectors, with overlapping responsibilities and capabilities [37].

The forms of decentralised governance that were presented above, although they offer useful information and highlight different aspects of organisation, remain theoretical. However, we want to propose a definition of self-governance that can be easily ‘translated’ to a formal model. Having that formal model we can then design and implement a system in which we can run experiments and evaluate the effectiveness.

The next Section provides a definition of Self-Governing System (SGS) while the remainder of this thesis provides mechanisms for integrating concepts and methods from various disciplines to facilitate effective self-governance and address associated challenges.

2.2.2 SGSs

Combining Ostrom's principles for sustainable management of 'commons' (corresponding to self-governing institutions sharing a common pool of resources) [38], Nowak's theory concerning the bidirectional flow of social influence [12] and Ober's work on democratic knowledge management [39], we provide a definition of SGS and we identify the eight features for sustainability together with the main challenges for effective self-governance.

A SGS consists of a group of otherwise independent and autonomous entities pursuing their collective (public) and personal (private) interests in the context of an over-arching, mutually-agreed, and mutable set of socially-constructed conventional social arrangements, by which we mean:

- social arrangements: any kind of convention, norm, policy, rule, contract or law for voluntarily prescribing acceptable behaviour in the context of the system;
- socially-constructed: a mutually-agreed product of interaction between the entities themselves; and
- conventional: unlike physical laws, the rules are breakable; indeed sometimes they have to be breakable to be mutable (i.e. modifiable) to demonstrate the need for modification or benefits of improvement.

It can be assumed that the entities have established channels for reliable communication (which might be achieved using a common language, or any other mutually understood way of signalling), but not necessarily that they have a shared set of congruent values and goals. Moreover, entities cannot be considered identical in competence, and are distributed over a number of different preferences and attributes (such as avoidance of tyranny or propensity to cheat).

Therefore critical features for sustainability of such systems are that the over-arching set of rules should specify:

1. Articles of (voluntary) association, for inclusion and exclusion of entities, and expectations of entities in the system (such as adherence to the rules, participation in the application of the rules).
2. Structures and procedures for facilitating the intended functions of the system.

3. Structures and procedures for run-time self-modification of the over-arching set of social arrangements.
4. Processes for knowledge aggregation (deliberation and decision-making) and knowledge alignment (collective action in the public interest).
5. Trustworthy gatekeepers and processes for reliable knowledge codification.
6. A non-repudiable means of monitoring compliance, reporting non-compliance and resolving disputes, with appealable disciplinary procedures.
7. Limits of and rights to self-determination, whereby the selection, modification, application and enforcement of the rules can be conducted within specified guardrails and with recognised rights.
8. The specific form of political organisation defining the polity, in particular determining the organisational behaviour with respect to external actors and authorities.

The specification, and in particular the operationalisation, of the over-arching set of rules faces many problems encountered in political science and political philosophy:

- sustainability: features of, and design principles for, self-governing institutions to ensure sustainability of self-organisation [38];
- fairness and justice: canonical principles of distributive justice expressed as legitimate claims in a sector context to ensure fair allocation of resources [40], [41];
- knowledge management: processes for knowledge aggregation, alignment and codification to ensure ‘correctness’ in deliberation and decision-making, emergence of expertise through social influence [12], and successful collective action [39];
- legitimate consent: the avoidance to tyranny through voluntary association, negotiation and operationalisation of articles of association, which are informed, meaningful and revocable [42];
- legitimate dissent: the tolerance of dissent when outcomes of current practices are diverging from a communities shared set of congruent values [43];
- iron law of oligarchy: resisting the tendency of any organisation, no matter how democratically founded, to degenerate into control by a few who run it in their own, rather than the common interest [44];

- unrestricted self-modification: the observation that unrestricted self-modification of a set of conventional rules tends to paradoxical rules, inconsistency or incompleteness [45];
- voting paradoxes: standard problems in social choice theory about mapping expressed preferences over ranked candidates into a specific choice or choices [46],[47];
- citizenship and rights: for example, the human right to human decision [48], which is diminished by the increasing automation (using “AI”) in classification and selection tasks, e.g. for job selection, policing and sentencing;
- dignity: an individual and collective value, which is reinforced if people are able to make meaningful and valued contributions to issues of public interest, and is undermined if people are deceived into making decisions that they would not have made had they been aware of full or accurate knowledge [49].

Accordingly, next we introduce the main components of a SGS, and we analyse the challenges mentioned above.

2.2.3 Components of a SGS

This Section introduces the main components of a SGS, which include MAS, agents, social network, and institution. In more detail, a MAS is a collection of diverse, heterogeneous, independent, individual, naturally (i.e. humans) or artificially intelligent (i.e. software) agents, each with unique abilities [50]. These collaborate and exchange information, over a social network, in pursuit of a common objective beyond the capabilities of a single agent. The agents might, also, identify a set of socially constructed rules that they can use to facilitate their coexistence and the accomplishment of the shared goal [51].

An *agent* corresponds to an autonomous entity that can perceive its environment, make decisions based on its observations, and take actions to achieve its goals, while interacting with other agents other intelligent agents, being artificial or human [52]. Each agent has each individual characteristics that describe their preferences, behaviour, and knowledge. Those agents are situated in a shared (multi-agent) environment, and they can have some social skills (i.e. communication, negotiation, cooperation). Moreover, the agents can use their social skills to interact with other agents from their *social network*.

A *social network* corresponds to a social structure comprising a set of social actors (i.e. agents),

sets of connections between those social actors, and the relations between those actors. A fundamental issue in social networks is their topology, that affects the interactions between the actors of it and the way that one influences another [11].

As discussed above, we are interested in deliberate forms of organisation and governance. Hence, we focus on MAS in which the group of individuals makes an agreement to form an institution. Specifically, Ostrom[38] defined an *institution* as a collection of operational rules that define who has the authority to make decisions within a certain context, what actions are permitted or restricted, and include directives that prohibit, allow, or mandate specific actions or outcomes.

2.2.4 Challenges of Self-Governance

2.2.4.1 Sustainability

Addressing challenges of sustainability of self-governing institutions, Ostrom developed her theory on ‘governance of the commons’ [38]. Specifically, she focused on the abilities of communities to self-organise and self-enforce sustainable management of collective action dilemmas. Ostrom focused on collective action dilemmas applied in Common Pool Resource (CPR) problems in which individuals collaborate to achieve a common goal that is to sustain a pool of resources.

Generally, in a collective action problem, it is more beneficial for the individuals to collaborate, than act selfishly, although this conflicts with the personal incentives of the individuals [11]. When investigating this type of problem, we emphasised how cooperation can be structured and sustained in decentralised systems, and specifically on how norms and rules enable agents to organise themselves without a central authority. Also, the main challenges in collective action problems are free-riding, coordination, and enforcement, which are often dealt with in political settlements [53].

Specifically, a CPR problem corresponds to a collective action dilemma in which individuals contribute to, and appropriate from a shared resource, while trying to maintain this resource. Although individuals want to maximise their utility in the short term, they need to decide on a sustainable collective strategy that would prevent resource depletion. Accordingly, individuals have to balance out the tensions between individual, possibly short-term, benefits and collective, usually long-term, goals leading to sustainability.

Through a range of case studies, Ostrom [38] developed a model for governing the commons that transcends traditional market or state solutions, instead situating the responsibility for collective action within society itself. This approach emphasises self-organising collectives that, over

centuries, have managed CPR through the development of self-organising collective relationships, which correspond to networks of relationships and shared values within a community that facilitate the resolution of both present and future collective action challenges.

Specifically, she identified eight fundamental characteristics in these enduring institutions which were then reformulated as a set of design principles for creating self-governing institutions. Those design principles are:

1. Clearly defined boundaries
2. Congruence between appropriation and provision rules and local conditions
3. Collective-choice arrangements
4. Monitoring
5. Graduated sanctions
6. Conflict resolution mechanisms
7. Minimal of right to organise
8. Nested enterprises (for CPRs that are parts of larger systems)

They can be used to enable institutions prevent free-riding, and maximise the satisfaction of individuals, while remaining sustainable. However, in SGSs facing problems of management of a scarce shared resource, satisfaction becomes relative to others. Specifically, an agent's level of satisfaction is relevant to what they were allocation in comparison to what other were allocated, which requires also considering of issues of fairness and justice.

2.2.4.2 Fairness and Justice

Fairness and just procedures have long been topics of great interest [54] that most people in society view it as important across various situations, such as interpersonal relationships, organisations, and consumer behaviour. In general, individuals evaluate the fairness of an event or situation based on the outcomes and the procedures used to determine them, as well as considering how they are treated by others in the process [55]. Fairness, inter alia, has been examined in the context of distributive justice. According to Rescher [40], there are seven primary methods throughout history for determining who “deserves” rewards or punishments: equal treatment, or distribution based on need, ability, effort, productivity, social utility, or supply and demand. Rescher argued that distributive justice involves determining which of these principles, or “canons”, apply to a given situation (referred to as legitimate claims) and deciding how to balance them when multiple canons are relevant, or how to resolve conflicts among them by assigning different weights to each.

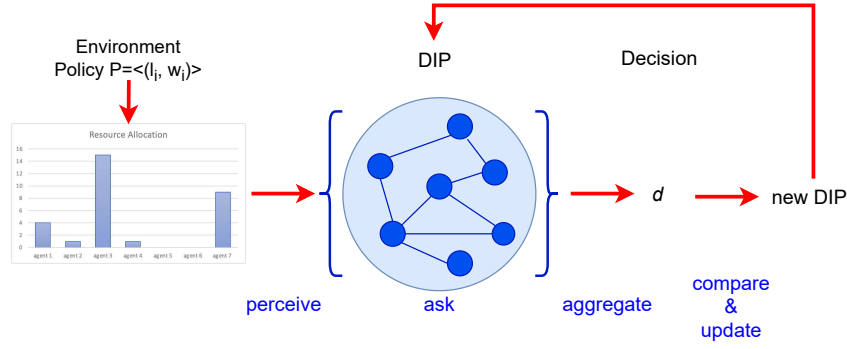


Figure 2.1: Subjective Value-Judgements performed by a DIP Unit

The combination of Ostrom’s institutional design principles for managing CPRs with Rescher’s framework for distributive justice through legitimate claims offers a structure for achieving both sustainability and fairness [56]. However, resource allocation in a CPR is not a one-time event but an ongoing process that involves value-based judgements, as shown in Figure 2.1. This figure depicts a DIP where agents are connected through a social network within a dynamic environment. In this context, the environment includes both factual elements, such as the available resources and distribution policies that are beyond the agents’ control, and institutional elements, such as the legitimate claims used to assess the distribution, which are within the agents’ control. Given the resources and the environmental policy (determined based on the legitimate claims and their relative significance), a resource distribution occurs, which can be observed by each agent in the DIP. In this way, agents can assess the fairness of the allocation by comparing the actual distribution (shaped by environmental policy) with their personal views on what it should be, based on their values and perceptions of fairness. This process can enable a group of autonomous agents to collectively answer questions of fairness of a policy, corresponding to a collective decision d , which can enable them to improve their mechanisms for self-governance in the future.

However, there are multiple ways in which individuals assessments can be used to affect the collective decision (which can correspond to an evaluation of the fairness of the resource distribution policy). The ways that a set of individual assessments can be mapped to a single output (i.e. collective assessment) have been extensively studied by Social Choice Theory [57]. However, the aim of this is not to review all the different mechanisms for achieving that, but we want to focus on how to combine the knowledge of the individuals to achieve effective self-governance. Hence, next we discuss a method of knowledge management that was used by classical Athenians and aimed to use the information that was spread across the individuals to effectively solve collective action problems. Additionally, we focus on an aspect of knowledge management and we present different approaches for aggregating knowledge to form collective assessments.

2.2.4.3 Knowledge Management

Knowledge Management (KM) is the process of enabling access to information for socially beneficial uses through individual decisions made within institutional frameworks, which support both innovation and learning. In his study of Athenian democracy, Ober [39] suggests that the effective resolution of collective action problems can be credited to the robust knowledge management methods developed by the Athenians, which included processes for knowledge alignment, knowledge aggregation, and knowledge codification. Specifically, he examines the institutional frameworks that support democratic knowledge management, including the use of social networks to gather information, public platforms to build shared knowledge, and open access to reduce transaction costs. He argues that when a government restricts the flow of information, democracy falters. While democratic participation and deliberation require state resources and social effort, [39] demonstrates that the benefits of a thoughtfully structured democracy significantly outweigh these costs.

Knowledge management, *inter alia*, involves knowledge aggregation. There are various methods for aggregating diverse knowledge into a single output but here we focus on four types. These include Independent Guess Aggregation (IGA), often referred to as the “wisdom of crowd” [58], a method based on expertise (EXP) [59], [60], another using a process of deliberation (DEL) [39], and one combining the task-related knowledge of different groups of experts known as Relevant Expertise Aggregation (REA) [61].

In IGA, each shares their preference and the final decision is determined based on a selection mechanism (e.g. a voting mechanism). Although this method increases participation, is inclusive and relatively economic, it assumes that there is no social interaction, and relies on the majority being correct. In contrast, EXP uses a group of experts to come up with a decision. This method is more likely to come up with an accurate, or correct, decision, but is expensive and might lead to groupthink.

Furthermore, DEL involves establishing a formal structure with rules for decision-making. While it promotes broader participation, it can be slow and expensive. Social interaction is vital in this method, and success often depends on persuasive communication, though domain expertise is not always essential. Moving on to REA, it is a method that brings together expert groups from different fields to rank options for solving a problem. Non-experts then express preferences for the experts’ recommendations. This method balances expert judgement with public input, aiming to address the tension between majority opinion and expertise seen in other methods.

Overall, different methods are more suitable for different situations (e.g., IGA for guessing games, EXP for complex tasks). However, in socio-technical systems, it is not sufficient to optimise against one objective, as would be analysed in Section 2.3.

2.2.4.4 Legitimate Consent & Dissent

In SGSs, collective agreement is essential, serving both as a foundational requirement and as a continuous principle for effective implementation. For instance, in Ober’s Demopolis thought experiment [42], after citizens are selected based on their preference for avoiding tyranny, they must establish consensus on fundamental rules regarding participation, legislation, and institutional entrenchment. This process exemplifies the democracy-by-design framework, which supports the development of decentralised platforms for public interest consultations [62].

A key issue in forming consensus is accommodating legitimate dissent, especially when established practices diverge from a community’s core values [39]. Suppressing dissent risks enforcing conformity, whereas knowledge diversity is crucial for aggregating collective insights [39]. Furthermore, varied perspectives are often necessary for innovation and adaptation to evolving environments; thus, failing to sustain diversity could adversely impact long-term effectiveness.

Additionally, Canevaro [20] contends that a deliberative system seeking consensus through democratic processes does not inherently require unanimity. Occasional recourse to majority decision-making, when consensus cannot be achieved, does not undermine its democratic nature.

In this context, legitimate dissent and non-compliance are not only expected but necessary within SGSs. There are instances where the norm itself may be sound, but those enforcing it may act improperly; conversely, there are situations where the norm itself may be flawed. In such cases, dissent and non-compliance are vital, as they enable correction, uphold integrity, and ensure the system’s resilience and adaptability [43].

2.2.4.5 Iron Law of Oligarchy

Another challenge of a SGS is to avoid its tendency to degenerate into a form of governance in which a privileged elite blocks proposals for the common good in order to protect their own self-interest [44]. Specifically, Michels argues that any organisation, regardless of how democratically was founded, is prone to engage in less democratic or even oligarchic forms of governance. In his book, he examines the forces that hinder the achievement of democracy, arguing that these

forces can be grouped into three categories: those rooted in the nature of the individual, those arising from the structure of the political system, and those stemming from the characteristics of the organisation.

2.2.4.6 Unrestricted Self-Modification

Suber [45] explores a logical contradiction that emerges when a legal system, such as a constitution, permits self-amendment, particularly when using its own amending clause to modify itself. The paradox lies in whether a system can authentically alter its foundational rules, including the rule enabling amendments, without undermining its own authority. When a constitution amends its own amendment rule, it simultaneously relies on and modifies the rule's authority, creating a contradiction in which it both preserves and transforms the original source of its power.

However, historical evidence shows that self-amendment is common, legal analysis affirms its lawfulness (even if some logicians claim it is self-contradictory), and philosophical analysis regards it as foundational to the concept of legality. The legal acceptability of self-amendment thus illuminates key aspects of legal reasoning and rationality, challenging the notion of immutability at any level of law.

2.2.4.7 Voting Paradoxes

In a SGS, a group of individuals has to come up with and agree on a set of socially constructed social arrangements that would affect their living. However, there are various ways of making a collective agreement on those rules, roles, structures, and procedures. While consensus seems to be the favourable approach, often it cannot be achieved. Thus, often voting mechanisms are used for forming a collective agreement.

There exists a great variety of voting mechanisms, which are based on different sets of rules, which include Plurality, Condorcet pairwise comparisons, the Borda count protocol, and the Hare or Coombs systems. Although in the two-option scenario they converge on the majority winner, in multi-option cases they may lead to different outcomes [46].

This observation was formally specified in Arrow's book on 'Social Choice and Individual Values' [47]. Specifically, he presented his theorem that suggests that in the absence of limitations on individual preferences or the neutrality of the constitution regarding viable alternatives, there is no social choice rule that can meet a reasonable set of criteria. This finding extends the voting

paradox, which demonstrates that majority voting can lead to unsatisfactory outcomes for the collective. As such, identifying the more appropriate mechanism for making collective decisions with respect to ‘matters’ (i.e. social arrangements) of common interest in SGSs remains one of the fundamental challenges of effective and sustainable self-governance.

2.2.4.8 Citizenship and Rights

The rapid rise of Generative Artificial Intelligence (GenAI) applications, such as ChatGPT and Bing, has highlighted a significant gap between technological progress and regulatory oversight. This situation raises the question of what ethical framework can effectively guide AI regulation. Tasioulas [48] suggests that a broadly Aristotelian ethical framework holds great potential and offers advantages over the utilitarian approach commonly adopted by many AI theorists and policymakers. Specifically, he examines the following two topics:

1. The role of work in a meaningful life and how advancements in AI impact work as a source of human fulfilment, questioning whether play might replace work in an AI-driven future.
2. The need for fundamental changes to our human rights framework in light of AI developments, particularly regarding the recognition of a human right to a human decision, which would reserve certain forms of decision-making exclusively for humans.

Therefore, during this period of digital transformation, effective self-governance necessitates identifying ways to integrate technology into society while ensuring the protection of human rights [2]. It is essential to develop regulatory approaches that not only accommodate technological advancements but also uphold ethical standards that prioritise human dignity and agency [8].

2.2.4.9 Dignity

In SGSs, dignity is both an individual and collective value, grounded in people’s ability to make informed, meaningful contributions to public issues. A dignified SGS operates within a Zone of Dignity (ZoD), which acts as a guardrail for policy options [49]. This zone delineates acceptable boundaries, avoiding extremes like excessive control or neglect of autonomy, and balancing the need for resource extraction (taxation) with respect for self-organization. By situating policies within a two-dimensional plane of control and taxation, the ZoD can identify points where systemic

configurations risk collapse—through exploitation, stagnation, devaluation, or fragmentation—if dignity is compromised.

However, recent AI advancements introduce new risks to civic dignity through deception, particularly as AI can disseminate disinformation that misleads individuals, influencing choices they might not make if fully informed. This deception compromises both personal dignity and the integrity of self-governance, as individuals are steered by inaccurate information, limiting their capacity for autonomous, informed decision-making.

To protect dignity in these hybrid systems, mechanisms must address the risks of disinformation and ensure transparency, allowing individuals to engage in governance processes free from manipulation. Sarkadi’s work on the resilience of SGSs to deceptive behaviour highlights the need for adaptive, meta-rule safeguards that detect and mitigate deception [63]. By countering deceptive influences, such mechanisms preserve civic dignity, ensuring that human-AI interactions support rather than erode society’s capacity for dignified self-governance, where individuals and communities engage with accurate information and full autonomy. This focus on combating deception within hybrid systems underscores the need to protect civic dignity, ensuring that human-AI interactions support, rather than undermine, a society’s capacity for dignified, informed self-governance.

2.3 Social Influence

Social influence is a key element affecting how individuals interact with each other. It refers to any deliberate or unintended effort to change another person’s thoughts or actions [64]. This influence can take many forms depending on the goals of both the influencer (being the source) and the one that is being influenced (being the target). While some forms of social influence can be beneficial for the individual, others may have negative impacts, regardless of factors like age or background [65]. In daily life, negative forms of social influence may appear as conformity [66], obedience [67], persuasion [68], compliance [69], or even manipulation. Conversely, social influence can also benefit individuals who seek out information [70]. In this case, it can be used as a tool for knowledge management in MASs. Thus, this Section focuses on the positive aspect of social influence and introduces a psychological theory, called RTSI, together with a computational model that uses RTSI to improve the performance of a Distributed Information Processing (DIP) unit corresponding to a self-governing MAS.

2.3.1 The Regulatory Theory of Social Influence

In RTSI [12], Nowak et al. emphasise the beneficial aspect of social influence, particularly how targets may actively seek sources to whom they can delegate tasks such as information gathering and processing. This theory posits that social influence is bidirectional: not only do sources try to influence targets, but targets actively seek sources by whom they wish to be influenced. Specifically, the latter occurs because targets want to gather information or information processing tools.

Accordingly, social influence enables the receivers of information to leverage others' (social) network, processing capacity or knowledge, maximising their access to information and information processing rules. In this way, a collection of autonomous entities (i.e. agents) situated in the same environment operate as a DIP unit. Overall, 'targets' use social influence as a tool for making decisions or judgements, deliberately choosing sources and topics that align with the expertise of the source and the topic of their interest.

2.3.2 RTSI for Collective Decision Making

Drawing inspiration from RTSI, Pitt et al. [71] leverage social influence to enable a group to form a quantitative assessment. In particular, they designed a computational model of RTSI to enable a group of N agents, spread in a grid, to form a collective agreement on the value of a signal which is represented by a float value ρ . Each agent, depending on their location in the grid, can observe a different value of that signal. Agents formed an individual decision on the value of the measurement either using their own knowledge or asking their social acquaintances. Running experiments on a multi-agent simulator, they showed that social influence can improve the system's accuracy in signal tracking. This was achieved through the emergence of expertise which leads to individual and collective improvements over time.

Additionally, in previous work [9], we explored how a group of N autonomous agents, interconnected over a social network SN , can organise themselves without a centralised authority, to form a collective agreement on a qualitative assessment C , while each agent has its individual opinion. This requires them to operate as a DIP unit that uses RTSI to improve their information processing. In more detail, we examined a scenario of CPR management problem in an economy of scarcity. In this scenario, agents contributed resources to a common pool while they also needed some resources to satisfy their needs. The resources allocated to the agents were decided based on a resource distribution policy $\mathcal{P}$ that was taking into account the agents' legitimate claims, as

specified in Rescher's theory [40] and presented in Table 6.1.

No.	Rank agent based on:	Rescher's Canon:
lc_1	times participated	effort
lc_2	times allocated $ri > 0$	equals
lc_3	average provisioned amount	contribution
lc_4	average requested amount	needs
lc_5	average allocated amount	equals
lc_6	normative satisfaction	equals
lc_7	times asked for evaluation	social utility
lc_8	times asked for rule	social utility

Table 2.1: The eight legitimate claims in LC with respect to the Rescher's Canons

The policy $\mathcal{P}$ was externally determined and comprised a set of legitimate claims $LC_{\mathcal{P}}$ each associated with a weight w_{lc} . Those weights represented the relative effect of that legitimate claim to the resource distribution, compared to the effect of the other legitimate claims. Each of those weights is represented by a float number having a value in $[0, 1]$ and they all sum to 1. Following this specification, an indicative policy was that presented by the equation 2.1.

$$\mathcal{P} = w_{lc_1} * lc_1 + w_{lc_2} * lc_2 + w_{lc_3} * lc_3 + w_{lc_4} * lc_4 + w_{lc_5} * lc_5 + w_{lc_6} * lc_6 + w_{lc_7} * lc_7 + w_{lc_8} * lc_8 \quad (2.1)$$

where $w_{lc_1} + w_{lc_1} + w_{lc_2} + w_{lc_3} + w_{lc_4} * + w_{lc_5} * + w_{lc_6} + w_{lc_7} + w_{lc_8} = 1$.

The agents in the population observed this distribution of resources and evaluate its fairness based on their individual rules. Specifically, the definition of the agents was an extension of presented in [72], and each agent $i \in A$, $A \in \mathcal{I}$ was defined in 3.2:

$$i = \langle attr, ijrtsi, SN, \mathcal{J}, LC \rangle \quad (2.2)$$

where $attr$ is a set of attributes that describes agent's coefficients, values and weights, $ijrtsi$ the framework used for the evaluation of the resource allocation method which combines own allocation according to legitimate claims with delegating the task to the social network, SN is the network that describes the social connections of the agent, $\mathcal{J}$ is the value that affects the assessment of the agent and is the same with the value of institution, namely fairness, and LC is the set of rules that have been given to the agent. Each agent owned a random set of legitimate claims LC_i which is a subset of the set of claims presented in Table 6.1. For instance, that could correspond to:

$$LC_i = w_{lc_1}^i * lc_1 + w_{lc_2}^i * lc_2 + w_{lc_7}^i * lc_7 = 0.1 * lc_1 + 0.1 * lc_2 + 0.8 * lc_7$$

Accordingly, individually, agents assessed the fairness of the policy based on their own legitimate claims. Also, agents could either assess the fairness of the policy by using their own legitimate claims (forming their individual opinion $o_{i,own}$) or by asking the opinion (i.e. the qualitative assessment) of their social network SN (forming opinion $o_{j,own}$ where j is the neighbour selected to be asked). The social network of the agents followed the structure presented in [73] and was generated according to the algorithm described by [74]:

$$SN = \langle N, m, \mu \rangle \quad (2.3)$$

where N is the number of agents, m is the number of nodes in the initial fully-connected network, which is also the size of the active node list (which is a list biased toward containing nodes with higher degrees), and the number of connections from new $(N - m)$ nodes to existing nodes, and μ the probability of attaching a new node to an active or deactivated node.

Furthermore, the agents in order to identify whether they wanted to use their own or their neighbours opinion, they considered the value of their self-confidence and their credence to their social network. In more detail, they were initialised to have the same self-confidence and credence to each of their neighbours in their social network, and those values were updated after each iteration of the algorithm. The individual opinions o_i of each agent i on the fairness of the policy were aggregated to form the collective assessment C . Then, agents updated their self-confidence sc_i and credence $c_{i,j}$ coefficients based on the deviation of their own assessment from the collective outcome, compared to their neighbours' deviation from the collective outcome. This process occurred iteratively for T epochs.

The scenario is described by Algorithm 1, which was used to design the SMARTSIS Simulator. Furthermore, the experimental results on the simulator demonstrated that RTSI facilitated the development and emergence of expertise, as a specific subset of the population was identified as experts, and their opinions or processing rules were frequently sought. However, this expert group varied over time, allowing for a range of opinions (i.e. assessments of fairness) to surface throughout the iterations. The experiments also supported Price Law [75], which states that half of an organisation's work is done by the square root of its population. In this case, the task of evaluating the resource distribution policy was delegated to the experts, representing the square root of the population. This improved the quality of the collective decision and decreased the computational costs.

Algorithm 1: RTSI for Forming a Qualitative Assessment

```

 $t = 0$ ;
Set  $T = \text{max epochs}$ ;
Initialise  $N$  agents:  $\forall i \in N. \text{initialise}(sc_i, c_i, LC_i)$ ;
Initialise  $R$  resources;
Generate Social Network  $SN$ ;
Specify Initial Policy  $\mathcal{P}$ ;
while  $t < T$  do
    Generation of Resources  $R$  & Distribution according to  $\mathcal{P}$ ; // agents' generate resources,
    provide them in the common pool and resources are distributed according to
    policy  $\mathcal{P}$ 
    for  $i \in N$  do
        Computation of Individual Opinion  $o_{i,own}$  based on  $LC_i$ ; // agents compute own
        assessment
        Select Neighbour  $j \in SN_i$ ; // select neighbour having the greatest value of
        credence
        RTSI for Opinions  $ask(j, o_{j,own})$ ; // agents ask their neighbours for an assessment
        Selection of Assessment  $o_i = o_{j,own} || o_i = o_{i,own}$ ; // agents select between self and
        network's opinion
    end
    Collective Assessment  $C = \sum_{i \in N} o_i$ ; // individual assessments are aggregated to form
    the collective
    for  $i \in N$  do
        Update Agent Parameters  $sc_i, c_i$ ; // update credence and self-confidence based on
        collective assessment
        RTSI for Rules  $ask(j, LC_j)$ ; // ask for rules if necessary
        Update Agent Rules  $LC_i$ ; // update sets of rule based on rules of neighbours
    end
    Store Systemic Parameters  $R, C$ ;  $t = t + 1$ ;
end

```

Apart from proposing an algorithm that enables a group of autonomous agents to self-organise as a DIP unit using RTSI to achieve knowledge aggregation and resource conservation, this work also identified eight criteria (accuracy, economy, coherence, diversity, specialisation, inclusivity, stability, and flexibility) that reflect systemic performance, often in tension with each other. Specifically, the experiments illustrate the dynamics of expertise and labour division, showing that a small percentage of agents often become transient experts, contributing most to the system's decision-making while still allowing for inclusivity and knowledge sharing. This balance is key to knowledge management leading to system sustainability in both cyber-physical and socio-technical systems.

2.3.3 From Social Influence to Adaptation

Social influence can inform a mechanism for knowledge management in self-governing MAS. In contrast to other methods of knowledge management, social influence, as described in RTSI, can balance out conflicting systemic drivers through the emergence of expertise. However, existing research [9], [71] is limited in using RTSI to enable a group to react to a stimulus of the environment.

Also, the developed model considered that the system's performance over time could be evaluated based on eight criteria, which formed four pairs of conflicting systemic drivers: accuracy vs. economy, stability vs. flexibility, inclusivity vs. expertise, and coherence vs. diversity [49]. To ensure effective functioning, the system needs to be capable of evaluating whether the balance between the conflicting pairs of drivers is upheld. This, in turn, requires the system to engage in

reflection [76] and reflective governance [49] to answer questions concerning their own performance. RTSI could form the basis for achieving reflection and adapting to an ever-changing environment. To achieve that, though, the system need to possess a set of metrics measuring its performance, as will be analysed in Section 4.10.

2.4 Learning

As observed in [9], a simple form of learning could be achieved using simple trust and self-confidence coefficients to enable agents to identify whether they should use their opinion or one coming from their social network to effectively react to their environment. However, this Section focuses on more elaborate forms of learning. Thus, it briefly introduces AI methodologies for enabling a group to develop awareness of its functioning and learn to improve it over time. Those include some concepts of (first-order) cybernetics, and ML.

2.4.1 Self-Regulation & First Order Cybernetics

Building on the concept of the DIP unit [12], this corresponds to a system consisting of autonomous components (i.e. agents) embedded in a dynamic environment and connected through a social network. Each agent is equipped with sensors to perceive the environment, while it also owns attributes that affect its reaction to environmental stimuli. The agents, however, have to produce a unified outcome, which requires them to interact and form collective agreements.

An external observer of the system (i.e. the DIP), would perceive it as a single unit, although it comprises multiple heterogeneous and autonomous entities. The DIP could be described as a self-regulated system. The latter corresponds to a (first-order) cybernetics system operating with an input, an output, and a negative feedback loop connecting the output with the input [16]. The first-order cybernetics system is a goal-directed system that is perceived as a black-box function to an observer outside the system. The feedback loop is the mechanism of the system that enables it to operate autonomously, by evaluating its performance and adapting its functionality [22].

In particular, a self-regulated system corresponds to an “adaptive organism” [16] which owns two feedback mechanisms. The one is the negative feedback from the output to the input, which was analysed above. This corresponds to the outer reflective loop that accounts for gathering information necessary for adaptation. The other is the inner reactive loop which leverages the

interaction between the agents to achieve success and enable the system to operate in the desired way. So, the inner feedback loop enables the system to effectively perform in the short term, while the outer feedback loop is responsible for adapting to ensure functionality in the long term.

In summary, the self-regulated system is a system of systems, comprising a regulated unit (corresponding to a group of agents) and a regulator (corresponding to a designated agent or a group of agents). M. Ashby [19] built on the Good Regulator Theorem, which suggests that an effective regulator needs to have a model of the system that it aims to regulate, and the Law of Requisite Variety, which suggests that a regulator must be more complex than the regulated system, and identified the sufficient and necessary conditions, referred as requisites, for ensuring ethical and effective regulation in a cybernetics system. Those requisites include truth, variety, predictability, purpose, ethics, intelligence, influence, integrity, and transparency. Overall, those requirements for effective and ethical regulation can provide a mechanism for learning to adapt to the environment and achieve self-improvement in socio-technical systems.

2.4.2 Machine Learning (ML) Overview

Machine Learning (ML) describes a system's ability to use empirical knowledge, i.e. data, to produce knowledge that it can transform into useful models. This allows for intelligent decision-making across various fields, including healthcare [77], manufacturing [78], education [79], and entertainment.

In the context of self-governing MASs, a method that can enable a system to make effective decisions in uncertain environments is Reinforcement Learning (RL) [80]. This methodology focuses on learning how to make decisions in the presence of limited information concerning the relationships between observations, actions, and outcomes. Specifically, it models the environment as a Markov Decision Process (MDP) defined as $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$ where $\mathcal{A}$ is a set of actions, $\mathcal{S}$ is a set of states, and P is the state transition probability function, R the reward function and γ is the discount factor. Accordingly, s corresponds to a state from the set $\mathcal{S}$, a corresponds to the action selected from the set $\mathcal{A}$, and $P(s, a)$ corresponds to the probability of a transition from state s_t to state s_{t+1} when an agent executes action a . Also, each time agents selects an action it receives immediate reward $r \in R$.

RL can be described as a methodology that uses function approximations to help agents select actions that maximise future cumulative rewards. This is achieved by defining an agent who interacts with the environment and tries to learn the policy π that maximises its expected return

$\mathcal{R}_t := \sum_{t=0}^{\infty} \gamma^t R_t$. In Q-Learning, this policy is learnt by defining an action-value function $Q : (\mathcal{S} \times \mathcal{A})$, which represents the value of action a in state s , and follows the Bellman equation [81]:

$$V(s_t) = \max_a (\mathcal{R}(s_t, a) + \gamma V(s_{t+1}))$$

According to the algorithm proposed in [17], the function Q is updated at each time step by:

$$Q(s_t, a_t) = Q(s_t, a_t) + \alpha(r_t + \gamma \max Q(s_{t+1}, a_t) - Q(s_t, a_t))$$

where α is the step-size and (s_t, a_t, r_t, s_{t+1}) is the transition from state s_t to s_{t+1} with reward r_t and γ is the learning rate.

Building on the Q-learning algorithm to enable learning in large scale, [82] defined a method for combining Q-learning with deep neural networks. Additionally, Deep Q-Learning Networks using deep neural networks for approximating the optimal state-action function Q was introduced in [83]. In this technique, a stochastic gradient descent algorithm was used to minimise the loss between the learnt network θ and the target θ^- . Furthermore, other variants of Q-learning have been introduced to deal with issues of overestimating the expected reward [84], or for generalising learning across different domains [85].

To model environments where multiple networked agents embedded in the same environment we can also use Multi-Agent Reinforcement Learning (MARL) approaches. MARL, structured as a Multi-Agent Markov Decision Process, addresses collective decision-making challenges. In MARL, agents learn to adapt to changes in the environment and the behaviours of other agents to achieve their goals. Depending on the context, agents may cooperate, compete, or form teams. MARL is particularly effective in enabling organisms within an ecosystem to learn and adapt, working together to maintain sustainability.

Additionally, recent advances in technology introduced another learning methodology, that of Generative Artificial Intelligence (GenAI). GenAI corresponds to a method for using empirical data to produce knowledge (as a result of the recombination of existing information). This methodology allowed for the creation of elaborate Large Language Model (LLM) capable of providing real-time answers to queries expressed in natural language. Also, GenAI, in the form of LLMs, was combined with a multi-agent simulator by Park et al. [86], who introduced a model for simulating human behaviours. This approach demonstrated an effective use of computational intelligence for simulating believable human-like behaviour in agents that can observe, plan, and reflect.

2.5 Innovation & Empowerment

As introduced in Chapter 1, the next generation of socio-technical systems features meaningful interactions between human intelligence (natural life, or NLife) and artificial intelligence (artificial life, or ALife) in a hybrid NA-Life system [23], [63]. This imposes a requirement for natural life to ensure that technology is integrated into the social system in a way that empowers humans and communities to move towards more sustainable hybrid (i.e. human-AI) realities. Also, this requires humans to innovate the ways they organise themselves, since current practices might not be sufficient to adapt to the new and unprecedented circumstances. Accordingly, this section discusses issues of innovation for human and community empowerment in hybrid SGSs.

In particular, it reviews power and empowerment definitions in different contexts and presents patterns of higher-order cybernetics. This is followed by an analysis of innovation and a remark on the self-reinforcing nature of empowerment and innovation in the context of SGSs.

2.5.1 Power & Empowerment

To support human and community empowerment in socio-technical systems, we need to investigate the concepts of power and empowerment in various different contexts, often contentiously [87]. In [88], a formal characterisation of institutionalised power was developed using modal logic. In this framework, an agent could be appointed to a role which would allow that agent to create facts of conventional significance through particular actions (mainly speech acts), which were treated as if the institution itself had performed them. Extending that framework, Pitt et al.[89] specified a model for procedural justice, achieving an equitable distribution of institutionalised power among agents in a multi-agent system.

Other forms of power, more relevant to empowerment in the context of self-governance, include sovereign power, which is described as a monopoly on sovereignty, bureaucracy, and charisma [90]; and constitutional power, which is related to the establishment and contextualisation of rules, roles, and structures, in an institution. Additionally, economic power has been defined as the accumulation of scarce resources and the influence that entails, while compute power has been related to the capacity to execute computationally intensive tasks with computers. The latter form of power, specifically, becomes increasingly important for empowerment or disempowerment, due to the vast adoption of ML technologies which are computationally expensive and can have a significant societal impact.

2.5.2 Higher Order Cybernetics

This Section introduces patterns of higher-order cybernetics which can be used for increasing human empowerment in hybrid SGSs. However, in this Section we will focus on the analysis of the different orders of cybernetics while in Chapter 6 we will demonstrate a mapping from cybernetics orders to levels of empowerment.

The Law of Requisite Variety [16] shows that there is potential for using higher order cybernetics patterns to enable different levels of participation and engagements of humans in the processes of self-governance. This Section develops the concepts of cybernetics, and specifically on the first order systems described earlier, and specifies the Law of Requisite Variety and introduces second to fourth order cybernetics systems.

The Law of Requisite Variety [16] specifies that a regulator has to be more complex, meaning that it should have a greater variety of responses, than the regulated system in order to effectively control it and respond to the potential disturbances. In other words, the regulator's variety of responses to the changes of the environment should be greater than the possible changes of the environment.

Earlier, this Chapter introduced first-order cybernetics systems corresponding to cybernetics of observed systems [91]. Although first-order systems are critical for achieving autonomy and self-organisation, higher order cybernetics patterns allow for higher level of control leading to greater empowerment. In more detail, second-order cybernetics are described as the cybernetics of observing systems [91], since they concern systems having an observer. This observer can affect the system, which provides a second feedback mechanism for adjusting the behaviour of the observed system.

Third-order cybernetics are concerned with systems having an observer that not only affects the system but is also affected by it. Thus, the observer becomes an active and interactive component of the system that evolves as the system evolve. This adds a new feedback mechanism, on top of the feedback from the output of the self-regulated system to the input, and the feedback from the observer to the system, that is the feedback from the system to the observer. The observer understands that they are parts of the system, and both the system and the observer acknowledge the existence of each other and the fact that they are co-dependent [92].

Fourth-order cybernetics involve self-regulated systems having an active and interactive observer while they are situated in a certain environment. Thus, the whole system, comprising

by the self-regulated system and the observer is contextualised [93], while it has the capability to become an observer of itself. This results to the development of self-awareness and provides mechanisms of self-healing and self-regeneration. This is achieved through the embodiment of knowledge, which corresponds to knowledge acquired through experience.

2.5.3 Issues of Innovation

Section 2.5.1 defined empowerment in the context of SGSs and Section 2.5.2 analysed higher-order cybernetics patterns that will be used (in Chapter 6) for leveraging different levels of empowerment. This Section explores innovation-related issues by reviewing related concepts, existing instances of innovation in the context of governance, and techniques for achieving that.

2.5.3.1 Related Terms & Types

Innovation and related terms, such as creativity, inventiveness, imagination, defamiliarisation, improvisation, recombination, and abduction, have been studied across various scientific disciplines, including psychology, arts, philosophy, economics, and management. Some definitions relate with the creation or generation of new outcomes, content, or knowledge, from the synthesis of existing sources, inputs, experiences, materials, or even fantastic scenes [94]–[100].

Innovation is characterised as endogenous, if the source of it comes from within the system through process of synthesis or recombination of existing knowledge, and exogenous, if the source comes the environment through processes of knowledge transfer and collaboration with others [101]. Additionally, based on the speed that innovation is happening, it is classified into incremental, additive, complementary and radical [102]. Those categories refer to whether innovation occurred by performing a slightly improving action repetitively, by exploiting existing resources resulting in improved results, by introducing something new which results in a new structure, and by performing a fundamental change in the operation.

2.5.3.2 Political Innovation

Since the focus of this work is innovation for self-governance, this looks into existing paradigms of political innovation. Starting from antiquity, classical Athenians used a process for knowledge management, involving knowledge aggregation, knowledge codification and knowledge alignment. Specifically, knowledge aggregation considered the gathering of information from diverse sources

including also tacit and common non-documented knowledge, aiming to trigger the generation of new ideas. One of those methods was the method for relevant expertise aggregation described earlier in this Chapter.

Knowledge codification was concerned with storing information and expertise in a form that can be accessible and readily utilisable. This aimed to avoid spending time and effort in re-obtaining existing knowledge and expertise gained from experience. Although it is an area of contention, LLMs could be considered an attempt of a technology for knowledge codification. Specifically, they constitute approximations of distilled human knowledge and have been described to reflect the collective unconscious [103] or the collective wisdom [104]. With human (ideally expert) supervision, there is a potential for LLMs to constitute mechanisms for knowledge codification leading to knowledge management and innovation.

Although there is a potential for LLMs, they have not managed to have a significant effect on self-organisation and political innovation. The reason for that is because many factors (e.g. community maturity, complexity, common knowledge, and social cohesion) affecting collective well-being are inter-dependent and hard to measure [12],[105]. For instance, social cohesion is a property describing the extent to which the social arrangements of a community support coherence, resilience, and robustness, relies on community maturity and complexity.

However, according to research on citizen science [106], this property could be fostered by involving people in the decision making with respect to processes affecting the collective. Thus, AI technologies could be used to support participatory consultation through the development of platforms for democratic deliberation. Specifically, Zarkadakis [107] asserts that AI can change the political landscape by supporting citizen assemblies and strengthening the social bonds.

2.5.3.3 Artificial Intelligence for Innovation

The previous Section captured briefly how ML approaches could be used to support political innovation. Accordingly, this Section broadens the scope to include other AI approaches that try to achieve innovation. Those involve ML techniques such as Retrieval Augmented Generation, counterfactual explanations, LLMs, Reinforcement Learning, and some methods combining ML with human expertise.

Aiming to foster innovation by supporting scientific research, Lala et al. [108] used Retrieval Augmented Generation to answer questions based on scientific literature while evaluating the

source's expertise. Though the system is focused on research assistance, it can also be used for retrieving information while factoring in the credibility of the source.

Another approach relevant to innovation is presented in [109], which uses Deep Neural Networks to create counterfactual explanations that provide visual insights for a given image. While it focuses on generating explanations, the technique of counterfactual explanation generation may be valuable for offering new perspectives and solutions.

Supermind Ideator [110], on the other hand, uses a LLM along with prompting and fine-tuning, combined with a user interface, to help users solve creative problems. By providing a selection of ideas, this system facilitates human-machine interaction and supports users in generating innovative ideas, though some may be irrelevant or risky.

Additionally, a framework introduced by Thomaz and Breazeal improves human-machine cooperation by examining how humans teach machines to perform tasks. This framework provides guidance for a RL agent [111], and while applied in a different context, it highlights the potential of Socially-Guided Machine Learning (SGML) to trigger novel ways of interaction through human-ML collaboration. A hybrid approach, combining human brainstorming with an ML agent querying an LLM for relevant content, is proposed in [112]. This model, though focused on resource-efficient information extraction, shows promise for innovation by allowing the ML agent to distil human knowledge and assist in brainstorming new ideas, particularly in political contexts.

Although there is great potential, existing methods do not fully address confirmation bias, as they tend to generate ideas aligned with the user's original thoughts or queries. While certain approaches, such as [110], propose that LLMs can offer unconventional ideas—often seen as a limitation—they may help reduce confirmation bias. Still, they cannot assess the novelty, relevance, effectiveness, and applicability of these ideas in specific contexts. Therefore, while these methods can contribute to innovation, they cannot function as stand-alone solutions.

2.5.4 The Interconnection between Empowerment & Innovation

The study of issues of innovation and empowerment in next-generation self-governing socio-technical systems suggests that those concepts are tightly related. Specifically, during digital transformation humans need to be empowered to innovate their social arrangements (i.e. the way they organise themselves) in order to adapt to the new conditions. At the same time, the capability of an individual and a collective to self-determine their social arrangements is the essence of empowerment.

Hence, to be empowered during the transition to the digital society, humans need to be empowered to innovate their empowerment.

Additionally, there is another dimension: due to the integration of AI technologies into human societies, it is observed that the asymmetries in the distribution of power are augmented [23], [24]. Those asymmetries can be mitigated by modifying the existing structures, rules, and roles, or in other words, by modifying our social arrangements. Therefore, humans need to be empowered to innovate their empowerment (corresponding to social arrangements) in order to maintain an equitable distribution of power during Digital Transformation. So, in the context of SGSs, power leads to innovation and innovation leads to empowerment.

2.6 Summary

Chapter 2 discussed how self-governance in hybrid socio-technical systems can be achieved through the combination of theories of social sciences (i.e. anthropology, economics, political science, psychology, and sociology) with AI technologies. This briefly reviewed:

- *Self-governance*: This introduced the main concepts and issues of self-governance, by describing some types of autonomous organisation and governance, providing a definition of SGSs, presenting the main components of those and briefly discussing the main challenges of self-governance.
- *Social influence*: This introduced RTSI [12] and discussed how this theory can be used as a mechanism for improving information processing of a distributed system.
- *Learning*: This introduced concepts of cybernetics and ML, and described how they can enable agents to adapt and enhance their performance in dynamic environments.
- *Innovation*: This introduced innovation and empowerment, in the context of self-governance, and discussed why both are vital, as well as interconnected, for human empowerment during the transition to a hybrid future.

To achieve sustainable self-governance, however, we need methodologies for effectively combining these elements. Accordingly, Chapter 3 explores how different tools and mechanisms enable us to design adaptive systems capable of managing problems of self-governance in complex socio-technical environments.

Chapter 3

Methodologies for Self-Governance in Socio-Technical Systems

Contents

3.1 Introduction	35
3.2 Socially-Inspired Agent-based Simulation	38
3.2.1 Socially Inspired Computing Methodology	38
3.2.2 Machine Reasoning	40
3.2.3 Agent-Based Modelling & Simulation	41
3.2.4 Overview of Architecture & Implementation	41
3.3 Hybrid Reasoning and Learning	44
3.3.1 Introduction	44
3.3.2 Hybrid Reasoning and Learning Methodology	44
3.3.3 Overview of Architecture & Implementation	46
3.4 Socially-Guided Reasoning and Learning	48
3.4.1 Introduction	48
3.4.2 Socially-Guided Machine Learning	49
3.4.3 Power-Sensitive Design	50
3.4.4 Socially-Guided Reasoning and Learning (SGRL)	52
3.4.5 SGRL Architectural Overview and Notes on Implementation	53
3.5 Summary	56

3.1 Introduction

Multi-Agent Systems (MASs) research has (at least) two forms, each of them using a different methodology and having different objectives. One form focuses on building *precise and testable*

models of human social systems [113]. The other focuses on building a *prototype for an actual cyber-physical system*. In the former case, a simulation environment with specific rules and processes is developed and experiments are executed to observe emerging behaviours that enable the researchers to have a better understanding of human social systems [114]. In the latter case, researchers design models comprising computational agents (such as sensors and processors) and specify algorithms for optimising the performance of the cyber-physical system.

However, to address problems concerning self-governance in the next generation of socio-technical systems, in this thesis, we combine these approaches. We use existing methodologies (such as Socially-Inspired Computing (SIC) [11], [115], Machine Reasoning (MR) [13], Agent-Based Modelling and Simulation (ABMS), Machine Learning (ML) [116], Generative Artificial Intelligence (GenAI) [117], Socially-Guided Machine Learning (SGML) [118]) as the basis, and enhance them to design systems that integrate humans and computational components. Furthermore, we bring together processes of social systems with Artificial Intelligence (AI) techniques to design fit-for-purpose algorithms for learning and innovation. Below, we present the incremental refinement of three methodologies to include machine reasoning, machine learning and human-machine innovation capabilities, resulting in the novel Socially-Guided Reasoning and Learning (SGRL) methodology for designing and implementing Self-Governing Systems (SGSs).

Firstly, we investigate whether a human mechanism for information processing (i.e. social influence [12]) can inform the design of an algorithm for collective decision making (i.e. enable a group to identify the processing rules producing an observed phenomenon). Therefore, we follow a Socially-Inspired Computing (SIC) methodology to design a computational system that operates as a model of a social system. We include also MR techniques to combine logic-based decision-making with socially-inspired, collaborative, problem-solving (i.e. collective decision making with respect to issues of self-governance). Thus, we develop a socially-inspired model capable of performing rule-based reasoning, which is implemented following the ABMS framework. This enables simulation with the model to observe the emergent collective patterns, and evaluate whether this model can support sustainable self-governance through effective knowledge management. Overall, in Chapter 4, we present a study that combines these three existing methodologies, namely SIC, MR, and ABMS, within a framework to explore whether our model (which draws inspiration from social influence) can solve a collective decision-making problem (i.e. development of explanatory adequacy) [10].

Secondly, we explore dynamic scenarios that require designing systems with the capability to react and reflect on their environment. This imposes a requirement for integrating ML techniques

to enable agents to reason and learn. We introduce the Hybrid Reasoning and Learning (HyRL) methodology which unifies SIC, MR, ML, and ABMS to a single HyRL framework. HyRL builds on the neurosymbolic AI paradigm which brings together neural networks with knowledge-guided symbolic reasoning to create flexible systems that can learn and generalise [119]. The HyRL methodology is used to develop two models: one for establishing the pathways for requisite social influence [18], and one for achieving sustainable consensus formation [21]. Those models are analysed in Chapter 5, which essentially explores whether the interplay between the SIC-MR-ABMS methodology and ML enables dealing with complex issues of self-governance.

Thirdly, we address ‘wicked’ problems [120] faced by hybrid SGS, which require integrating humans in our computational models, to achieve human empowerment through Power Sensitive Design (PSD). As such, we enhance the HyRL by incorporating GenAI capabilities and by using an SGML framework to involve human user(s) that steer the system in order to achieve Power Sensitive Operationalisation (PSO). This results in a novel innovation-support methodology, called Socially-Guided Reasoning and Learning (SGRL), that leverages the co-production between human-AI to support human empowerment in hybrid Self-Governing System (SGS). An implementation of the latter is presented in Chapter 6.

This Chapter describes the three incremental advancements of our methodology as follows:

1. SIC-MR-ABMS: combining the methodologies of SIC, MR, and ABMS we design and implement socially inspired algorithms for knowledge management. The synthesis of those allows to agents to simulate human behaviours, reason about their environment and react to it.
2. (SIC-MR-ABMS)+ML = HyRL: extending the previous framework to include ML techniques, similar to the neurosymbolic AI approach, we develop hybrid algorithms for effective self-regulation and sustainable collective decision making. This advancement enables the agents to not only reason about their environment but also learn from the interaction with it and improve their own and the collective’s performance.
3. HyRL + SGML + GenAI = SGRL: enhancing the previous framework to integrate humans, following an SGML approach, and GenAI, we introduce a novel methodology, called SGRL, for human-machine innovation and community empowerment through PSO. This way, we design systems that are able to reason, learn, while they can also innovate in order to adapt to an environment that changes in unexpected and unprecedented ways. At the same time, we not only follow a hybrid research methodology as in HyRL (i.e. it uses diverse AI techniques), but also we design hybrid systems comprising naturally (i.e. humans) and artificially

intelligent (i.e. AI agents) components.

Hence, this Chapter is structured as follows. Section 3.2 examines the SIC-MR-ABMS methodology, which is followed by the presentation of the HyRL methodology in Section 3.3. Then, Section 3.4 discusses the need for PSD, and introduces a novel SGRL methodology for innovation-support and PSO. Finally, this Chapter concludes with a discussion of the application of the three methodological advances with the following Chapters of this thesis.

3

3.2 Socially-Inspired Agent-based Simulation

This Section introduces our approach for investigating problems and phenomena observed in social systems, such as collective action, knowledge aggregation, collective decision-making, and social influence. This combines three existing methodologies (i.e. SIC [11], MR [13], ABMS [15], [121]) to design and implement systems that are capable of performing rule-based self-organisation. Running experiments using these models, we can understand better human societies and also explore whether the social processes used for designing our agent-based model are effective in solving problems of cyber-physical systems. Accordingly, this introduces:

- *socially-inspired computing*, which is used to design artificially intelligent computational models that are effective on performing rule-based operations and solving problems through collective action;
- *machine reasoning*, which corresponds to an AI technique that supports SIC and enables systems perform planning and inference tasks;
- *agent-based modelling and simulations*, which allows for developing platforms capable of supporting the experimentation on SIC and enable us observe emerging behaviours of the system under specific starting conditions and specific recurrence of particular events.

3.2.1 Socially Inspired Computing Methodology

The rapid advances of technology, and specifically in the field of AI, resulted in a new model of socio-technical systems. This model involves humans and software artefacts co-existing and co-evolving in a complex environment. This results in the development of novel tools and methodologies, including biologically-inspired computing and sociologically-inspired (or socially-inspired) computing. The

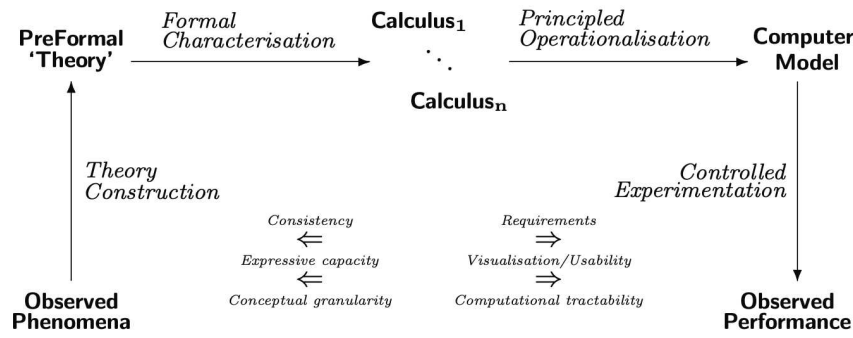


Figure 3.1: The socially-inspired computing methodology [11]

former investigates how biological systems can inspire new computing paradigms which can be used to optimise the performance of cyber-physical systems but also to enable us to develop a better understanding of complex biological systems [122]. Borrowing concepts of biologically-inspired computing, the latter corresponds to a synthetic method which uses theories of social sciences to address problems of complex socio-technical systems.

SIC is targeted at the study of collective action and self-organisation through rule-based operations. This focuses on the intersection of social sciences and AI and aims to identify the ways that social models can be used to design adaptive algorithms which can enable a group of autonomous, heterogeneous, independent or co-dependent, and interconnected artificial agents to maintain a balance between individual and collective objectives [115]. Accordingly, social metaphors are used to design systems in which complex patterns of behaviour and interactions can emerge.

Therefore, we follow a SIC methodology to study issues of self-governance in socio-technical systems for two reasons. Firstly, it enables complex social patterns to emerge which allows for identifying the rules and structures which are capable of mitigating complex adaptive socio-technical systems problems. For instance, Garbiso et. al [123] used a model of justice to design a self-adaptive clustering algorithm which manages to connect vehicles in a fair and socially acceptable way. Secondly, human processes and problem-solving mechanisms can be used to optimise cyber-physical systems. For instance, Li et. al [124] used concepts of trust to design an algorithm for information transmission in wireless sensor networks.

This methodology comprises three steps [11] which are summarised in Table 3.1. The first step involves *developing a theory* by defining, typically in natural language, the terms and principles used to describe the observed social phenomena. The second step of formal characterisation transitions from the primarily informal representations of the source theory to *formal representations*, expressed usually in a formal language. The final step involves *systematic, controlled experimenta-*

tion with a computational model of an artificial system derived from the formal representation. This step focuses on *translating the computational framework into a model* of the artificial system, where algorithmic intelligence is embedded within identifiable functions and processes. While this step may also involve selective or approximate modelling, it remains principled by clearly identifying and justifying the selections and approximations made.

Step	What	How
1 st	theory development	natural language
2 nd	formal representation	formal language
3 rd	computational model	code

Table 3.1: SIC steps

An example of applying the SIC methodology for solving problems of collective action is described in [11]. In the first step, Ostrom [38] defined eight principles for sustainable self-governance in the context of institutions addressing problems of Common Pool Resource (CPR) management. Pitt et. al [125] reformulated these principles and specified a formal model in Event Calculus [126]. In the last step, the formal model was translated into a MAS that was used as a testbed for running simulations and the effectiveness of this approach in dealing with issues of sustainability in norm-governed multi-agent electronic institutions [125].

3.2.2 Machine Reasoning

Machine reasoning refers to computational programs that use rules, logical representations, probabilistic reasoning, and symbolic reasoning to draw conclusions from facts and observations [127]. This allows designing machines capable of making to make inferences in a comparable fashion with humans [128]. In more detail, it enables software artefacts to perform abstract thinking and apply human-like common sense to produce explainable insights. This makes it feasible to combine sophisticated domain knowledge (in the form of context-specific facts provided by humans) with mathematical tools in such a way that enables machines to understand such facts and learn on their own.

Overall, we use MR to support SIC since it enables machines to perform reasoning and planning tasks through representations according to rules and calculus. So, we use them to map known theories from sociology, psychology, political sciences, and philosophy, to rules and concepts that can be understood by a computational system. Additionally, we develop models simulating human social systems properties and enhance them with capabilities that enable complex patterns and

behaviours to emerge.

3.2.3 Agent-Based Modelling & Simulation

Conventional simulations use computational models to develop a better understanding of the dynamics of a system and predict its behaviour, in cases that those dynamics could not be observed otherwise [14]. Such approaches mainly model the environment to comprise reactive agents that behave according to a set of condition-action rules. However, this results in robust systems which have limited expressiveness.

In contrast, ABMS study systems comprising autonomous, interacting, cognitive or deliberate agents the behaviours of which are based on the way they represent their knowledge about the environment as well as their experiences from past interactions. This makes those systems flexible and adaptable to the changes while allowing for observing more complex behaviours. Accordingly, ABMS can be applied to a wide variety of (complex) problems. Also, agent-based models are translated to software artefacts which can be reproduced and transferred to other applications. Another feature of agent-based models is that they can inform human decision making with respect to practical problems [15], [121]. Due to those three main characteristics, they have been applied to many different areas, including human social, physical, and biological systems. As such, in our work, we use ABMS as a way to implement our SIC methodology and run experiments.

Furthermore, modelling comprises of four main steps. First, it requires identifying the *agents* and their *behaviour*. Second, follows the step of the specifying the *environment* which affects their perceptions and triggers their actions. Finally, the third step is to describe the mechanisms for their *interactions* which might be achieved through communication or just observation of others' actions [14], [15]. In particular, the interactions of agents are affected by the connections of agents, the reason for their actions, and the instant that they occur. Having defined the three main elements of ABMS (i.e. agents, environment, interactions), the last step is the *implementation* of the system, which is described below.

3.2.4 Overview of Architecture & Implementation

There is a wide variety of toolkits that can support ABMS, such as NetLogo, MATLAB, Mathematica, Swarm [129], Repast [130], and Java [131], however in this thesis, we develop our agent-based models in Python. This is because it is an open-source, cross-platform, and accessible program-

ming language, with multiple available libraries, such as **numpy**, **scipy**, **matplotlib**, and **networkx**, that enable data manipulation and visualising the simulation results. Also, those results can be reproduced and validated by other researchers of the community.

3

In more detail, for our SIC agent-based models we implement the MAS using a modular framework for agent-based modelling, called **mesa** [132]. The operation of the system is as follows. Firstly, the system initialises the model by describing the available processes and knowledge, setting the values of the independent and dependent experimental parameters, generating the group of agents, with their properties (using **math** and **random** libraries) and their social network which is a Small-World-Scale-Free (SWSF) network using **networkx** library (which according to literature resembles real-life networks [74]). Additionally, it includes the agents in the schedule, and describes their behaviour as well as the transition dynamics of the system according to knowledge of social sciences (i.e. following the SIC methodology implemented using the ABMS framework).

Then follows the main loop which is implemented n times (the number n is predetermined by the designer of the system and is specified before the initialisation of the system). This executes the general functions (using python libraries such as **math**, **scipy**, **numpy**, and **statistics**) that are relevant with what happens in the environment and the way that the model changes over the time (such as resource generation, punishments, sensing of agents). Also, this triggers the step functions of the agents that are relevant with the specific tasks that agents perform (such as sensing and messaging), and stores all the variables. After the iterative execution of the central-loop for n -rounds the results are shown to the user (i.e. system designer), using the **matplotlib** and **seaborn** libraries, who can observe emerging phenomena and behaviours under different initial conditions. The python libraries used for the implementation of the system are presented in Table 3.2.

Library Name	Purpose	Relevance
<i>mesa</i>	scheduling, MAS, data collection	ABMS
<i>networkx</i>	social network	SIC & ABMS
<i>random</i>	init of agent params	SIC & ABMS
<i>scipy</i> , <i>math</i> , <i>numpy</i> , <i>statistics</i>	mathematical operations	mainly MR & SIC
<i>matplotlib</i> , <i>seaborn</i>	visualisation	ABMS & SIC

Table 3.2: Library Usage in SIC-ABMS

Algorithm 2: Indicative Control Flow of SIC-ABMS (based on RTSI from Section 2.3)

```

 $t = 0$ ;
Set  $T = \text{max epochs}$ ;
Initialise  $N$  agents:  $\forall i \in N. \text{initialise}(sc_i, c_i, LC_i)$ ;
Initialise  $R$  resources;
Generate Social Network  $SN$ ;
Specify Initial Policy  $\mathcal{P}$ ;
while  $t < T$  do
    Generation of Resources  $R$  & Distribution according to  $\mathcal{P}$ ;
    for  $i \in N$  do
        Compute Opinion  $o_{i,own}$ ;
        Select neighbour  $j \in SN_i$ ; Ask for opinion ( $ask(j, o_{j,own})$ );
    end
    Select Opinion  $o_i = o_{j,own} || o_i = o_{i,own}$ ;
    Aggregate Individual Opinions  $o_i$ ;
    Form Collective Assessment  $C = \sum_{i \in N} o_i$ ;
    for  $i \in N$  do
        Update of Individual Parameters  $sc_i, j_i$ ;
    end
    Store Systemic Parameters  $R, C$ ;
     $t = t + 1$ ;
end

```

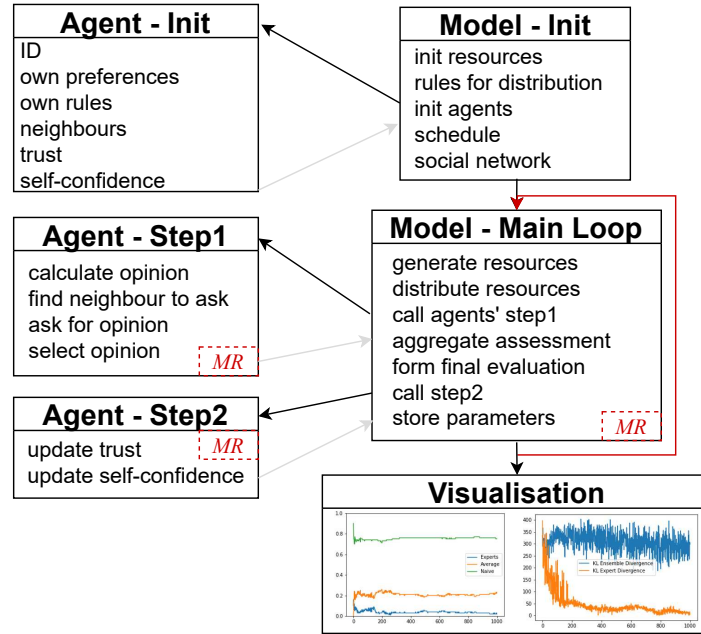


Figure 3.2: SIC-ABMS Architectural Overview

For instance, let us assume that we want to develop a model using social influence for forming a collective qualitative assessment, as presented in Section 2.3. The control flow of the system is described in the Algorithm 2. Finally, an indicative overview of the system is the one illustrated in Figure 3.2. This shows the implementation of the system according to the ABMS approach, designed based on the SIC methodology and using MR techniques to support that.

3.3 Hybrid Reasoning and Learning

3.3.1 Introduction

To explore more complex problems which require the agents not only to *reason* about their environment and *react* to it, but also to *learn* from the interaction with the environment, *reflect* on it, and *adapt* their behaviour, we enhance the methodology to include ML techniques, similarly to the neurosymbolic AI paradigm.

Specifically, we use MR to approximate reasoning and ML to approximate perception [116], which both correspond to capabilities of human cognition [128]. As discussed in Section 2.4, MR can effectively perform reasoning tasks while ML has been shown to be proven proficient in approximating perception. The latter is due to the fact that ML handles large volumes of data and uses mathematical tools to identify hidden patterns in data which are essential for predicting future outcomes.

Moreover, literature on various ecosystem models argues that hybrid models combining various techniques have proved more effective in adaptively managing complex environments [133], [134]. Accordingly, we extend the SIC-MR-ABMS approach by integrating ML. This results in the HyRL methodology which corresponds to a two-phase process seamlessly iterating over an MR and an ML operation, as analysed below.

3.3.2 Hybrid Reasoning and Learning Methodology

HyRL methodology is a cyclic two-phase process applied in a MAS [134]. Those phases correspond to *self-determination*, which enables the agents of the model to form explicit representations of their environment and reason about them, and *learning*, which empowers agents to reflect on their environment, learn from experiences, and adapt to a non-fixed parameter space. The former phase uses MR techniques to integrate concepts and theories from social sciences while the latter uses ML algorithms to learn from sequences of decisions (i.e. actions) and outcomes, and self-improve.

The *self-determination phase* is essential for enabling agents to explicitly represent and reason about their environment and select their actions. However, this requires understanding the system and modelling it such that it can simulate various possible behaviours. This can be facilitated by the study and application of theories from political sciences, sociology, and psychology. During this

phase, agents interact and influence one another, which results in the emergence of new norms and behaviours. Agents also make inferences reason about their environment using MR techniques.

After the self-determination phase, in which the population is updated, agent characteristics such as credence, self-confidence and preferences are adjusted based on their new experiences, and the environment changes (for instance resources are generated and consumed, and agents communicate), agents need to store these changes in their memory and reflect on them. To effectively deliberate on experiences and self-improve, they need learning capabilities which are enabled in the learning phase.

In HyRL, the selected ML technique for supporting the *learning phase* is that of RL, which was discussed in Chapter 2. Specifically, we propose to model the environment as a Partially Observable Markov Decision Process (POMDP), where agents learn to make both individual and collective decisions that foster sustainable self-organisation. Key properties and activities of the system are captured in the observation space, while the state space for each agent includes relevant environmental characteristics that assist decision-making (such as agreements on SAs corresponding to rules, roles, and structures).

The action space encompasses the possible actions and decisions agents can take (e.g. messaging, voting, moving). The reward function and learning objectives are designed to encourage long-term sustainability through adaptation. The HyRL methodology is designed to support both (single agent) Q-learning, and multi-agent Q-learning. In the second case, the agent definition is modified (compared to the one presented in Section 2.3) follows:

$$i = \langle attr, SN, LC, Q, a, s \rangle \quad (3.1)$$

where *attr* is a set of attributes that describes agent's coefficients, values and weights, *SN* is the network that describes the social connections of the agent, *LC* is the set of rules that have been given to the agent, *Q* is the q-table of each agent, *a* corresponds to agent's action, and *s* is the state of that agent.

Overall, this combined with the necessary information and knowledge enables agents to adapt their behaviour based on the outcomes of their actions, plan for future outcomes based on expected short- and long-term rewards, and maintain sustainability. After adapting to the 'new' environment, another phase of self-determination occurs which is followed by another learning phase, and so on.

3.3.3 Overview of Architecture & Implementation

The methodology described in Section 3.2 is enhanced to include variables, processes, as well as python libraries to support ML. Specifically, we include the `gym` and `stable_baselines3` libraries for performing RL operations. HyRL is designed to support both RL operations (performed in the main loop) and MARL (performed in the agents' step functions). Algorithm 3 shows the enhanced (i.e. HyRL) version of the control flow presented in Section 3.2 (i.e. Algorithm 2). The main enhancements are performed to allow for learning from past experiences and modifying individual and collective behaviour. For instance, in the resource distribution scenario, after forming individual and collective assessments, agents can modify their policy (for resource distribution) to improve satisfaction, remain economic, or maintain sustainability.

The control flow indicates in blue the parts that are added to support RL, and in green the parts that are added to support MARL. Q, A, S refer to the Q-table, the action, and the state spaces in the RL setting, while Q, a, s refer those for the MARL setting.

Algorithm 3: Indicative Control Flow of HyRL - based on algorithm from Section 3.2

```

 $t = 0$ ;
Set  $T = \text{max epochs}$ ;
Initialise  $N$  agents:  $\forall i \in N. \text{initialise}(sc_i, c_i, LC_i)$ ;
Initialise  $R$  resources;
Initialise  $Q$  table:  $\forall Q_{a,b} \in \{\text{state}, \text{action}\}. \text{set } Q_{a,b} = 0$ ; HyRL enhancement
Generate Social Network  $SN$ ;
Specify Initial Policy  $P$ ;
while  $t < T$  do
    Resource Generation & Distribution;
    for  $i \in N$  do
        Compute Opinion;
        Ask for Opinion;
        Select Opinion;
    end
    Aggregate Individual Opinions;
    Form Collective Assessment;
    for  $i \in N$  do
        Update of Individual Parameters;
        Update Q-Values; HyRL enhancement
    end
    Update Q-values and Modify Rules; HyRL enhancement
    Store Systemic Parameters;
     $t = t + 1$ ;
end

```

For instance, if the algorithm we implement corresponds to 3, an overview of the architecture of the system would be that illustrated in Figure 3.3. In this Figure, the blue parts correspond to the single RL approach of the algorithm, and the green to the MARL respectively. In either of the two scenarios, the parts that overlap with those presented in Algorithm 2 and Figure 3.2 correspond to the MR techniques designed according to the SIC methodology and implemented as per ABMS, while the rest of the parts (specified in green or blue) correspond to the ML parts

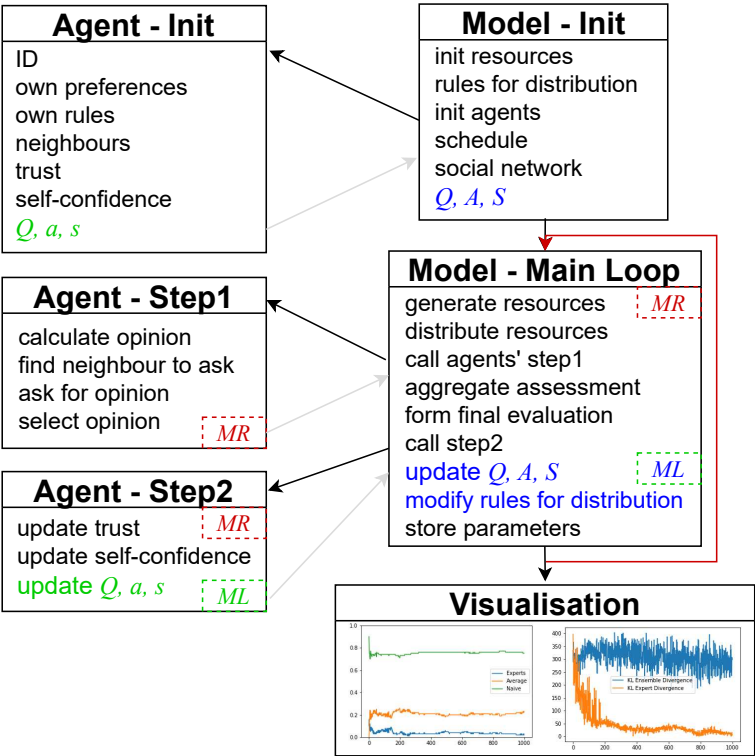


Figure 3.3: HyRL Architectural Overview

support that.

3.4 Socially-Guided Reasoning and Learning

3.4.1 Introduction

SIC methodology leverages rule-based operations, symbolic representations, and logic to develop systems that perform *inference and planning*. HyRL methodology integrates ML functionalities in SIC to enable systems to *reason* and also *learn* from interacting with an environment.

However, the increased hybridisation of our societies (i.e. the integration of AI actors in the social systems [63]), which is caused by the rapid advances in technologies, imposed a requirement for humans to modify the way they organise themselves and their societies. At the same time, the advancements of technology demonstrated the potential for using AI to support humans in innovating their solutions and adapting to this new landscape.

Hence, we propose utilising the interaction between human intelligence and AI to enable innovation and effectively support adaptation to unprecedented conditions. Specifically, we suggest leveraging the human-AI co-production to extend the HyRL framework to perform not only *reasoning* and *learning* but also *innovation*. It should be noted that the effects of human machine interaction are bilateral; not only the machine changes its behaviour in response to human input, but also humans might modify their behaviour as a result of the interaction with the machine [135]. Although both types of effects are important, in this work we focus on the former while the later remains a subject of future work.

Here, we are interested in innovation of Social Arrangements (SAs) for supporting human deliberation and empowerment, we need a human/socially-driven method. To achieve that, we draw inspiration from the Socially-Guided Machine Learning (SGML) methodology, which was briefly introduced in Section 2.5. However, to design systems capable of supporting human empowerment, we need to specify a methodology aimed at human empowerment. To that end, we propose Power Sensitive Design (PSD) [23], a new instance of Value Sensitive Design, in which the value that we optimise against is the equitable distribution of power in hybrid systems.

Finally, we extend the HyRL methodology to combine concepts of SGML and bring together human users(s), MAS and GenAI, to propose a methodology for designing and implementing

SGSs. This results in a novel methodology, called SGRL, for innovation support and PSO, leading to human empowerment through human-machine innovation.

3.4.2 Socially-Guided Machine Learning

SGML methodology focuses on learning from social interactions to enable machines to learn from (non-expert) human users [118]. It should be clarified that this approach does not coincide with Reinforcement Learning from Human Feedback (RLHF). RLHF [136], [137] is a methodology which updates the parameters of an RL model based on the feedback of non-experts aiming to align the model with human preferences. In contrast, SGML intends to enable humans to guide the RL agent through a feedback channel [138]. Rather than imitating human preferences, SGML encourages collaboration between humans and AI to jointly enhance the overall performance of the human-AI system.

This is because the motivation for developing this method is the realisation that in an ever-changing environment, it is not attainable to prescribe all the knowledge and capabilities that would be necessary for machines to support long-term decision-making. Hence, there is a need for a methodology that enables non-expert users, where non-expert might not only refer to the specific task but also to the human-machine interaction (e.g. users might be experts on the problem that they try to solve with the machine, but they are not familiar with how the machine should be treated to demonstrate the desired behaviour) to guide the machine. Therefore, machines should be designed with the capability to align their behaviour in response to the behaviour of humans. Overall, this aims for a two-way collaboration between humans and machines that supports human empowerment.

Although we share the same motivation with the designers of SGML, the application domain differs. SGML's objective is to combine human and machine capabilities to facilitate human-robot interaction. However, our aim is to support human-machine innovation in scenarios related to innovation for community empowerment. This requires, first, having a methodology to design systems that can support empowerment, and, second, having a systematic way to facilitate the collaboration of (one or more) user(s) with machines to achieve innovation. The first is accomplished by the PSD methodology, which is presented in the next Section, while the second is managed by the operationalisation of PSD through the SGRL methodology.

3.4.3 Power-Sensitive Design

After establishing the concept of empowerment in Section 2.5 and outlining its measurement at both individual and collective levels, we now align this understanding with the design of a system aimed at empowering individuals and communities. This becomes increasingly critical since the hybridisation of our societies (leading to the next generation of socio-technical systems) carries the risk of introducing novel forms of asymmetrical power distribution, which may lead to community disempowerment. Such disempowerment can stem from several sources: unequal power distribution within the socio-technical system as a legacy of previous social structures [44]; the intensification of existing inequalities through the application of AI as a proxy [139]; and the widespread adoption of AI technologies, which may inadvertently result in the surrender of human agency [140].

In human-AI systems, power may be concentrated within humans, AI, or shared between the two, thereby creating new forms of asymmetrical power relations. One type involves the distribution of power among human agents, a pattern well-documented in existing literature [90]. However, this asymmetry is evolving into a new paradigm where humans leverage AI to consolidate the power of others.

In this emerging paradigm, a segment of the human population, possessing control over technology or the ability to restrict access to it, diminishes the collective capacity for empowerment [22]. By effectively controlling AI, this group gains access to data produced by others and dominates the flow of information. Furthermore, by limiting access to AI, they can manipulate the dissemination of false information and obscure inconvenient truths. In both cases, power becomes concentrated within this subset, enabling them to exert disproportionate influence over collective decision-making.

Another form of disempowerment arises from the inadvertent delegation of power to AI [141]. The widespread adoption of effective AI tools has led to a pervasive outsourcing of tasks to AI, which undermines human critical thinking. Consequently, power increasingly shifts towards AI, which gives to AI a regulatory role within the socio-technical system.

Thus, there is a need to specify a design methodology that allows humans to coexist with digital agents in a manner that enhances human empowerment. We propose PSD [23], a variant of Value-Sensitive Design [142], wherein the qualitative value to be optimised is (the equitable distribution of) power. This specification is being translated into functional requirements for PSD, addressing two key objectives:

- maximising the power of naturally intelligent agents (i.e. humans) while restricting the power of artificially intelligent agents (i.e. AI) to control only processes that support human empowerment, and
- maximising the equitable distribution of power among humans.

To fulfil these requirements, it is essential to understand the capabilities of both Artificial Life (ALife) and Natural Life (NLife) components and to allocate roles accordingly. Starting with ALife, it excels in leveraging data and existing knowledge for predictive modelling, analysing high-dimensional data, and identifying complex patterns within n-dimensional spaces, as well as providing sophisticated visualisations. In contrast, NLife, represented by humans, possesses innate abilities such as applying common sense reasoning, understanding broad contextual information, rapidly switching between tasks, and adapting to unfamiliar or unexpected situations. Moreover, humans are inherently creative, empathetic, and able to incorporate social, ethical, and moral considerations when making complex ethical judgements. Furthermore, humans exhibit transformational creativity, a capability that can be harnessed to effect meaningful and lasting changes within the NA-Life system [143].

Given these distinct strengths, the role of ALife is to support decision-making by offering insights into past, present, and future scenarios, while the role of NLife is to integrate these insights with lived experience and the collective knowledge within social networks in order to make informed and responsible decisions. ALife's predictive abilities can be utilised to safely explore alternative solutions, project their potential trajectories, and select the most favourable options. This form of practical reasoning [144], when combined with human expertise, enhances decision-making within the NA-Life system (i.e. system having artificially and naturally intelligent components). Additionally, this capacity can be used to test counterfactuals—examining what could have happened under different choices in the past—and thus improve decision-making for the future.

ALife's effectiveness in identifying patterns across multiple dimensions and producing descriptive graphs, text, and audio can also provide NLife with valuable insights into the present. This enables NLife to perceive reality through different lenses, envision alternative perspectives, and identify power asymmetries within the NA-Life system. Ultimately, NLife can utilise the capabilities of ALife to learn from past experiences, understand current conditions, and take early action to avoid unintended outcomes.

Finally, the conclusive decision must rest with NA-Life for two key reasons. First, NLife must

remain at the centre of the NA-Life system, maintaining control over the “state of affairs” that directly affects them. Second, humans possess unique qualities such as common sense reasoning, lived and embodied knowledge, expertise, contextual understanding, the ability to synthesise diverse concepts, and the capacity for moral judgment.

However, we argue that when we develop systems that involve humans [48], it is not sufficient only to design them according to some standards, but we also need to ensure that those standards (i.e. values) would be maintained during the system’s operation [37]. In other words, apart from having a methodology for designing systems to empower humans (i.e. PSD), we also need a mechanism for operationalising this approach, which we refer to as Power Sensitive Operationalisation (PSO). Next, we introduce SGRL which corresponds to a framework for implementing systems that achieve innovation through the co-production between humans and machines. This, also, corresponds to a mechanism for achieving PSO through the support of innovation for empowerment.

3.4.4 Socially-Guided Reasoning and Learning (SGRL)

We contend that hybrid societies, comprising humans and AIs, necessitate tools that support innovation aiming for human empowerment. In this context, innovation demands the ability to anticipate the unexpected, generate novel ideas and processes (corresponding to SAs) for adapting to new conditions, and the capability to imagine the effects of the application of those novelties to the socio-technical system. We propose to engage in an SGML approach and enhance it with reasoning and learning capabilities, to create a safe space for exploring plausible solutions and an evidence-base for evaluating those solutions in short-term and long-term. As such, we include in the agent-based simulator, corresponding to a MAS, presented in Figure 3.3, a GenAI API, and human users(s). Those three ‘actors’ (i.e. Multi-Agent System (MAS), GenAI, humans) communicate through an interactive interface as described in the previous Section and illustrated in Figure 3.4.

Specifically, we introduce the SGRL as a methodology to support the generation and the evaluation of ideas, corresponding to new ways of solving problems in human-AI societies [24]. To achieve that, we synthesise distilled human knowledge and expertise with the combinatorial capabilities of GenAI and the envisioning capacity of Socially-Inspired MASs. This is because GenAI is capable of connecting information coming from diverse sources in unexpected ways that can boost human creativity and capacity to conceive new ways to solve problems and shape knowledge [145]. Also, MAS effectively demonstrates, often unexpected, emergent behaviours and phenomena [146]

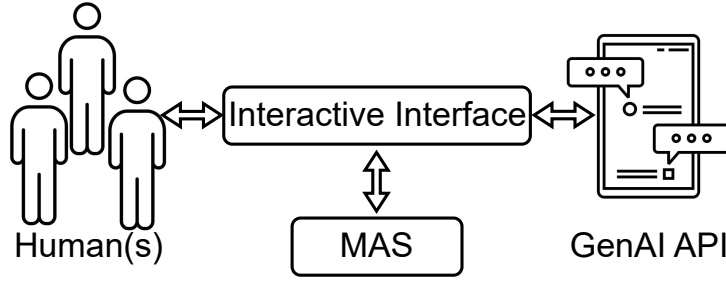


Figure 3.4: SGRL system comprising human users, a multi-agent simulator, a GenAI agents, and an interactive interface

that can empower humans to envision future trajectories of alternative solutions. The next Section presents the overview of the SGRL architecture and provides guidance on the implementation. Note that in Chapter 6 follows a more elaborate discussion on the implementation of the SGRL approach.

3.4.5 SGRL Architectural Overview and Notes on Implementation

We present the overview of the architecture of the SGRL that extends HyRL methodology to achieve innovation in order to support human empowerment during the transition to the digital society. Specifically, recognising that innovation involves generating novel ideas (i.e. policies linked to Social Arrangements (SAs)) and assessing these ideas based on their short- and long-term impacts on various stakeholders as well as the wider community, we propose a collaborative approach combining human and AI strengths, as presented in Figure 3.4.

The overall system comprises a SIC (multi)-ABMS system, a GenAI component, and (a) human user(s). To support this methodology the definition of the agent corresponds to the following:

$$i = \langle attr, SN, LC, cd \rangle \quad (3.2)$$

where *attr* is a set of attributes that describes agent's coefficients, values and weights, *SN* is the network that describes the social connections of the agent, *LC* is the set of rules that have been given to the agent, and *cd* correspond to the agent's cognitive dimensions of empowerment. Also, the environment is extended with the collective cognitive dimensions *CD* of empowerment and the community health *CH*. Thus, the implementation of the system requires performing the operations described in the control flow in Algorithm 5. This involves operations of the control flow presented in Algorithm 4 which is based on the HyRL algorithm presented in Section 3.3.

Additionally, this requires installing some additional libraries to support the human-computer interaction (i.e. `keyboard` library), and to integrate the GenAI API (i.e. `openai` library). Note that the enhancements aiming to support the SGRL methodology are marked in red in Algorithm 4.

This methodology begins with system initialisation (i.e. initialisation of the MAS). Once initialisation is complete, the MAS is run for a specified number of rounds (following the steps described in Algorithm 4), after which results from the first iteration are presented to the user. Then, the user draws on their knowledge, experience, and intuition to decide if current conditions are satisfactory or if they wish to introduce a new SA. If the user sees no need for changes, the existing policy is retained, and the system progresses to the *application* phase. Otherwise, the user decides whether they are confident or curious enough to propose a new SA themselves, or if they would prefer assistance from GenAI. In the former case, the user submits the new SA; in the latter, the GenAI API is queried for an SA, utilising both general internet-sourced knowledge and environment-specific insights from previous framework iterations.

Algorithm 4: Indicative Control Flow of MAS in SGRL - based on control flow in Section 3.3

```

1:  $t = 0$ ;
2: Set  $T = \text{max epochs}$ ;
3: Initialise  $SA$ ;
4: Initialise  $N$  agents:  $\forall i \in N. \text{initialise}(sc_i, c_i, cd_i)$ ;
5: Initialise  $R$  resources; SGRL enhancement
6: Initialise  $CD$  collective dimensions of individual empowerment; SGRL enhancement
7: Initialise  $CH$  community health properties; SGRL enhancement
8: Initialise  $Q$  table:  $\forall Q_{a,b} \in \{\text{state}, \text{action}\}. \text{set} Q_{a,b} = 0$ ; HyRL enhancement
9: Generate Social Network  $SN$ ;
10: while  $t < T$  do
11:   Apply  $SA$ ;
12:   for  $i \in N$  do
13:     Calculate Individual Cognitive Dimensions  $cd_i$ ; SGRL enhancement
14:   end for
15:   Aggregate Individual Cognitive Dimensions;
16:   Form Individual Empowerment  $CD = \sum_{i \in N} cd_i$ ; SGRL enhancement
17:   Calculate Community Health  $CH$ ; SGRL enhancement
18:   for  $i \in N$  do
19:     Update of Individual Parameters;
20:     Update Q-Values;
21:   end for
22:   Update Q-values and Adjust Behaviour;
23:   Store Systemic Parameters;
24:    $t = t + 1$ ;
25: end while

```

Following the SA selection, either by keeping the current arrangement or proposing a new one through user or GenAI input, the system proceeds in the iterative application of that SA in the MAS (according to Algorithm 4). The results of the application corresponding to metrics capturing the transition dynamics as well as the final simulator state are transformed into user-friendly graphs. Those results could be also enhanced with advanced tools to present information

Algorithm 5: SGRL Control Flow

```

1: reset = false.
2: for m rounds do
3:   Run Algorithm 4.
4: end for
5: Get results from Algorithm 4 for visualisation.
6: Show visualisation to the user.
7: while not(reset) do
8:   System asks user whether they want to change SA.
9:   if User decides to give a new SA then
10:    User gives the SA.
11:    The new SA is given to the system and it is stored.
12:   else if User decides to use GenAI then
13:    System calls GenAI API to ask for a new SA.
14:    The new SA is given to the system and it is stored.
15:   else
16:    Existing situation in MAS is maintained.
17:   end if
18:   for m rounds do
19:    Run Steps 10 to 25 of Algorithm 4. // The MAS should not be initialised again
20:   end for
21:   Get results from Algorithm 4 for visualisation.
22:   Show visualisation to the user.
23:   reset = consult(user, bootstrap)
24: end while

```

visually, textually, audibly, or even via augmented reality, empowering the user to explore SA effects without experiencing them directly (similar to people preferring certain political systems without living under them [90]).

Next, the user evaluates the SA's effectiveness, providing feedback on whether it was beneficial, neutral, or detrimental, and decides whether to continue the simulation or re-initialise the system. If the user chooses to proceed, the system returns to the change query where the user assesses if further change is needed. Alternatively, the system may be re-initialised, with data from the current iteration feeding into the GenAI, enabling it to suggest improved or more contextually relevant policies (i.e. SAs) in the future.

In either case, data from the final iteration is stored, enabling future comparisons and aiding decision-making. This distinction between system termination (closing the program) and re-initialisation offers two primary advantages. Firstly, it allows both the system and user to revisit past choices, facilitating more informed decisions and possibly avoiding suboptimal, seemingly stable SAs. Secondly, it enables recovery from unfavourable situations and the chance to restart while maintaining awareness of the choices that led to previous states. Overall, an overview of the which SGRL architecture is given in 3.5.

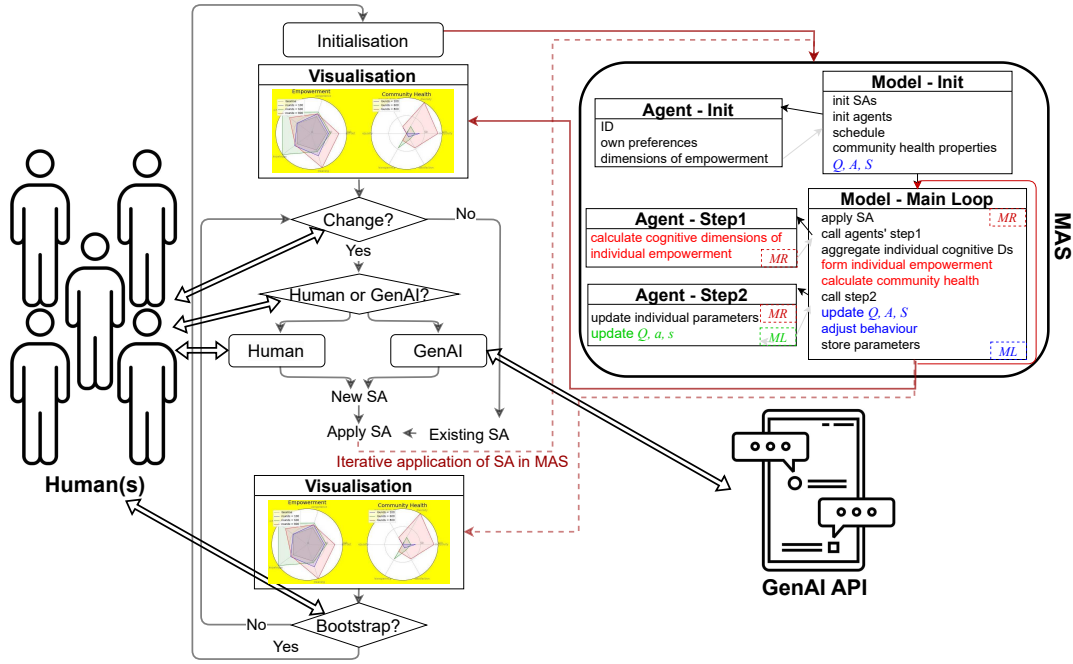


Figure 3.5: SGRL for PSO Architectural Overview

3.5 Summary

Chapter 3 outlined the evolution of the research methodology of this thesis, comprising three main refinements. The first instance corresponds to a methodology for designing and simulating socially-inspired socio-technical systems with MR capabilities, referred as SIC-MR-ABMS. The second draws inspiration from neurosymbolic AI, and investigates a hybrid approach for designing socially-inspired socio-technical systems which combines MR and ML techniques, referred as HyRL. This is followed by the last advancement, which extends the HyRL framework, to include human user(s) (following the SGML approach), and a GenAI API. The combination of those within a unified tool corresponds to the novel methodology for human-machine innovation and PSO, referred as SGRL.

Consequently, the methodological contributions of this thesis are the following, while the enhancement of the methodology is illustrated in Figure 3.6:

- a mechanism for simulating social systems having *reasoning* capabilities, which allows developing a better understanding of human societies and using social processes of decision-making to engineer cyber-physical systems that effectively perform their computational tasks,
- a process for extending the design of a socially-inspired *reasoning* system to include also

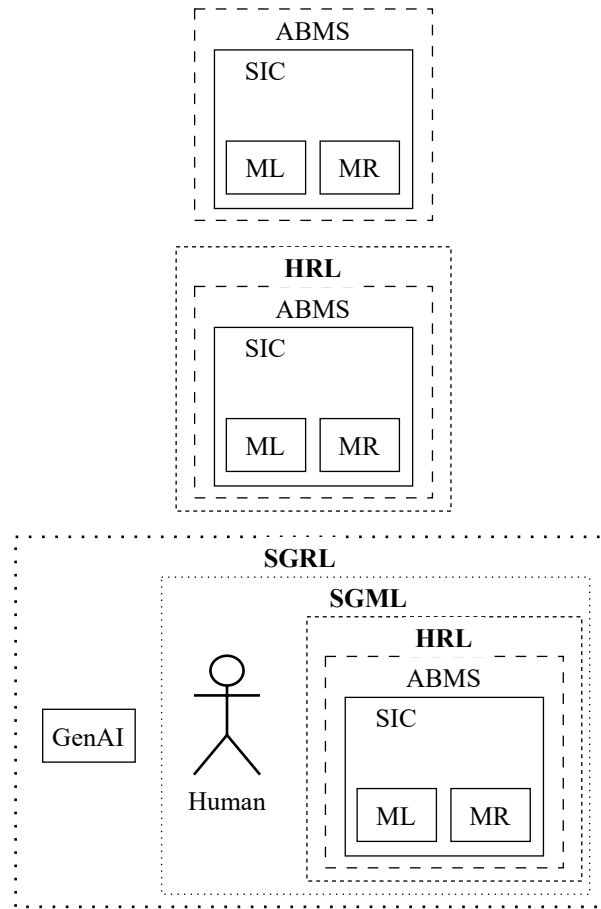


Figure 3.6: The incremental refinement of the research methodology

learning capabilities which allows for developing cyber-physical systems competent to adapt to complex environments while also it can support human decision-making, and

- a framework for interleaving human domain knowledge and lived expertise with GenAI and MAS to equip human mind with imagination to conceive alternative solutions, capacity to envision the future trajectories of those novel solutions, and capability to evaluate them and select the ones that would support human flourishing through (human-machine) *innovation*.

Those three methodological contributions are applied as following:

- Chapter 4 uses the SIC-MR-ABMS methodology to design socially inspired computational models of systems that encompass reasoning capabilities achieved through rule-based operations.

- Chapter 5 uses the HyRL methodology to equip agents with reasoning and learning capabilities achieved through the use of MR and ML algorithms.
- Chapter 6 uses the SGRL methodology to design systems capable of achieving innovation through the co-production between humans, GenAI, and MAS.

Specifically, Chapter 4 introduces a framework for using social influence to develop explanatory adequacy in socio-technical systems.

Chapter 4

Explanatory Adequacy and Reflection

Contents

4.1 Introduction	60
4.2 Motivation	61
4.3 Problem Specification	62
4.3.1 The Problem of Developing Explanatory Adequacy	62
4.3.2 Evaluation Criteria	65
4.3.3 Experimental Conditions	66
4.4 Formal Specification	68
4.4.1 The Environment	68
4.4.2 The Agents	68
4.5 The <i>RTSI+</i> Algorithm for Explanatory Adequacy	69
4.6 Experimental Setup	70
4.7 Experimental Hypothesis and Results	71
4.7.1 Experiment 1: Static and dynamic populations	72
4.7.2 Experiment 2: Dynamic population, progressive addition, non-persistence	73
4.7.3 Experiment 3: Dynamic population, progressive addition, persistence	75
4.7.4 Summary of Results	77
4.8 Conclusions and Discussion on Explanatory Adequacy	78
4.9 From Explanatory Adequacy to Reflection	80
4.10 Metrics for Reflection	81
4.10.1 Problem Specification	81
4.10.2 The Metrics	82
4.10.3 Risks	87
4.10.4 Conclusions on Metrics for Reflection	90
4.11 Summary & Discussion	90

4.1 Introduction

Chapter 2 showed how social influence, as described in Regulatory Theory of Social Influence (RTSI) [12], can be used as a mechanism to improve the information processing of a distributed group of autonomous agents with respect to the value of a signal (either corresponding to a quantitative value or a qualitative assessment).

4

Chapter 4 uses the algorithm presented in Section 2.3 as the basis to develop a new algorithm, called *RTSI+*, which extends the one presented by:

- changing the problem: instead of agreeing on a signal, the group of individuals, operating as a Distributed Information Processing (DIP) unit, has to agree on the processing rules that generate that signal, and
- adding a preliminary form of reflection: the individuals (i.e. computational agents) have to reflect on their environment, and modify their knowledge to match the *ground truth knowledge*.

In particular, Section 4.3 introduces this meta-problem of identifying the processing rules generating a signal, the evaluation criteria based on which we are going to evaluate the explanatory adequacy of the Distributed Information Processing (DIP) as well as the initial conditions in the DIP we are interested in exploring, Section 4.4 specifies the environment and the agents, and Section 4.5 presents the novel *RTSI+* algorithm, which uses Machine Reasoning (MR) techniques, for enabling the group to develop explanatory adequacy (i.e. identify the rules producing a signal). This is followed by the analysis of the experimental design in Section 4.6 and the discussion of the experimental hypothesis and results in Section 4.7. After highlighting the contributions of this work, this Chapter discusses the need for reflection.

In particular, Section 4.10 investigates how RTSI can be used to enable a Multi-Agent System (MAS) operating as a DIP to not only *react* to the environment but also *reflect* on it and modify itself. Hence, this Section offers a set of metrics for evaluating the performance of a DIP with respect to the main dimensions of interest which correspond to systemic drivers. Finally, this concludes with a discussion of the challenges in the definition and application of metrics and demonstrates some examples of their misuse.

4.2 Motivation

A common challenge for an organisation of autonomous agents, connected via a social network but without a central authority, is aggregating individual measurements, judgements, votes, opinions, or preferences into a collectively agreed-upon output. This issue arises in both social systems (e.g., jury trials or deliberative assemblies) and cyber-physical systems (e.g., sensor networks or cybernetic systems) and corresponds to a problem of collective decision-making, which was introduced in Chapter 2.

One instance of this problem is forming a collective agreement on the value of a signal. Section 2.3 analysed how social influence, as expressed by RTSI, can be used as a mechanism for enabling a group to effectively track that value (i.e. a quantitative measurement or a qualitative assessment). Another instance of this problem is to explain an observed phenomenon (i.e. or a signal) by determining the processing rules that produce it. This is referred as a problem of developing explanatory adequacy.

This corresponds to a meta-problem in which rather than producing a collective decision through consensus or voting, the group has to combine its diverse knowledge to formulate a plausible explanation for the causes (corresponding to a set of processing rules) of a signal. Additionally, this requires the agents to perform a preliminary form of *reflection* on the environment in order to adapt to the knowledge of the others and modify their own processing rules to match the *ground truth knowledge*.

Here, we address this meta-problem by proposing a new mechanism for knowledge aggregation based on RTSI. Specifically, this shows how social influence can facilitate distributed information processing and enable a group to combine partial knowledge to form an accurate explanation. This, also, corresponds to a process that leverages social influence to address issues of self-governance (i.e. knowledge management, fairness and justice), which were discussed in Section 2.2.4.

Going back to the example of the co-housing project that was presented in Section 1.2, we can imagine the group of individuals and devices that are situated in this shared-housing project to be subject to a distribution of electricity. The policy of distribution is decided by a dedicated agent, who does not disclose the information regarding it. In that case, agents (humans or computational components) might use social influence to explain the policy applied to them, by identifying the information processing rules comprising it. As such, individuals might try to distinguish from their social network the agents having more knowledge regarding the policy applied to them, and

exchange information processing rules in order to be able to explain the policy of resource distribution that is applied to them. This way, the group uses the partial knowledge of the individuals to develop explanatory adequacy (i.e. be able to identify the rules that cause the distribution of electrical energy that they observe).

4

4.3 Problem Specification

This Section introduces the problem of developing explanatory adequacy, the evaluation criteria and a set of different initial conditions in the DIP that we are going to explore in our experiments.

4.3.1 The Problem of Developing Explanatory Adequacy

This work looks into the problem of explanatory adequacy [10]. In particular, it investigates how a group of autonomous, interconnected agents having partial knowledge (which is formed by a subset of information processing rules), can collectively produce an accurate explanation for the rules causing a commonly observed output. To achieve that, they can use social influence, as expressed by RTSI, to exchange information processing rules, corresponding to explanations for the observed signal. In other words, instead of identifying the value of a signal, the group has to converge on the rules that produce that signal by iteratively exchanging information processing rules and reflecting on the accuracy of their explanations in previous epochs. Although this work addresses generally the problem of developing explanatory adequacy, we ground this to the specific problem of resource distribution in a CPR in order to evaluate and measure the effectiveness of the proposed algorithm.

In particular, the group of autonomous agents is situated in an environment in which a process $\mathcal{P}$ produces an output by applying a set of n processing rules to a given input. $\mathcal{P}$ corresponds to a process of resource distribution in CPR, where N agents contribute to and appropriate from a shared resource. $\mathcal{P}$ comprises a set of rules (corresponding to legitimate claims as presented in Table 6.1 in Section 2.3.2) that affect the distribution and each of those rules is associated with a weight (selected from $[0, \dots, 1]$). The weights demonstrate the effect that each rule has to the final resource distribution (e.g. a rule associated with a weight 0.2 affects less the distribution compared to a rule having weight 0.7).

Hence, the input to that process is the contribution of agents, their needs, and their past allocations of resources, while the output is a certain distribution of resources as presented in

Table 4.1. Agents observe the output of $\mathcal{P}$ (e.g. a resource distribution) and try to identify the rules producing this distribution.

Agent	# resources
agent 1	r_1
agent 2	r_2
...	...
agent N	r_N

Table 4.1: Output of $\mathcal{P}$: a specific allocation of resources in a certain round.

The set of rules together with their weights define abstractly the *ground truth knowledge* $\mathcal{K}$ given by:

$$\mathcal{K} = \{(r_i, w_i) \mid i \in [1..n] \wedge \sum_{i=1}^n w_i = 1.0\}$$

while we use the symbol by $\mathcal{K}_r$ to refer to the set of rules in $\mathcal{K}$ (without the weights). For instance, ground truth knowledge might be described by Equation 4.1:

$$\mathcal{K} = w_{lc_1} * lc_1 + w_{lc_2} * lc_2 + w_{lc_3} * lc_3 + w_{lc_4} * lc_4 + w_{lc_5} * lc_5 + w_{lc_6} * lc_6 + w_{lc_7} * lc_7 + w_{lc_8} * lc_8 \quad (4.1)$$

Additionally, the partial knowledge of an agent i is represented by $\mathcal{K}^i$ and corresponds to a subset of m rules of $\mathcal{K}_r$, $m \leq n$. Each agent associates an arbitrary weight with each of its rules, while the sum of the weights in the knowledge of an agent is equal to 1.0. Thus, the knowledge of an agent i is given by:

$$\mathcal{K}^i = \mathcal{K}^{(self)} = \{(r_1, w_1), \dots, (r_m, w_m)\} \quad \text{such that} \quad \forall i, 0 \leq i \leq m. r_i \in \mathcal{K}_r$$

Note, that if $i = 0$, then the agent has an empty knowledge base.

The group of N agents has to self-organise as a DIP unit and collectively agree on a set of rules, each associated with a weight, corresponding to their explanation. This collective explanation reflects the knowledge of the group, denoted by $\mathcal{K}^{DIP}$, and is formed by aggregating the partial knowledge of individuals, and is illustrated in Figure 4.1.

To find the solution to process $\mathcal{P}$ as parametrised by $\mathcal{K}$ they have the following mechanisms:

- ask their neighbouring agents in their social network to share one or multiple processing rules (from their own partial knowledge);

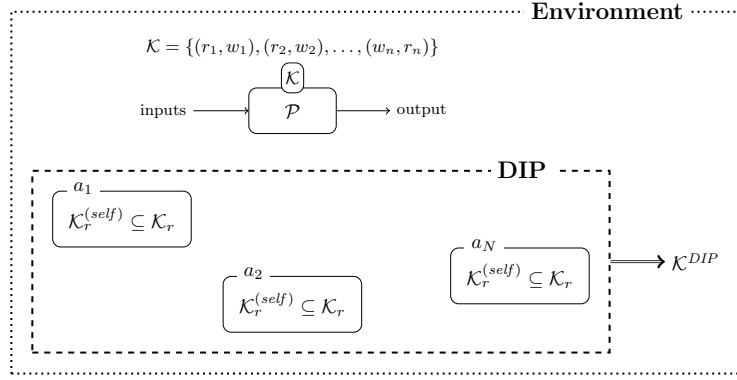


Figure 4.1: The problem of developing explanatory adequacy

- receive feedback from the environment on the quality of the collective knowledge and their own contribution, which is used to update their self-confidence on their rules and the credence to their neighbours' knowledge; and
- discover new knowledge by interacting with new agents joining the system.

Hence, we determine which initial conditions of the DIP produce what different epistemic condition on the individual knowledge bases, i.e. the similarity of $\{K^i \mid i \in N\}$, which we identify as either diversity, incongruence, or stagnation, and what is the explanatory adequacy of K^{DIP} , i.e. the (dis)similarity of K^{DIP} to K in each of those situations.

Accordingly, there exist three types of knowledge:

- K : the *ground truth knowledge*,
- K^{DIP} : the aggregated knowledge of the DIP, and
- K^U : the “knowledge potential” corresponding to the epistemological limit on what it is possible for an agent to know, because this knowledge exists somewhere in the DIP.

This is not a one-shot or static problem, as the composition of the DIP unit is dynamic (i.e. agents may leave or join), and consequently the available knowledge constantly changes. As a result, expertise within the group is temporary, and agents with valuable knowledge or skills may be lost. There are two key perspectives on the collective: from a functional standpoint, the DIP acts as a unit that pools knowledge and identifies expertise to achieve a shared objective (cf. [147]); from a societal perspective, the group leverages social influence to shift attitudes about a common issue (cf. [68]). Consequently, the terms ‘experts’ or ‘specialists’ in the DIP and the sources of

influence in RTSI are used interchangeably, with knowledge being equated to the processing rules. Similarly, the terms DIP, community, and collective are treated as synonymous.

4.3.2 Evaluation Criteria

Furthermore, to investigate the problem of developing explanatory adequacy in a situation in which experts and knowledge are temporal and often not easily accessible, it is necessary to identify *which experimental conditions on the DIP* are sufficient for enabling the group to *develop accurate explanations*. However, to evaluate the impact of the different experimental conditions, we have to measure the following properties of the DIP:

- the epistemic condition in the knowledge bases of the individuals (i.e. similarity between $\mathcal{K}^a | a \in N$)
- the explanatory adequacy of the collective knowledge in the DIP (i.e. similarity between $\mathcal{K}^{DIP}$ and *ground truth knowledge* $\mathcal{K}$).

To assess the *epistemic condition* and *explanatory adequacy* of the DIP under different initial conditions, we need a metric that evaluates diversity across two dimensions:

- Epistemic diversity: This measures how different the agents' knowledge bases are from one another, calculated as $\sum_{i=1}^N \sum_{j=1}^N \mathbf{diff}(i, j)$; and
- Explanatory adequacy: This gauges how much the DIP's knowledge diverges from the *ground truth knowledge*, given by $\mathbf{diff}(\mathcal{K}^{DIP}, \mathcal{K})$.

Various metrics can measure diversity, such as Euclidean Distance, Manhattan Distance, or KL divergence, but cosine similarity is used here because it focuses on the orientation rather than the magnitude of the vectors representing the agents' knowledge and the *ground truth knowledge*. Specifically, cosine similarity measures the cosine of the angle between two non-zero vectors, $\mathbf{A}$ and $\mathbf{B}$ in $\mathbb{R}^n$, and is given by the Equation 4.2:

$$\cos_sim(\mathbf{A}, \mathbf{B}) = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}} \quad (4.2)$$

After establishing the metric to evaluate both collective and individual performance, two groups are defined: participants, which include all agents selected to participate in the next epoch, and

specialists, a subset of participants who serve as sources of processing rules. In this context, if agent i asks for processing rules agent j , and j asks agent k , then specialist is considered the agent k which constitute the actual source of influence. These agents don't necessarily possess expertise but are valued by their network for their contributions.

Initially, all agents give equal credence to their network peers regarding processing rules. Each agent starts with different knowledge, and as they exchange information, the credibility others assign to them is updated based on the utility of the knowledge they provide.

4

To evaluate the agents' ability to explain the *ground truth knowledge* (i.e. explanatory adequacy), the cosine similarity between agents' knowledge bases $\mathcal{K}^i$ and the *ground truth knowledge* $\mathcal{K}$ is computed, denoted as $\mathcal{CE}_1$. The same metric is applied to specialists, denoted as $\mathcal{CS}_1$. Additionally, to observe how knowledge distribution and diversity evolve through the exchange of processing rules (epistemic diversity), we calculate the pairwise cosine similarity of participants' knowledge bases, denoted as $\mathcal{CE}_2$, and the same for specialists, denoted as $\mathcal{CS}_2$. The calculation of $\mathcal{CE}_1$ and $\mathcal{CE}_2$ is given by Equations 4.3 and 4.4 respectively, and the calculation of $\mathcal{CS}_1$ and $\mathcal{CS}_2$ can be computed by substituting *participants* with *specialists* on those equations.

$$\mathcal{CE}_1 = \frac{\sum_{i=1}^{participants} \cos_sim(\mathcal{K}^i, \mathcal{K})}{\sum_{i=1}^{participants} 1} \quad (4.3)$$

$$\mathcal{CE}_2 = \frac{\sum_{i=1}^{participants} \sum_{j=1, j \neq i}^{participants} \cos_sim(\mathcal{K}^i, \mathcal{K}^j) - \sum_{i=1}^{participants} 1}{(\sum_{i=1}^{participants} 1)^2 - \sum_{i=1}^{participants} 1} \quad (4.4)$$

4.3.3 Experimental Conditions

We aim to observe under which conditions the DIP manages to develop explanatory adequacy, here we specify a set of different initial conditions in the DIP that we are going to explore in our experiments. Specifically, to explore various scenarios, we introduce two independent experimental variables, $\mathcal{F}$ and $\mathcal{R}$. $\mathcal{F}$ controls the rate of population turnover affecting also the rate at which knowledge changes, while $\mathcal{R}$ imposes a constraint on the *RTSI+* algorithm, affecting how agents share processing rules and influence each other.

The DIP operates over a sequence of T epochs, with agents being added or removed at intervals of $t < T$ epochs (except in the static condition). In each epoch, the DIP generates and evaluates $\mathcal{K}^{DIP}$ against $\mathcal{K}$. Hence, the variable $\mathcal{F}$ can take one of eight possible values:

- *static*: The population consists of N agents who remain active throughout all T epochs. Each agent i is initialised with knowledge $\mathcal{K}^i$ which is a subset of the *ground truth knowledge* $\mathcal{K}$.
- *dynamic*: The population consists of N agents, with a new generation of $\frac{N}{10}$ agents joining and $\frac{N}{10}$ agents leaving every t epochs. Agents are initialised with a subset of the eight processing rules.
- *restart*: The population consists of N agents, with $\frac{N}{10}$ new agents joining and $\frac{N}{10}$ leaving every t epochs. The first generation of agents starts with knowledge limited to the rules $\{(r_1, w_1), (r_2, w_2)\}$. Subsequent generations are initialised with either a new rule or no rule, such as $\{(r_3, 1)\}$ or an empty set. After the initialisation of the generation having knowledge $\{(r_8, 1)\}$ or $\{\}$, each future generation a is initialised with knowledge $\mathcal{K}^a$ being any subset of $\mathcal{K}$.
- *iterate*: The population of N agents evolves every t epochs, with $\frac{N}{10}$ new agents joining and the same number leaving. The first generation starts with $\{(r_1, 1)\}$, the next with $\{(r_2, 1)\}$, and so on. Each new generation only knows the latest rule, and accessing previous knowledge requires interaction with others.
- *add rapid*: The population consists of N agents, and every t epochs $\frac{N}{10}$ new agents join and $\frac{N}{10}$ leave. The 1st generation of N agents is initialised so that each agent's a knowledge is a subset of $\{(r_1, w_1), (r_2, w_2)\}$. The next generation knows what their ancestors knew and a new processing rule, so the knowledge of each new agent a is a subset of the rules $\{(r_1, w_1), (r_2, w_2), (r_3, w_3)\}$, and so on. So, new knowledge is progressively added to the population through the new generations.
- *add slow*: Similar to the *add rapid* scenario, but new generations are introduced every $t * 1.6$ epochs instead of every t epochs, slowing the pace of knowledge accumulation.
- *add rapid/slow long*: New generations are added every $t/t * 1.6$ epochs as per *add rapid/slow*, but the simulator runs for $T * 2$ epochs to observe long-term effects.

The independent experimental parameter $\mathcal{R}$ this can get two values affecting the communication of the agents in the following way:

- *max*: The sources of influence share with their neighbours one processing rule, together with their weight, and the rule corresponds to the one having the greatest weight in their knowledge base.

- *all*: The sources share all their rules and their weight, giving full visibility to their neighbours (e.g. targets) of what they know.

4.4 Formal Specification

4

From the description above we observe that there is a group of agents situated in the environment interacting according to a process based on RTSI. This Section provides a specification of the environment and the agents. Based on this specification, the next Section will introduce a new algorithm that uses RTSI to enable a group develop explanatory adequacy.

4.4.1 The Environment

The environment $\mathcal{E}$ comprises a group of N agents connected over a social network SN which face the requirement to determine the processing rules and weights (i.e. $\{(r_1, w_1), (r_2, w_2), \dots\}$) existing in $\mathcal{K}$ which represents the *ground truth knowledge*. The social network $\mathcal{G}$ consists of N nodes each corresponding to an agent, and is a Klemm-Eguiluz network [73]. We select this network topology because it resembles real-world networks, combining characteristics of both scale-free [148] and small world [149] networks, which include a power-law degree distribution, a high clustering coefficient, and a small average shortest path. The network consists of N nodes (agents), initialised with m_G fully connected nodes, and has a probability μ of adding an edge to one of these fully connected nodes.

4.4.2 The Agents

Each agent i of the group is defined as:

$$i = \langle SN, \mathcal{K}^i, sc_i, cr_i, a, b \rangle$$

where SN_i is the agent's social network (set of connected neighbours), $\mathcal{K}^i$ is its partial knowledge, which comprises a subset of the rules of the *ground truth knowledge* (accompanied by their weights), sc_i is a measure of self-confidence of its knowledge with respect to the knowledge of its neighbours, to whom it assigns credence $cr_{i,j}$ for each agent $j \in SN_i$ (cf. [66]). For each agent, its credence to its neighbours is represented by a ordered list cr_i in descending order, and is updated each time agent i asks a neighbour j for knowledge. Specifically, this reflects how agent i believes its

neighbour j approximates the *ground truth knowledge* $\mathcal{K}$. Each agent has also two reinforcement coefficients a, b which define the rate of change of self-confidence and credence to the neighbours after each iteration of the algorithm, corresponding to an epoch, as described below.

4.5 The *RTSI+* Algorithm for Explanatory Adequacy

The proposed algorithm operates iteratively over T epochs, and in each epoch, the participating agents follow the steps outlined in Algorithm 6. At the start of each epoch, a subset of agents N is randomly selected to participate, with $participants \subseteq N$. The objective of the agents is to approximate (first individually and then collectively) the complete *ground truth knowledge* $\mathcal{K}$ as closely as possible, even though they possess partial knowledge.

During the epochs, each agent seeks out sources (e.g. other agents) within the DIP that can provide the processing rules to achieve the best approximation of the *ground truth knowledge* $\mathcal{K}$. Agents can access the knowledge of the DIP by querying neighbouring agents. Thus, in each epoch, agents iterate through their social network SN_i , following the order of credence cr_i , to identify a neighbour to ask for processing knowledge.

After finding a neighbour j , agent i asks agent j which is the similarity between their individual knowledge and the *ground truth knowledge* (e.g. $\mathcal{S}^j = \cos_sim(\mathcal{K}^j, \mathcal{K})$). At the same time, agent i computes the similarity of their own knowledge ($\mathcal{S}^i = \cos_sim(\mathcal{K}^i, \mathcal{K})$). If the neighbour provides a better approximation of the *ground truth knowledge*, the agent i asks j for their processing rules.

Then there are two alternative flows of the algorithm. If the value of the independent experimental parameter $\mathcal{R}$, which affects the communication of the agents and would be properly introduced below, is *max*, then the agent i either receives the most weighted processing rule from the neighbour j together with its weigh (e.g. the rule being the most important in j 's knowledge). Otherwise, i receives all the set of the processing rules of j together with their weights.

Next, the agent i integrates this new knowledge $\mathcal{K}^j$ into its own knowledge base $\mathcal{K}^i$. Following this, the agent undergoes a reflection process, updating its credence to the selected neighbour cr_{ij} and adjusting its self-confidence sc_i , depending on whether it can now explain the *ground truth knowledge* more accurately than its neighbour.

As such, the group agrees on a knowledge $\mathcal{K}^{DIP}$ which is the outcome of averaging over the

participant's weights of rules, as depicted in 4.5 and 4.6. Moreover, the DIP knowledge is defined as in 4.7.

$$w_{r_i avg} = \frac{\sum_{j=1}^{participants} w_{r_i j}}{\sum_{j=1}^{participants} j} \quad (4.5)$$

$$w_{r_i c} = \frac{w_{r_i avg}}{\sum_{i=1}^n w_{r_i avg}} \quad (4.6)$$

$$\mathcal{K}^{DIP} = \{(r_1, w_{r_1 c}), \dots, (r_r, w_{r_r c})\} \quad (4.7)$$

4

Algorithm 6: RTSI+ for knowledge seeking

```

t = 0;
Set T = max epochs;
Initialise N agents:  $\forall i \in N. initialise(sc_i, c_i, \mathcal{K}^i)$ ;
Generate Social Network SN;
Specify Ground Truth Knowledge  $\mathcal{K}$ ;
Specify Experimental Parameters  $\mathcal{F}, \mathcal{R}, a, b$ ;
while t < T do
    Randomly Select participants  $\subseteq N$ ;
    for i in participants do
        j = selected neighbour from network;
         $S^i = cos\_sim(\mathcal{K}^i, \mathcal{K})$ ; // measure similarity between individual knowledge
        and ground truth
         $S^j = cos\_sim(\mathcal{K}^j, \mathcal{K})$ ; // measure similarity between selected neighbour's
        knowledge and ground truth
        if  $S^i < S^j$  then
            else
                If neighbour's knowledge greater than own, enhance individual knowledge,
                otherwise keep as is
            end
        if  $\mathcal{R} = max$  then
             $\mathcal{K}^i = \{(r_x, w_{r_x}) | (r_x, w_{r_x}) \in \mathcal{K}^j \wedge \neg \exists (r_y, w_{r_y}) \in \mathcal{K}^j. w_{r_y} > w_{r_x}\}$ ; // integrate
            neighbour's rule with greater weight
        else
             $\mathcal{K}^i = \mathcal{K}^j$ ; // include neighbour's knowledge base in own knowledge
        end
        Maintain own knowledge  $\mathcal{K}^i$ ;
        if  $S^i > S^j$  then
             $sc_i = sc_i + a * (1 - sc_i)$ ; // increase self-confidence
             $cr_{i,j} = cr_{i,j} - b * cr_{i,j}$ ; // decrease credence to neighbour
        end
        if  $S^{(self)} < S^{\cup}$  then
             $sc_i = sc_i - b * sc_i$ ; // decrease self-confidence
             $cr_{i,j} = cr_{i,j} + a * (1 - cr_{i,j})$ ; // increase credence to neighbour
        end
    end
    t = t + 1;
end

```

4.6 Experimental Setup

Having provided the formal specification of the system and defined the initial conditions and the metrics to evaluate their impact in the model, the next step is to develop a multi-agent simulator to run experiments on. This was developed in Python3, building upon the system introduced in [9]

by adding the exchange of processing rules (the system could initially support only the exchange of opinions).

In particular, we follow an Agent-Based Modelling and Simulation (ABMS) approach and we design a multi-agent simulator to explore the relation between initial conditions, epistemic condition in the DIP, and explanatory adequacy. This is designed based on an Socially-Inspired Computing (SIC) methodology that combines a theory of social influence (i.e. RTSI) with MR techniques. This is achieved following the methodology described in Section 3.2. This approach enables us to observe emerging behaviours that can be useful for understanding better social interaction but also for improving the operation of cyber-physical systems.

Furthermore, Table 4.2 outlines the simulator's parameters for the agents and the *RTSI+* algorithm, specifying either fixed values used in the experiments (e.g., the number of agents, reinforcement coefficients) or a range of values for those assigned randomly (e.g., individual agent knowledge bases). Since the parameter space is too large to explore all the combinations of parameters values, we fix some of them to produce stable behaviours, and modify some to investigate if some specific behaviours (such as explanatory adequacy, or systemic stability) emerge.

Symbol	Description: Factor of Agent i	Initial Condition/Range
N	network of agents	100
m	total number of edges	$\frac{N}{10}$
μ	number of edges to 'active' agents	0.75
$participants$	agents participating in the next epoch	$\frac{N}{2}$
$\mathcal{K}^i$	individual knowledge base	$\{(r_1, w_1), \dots, (r_m, w_m)\}, \forall i, 0 \leq i \leq m, r_i \in \mathcal{K}_r$
r_k	processing rule k	$k = \{1, 2, \dots, r\}$
w_{r_k}	weight of rule r_k	$0 \leq w_{r_k} \leq 1$
sc_i	self-confidence for similarity of knowledge	0.5
a, b	self-confidence & credence reinforcement coefficients	0.1, 0.1
$\mathcal{G}_j$	social network	1 to N agents
CR_i	ordered list of credence to social network	list length from 1 to $ SN_i $
$cr_{i,j}$	credence to agent j	$0 \leq cr_{i,j} \leq 1$
$\mathcal{S}^i$	cos_sim between self and <i>ground truth knowledge</i>	$0 \leq \mathcal{S}^i \leq 1$
$\mathcal{S}^j$	cos_sim between knowledge of agent (neighbour) asked and <i>ground truth knowledge</i>	$0 \leq \mathcal{S}^j \leq 1$
r_j	rule proposal of (neighbour) agent asked	$r_k \in \mathcal{K}_r$
w_j	weight proposal of (neighbour) agent asked	$0 \leq w_{\cup} \leq 1$
$\mathcal{K}^j$	knowledge proposal of (neighbour) agent asked	$\{(r_1, w_1), \dots, (r_m, w_m)\}, \forall i, 0 \leq i \leq m, r_i \in \mathcal{K}_r$
$\mathcal{K}^{DIP}$	collective knowledge	$\{(r_1, w_1), \dots, (r_m, w_m)\}, \forall i, 0 \leq i \leq m, r_i \in \mathcal{K}_r$
$w_{r_i avg}$	average weight of rule r_i	$0 \leq w_{r_i avg} \leq \frac{N}{2}$
$w_{r_i c}$	normalised average weight of rule r_i	$0 \leq w_{r_i c} \leq 1$

Table 4.2: Simulator Parameters and Variables

4.7 Experimental Hypothesis and Results

This Section outlines three experiments aimed at exploring how different initial conditions in the DIP affect the resulting epistemic state and how well that state explains the *ground truth knowledge*.

The experiments vary the variables $\mathcal{F}$ and $\mathcal{R}$ under the initial conditions specified in Table 4.2, with $T = 5000, t = 300$:

- Experiment 1: A static and a dynamic population, both having complete, fixed knowledge (i.e. all knowledge is available from the first epoch of the simulation).
- Experiment 2: A dynamic population where knowledge does not persist, and new generation have either heterogeneous knowledge (restart) or homogenous (iterate).
- Experiment 3: A dynamic population where new knowledge is progressively added (with different rates), and previously discovered knowledge is retained.

For the above conditions, our hypotheses are the following:

- Experiment 1: Our hypothesis is that static populations would stagnate because of the lack of epistemic diversity, and the lack of change in the population dynamics, while dynamic populations, although epistemically incongruent, could conditionally develop explanatory adequacy.
- Experiment 2: Our hypothesis is that the heterogeneity of knowledge in the population, depending on the conditions, would either result in stagnation and, consequently, explanatory inadequacy, or lead to epistemic diversity and explanatory adequacy.
- Experiment 3: Our hypothesis is that the progressive addition of knowledge would result in ‘controlled’ epistemic diversity, leading to high explanatory adequacy.

We execute these experiments to explore the effects of the different initial conditions to the epistemic condition in DIP, as well as its explanatory adequacy.

4.7.1 Experiment 1: Static and dynamic populations

In this set of experiments, we compare the system dynamics when $\mathcal{F}$ is static and dynamic, and agents initialised with various combinations of available processing rules. Figure 4.2 shows the evolution of knowledge within the agents and specialists under different conditions. The first column displays the results for the static case, while the second shows the dynamic case, and the rows correspond to different means of communication (first row corresponding to sharing only one rule, and second to the whole knowledge base). The black line represents $\mathcal{CE}_1$, indicating how well

the DIP tracks the *ground truth knowledge*, while the red line, based on $\mathcal{CS}_1$, shows how well the specialists track the *ground truth knowledge* $\mathcal{K}$. The green $\mathcal{CE}_2$, and blue $\mathcal{CS}_2$, lines reflect the diversity of knowledge within the collective and specialists, respectively, indicating the degree of similarity or dissimilarity between the knowledge bases.

In the static condition with $\mathcal{R} = \text{max}$, agents' knowledge tends to converge, as the group is primarily influenced by specialists promoting a single processing rule. In contrast, when $\mathcal{R} = \text{all}$, where sources share all their knowledge, the similarity within the community and among specialists decreases. However, both cases show low accuracy in tracking the *ground truth knowledge* (black and red lines), as the static population consistently relies on the sources credited early in the simulation, regardless of whether those sources retain relevant knowledge. *This results in stability but also stagnation, with agents not significantly improving their processing capacity despite having access to all the knowledge. This indicates that although there exists congruence, the group does not interact and therefore they do not manage to develop explanatory adequacy.*

In the dynamic condition, both the *max* and *all* settings show improvements in how agents explain the environment, corresponding to the *ground truth knowledge*. In the *max* case, the fluctuating green and blue lines suggest continuous changes in both specialist and community knowledge, indicating that the system generates different epistemes and shifts from one temporary equilibrium to another, with varying control variables (cf. [62]). In the *all* case, specialists are well-aligned with the *ground truth knowledge* (i.e. high explanatory adequacy), but other agents struggle to assimilate this information, leading to incongruence in the DIP's collective knowledge. *Although the knowledge of the group is incongruent, they manage to produce an accurate explanation for the observed phenomenon.*

4.7.2 Experiment 2: Dynamic population, progressive addition, non-persistence

In this experiment, we observe how the system operates with a dynamic population, where specialists are harder to identify. New knowledge is progressively introduced by incoming agents, but it is non-persistent, meaning new agents only bring newly “discovered” knowledge while existing knowledge might not persist if not shared (because agents having it might leave).

Figure 5.10 presents the results for the *restart* condition in the first column and *iterate* in the second. In the *restart* scenario, the first generation starts with two rules, with subsequent

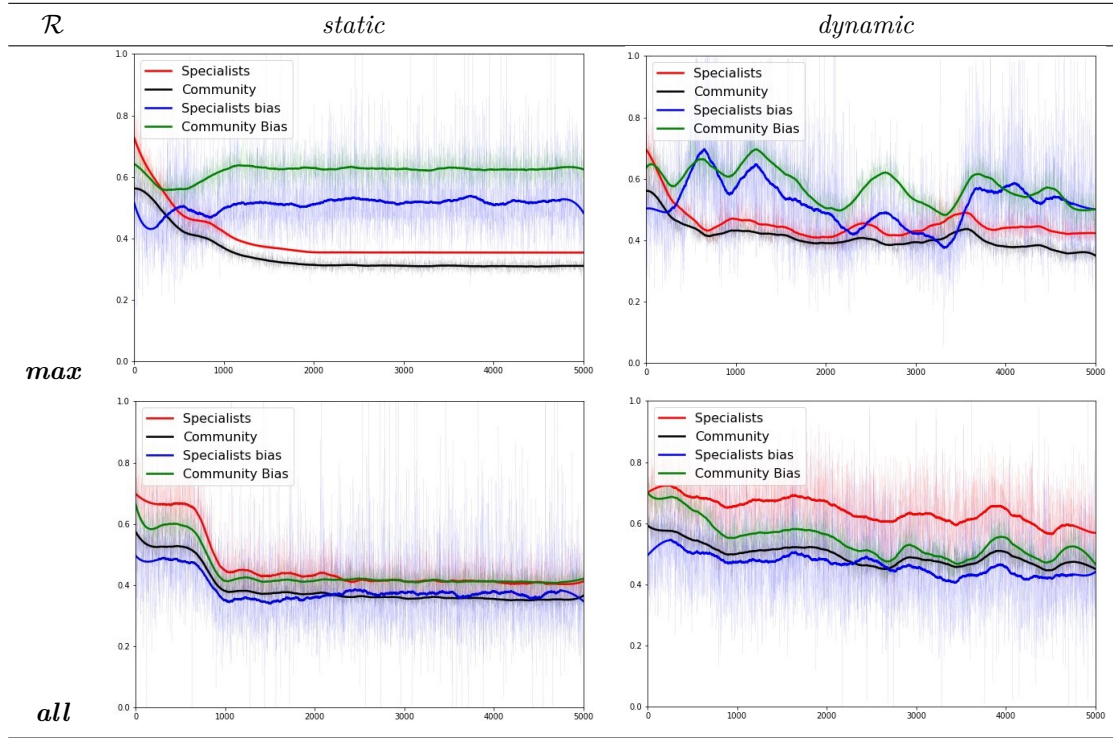


Figure 4.2: Experiment 1: Knowledge dynamics for static and dynamic population, showing that static populations are congruent but stagnate and remain explanatory inadequate, while dynamic populations can develop explanatory adequacy.

generations introducing one additional rule until the 8th generation has access to all rules. When $R = \text{max}$, there is high similarity in processing rules across the community, as agents are given just one rule from the sources. This effect is less pronounced when $R = \text{all}$, where agents quickly assimilate new knowledge. When new generations have only one processing rule, agents perceive little opportunity to learn from others, leading to limited knowledge (low similarity with the *ground truth knowledge*). However, after epoch 2100, when *all processing rules become available to new generations*, there is a notable *increase in knowledge* for both the community and specialists, as agents with varying levels of knowledge seek out missing information from others.

The *iterate* condition makes it even more evident that agents are not requesting processing rules because they perceive others as having similar knowledge. In both $R = \text{max}$ and $R = \text{all}$, most agents exhibit similar levels of knowledge (i.e. similar alignment between their own processing rules and rules in the *ground truth knowledge*), as they only possess either zero or one processing rule. As a result, only agents with an empty knowledge base actively seek knowledge. This discrepancy between agents with no knowledge (no processing rules) and those with one processing rule explains the differences in the similarity of knowledge between sources and the community (as indicated by the blue and green lines). Additionally, fluctuations in the collective and specialists' knowledge

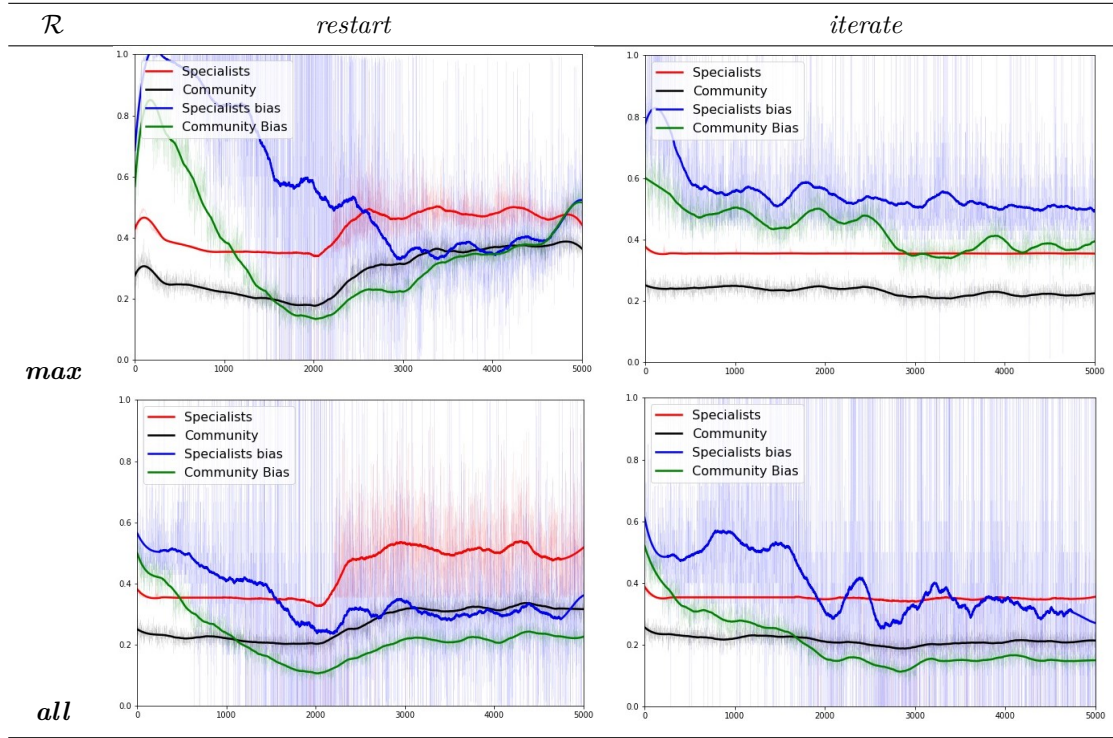


Figure 4.3: Experiment 2: Knowledge dynamics for DIP with non-persistent knowledge showing that there is a potential for the DIP to develop explanatory adequacy as long as the knowledge is not lost permanently

are caused by the randomness in selecting agents to join or leave the network (e.g. depending on whether agents with one or no rules are removed, the explanatory adequacy of the DIP is increased or decreased).

4.7.3 Experiment 3: Dynamic population, progressive addition, persistence

In this experiment, we examine the dynamics of the system under progressive knowledge addition, where new agents may introduce any previously discovered knowledge.

Figure 4.4 shows the results for $\mathcal{CE}_1$, $\mathcal{CS}_1$, $\mathcal{CE}_2$, and $\mathcal{CS}_2$ in the *add rapid/slow (long)* scenarios, with $\mathcal{R}$ set to *max* in the first row and *all* in the second. In the *add rapid* and *add rapid long* conditions, where knowledge is added quickly, there is only minor improvement in both the short and long term. For both *max* and *all*, the specialists' and community's knowledge remains low (represented by the red and black lines) because new rules are not fully assimilated.

In contrast, the slow addition of knowledge (*add slow* and *add slow long*) leads to greater improvement in terms of explanatory adequacy. Across all scenarios, when $\mathcal{R} = \text{max}$, different

epistemes emerge, as sources share portions of their knowledge, causing both the community and specialists to develop varying beliefs over time, and being epistemically incoherent. This observation is aligned with Foucault’s claims in his book ‘The Order of Things’ [150], about *epistemes*, which correspond to different sets of knowledge that prevail over certain periods.

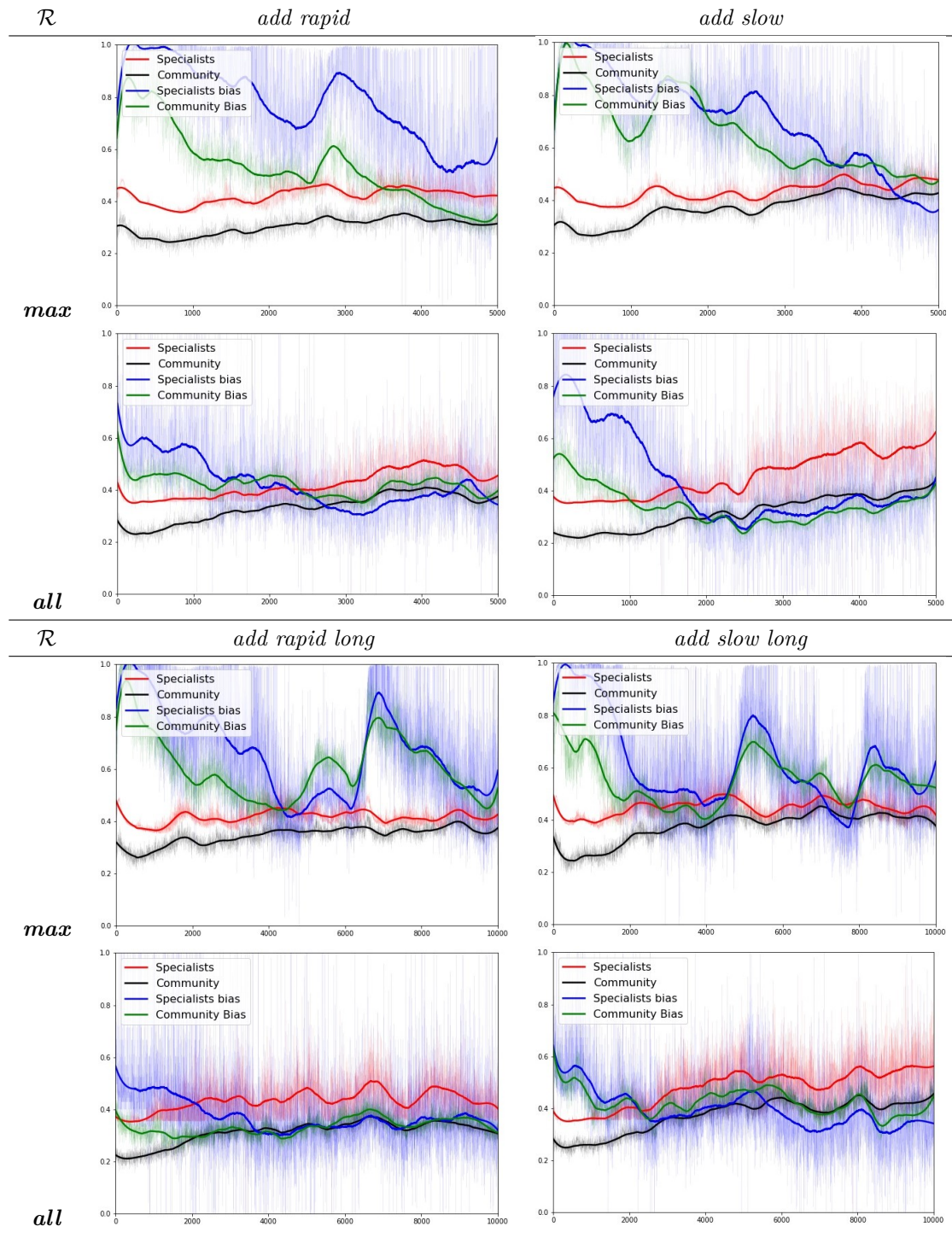


Figure 4.4: Experiment 3: Knowledge dynamics for progressively added knowledge showing that the slow addition of new knowledge fosters developing explanatory adequacy

4.7.4 Summary of Results

Exp	$\mathcal{F}$	$\mathcal{R}$	Epistemic Condition	Explanatory Adequacy
1	static	max	epistemic stagnation	$\mathcal{K}^{DIP} \not\cong \mathcal{K}$
1	static	all	epistemic stagnation	$\mathcal{K}^{DIP} \not\cong \mathcal{K}$
1	dynamic	max	epistemic incongruence	conditionally $\mathcal{K}^{DIP} \cong \mathcal{K}$
1	dynamic	all	epistemic incongruence	conditionally $\mathcal{K}^{DIP} \cong \mathcal{K}$
2	restart	max	epistemic diversity	$\mathcal{K}^{DIP} \cong \mathcal{K}$
2	restart	all	epistemic diversity	$\mathcal{K}^{DIP} \cong \mathcal{K}$
2	iterate	max	epistemic stagnation	$\mathcal{K}^{DIP} \not\cong \mathcal{K}$
2	iterate	all	epistemic stagnation	$\mathcal{K}^{DIP} \not\cong \mathcal{K}$
3	add rapid	max	epistemic incongruence	conditionally $\mathcal{K}^{DIP} \cong \mathcal{K}$
3	add rapid	all	epistemic incongruence	conditionally $\mathcal{K}^{DIP} \cong \mathcal{K}$
3	add slow	max	epistemic diversity	$\mathcal{K}^{DIP} \cong \mathcal{K}$
3	add slow	all	epistemic diversity	$\mathcal{K}^{DIP} \cong \mathcal{K}$

Table 4.3: Summary of experimental results

The initial motivation for this work was to investigate the impact of the different configurations of the DIP to its epistemic condition, as well as to its explanatory adequacy. Hence, Table 4.3 combines the results of the different experiments by specifying the values of the independent experimental parameters $\mathcal{F}$ and $\mathcal{R}$ in the first two rows and determining the epistemic condition and the explanatory adequacy in the two following. Starting with the *static* configuration, a static population shows high knowledge similarity, suggesting congruence (i.e. a determinant of epistemic coherence), but knowledge exchange does not occur across the network. This corresponds to a state of epistemic stagnation which prevents system improvement and limits its adaptation to a changing environment. Additionally, the knowledge of the DIP is dissimilar to the *ground truth knowledge*, which demonstrates that the explanatory adequacy (i.e. the capability of the DIP to accurately explain the environment by tracking the *ground truth knowledge*) of the population under those experimental conditions remains limited.

Conversely, dynamic populations that have access to all knowledge from the outset (*dynamic*) display greater diversity, shifting between periods of high and low congruence (i.e. congruence is temporal). Moreover, although they can relatively accurately model the environment at times, only the specialists when they exchange all their processing rules are capable of producing relatively adequate explanations (of which the accuracy degrades over time).

Moreover, when agents perceive their knowledge to be similar to others' (i.e. $(\mathcal{S}^{(self)} \approx \mathcal{S}^U)$,

they prefer using their knowledge instead of using social influence. For instance, in the *restart* condition, where agents start with only one processing rule, communication of knowledge is minimal during early epochs. A similar effect occurs in the *iterate* condition, where collective knowledge stagnates and fragments. This suggests that when agents assume their peers cannot offer useful knowledge (due to perceived similarity), knowledge remains static and they fail to model the causes of the observed signal.

4

This implies that epistemic diversity and systemic evolution (in terms of accuracy) rely on both knowledge differentiation and agents' recognition of this difference. Although this diversity could be perceived as a negative situation, it can potentially lead to positive outcomes (i.e. high explanatory adequacy) if the agents acknowledge that they can benefit from the exchange of knowledge. However, in the *restart* scenario, once all knowledge becomes available (after 2500 epochs), the collective quickly improves its explanatory adequacy, producing knowledge that more closely approximates the *ground truth knowledge*. Notably, certain sources, identified as specialists by the collective, experience a significant rise in knowledge utility, indicating that those identified as experts are also more likely to assimilate new knowledge.

Finally, the rapid introduction of new processing rules promotes diversity, but agents do not have sufficient time to accept and assimilate new knowledge. As a result, short-term congruence can lead to long-term incongruence, hindering the DIP's ability to consistently identify the *ground truth knowledge*. However, when agents share all their knowledge (*all*), the population becomes more congruent, or in other words epistemically coherent, and shows potential for evolution, as it is better equipped to provide a coherent (i.e. more accurate) explanation of the *ground truth knowledge*.

4.8 Conclusions and Discussion on Explanatory Adequacy

In summary, the work presented above makes the following contributions:

- It defined the problem of explanatory adequacy in self-organising MASs, where agents use their social network to aggregate incomplete or inconsistent knowledge to explain an observed phenomenon.
- It introduced a novel socially-inspired algorithm with MR capabilities, called *RTSI+*, incorporating bilateral influence between targets and sources, along with the exchange of information processing rules.

- Based on the *RTSI+* and following an ABMS approach, it developed a simulator for enabling a DIP to develop explanatory adequacy, through MR.
- It conducted a series of experiments to investigate how the initial conditions of the DIP affect the knowledge dynamics in the population (in terms of epistemic condition and explanatory adequacy).
- Through the analysis of experimental results, it showed that *RTSI+* can enable epistemic coherence (when agents acknowledge that they can benefit from social interactions).
- Overall, it proposed a mechanism that uses social influence for effectively managing the DIP's knowledge, through the exchange of processing rules, and dealing with some of the fundamental issues of self-governance (i.e. knowledge management, fairness, and justice).

In particular, the experiments highlighted the necessity of diversity, a willingness to learn, and good intentions for achieving systemic self-improvement through epistemic evolution in dynamic socio-technical and cyber-physical environments. First, systemic improvement through epistemic evolution requires diversity (i.e. epistemic incoherence). Specifically, if the group perceives this epistemic incoherence, it can use *RTSI* to become more coherent and develop explanatory adequacy. In more detail, we showed that DIPs composed of agents with similar knowledge were unable to enhance explanatory adequacy. It is crucial to preserve knowledge within the network, as it might become useful later. Moreover, agents should not only possess diverse knowledge but also recognise and understand this diversity, and be able to identify reliable sources of expertise. This is also validated by Ober's claims on knowledge management [39], and specifically the fact that both common knowledge and diversity is required for self-improvement and innovation. This is because knowledge alignment and knowledge aggregation are integral parts of knowledge management, and common knowledge is required for alignment while diversity contributes to effective knowledge aggregation.

Second, systemic evolution relies on individuals being willing to learn. For epistemic evolution to occur, agents must make the effort to ask questions, provide answers, and assimilate the information they receive. Thirdly, *RTSI*'s ability to address the explanatory adequacy problem depends on both sources and targets having good intentions, supported by popular legitimacy. For instance, DIPs may encounter deceptive agents, where a knowledge source manipulates targets or distorts collective knowledge for personal gain. Similarly, denialist groups, whose flawed knowledge is resistant to evidence or argument, can undermine system performance. Such scenarios require institutional rules and legitimacy to regulate behaviour and ensure accountability.

Overall, this work provides a mechanism for fostering knowledge management through the use of social influence. This is because the group although has impartial knowledge, through interactions, it manages to increase its understanding of the environment and develop explanatory adequacy. Additionally, *RTSI+* could be considered an approach for enabling a group to deal with issues of fairness and justice. Specifically, the individuals try to explain an observed phenomenon instead of directly forming an opinion on it. Also, using social influence they can update their knowledge and develop a better understanding of the environment. If we assume that the observed phenomenon is a resource distribution, the attempt of agents to identify the processing rules resulting in it gives them the capability to understand whether the process is fair and just, and act accordingly.

4.9 From Explanatory Adequacy to Reflection

The work presented above builds on RTSI algorithm for signal tracking [9] and modifies it to solve a different problem. Instead of using social influence to agree on the value of a signal, the group of agents has to self-organise to agree on the processing rules producing a signal that they collectively observe. This requires managing their knowledge but also performing some kind of reflection on their environment, adapt to the new knowledge that is introduced through the new agents in the population and manage to form a coherent explanation that approximates the *ground truth knowledge*.

Additionally, the process of introduction of new knowledge corresponds to a first instance of exogenous innovation, or knowledge transfer. Specifically, the experimental findings that demonstrate that there needs to be epistemic diversity to achieve adaptation and self-improvement ground the need for innovation, which will be extensively analysed in Chapter 6. Moreover, the fact that the quick addition of knowledge results in a population being inept or unwilling to enhance it provides some insights related with the timing and specifically how it can affect the properties of the system [151] as well as how individuals distinguish between matters requiring slow and deliberate processing and others that occur fast or based on intuition and emotion [152].

Although under some initial conditions the population manages to develop explanatory adequacy and adapt to the new knowledge, this adaptation is limited since the environment is fixed. Moreover, innovation occurs only from the introduction of new knowledge and it is not generated from inside. Accordingly, in the next Chapters we would look into issues of learning (in order to adapt in a non-fixed parameter space) and to achieve endogenous innovation. However, a prereq-

quisite for achieving those is to design systems capable of evaluating their own performance and identifying whether they are getting better or not. As such, next Sections investigate metrics, and specifically how a DIP's performance can be assessed with respect to a set of conflicting pairs of drivers.

4.10 Metrics for Reflection

4.10.1 Problem Specification

Previous work examined scenarios in which a group of agents manages to self-organise as a DIP (i.e. RTSI) and *react* to the environment by determining the value of a signal [71], forming a qualitative assessment with respect to the fairness of a resource distribution [9] or identifying the processing rules causing an observation [10]. All those cases correspond to reactive-RTSI models.

Moreover, the experimental findings on the study of RTSI for signal tracking [9] presented in Section 2.3 shed light on the requirement for the DIP to evaluate the system's performance against multiple criteria, corresponding to systemic drivers. In particular, four conflicting pairs of drivers were identified, which include accuracy vs. economy, stability vs. flexibility, inclusivity vs. expertise, and coherence vs. diversity, and are illustrated in Figure 4.5.

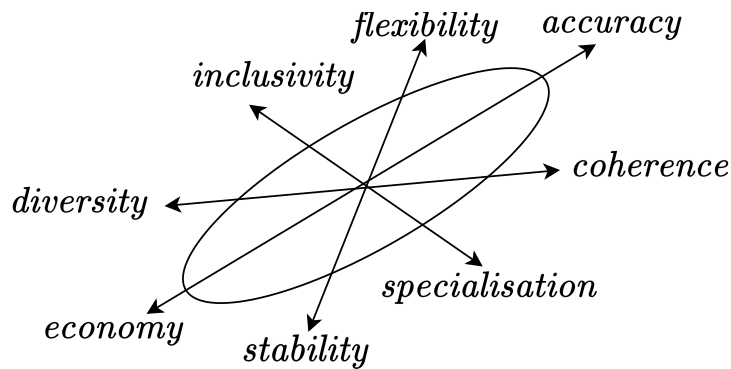


Figure 4.5: The drivers of the DIP

To determine whether this balance is maintained, the DIP must engage in reflection [153] and reflective governance [154] to address questions it poses about itself. Therefore, the DIP needs a set of informative metrics to evaluate its performance across eight key “dimensions”, along with a process to deliberate on these evaluations. This deliberation could be achieved using the RTSI framework, effectively transforming the DIP into a cybernetic system [16] with both reactive and reflective components, as depicted in Figure 4.6.

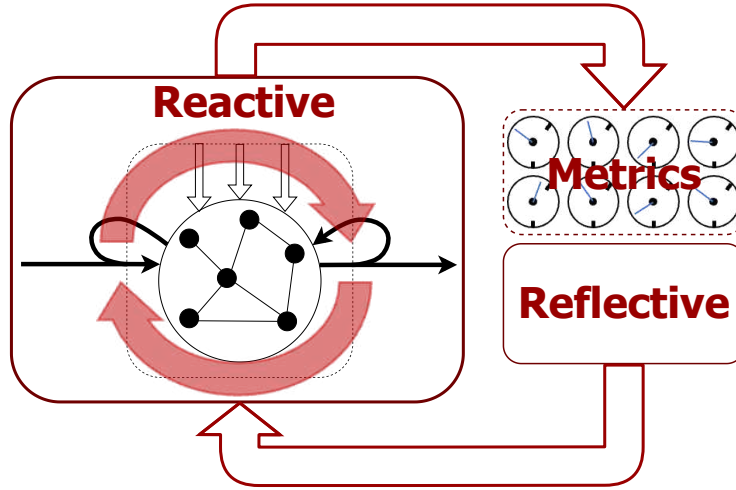


Figure 4.6: The DIP as a cybernetic system

Hence, this work extends on previous work on reactive-RTSI [9], [10] by proposing metrics for achieving reflective-RTSI. In particular, it maps systemic drivers, which concern dimensions of the system's (i.e. the DIP) performance, to metrics. The values of those metrics can be given as input to the DIP, in the next iteration, and enable it improve its own performance.

4.10.2 The Metrics

As discussed in previous Sections, the performance of a group, acting as a DIP, can be evaluated in terms of accuracy, economy, flexibility, stability, coherence, diversity, inclusivity, and specialisation. *Metrics* can be defined to evaluate the performance of a DIP in terms of these eight *dimensions* which correspond to conflicting *systemic drivers*. The *purpose* of each of them is described below.

Starting with stability, this metric assesses whether agents can operate effectively regardless of changing conditions. Additionally, the DIP needs to evaluate whether there are multiple clusters of opinion in the population and how much they diverge. Thus, a coherence metric based on opinion similarity and divergence between groups is necessary [12].

To avoid groupthink and the formation of cliques, agents should be able to evaluate the distribution of the sources in the population and how they, and their opinions, diverge change. Also, to have the capability to provide an accurate output, the DIP should delegate most tasks to experts [9]. At the same time, it should ensure that all agents participate in the collective process. Moreover, to function effectively, the DIP's distribution of work should follow Price Law, where half the work is done by the square root of the population (i.e. the cognitive task experts), while the rest is handled by the remaining agents. To assess the operation of the DIP in these areas, metrics

for specialisation and diversity, based on expertise and influence, should be incorporated into the system design. Additionally, the DIP must not only be stable but also flexible enough to adapt to changing conditions. Balancing the ability to change with the ability to remain consistent represents a fundamental conflicts. Overall, the dimensions of interest in the context of self-organising MASs together with the corresponding metrics and their purpose are described in Table 6.1, and analysed below.

Dimensions	Metrics	Purpose
Accuracy	Performance	Accuracy with respect to expense
Economy	Expense	Processing Cost
Coherence	Cohesivity	Network clusters' similarity and divergence
Diversity	Influence	Distribution of sources
Flexibility	Flexibility	Adaptation to the different conditions
Stability	Stability	Effective functionality
Inclusivity	Participation	Equal participation
Specialisation	Expertise	Emergence of expertise

Table 4.4: Multi-agent dimensions mapping to metrics

In the rest of this section, we propose some metrics for achieving reflective-RTSI.

4.10.2.1 A Metric for Performance

Aiming to achieve reflective RTSI, we define a set of metrics for assessing the fairness of a process of resource distribution in a CPR problem. This is based on previous work on reactive RTSI [9], [10] that was discussed earlier in this Chapter and in Chapter 2. In this context, the performance metric focuses on assessing the accuracy of the DIP's collective opinion on the fairness of lived experiences, in relation to the resources expended to achieve that accuracy. It considers the importance of obtaining the correct answer regarding the fairness of resource distribution. This involves a subjective judgement about how much effort is justifiable to achieve an acceptable level of accuracy, given the significance of arriving at the "correct" answer.

The metric is intended to inform agents about whether they have used too many resources to answer a relatively unimportant question, too few resources to get a sufficiently accurate answer for an important issue, or if achieving an acceptable level of accuracy is unattainable regardless of the resources spent.

To capture this, the performance metric Q_t can be defined as the accuracy of the collective answer acc_t , divided by the proportion of resources used ($sumres_t$) relative to the total available

resources ($allres_t$), and weighted by the relative significance σ at epoch t :

$$\mathcal{Q}_t = \sigma\left(\frac{acc_t}{sumres_t/allres_t}\right) \quad (4.8)$$

4.10.2.2 One and a Half Metrics for Influence

4

The influence metric focuses on measuring how many agents are seeking to delegate tasks to others (sources) rather than performing them themselves, and how many agents are acting as sources of influence. This helps to determine whether trust is being placed in experts, which could indicate resource optimisation by assigning cognitive tasks to more knowledgeable agents. However, it could also lead to potential risks like groupthink or exclusion of certain agents.

The influence metric $\mathcal{I}_t$ can be defined as the ratio of agents using the work (e.g. opinion, rules, answer) of others ($targets_t$) to the agents doing the work (sources, $sources_t$) at epoch t :

$$\mathcal{I}_t = \frac{targets_t}{sources_t} \quad (4.9)$$

If $\mathcal{I}_t$ equals 1, it means the number of agents asking for help is equal to the number of agents providing it. If $\mathcal{I}_t > 1$, more agents are seeking assistance than those offering it.

Additionally, to observe whether certain sources are influencing a disproportionate number of agents, we can define an auxiliary (Price) metric, $\mathcal{I}_t^{(max)}$, which identifies the proportion of sources in the DIP that influence half of the total number of agents.

4.10.2.3 A Metric for Cohesivity

Cohesivity is aimed at measuring the divergence of processing rules within a community. This metric provides feedback to agents about the connections between clusters from both a network perspective (i.e. agents within highly interconnected groups, as described by [155]) and a societal perspective (i.e. agents using similar processing rules).

Cohesivity $\mathcal{C}_t$ is defined as the ratio between the average divergence of processing rules within clusters ($\bar{d}_{in}$) and the average divergence between agents within a cluster and those outside the cluster ($\bar{d}_{out}$):

$$\mathcal{C}_t = \frac{\bar{d}_{in}}{\bar{d}_{out}} \quad (4.10)$$

The metric $\bar{d}_{in}$ measures the average divergence of opinions or processing rules, among agents

within the same cluster (e.g., among agents in cluster X or cluster Y), averaged across all clusters. In contrast, $\bar{d}_{out}$ measures the average divergence between agents in a cluster and those outside it (e.g., divergence between agents in cluster X and those not in X), also averaged across clusters.

For the system to remain cohesive and avoid fragmentation or conflict, both $\bar{d}_{in}$ and $\bar{d}_{out}$ should be kept low, indicating low divergence and a degree of consensus. However, to encourage innovation and adaptability, $\bar{d}_{in}$ should be slightly lower than $\bar{d}_{out}$, allowing for agreement within clusters while maintaining some diversity between different clusters.

4.10.2.4 A Metric for Expense

In MASs, there are two main types of costs: one is the cost of communication (gathering information from the social network), and the other is the cost of cognitive processes (finding the information). While communication can be expensive in large-scale networks, it is assumed that in small- to medium-scale networks, the cognitive task cost, $cost_{cogn}$, is higher than the communication cost, $cost_{com}$.

The total expense $\mathcal{E}_t$ of the DIP can be defined as the sum of the number of agents delegating tasks (*targets*) multiplied by the communication cost ($cost_{com}$) and the number of agents performing cognitive tasks (*sources*) multiplied by the cognitive cost ($cost_{cogn}$), as shown below:

$$\mathcal{E}_t = targets * cost_{com} + sources * cost_{cogn} \quad (4.11)$$

4.10.2.5 Two Metrics for Expertise

Expertise refers to sources of influence that either attempt to persuade and modify others' beliefs or are recognised by others as those who can help achieve the collective goal. From a functional perspective, the DIP needs to identify expertise to delegate tasks to those capable of providing the correct answer. This improves the quality of the final outcome by achieving accurate opinions and conserving resources through delegation. Additionally, expertise must evolve over time to adapt to dynamic environments where knowledge changes.

Two possible metrics for expertise can be identified: short-term expertise (E^s) and long-term expertise (E^l). Short-term expertise (E^s) reflects expertise over a short period, measuring the number of agents who have been sources for at least m . Ideally, short-term expertise (E^s) should

remain high, indicating stable expertise in the short term, while long-term expertise (E^l) should remain low, showing that expertise adapts and changes over time.

4.10.2.6 A Metric for Participation

4

This metric aims to assess both the quality and quantity of participation within the system. It tracks how many agents contribute to the collective decision-making process by offering an opinion (quantity of participation), how many consistently delegate the task to others, how many always take on the task themselves (which might suggest either a refusal to consider alternative opinions or an attempt to dominate the collective choice), and how many fail to align their own actions with the collective decision.

In dynamic systems, where change is essential, agents must be willing to adjust their strategies—alternating between consulting the network and relying on their own judgement—and adapt their beliefs and processing rules. They should seek input from a variety of agents to expose themselves to different perspectives, and they need to participate actively, recognising the value of collective decision-making.

By understanding this metric, agents can gauge whether the collective outcome truly reflects the majority’s opinion and determine whether adjustments are needed to foster broader participation or to address those who do not engage in constructive behaviours.

4.10.2.7 A Metric for Stability

The challenge of assessing stability, as outlined by [16], lies in the fact that stability is a property of a field and not of a material body. Based on that argument, Ashby provides a definition of stability: “given the field of a state-determined system and a region within that field, the region is considered *stable* if the lines of behaviour from all points in that region remain contained within it”. This implies that we are focused on trajectories in a space rather than merely on points. However, Ashby also pointed out that altering certain parameters can change the field, which stability relies upon. Consequently, DIP units are more likely to be *quasi-stable* [72], [156], experiencing phases in their trajectories where they are stable, followed by periods of instability, before returning to a stable state but with different values for their control variables.

4.10.2.8 A Metric for Flexibility

Flexibility stands in contrast to stability, and there are associated challenges in quantifying flexibility. Specifically, flexibility encompasses both a functional and a normative dimension: a DIP can change (functional) and it *ought to* be able to change (normative). This distinction allows us to differentiate between a DIP that has “stuck” due to factors like path dependency [157] (where current actions are limited or constrained by past choices) and a quasi-stable DIP that is undergoing a phase of relative “calm”.

However, defining metrics for balancing out stability with flexibility is very challenging in systems that include components with awareness, intelligence, or reflective capabilities. Accordingly, next Section explores some issues related to the application and interpretation of these metrics.

4.10.3 Risks

Metrics influence decisions in various ways [158], offering crucial information and enhancing problem understanding, often serving as criteria for decision-making [159]. While metrics are vital for systemic self-improvement, challenges such as improper metric selection, misinterpretation, and misuse frequently arise in cyber-physical institutions.

Initially, defining metrics can be difficult since they need to be linked to system characteristics. Adopting a metric from one system does not guarantee it will provide the same insights into another. Moreover, metrics should accurately capture the information they intend to represent. Since metrics directly impact decisions and actions, it is crucial to design them to encompass not just a single parameter but also related information. This can be achieved by incorporating auxiliary or multiple metrics. For example, if a situation is influenced by a , b , c , and d , choosing a metric that only considers a and b may lead to solutions for those two factors while inadvertently exacerbating c and d .

Another challenge in interpreting metrics is their potential for self-referentiality. A metric’s definition might depend on the outcomes of other metrics or a combination of them. Thus, metrics are interconnected, and analysing their results to inform actions requires considering all relevant metrics.

Furthermore, the first step in understanding a system is to identify appropriate metrics that describe its current state. The next step toward achieving organisational goals, such as sustainability or balancing various incentives, is to incorporate *meta*-metrics that indicate the rate of change

within the system and provide insights into the temporal evolution of observations. However, defining *meta*-metrics is challenging, and analysing them can be even more complex, especially for internal observers (e.g., agents). Consequently, systems often struggle to adapt and maintain sustainability because their metrics primarily reflect short-term effects.

Although metrics are crucial, organisations frequently fall victim to an obsession with them, focusing more on metrics than on identifying the right ones that yield valuable insights. This fixation can lead individuals to invest excessive time and effort in measuring performance, causing them to overlook essential factors. Thus, it becomes challenging to determine the minimum number of required metrics that clearly convey desired information while avoiding the pitfalls of metric obsession.

Identifying the appropriate metric is only part of the challenge; finding the right way to utilise it is equally important. Simply defining a set of metrics does not guarantee sustainable self-improvement. System agents must have access to these metrics, be able to interpret them, and be willing to adjust their behaviours and policies based on the feedback provided by those metrics. To address potential issues in metric definition, interpretation, and application, it is essential to periodically revise metrics, even if the system's policies remain unchanged. This is important because, if a metric is initially flawed, modifying it can help obtain the intended information. Additionally, if a metric becomes a target, altering it may mitigate unintended consequences. In dynamic institutions with evolving populations, changing metrics is necessary to reflect shifts in the DIP, knowledge, and needs and practices.

Having described the challenges related with metrics, we can consider also some phenomena and examples of misuse of metrics, namely Goodhart's Law, vanity metrics, challenges in metric interpretation, faux leagues, and social credit systems.

The tyranny of metrics [160] refers to the phenomenon where an overemphasis on metrics for performance evaluation can actually impair performance itself. This can be seen in various contexts, from trivial examples to serious consequences. For instance, football players may avoid making risky passes that could lead to goal-scoring opportunities in order to maintain a high pass-completion percentage. Similarly, lecturers might shy away from teaching challenging material to achieve better student evaluations, ultimately resulting in lower student comprehension and skills. In more critical cases, surgeons might decline to perform risky but potentially life-saving procedures to preserve their operation survival rankings.

Goodhart's Law [161] formalises the notion that when a measure is used as a target, it loses its

significance. Additionally, there is the quasi-quantum effect on human behaviour, where awareness of being measured can influence performance, much like the Hawthorne Effect, where individuals alter their behaviour simply because they know they are being observed. Goodhart's concerns originated in the context of monetary supply in economics, while Asimov's fictional psycho-history suggested that subjects should remain unaware of being analysed for predictions to hold true. These effects manifest across various domains. For instance, the h -index, used to gauge scientific productivity, was initially a useful measure until academics recognised that their appointments and promotions were influenced by this metric. This led to behavioural shifts, such as academics emphasising citation counts over actual scientific contributions, forming citation clubs of mutual referencing, and undermining the peer review process essential to the scientific method.

Another issue arises when a metric devolves into what is known as a vanity metric [162]. This refers to a statistic that may appear impressive but holds little relevance for those who are genuinely concerned with performance. The h -index arguably illustrates this point, as a high h -index may only impress those who take pride in their own high scores. Consequently, what a metric signifies can change over time. The h -index, once a measure of scientific productivity, now often reflects network centrality and longevity [163], indicating that certain individuals are cited more frequently simply because of their popularity and tenure in the field (the network effect).

Thus, a metric may not only fail to measure what is of real interest, but it may also overlook significant aspects. For example, previous research aimed to quantify "civic participation" in a self-organising MAS [62], which was defined as the collective responsibility for making, adjudicating, and enforcing rules. While it was possible to identify outliers for excessive participation, it remained unclear whether this stemmed from network centrality or a reluctance to relinquish power. Similarly, under-participation could result from network isolation or an aversion to taking responsibility, yet it could not be detected as an outlier using the statistical methods employed in analysing participation metrics.

Using metrics for comparative rankings, such as "league tables", raises another concern. Unlike pairwise comparisons, which offer meaningful rankings over time (as seen in sports), ranking institutions like universities based on dubious metrics fails to provide valuable insights. These rankings do not serve a clear purpose, creating a cycle of hollow accolades and criticisms as institutions fluctuate in status, often with those in power taking credit for successes while blaming others for failures.

Finally, socially constructed metrics of credence among peers, such as those examined in studies

on expertise emergence in knowledge aggregation, could potentially serve as a basis for social credit systems if aggregated. In such systems, rewards are distributed to those deemed deserving, while penalties are imposed on those considered unworthy. This can lead to numerous unintended and harmful consequences, including a tyranny of merit [164] and the suppression of dissent or disobedience [43]. These dynamics expose the gap between the intentions behind rules and their application, especially when a ruler enforces rules incorrectly or changes them retrospectively after violating them—illustrating the pettiness often exhibited by would-be tyrants throughout history.

4.10.4 Conclusions on Metrics for Reflection

Overall, this examined DIPs from a cybernetic systems perspective, focusing on both reactive and reflective components [16]. Specifically, Section 4.2 focused on the implementation of the reactive part, while this investigated the use of RTSI for managing the reflective part [154]. To support this, metrics are required.

Drawing from a survey of metrics in MASs, several metrics were proposed to assess a DIP based on eight performance criteria, organised into four pairs of conflicting systemic drivers. These metrics help maintain a balance between drivers and support intelligent components in generating reflective action plans. This work also noted some risks involved when relying on metrics, by highlighting some challenges for properly identify what and how to measure, as well as some examples of misuse.

Overall, this work offers insights on self-governance challenges, and specifically it can facilitate knowledge management, help resisting to oligarchic tendencies, and preserving dignity. The former because by measuring specialisation, accuracy, diversity, flexibility, and coherence it enables the group to control the way they exchange knowledge. On avoidance of the iron law of oligarchy, this is because by measuring aspects such as inclusivity, the group develops an understanding of the power dynamics and can proactively act. Finally, by taking into account multiple systemic drivers and making individuals aware of the performance of the system with respect to those, it can support dignity by allowing them making well-informed decisions.

4.11 Summary & Discussion

Chapter 4 specified an algorithm, called *RTSI+*, that uses social influence to enable a self-organising MAS to develop *explanatory adequacy*. This allowed the group to *reflect* on its functioning in order

to adapt to the environment and self-improve.

Firstly, it explored under which initial conditions the group manages to use social influence (by exchanging information processing rules) to form an accurate explanation for the processing rules producing a commonly observed phenomenon. This corresponded also to some kind of preliminary reflection on the environment, in the sense that agents had to modify their knowledge (i.e. their own processing rules) based on the knowledge of their peers and the *ground truth knowledge*. Although the definition of rules was abstract, such a tool could be used to enable us to understand how humans might explain the rules applied to them (for instance laws about taxation, or retirement). At the same time, it could be used to support decision-making with respect to such policies. For instance, policymakers could test different policy schemes in simulation and identify which might be more easily accessible and explainable to the end-users. This way they could select the ones that increase the individuals' satisfaction because they recognise that they are being treated fairly, or equally.

Secondly, Section 4.10 introduced a set of metrics for informing a SGS about its performance in terms of multiple criteria. Those metrics allowed the group to use social influence, as expressed by RTSI, to react to the environment but also to reflect on it and modify its behaviour. This way, the group was capable of performing two kinds of distributed information processing: one on its environment (e.g. agree on the value of a signal, or on the processing rules producing that signal), one on itself (e.g. adapt its behaviour to match the knowledge of the environment). Furthermore, it discussed some issues related to misinterpretation and misuse of metrics, and highlighted the fact that both the definition of metrics and the acknowledgement of the risks involved when using metrics are prerequisites for achieving meaningful reflection in SGSs.

However, using social influence, enabled through machine reasoning techniques, and identifying the appropriate metrics for evaluating performance might not be sufficient for adapting to a complex dynamic environments. Instead, the group might have to learn from its experiences, modify itself, and self-improve. Chapter 5 touches upon these issues and proposes the use of theories of social sciences together with machine reasoning and machine learning techniques, through an Hybrid Reasoning and Learning (HyRL) approach, to enable effective self-regulation and sustainable collective decision-making in self-governing socio-technical systems.

Chapter 5

Adaptation using Hybrid Systems

Contents

5.1 Introduction	93
5.2 Effective Self-Regulation	95
5.2.1 Motivation	95
5.2.2 Problem Specification	96
5.2.3 The <i>4voices</i> Algorithm for Self-Regulation	98
5.2.4 Experimental Design	103
5.2.5 Experimental Results	105
5.2.6 Conclusions on Effective Self-Regulation	110
5.3 Sustainable Consensus Formation	111
5.3.1 Motivation	111
5.3.2 Political Sciences Insights	113
5.3.3 Problem Specification	113
5.3.4 Θ -learning Algorithm for Sustainable Consensus Formation	114
5.3.5 Experimental Design	117
5.3.6 Experimental Results	120
5.3.7 Conclusions on Sustainable Consensus Formation	128
5.4 Managing Sustainability in Self-Governing Socio-Technical Systems	130
5.4.1 Motivation	130
5.4.2 Evaluation & Control of Sustainability	130
5.4.3 HyRL Systems For Uncertainty & Sustainability	132
5.4.4 Conclusions on Metrics for Sustainability	133
5.5 Summary & Conclusions	133

5.1 Introduction

Chapter 3 introduced Hybrid Reasoning and Learning (HyRL) methodology for integrating Machine Reasoning (MR) and Machine Learning (ML) algorithms in a system to allow for reaction

and reflection, leading to self-improvement. Chapter 5 presents two applications of this methodology that leverage principles from psychology, political sciences, cybernetics, and Reinforcement Learning (RL) to enable a system to achieve *effective self-regulation* through the establishment of the pathways for requisite influence, to *sustainably form collective decisions* by learning when to make compromises and when to express dissent.

Section 5.2 describes a novel psychologically-inspired *4voices* algorithm encompassing properties of learning and reasoning, using HyRL. Particularly, this combines concepts from psychoacoustics [27] with first order cybernetics [19] and RL [17]. The proposed model enables the units of a regulated system to identify which of the inner voices to attend to achieve a collective goal. Also, it enables an intelligent regulator, which uses Q-learning [165], to maintain attention to the collective expression of the regulated unit and adjust the regulatory mechanism to maintain stability. The results of applying this algorithm to a problem of continuous monitoring and control show that the *4voices* algorithm enables the self-regulated system to establish the pathways for requisite social influence, maintain stability and achieve ethical self-regulation.

Section 5.3 presents Θ -learning, an algorithm inspired by classical Athenian democratic deliberation, which uses HyRL to combine *thorybos* (citizen vocal expressions to policy options) [20] with Multi-Agent Reinforcement Learning (MARL). This manages to extract a productive signal out of something (i.e. *thorybos*) that would otherwise be disregarded as noise. Also, this highlights that a majority decision rule might be acceptable, if consensus is infeasible or ineffective, as long as the underlying principle is meaningful democratic deliberation. The experimental results show that the SGS manages to maintain a balance between legitimate consent and legitimate dissent (which correspond to two fundamental issues of self-governance as described in Section 2.2.4) and achieve a sustainable consensus formation through democratic deliberation.

Additionally, Section 5.4 analyses how the HyRL methodology enables the maintenance of sustainability within next generation socio-technical systems, underscoring the role of technology as a key component in balancing ecological, social, and economic interactions. In particular, it argues that HyRL-based systems can achieve sustainability when technology, environmental, and socio-political aspects co-exist and co-evolve in a common system. After identifying sustainability in the context of contemporary socio-technical systems, it describes how such systems can support sustainable self-governance while acknowledging their limitations.

5.2 Effective Self-Regulation

5.2.1 Motivation

Self-governance or self-regulation requires the capability of the system to decide on the policy that would govern it ethically and satisfactorily. According to Asbhy's Good Regulator Theorem [19], [166] concerning ethically regulated systems, effective regulation necessitates establishing the pathways to requisite influence. To achieve that the agents in the system have to be able to track the policy changes, evaluate its effects, communicate amongst themselves to collectively agree on the performance of that policy and influence the policy based on their collective assessment. Essentially, this corresponds to a self-regulated system having a designated agent acting as a regulator and a group of agents being the regulated units, as shown in Figure 5.1.

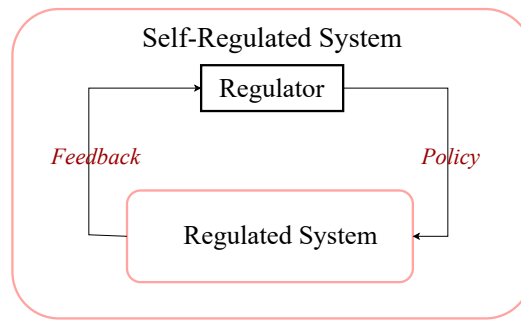


Figure 5.1: A Self-Regulated System

Since this thesis aims to propose mechanisms that facilitate human deliberation concerning matters of governance and self-regulation, we propose to model the system using insights from psychology together with concepts from cybernetics and Reinforcement Learning (RL). Accordingly, this Section investigates how those can be brought together through the HyRL methodology to achieve effective self-regulation. In particular, it uses MR techniques to model a system inspired by Nowak's theory of social influence [12], Fernyhough's theory about 'inner speech' [27], and Driver's concept of 'cocktail party effect' concerning the activation of inner voices by selective auditory attention [167]. Those concepts are combined with RL to enable the regulator to learn from the expressions of the regulated system.

Overall, this corresponds to a hybrid system that involves a psychologically inspired machine reasoning component (corresponding to the Socially-Inspired Computing (SIC)-Agent-Based Modelling and Simulation (ABMS) part of the HyRL methodology) and a RL component (corresponding to the ML part of the HyRL methodology). The operation of that system is formally specified

in the *4voices* algorithm which achieves timely feedback, responsive control, and stability through requisite social influence.

Before describing the formal specification of the problem, we will ground this back to the shared housing community example. Here, we might imagine residents being subject to a policy of resource distribution, but this time agents do not try to explain the policy, but instead they have to express their opinion in order to affect it. Also, residents might receive different noise signals from their social network, while they also they hold their own opinion on that policy. Agents, however, face the requirement to distinguish between the voices of influence (which might correspond to the voice of the experts, the voice of their trusted sources, the noise of the community, or even their own opinion), and select the one to form their individual expression. This expression would then be combined with the expressions of the others agents in the system and together they would form the collective expression. This collective expression would be attended by the regulator who would adjust the policy in such a way that the satisfaction of the individuals in this community are satisfied. However, the decision on the selection of voice as well as the selection of policy is not one off. Instead agents can learn from the past and improve their decisions in the long-term.

5

5.2.2 Problem Specification

This work uses the HyRL to combine concepts from cybernetics, concerning issues of ethical regulation [19], with psychoacoustics, and RL to establish the pathways to requisite influence. In particular, we outline the pathways to requisite social influence as:

- the expression of the regulator to the regulated units,
- the attention of the regulated units to the regulator,
- the expression of the regulated units to the regulator,
- the attention of the regulator to their feedback, and
- the social influence in the social network of the regulated units.

We outline the pathways to achieving requisite influence within a self-regulated system. This involves the communication from the regulator to the regulated units and their responsiveness to the regulator, the communication from the regulated units to the regulator and its attention to their feedback, and the social influence within the regulated units' social network. We present this

as the “generic problem” and apply it to a specific case of continuous monitoring and control. Note that this approach is not necessarily an “optimal” solution to the specific problem; rather, we are establishing a test environment to examine stability and requisite social influence in self-regulated systems.

5.2.2.1 The Generic Problem

We build upon Ashby’s framework **Ash2020** for a self-regulated system by introducing the psychologically inspired model of attention and expression to the regulated units. This model incorporates pathways for requisite influence, individual and collective learning, as illustrated in Figure 5.2. In

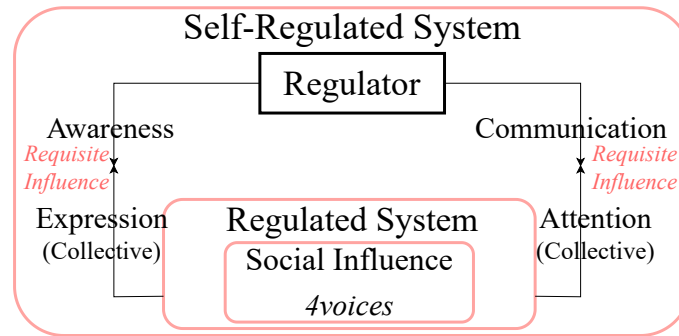


Figure 5.2: A Self-Regulated System with Requisite Social Influence

line with Ashby’s concept, on the one hand, the regulated units must form a collective expression to inform the regulator about the effects of the policy changes. On the other hand, the regulator has to attend their expression and effectively change the policy. This establishes an iterative process: the regulator selects a policy, applies it, and communicates the change to the regulated units. The units, in turn, attend to the change, listen to their inner voices of social influence, identify which one to select to form their individual expression, and the individual expressions are aggregated to form the collective expression which is communicated back to the regulator and informs it about the effects of the change.

Overall, this addresses the problem of effective self-regulation by designing a psychologically-inspired agent-based model. In this model, the group (regulated system) learns in which of the voices to attend, in order to form an expression that informs the regulator about the actual impact of the policy change, so that the future policy choices would support the regulated units’ needs. To achieve that each individual in the group uses social influence to improve its cognitive processing and offer a meaningful input that can support the self-regulation of the system. Although social influence raises concerns about manipulation, persuasion [68], and conformity [66], as discussed in

Section 2.3, we focus on the positive aspects and we consider that a tool that can support deliberation by improving information processing and fostering collective decision making, as depicted in [12].

5.2.2.2 The Specific Problem

We assess the effectiveness of our psychologically-inspired model in a specific context of continuous monitoring and control: job scheduling in cloud computing. Here, operates in defined epochs, and within each epoch, each agent i in the regulated system assigns a job j_i to the cloud and receives feedback s_i on whether the job will be queued for processing in the next epoch t . This scenario was chosen for its dynamic, non-deterministic nature, involving interconnected, diverse units that can influence one another. The preferred policy depends on context, preferences, and adaptability, and its performance can be reasonably easily measured with specific metrics.

In this setup, the self-regulated system consists of N agents, connected in a social network SN , that engage with a cloud system. During each epoch t , each agent i assigns a job with size $j_i \in \{1, J\}$ and urgency $u_i \in \{1, U\}$ to the cloud system, where J and U are the maximum job size and urgency, respectively. The regulator determines the policy $\mathcal{P}$ which is equal to the number of jobs m to process in the upcoming epoch as depicted in Equation 5.1. Jobs are then ranked by urgency (highest first) and size (smallest first); the top m jobs are selected for processing, and agents are informed.

$$\mathcal{P} = m, m \in \{1, N\} \quad (5.1)$$

Processing these jobs incurs a total cost C , calculated as the sum of the costs J for each job size j_i plus a fixed cost F . This total cost is divided equally among the m agents whose jobs were processed, resulting in an individual cost c_i . The average cost, $\overline{C}$, is the total cost divided by the number of jobs processed m . Additionally, each job is processed with a delay q_i , and the average delay $\overline{Q}$ is the sum of the delays for the m processed jobs, divided by m : $\overline{Q} = \frac{\sum_{k \in \{1, \dots, m\}} q_k}{m}$. The maximum delay, Q_M , represents the sum of delays for N jobs (if all were accepted and processed in order), divided by N : $Q_M = \frac{\sum_{k \in \{1, \dots, N\}} q_k}{N}$.

5.2.3 The *4voices* Algorithm for Self-Regulation

The operation of the system is described in the psychologically inspired agent-based *4voices* algorithm. This comprises an MR component in the regulated system and an ML component in the regulator. The former describes the four voices to which each agent is exposed to (presented in

Section 5.2.3.1), the process according to which agents select which of those to attend to form an expression (presented in Section 5.2.3.2), and the mechanism they use to improve the selection of voice based on past experiences (presented in Section 5.2.3.3). The latter is formed by the process of learning to attend the expression of the regulated system (presented in Section 5.2.3.4). Finally, the synthesis of those is formally specified in high level by the *4voices* algorithm (presented in Section 5.2.3.5) which is structured into three main components: problem-specific processes, processes for attention and voice selection, and learning processes.

5.2.3.1 Voices of the Regulated System

Inspired by psychological research indicating that individuals often engage in an internal monologue, known as “inner speech” or “inner voice”, when facing decision-making challenges [27], our model incorporates this multifaceted inner dialogue. Drawing on ergonomic studies that identify different interpersonal distances —intimate, personal, social, and public [168]— we define four distinct voices. These include an individual’s ‘own voice’ based on personal experience [27], the ‘expert voice’ rooted in credited knowledge or expertise [169], the ‘foreground noise’ representing inputs from the individual’s social network [12], and the ‘background noise’ that reflects the broader cultural or community context [170]. Also each agent has an attention to each of those voice that dynamically changes. So, each agent i at each time step attends the following four inner voices:

$$\langle n_i^o, n_i^e, n_i^f, n_i^b, \rangle$$

which correspond to the ‘own voice’, ‘expert voice’, ‘foreground noise’, and ‘background noise’, to which has corresponding attentions:

$$\langle att_i^o, att_i^e, att_i^f, att_i^b, \rangle$$

The calculation of those is described below, in turn, while the attentions and their update is analysed in Section 5.2.3.2.

Applying the *4voices* algorithm to job scheduling, each agent i forms an opinion depending on factors like delay, cost, urgency, and whether their job is accepted for processing. This opinion is represented by the agent’s own voice n_i^o . Specifically, if agent i ’s job is accepted, the absolute value of its own voice n_i^o represents the quality of service divided by the cost, with a negative value to indicate dissatisfaction as higher noise and satisfaction as lower noise. The relationship is captured

by Equation 5.2:

$$n_i^o = -\frac{Q_M - q_i}{c_i} \quad (5.2)$$

If the job is rejected, n_i^o remains unchanged from the prior epoch. In addition to their own voice, agents are influenced by three other types of voices: background noise n_i^b (reflecting community feedback), foreground noise n_i^f (social network influence), and the expert voice n_i^e (providing system-level insights).

The expert voice n_i^e , accessible to agent i , mirrors the actual system state and is calculated based on average service quality in relation to individual cost, combined with an urgency factor $\mathcal{U}_f$. This factor represents the balance between broad satisfaction (accepting all jobs, leading to higher delays) and prioritising urgent tasks (resulting in lower delays). The urgency factor $\mathcal{U}_f$ is computed as follows:

$$\mathcal{U}_f = \frac{\sum_{k \in \{1, \dots, m\}} u_k - \sum_{k \in \{m+1, \dots, N\}} u_k}{\sum_{k \in \{1, \dots, N\}} u_k}$$

Equation 5.3 defines the expert voice, with the first term representing objective service quality and the second term representing urgency, adjusted by a normalising constant u_c .

$$n_i^e = -\frac{Q_M - \bar{Q}}{\bar{C}} + (\mathcal{U}_f * u_c) \quad (5.3)$$

Here, the term “expert” denotes agents who provide an overview of the system’s status, based on comprehensive rather than individual insights.

Additionally, the interconnected units within the regulated system may be affected by the opinions (or noise) of their peers, or they might seek out others’ opinions when they lack confidence in their own judgements (or voices), as illustrated in Nowak’s RTSI theory [12]. To capture the impact of the social network, as analysed in Section 4.3, we incorporate foreground noise into our model’s voices. This foreground noise represents the input from the social network SN_i of agent i . Specifically, each agent i is connected to at least one other agent and can be influenced by them. The social network SN is modelled as a Small-World Scale-Free [73] network, similarly to the one used in [10], and analysed in Section 4.4.

In line with the formal definition of social influence in [9] and earlier in Section 2.3, each agent i assigns a credence value $cr_{i,j}$ to each of its neighbours j within its social network SN_i . As a result, each agent i ranks its neighbours in descending order of credibility. The foreground noise reflects the voice generated by agent j , which is chosen based on the following method: agent i picks the first agent from the list CR_i , and with probability $p - \varepsilon_{net}$, it utilises its own voice to

create the foreground noise. If not, agent i selects a random agent. This randomness in agent selection prevents any agent from being continuously influenced by the same individual. Thus, the foreground noise perceived by agent i is the voice of agent j from i 's social network, represented by Equation 5.4:

$$\begin{aligned} n_i^f &= n_j^o | \max_{j \in SN_i} (cr_{i,j}), & w.p. (p - \varepsilon_{net}) \\ &= n_k^o | k \in random(SN_i), & w.p. \varepsilon_{net} \end{aligned} \quad (5.4)$$

Agents also encounter foreground noise n_i^f , which reflects peer influence. Each agent i is part of a social network SN , structured as a Small-World Scale-Free network [73]. Each agent has a degree of trust, or credence $cr_{i,j}$, toward other agents j within their social network. The foreground noise is then generated based on the most trusted agents in i 's network or a random agent, determined by probability $p - \varepsilon_{net}$.

The background noise n_i^b for agent i reflects the broader community voice, defined as the average of the own voices from a random sample of size $\nu = \frac{N}{5}$ agents, as shown in Equation 5.5.

$$n_i^b = \frac{\sum_{i \in \{1, \dots, \nu\}} n_i^o}{\nu} \quad (5.5)$$

5.2.3.2 Voice Selection for Individual & Collective Expression

Each agent i must choose among the four inner voices ($n_i^o, n_i^e, n_i^f, n_i^b$) to form its unique expression, e_i which equals one of the four voices. This choice is based on comparing the value of the coefficient of attention which is given to each voice. Specifically, the attention of each agent i to its own voice is att_i^o , to the experts' voice is att_i^e , and att_i^b and att_i^f to the background and foreground noises, respectively. During each epoch, the agent compares the values of the attention and selects the voice with the highest attention. For example, if $\max(att_i^o, att_i^e, att_i^f, att_i^b)$ is att_i^o , then agent i chooses its own voice, setting $e_i = n_i^o$.

To prevent agents from becoming biased towards a single voice early on, we introduce a stochastic parameter, similar to that used in foreground noise selection. Specifically, with probability $p - \varepsilon_{voice}$, the voice with the highest attention is chosen, while with probability ε_{voice} , a voice is chosen at random. Initially, all four attention values are set equally, ensuring that each agent starts in a neutral position, with attention levels adjusting over time based on experience.

By accessing its own voice, the foreground noise, background noise, and expert voice, each

agent i selects among these voices to shape its individual expression e_i . The collective expression, $\mathcal{V}$, aggregates the individual expressions of all agents and is given by Equation 5.6:

$$\mathcal{V} = \sum_{i \in \{1, \dots, N\}} e_i \quad (5.6)$$

5.2.3.3 Learning to Attend to Voices

5

The attention update process for each agent relies on a reinforcement coefficient c , which dictates the magnitude of positive or negative updates in each epoch. Depending on the experimental parameter *update*, an agent will either learn to focus on the voice closest to the expert opinion (*update* = ‘Exp’), the voice most similar to the group average (*update* = ‘Col’), or the voice with the most immediate positive effect (*update* = ‘Ind’), aiming for a lower own voice (indicating higher satisfaction) in the next epoch.

In the first scenario, the agent reinforces attention to the voice closest to the expert voice by increasing its attention by c and penalises the voice that deviates the most by reducing attention by c . If an expert is accessible to the agent, attention to the expert voice is increased; otherwise, one of the other voices receives increased attention. In the second case (*update* = ‘Col’), the agent increases attention to the voice that most closely aligns with the average group expression (calculated as $\frac{\mathcal{E}}{N}$) and decreases it for the voice that deviates the most. Lastly, with *update* = ‘Ind’, the agent bases its updates on comparing each voice in epoch t to its own voice in $t + 1$. The agent calculates values such as $(n_i^o)^t - (n_i^o)^{t+1}$ and similarly for other voices, reinforcing the attention for the smallest difference by c and decreasing it for the largest difference.

This updating mechanism allows agents to gradually learn which voices to prioritise based on the value of the information they provide in assessing the regulator’s actions. Through this process, agents adjust their attention in line with their chosen goal (dictated by the *update* parameter). Consequently, if agents value alignment with expert perspectives, they will focus more on the expert voice when forming their individual expression. This model echoes the confirmation bias commonly observed in human societies, where individuals often favour opinions that align with their beliefs over potentially more objective viewpoints.

5.2.3.4 Learning to Attend to Expressions

In line with the Law of Requisite Complexity [171], we design a regulator that has higher complexity than the regulated system to achieve effective regulation and maintain stability. Instead of simple reinforcement coefficients, we utilise RL, an ML mechanism introduced in Chapter 2.

Particularly, we employ a variant of Q-learning algorithm [165] with Deep Q-Learning Network (DQN) [172], which combines Double Q-Learning [84] to prevent overestimating with Dueling Networks [85] to generalise learning across different actions without changing the underlying RL algorithm. This is implemented using the `stable_baselines` library [173].

The regulator's action space, represented as an $N \times 1$ matrix, consists of a vector of all possible $m_t \in \{1, N\}$ choices. The state space is an $N \times 3$ matrix with three vectors: job size j_i for each agent, job urgency u_i , and a signal s_i indicating whether the agent's job was accepted or rejected. The reward is the negative of the collective expression, $\mathcal{R} = -\mathcal{E}$, incentivising a low collective expression (high satisfaction) and penalising a high expression (dissatisfaction). This DQN framework allows the regulator to handle greater complexity than individual agents, enabling more responsive adjustments to agent feedback.

Training runs for 10,000 epochs, divided into episodes of 2,000 epochs, with learning commencing after epoch 1,000. The exploration constant ε starts at 1 and is gradually reduced to 0.1 during training.

5.2.3.5 The *4voices* Algorithm

Based on the previous description, the psychologically inspired model is illustrated in Figure 5.3 and its operation is described in the *4voices* Algorithm 7. The parts highlighted in green pertain to the specific problem of job scheduling in the cloud, while the blue parts correspond to the four voices of the social influence in the regulated units, and the purple parts relate to the learning process.

5.2.4 Experimental Design

This Section discusses the design of our experimental framework and presents the experimental questions that the two experiments aim to answer. Those investigate the following two questions:

Algorithm 7: The *4voices* algorithm

```

 $t = 0$ ;
Set  $T = \text{max epochs}$ ;
Initialise  $N$  agents:  $\forall i \in N. \text{initialise}(e_i^o, e_i^e, e_i^f, e_i^b, att_i^o, att_i^e, att_i^b, att_i^f, j_i, u_i, c_i, q_i)$ ;
Generate Social Network  $SN$ ;
Specify  $\mathcal{P}$ ;
Initialise Collective Expression  $\mathcal{V}$ ;
while  $t < T$  do
  Regulator Specifies Policy  $\mathcal{P}$ ;
  for  $i \in N$  do
     $\perp$  Generate Job  $j_i$ ;
  Job Scheduling According to Regulator's Policy  $\mathcal{P}$ ;
  for each agent  $i, 1 \leq i \leq N$  do
    Attend Own Voice  $e_i^o$  (Equations 5.2);
    Attend Expert Voice  $e_i^e$  (Equation 5.3);
    Attend Foreground Noise  $e_i^f$  (Equation 5.4);
    Attend Background Noise  $e_i^b$  (Equation 5.5);
    Voice Selection for Expression (Section 5.2.3.2);
  Collective Expression  $\mathcal{V}$  (Equation 5.6);
  Regulator: Learns to Attend to Expression (Section 5.2.3.4);
  Regulator Changes Policy  $\mathcal{P}$  based on Collective Expression;
  for each agent  $i, 1 \leq i \leq N$  do
     $\perp$  Learn to Attend to Voices (Section 5.2.3.3);
   $t = t + 1$ ;

```

- Experiment 1: Does the *4voices* algorithm manage to establish the pathways for requisite influence, i.e. attention/expression both ways between regulator and regulated units?
- Experiment 2: Under which experimental conditions is stability maintained?

Although this section investigates a complex system in which the dynamics cannot be predicted in advance, the consideration of the theories of cybernetics [171], [174] and learning [17] enables us to form the following hypotheses:

- Hypothesis 1: The pathways for requisite influence are established when the regulated units prioritise the collective good (i.e. ‘Exp’, ‘Col’) rather than their own interests (i.e. ‘Ind’).
- Hypothesis 2: Systemic stability is maintained if the pathways for requisite influence are established and the regulator has higher complexity than the regulated units (‘RL’). The latter is expected because of based on the Laws of Requisite Variety and Complexity [174],[171].

To explore whether our hypotheses are validated, we follow an ABMS approach to design a multi-agent simulator, with reasoning and learning capabilities, according to the SIC methodology. In particular, in order to conduct those experiments we first have to finetune the system and set the independent parameters to the values that can allow observing emerging behaviours in a relatively

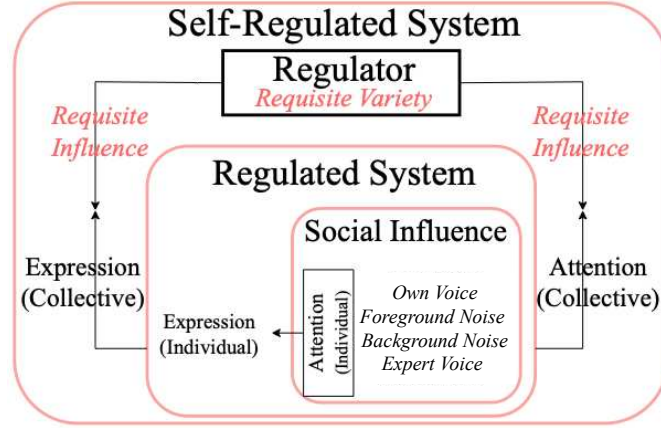


Figure 5.3: The psychologically inspired model

short amount of time. Hence, we set $\varepsilon_{net} = \varepsilon_{voice} = 0.3$, $c = 0.01$, $u_c = 10$, $J = U = 10$, $F = 500$ and $N = 100$, $\mu = 0.5$, $m_{SN} = \frac{N}{10}$. Also, for the initialisation of the system, all the voices $n_i^o, n_i^e, n_i^f, n_i^b$ for all the agents are set to the value of 0, the credence CR_i to all the agents from the social network is set to 0.5, and the size of jobs j_i and the urgency of those u_i are randomly selected for each agent. Moreover, we vary the experimental parameters and observe if a set of initial conditions is sufficient to establish the pathways for requisite influence, or leads to systemic stability. The experiments were conducted multiple times, and the graphs display the results averaged over 10 runs. In contrast, the second experiment shows the results from a single representative run to emphasise the emerging behaviour.

5.2.5 Experimental Results

5.2.5.1 Experiment 1: Pathways for Requisite Influence

This experiment aims to observe whether the *4voices* algorithm creates pathways for necessary influence. Figure 5.4 displays the average and standard deviation of the individual expressions of the agents (on the y-axis) across different epochs (on the x-axis), for various types of attention updates shown in different columns, as well as for both a random and a reinforcement learning (RL) regulator, presented in separate rows. To complement the individual perspectives shown in Figure 5.4, Figure 5.5 offers a collective perspective, specifically illustrating the policy chosen (i.e. the number of agents to be included in the next calculation) by the regulator on the y-axis for each epoch on the x-axis. The colours of the dots indicate the volume of noise, with blue representing relatively low volume and red indicating comparatively high volume.

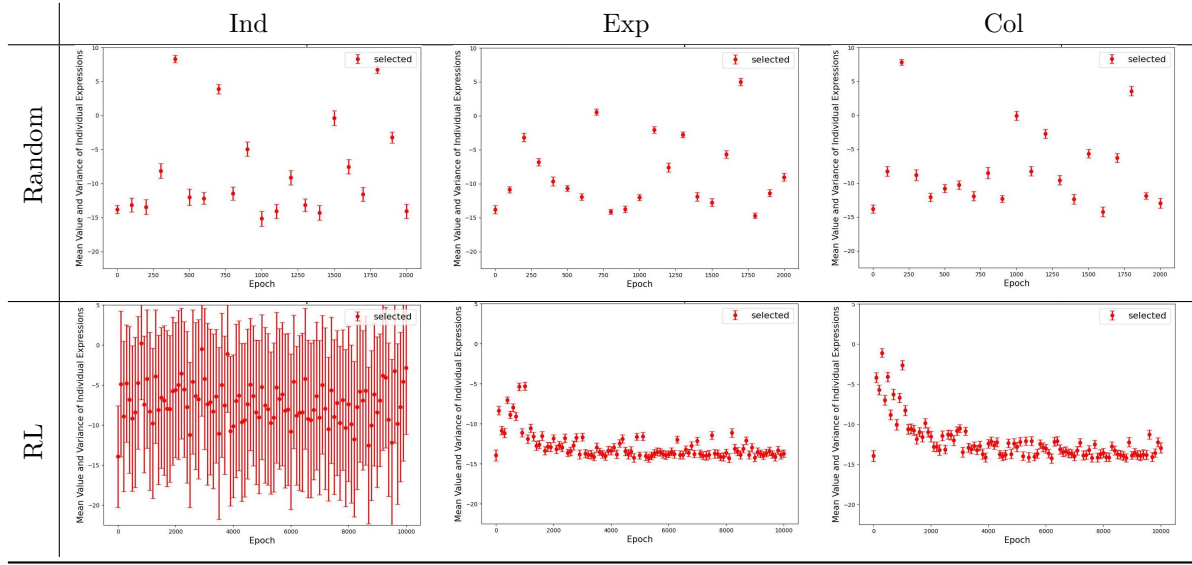


Figure 5.4: The mean value and the standard deviation of the individual expressions of the regulated units under different initial conditions depicting that requisite influence, corresponding to comparatively low expression, is established when the regulator determines the policy using RL - Results averaged over 10 runs

From Figure 5.4, it is evident that when the regulator randomly selects a policy, the agents' noise varies from epoch to epoch, reflecting their responses to the chosen policy. In contrast, when the regulator determines the policy using a reinforcement learning algorithm, the mean value of the agents' noise decreases for the 'Exp' and 'Col' types of attention updates. This indicates that the agents are responsive to the regulator's expressions and effectively adjust both their individual and collective expressions based on whether the selected policy is favourable or not.

The same trend can be observed in Figure 5.5, which shows the collective expression related to the regulator's chosen policy. Here, we see that when the policy includes many agents (more than 40), the agents' noise decreases. Conversely, when the regulator selects only a few agents, the group expresses dissatisfaction by increasing the noise volume. Additionally, when the agents' attention updates are based on 'Exp' or 'Col', their expressions reflect their satisfaction or dissatisfaction, creating pathways for necessary influence from the regulated units back to the regulator. This is evident in the third and fourth columns of Figure 5.5, where the regulator identifies a range of policies preferred by the agents and, after several iterations, selects policies only from the 50-80 agents interval. This is because when the service accepts from 50 and up to 80 agents, it can maintain a high quality for a 'reasonable' cost. Accepting more than 80 would diminish the quality of service, while accepting less than 50 would result in a significant cost.

Overall, these observations indicate that *pathways for requisite influence from the regulator to*

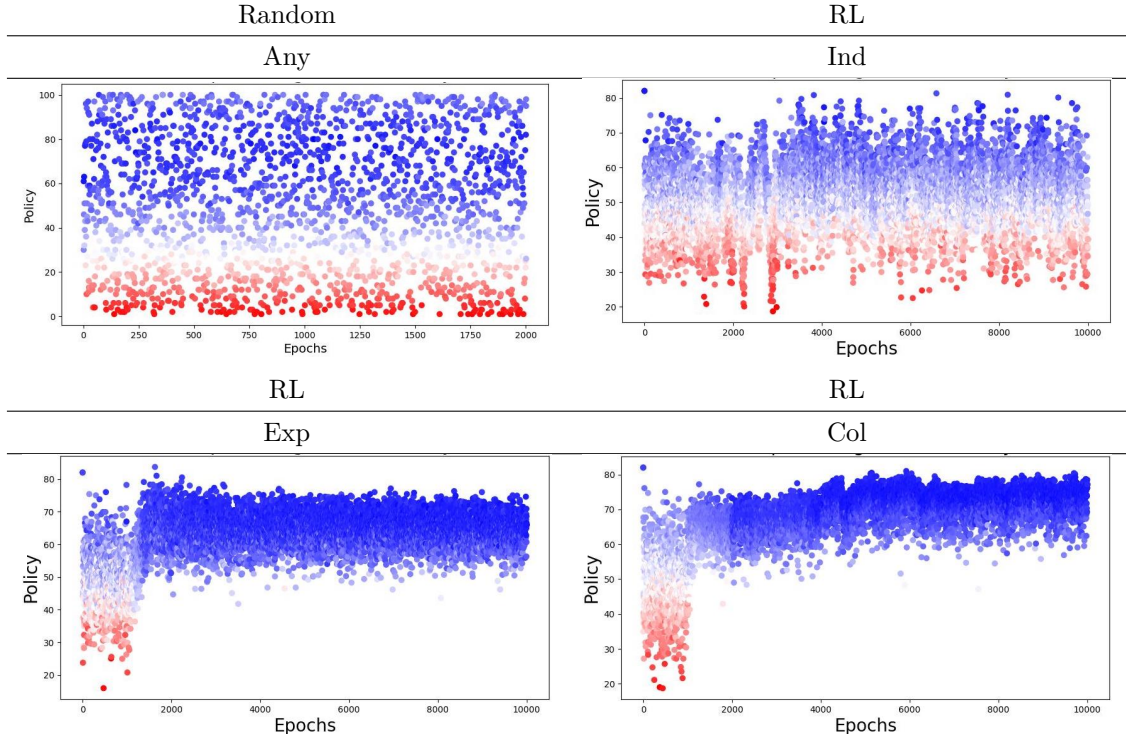


Figure 5.5: Collective expression of the regulated units with respect to the choice of policy of the regulator under different initial conditions depicting that requisite influence, corresponding to comparatively low expression (blue), is established when the regulator determines the policy using RL - Results averaged over 10 runs

the regulated units are established regardless of the conditions, while pathways from the regulated units back to the regulator are only established for the ‘Exp’ and ‘Col’ types of updates. This is significant as it allows the collective (the regulated system) to influence the regulator’s learning algorithm by providing data that shapes policies to help them achieve their goal of maximising collective satisfaction, similar to the findings in [175].

5.2.5.2 Experiment 2: Initial Conditions Resulting in Systemic Stability

Additionally, examining the graphs from a single run in Figures 5.6 and 5.7, we observe a trade-off between diversity and congruence. Specifically, as we transition from the individual voice and foreground noise (selected during the ‘Ind’ update) to the expert voice and background noise (chosen when the attention update is set to ‘Exp’ or ‘Col’), the standard deviation of individual expressions decreases, indicating lower diversity. Furthermore, the results in the second row of Figure 5.7 reveal that the experts’ noise remains relatively stable as long as the selected policy falls within the 40-80 agents range. In contrast, in the third row, we observe a limited range of satisfactory policies in the ‘Col’ case, which is associated with the network properties and the

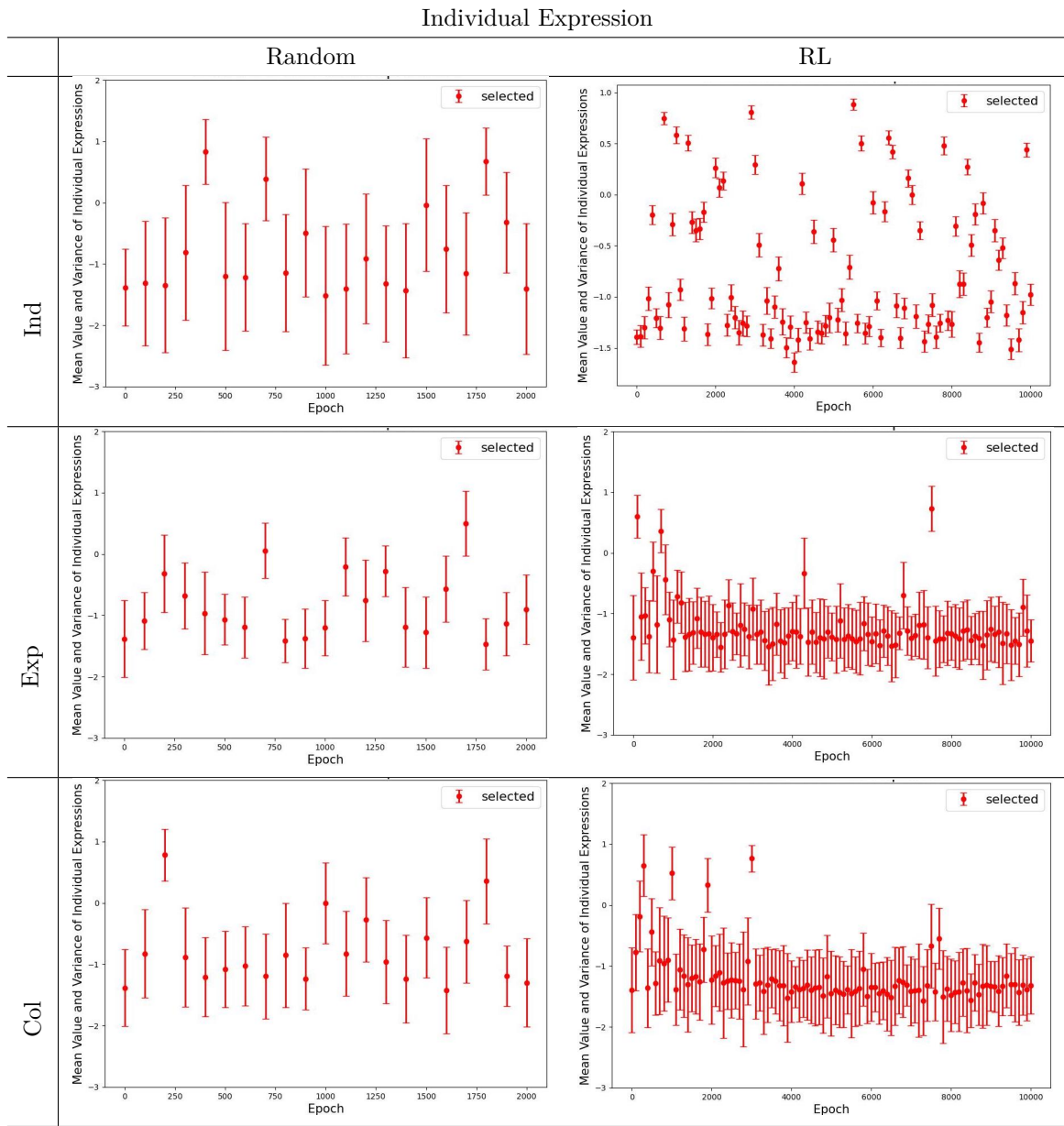


Figure 5.6: The mean value and the standard deviation of the individual expressions under different initial conditions showing that stability is maintained when the regulator determines the policy using RL and the regulated units aim to learn to attend the experts voice or the background noise (i.e. voice of the community) - Results of a single run

agents' influence on each other's expressions.

This experiment aims to determine which initial conditions lead to systemic stability, a vital factor in addressing sustainability issues in socio-technical systems. Combining the insights from both experiments, we conclude that *systemic stability is achieved when the regulator uses reinforcement learning to select policies and the agents update their attention based on the divergence from the experts' voice or collective decision* (when the update is 'Exp' or 'Col'). Referring back to the

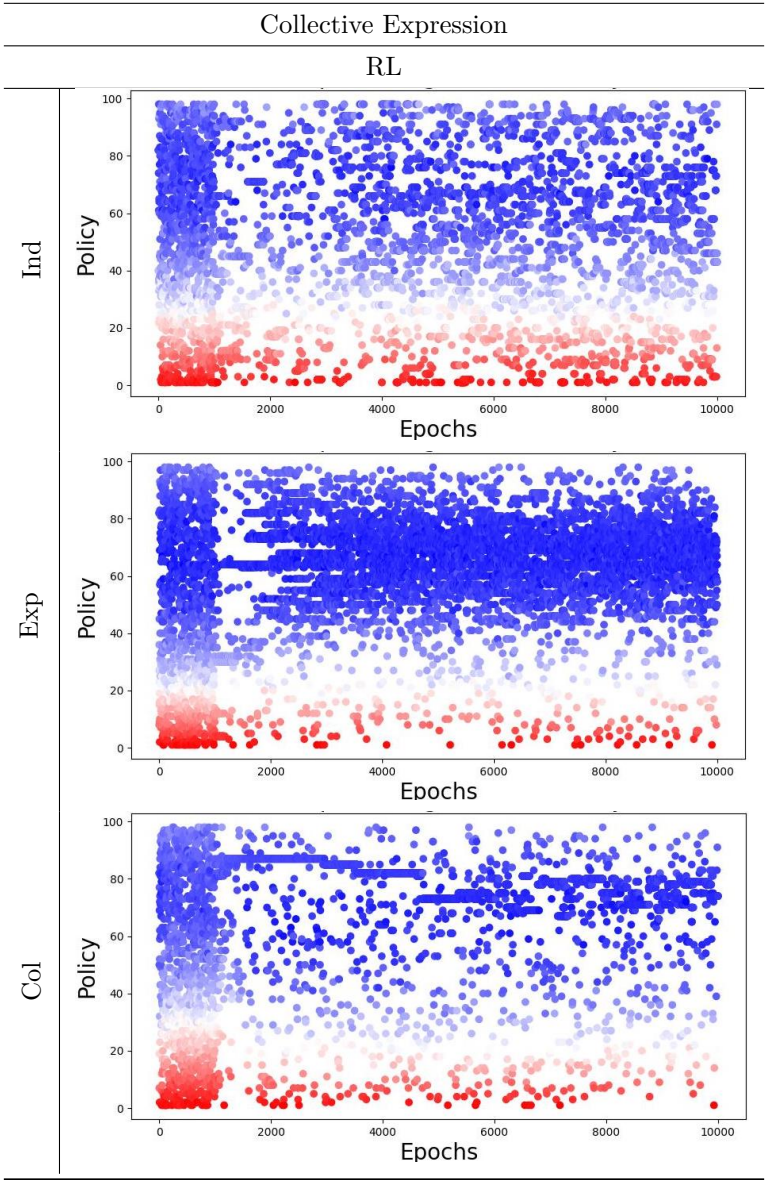


Figure 5.7: The policy selection and the corresponding collective expression under different initial conditions showing that stability is maintained when the regulator determines the policy using RL and the regulated units aim to learn to attend the experts voice or the background noise (i.e. voice of the community) - Results of a single run

third and fourth columns of Figure 5.5, we note that after sufficient iterations, the regulator identifies the suitable range of policies that result in comparatively low noise levels. This contributes to systemic stability, as indicated by the consistently low noise volume represented by the blue colour.

5.2.6 Conclusions on Effective Self-Regulation

Overall, this Section addressed two related problems in self-regulated systems. Firstly, ensuring that pathways for requisite influence exist, in the form of awareness (of the regulator) to the expression (of the regulated units), in the expression (by the regulator) and attention (by the regulated units) to that expression, and in the form of social influence in the social network of the regulated units. Secondly, to identify under which initial conditions the self-regulated system can maintain stability, and so establish what could be called *requisite social influence*. Moreover, our hypotheses were validated by the experimental findings, and are also aligned with the Laws of Requisite Variety and Complexity [171], [174].

To address those issues, we synthesised ideas from psychoacoustics, cybernetics, and ML, in an HyRL framework, to address the dual problems of requisite influence and stability in continuous monitoring and control of a dynamic and non-deterministic system. Specifically, the contributions were:

- based on ideas from opinion formation, social psychology and psychoacoustics, to introduce a model comprising four voices of social influence in the regulated units, which are the own voice, expert voice, foreground noise and background noise; social influence;
- to specify the *4voices* algorithm, which combined computation of a signal value for each of the voices, with reinforcement of attention based on the experience from past interactions with the voices for the regulated units to learn which voice to pay attention to, and Deep Q-learning for the regulator to learn to attend the expressions of the regulated units and adjust its actions (corresponding to policy selections); and
- experimental results which show that for control processes in dynamic and non-deterministic systems, the *4voices* algorithm establishes both the required relational complexity and the pathways for requisite influence, so that systemic stability is maintained,
- to provide a psychologically inspired mechanism for maintaining sustainability, effectively managing knowledge, irresistibly modifying rules, avoiding voting paradoxes and oligarchic tendencies, and maintaining dignity, which all correspond to known issues of self-governance.

Elaborating on the last contribution that relates the *4voices* algorithm with common problems of self-governance, first, it supports sustainable self-regulation (or, in other words, self-governance) though the dynamic adaptation of policy based on the collective opinion on that policy. Also, it

facilitates knowledge management because agents learn how to manage knowledge expressed by each voice, and select the one that seems more suitable for informing their expression. At the same time, the regulator manages to use the knowledge expressed by the units and enhance its own knowledge such that the policy is modified in the desired way. Furthermore, the fact that no matter what policy is chosen by the regulator, it can change based on the expressions of the units corresponds to an instance of a mechanism that allows unrestricted self-modification of rules.

On the voting paradoxes, the process does not include any voting mechanisms but allows to units to form their collective expression, and hence, it is not prone to be subject to such paradoxes. Additionally, the fact that the regulated units, each individually and all as a collective, express their satisfaction with respect to the regulating mechanisms, and this is attended by the regulator, is a step towards the avoidance of tyranny and oligarchic tendencies. Finally, given that the individuals can make informed decisions (by attending to the voices) on the effectiveness of the policy, can express their opinion, and this expression is attended and results in a new policy, gives them the power to control the states of affairs that directly affect them. In particular, if this approach is followed in a system where the agents of the regulated unit are humans, and the regulator can be another human being or an AI, the fact that they can express their decision, and that expression is attended, is an evidence of preservation of human dignity.

Beyond this, according to Ashby [19], an ethical regulator is required not only to reach reliable decisions from potentially unreliable evidence, but also to evaluate the consequences of its decisions, which raises issues of transparency and accountability in the regulator. Correspondingly, though, an ethically regulated unit should try to provide information to the best of its knowledge and belief. The future challenge lies in ensuring ethical behaviour with respect to *values* and innovating those values, especially in systems with multiple stakeholders with different priorities and preferences with respect to those values. Such issues of innovation of values and rules would be analysed further in Chapter 6.

5.3 Sustainable Consensus Formation

5.3.1 Motivation

A fundamental problem of SGSs is to reach agreements on the values, rules, and structures, (i.e. corresponding to social arrangements) and to maintain those agreements. Social choice theory [176] is concerned with investigating how individual preferences can be mapped into collective decisions,

and the study of it has resulted in multiple different voting mechanisms, such as plurality, runoff, Borda count, instant runoff, and approval. Those mechanisms expose different characteristics (such as universality, anonymity, neutrality or positive responsiveness) that might make them more effective under different conditions, and they can produce different results (even from the same votes). However, we need a mechanism for making sustainable sequential agreements between a group of individuals, and we propose to use past interactions (corresponding to votes) to achieve this.

5

Hence, we propose to draw inspiration from classical Athenian democracy, and specifically a process called *thorybos*, to detect emergent consensus and design a mechanism for sustainable collective self-governance through meaningful democratic deliberation. In this process, citizens of Athens were expressing vocally their opinion to policy options, and the moderators of the process were trying to identify whether an agreement was reached. Based on that process, we design a mechanism that leverages past interactions of agents to enable agents to identify when to consent and when to express dissent. This is achieved following an HyRL methodology and targets some of the main challenges of self-governance that were introduced in Section 2.2.4, namely legitimate consent, legitimate dissent, knowledge management and sustainability.

Grounding those to the example of the co-housing, residents might attend a new policy for distributing energy between households and common areas, and they can vocally express their opinions on that policy. For instance, one might propose a policy that distributes resources proportionally to the needs. However, many individuals might react negatively by object, loudly. Then another might suggest to distribute resources based on priority (e.g. who asked for energy first, second, and so on). This might result in an even louder objection, which would require another policy to be expressed, since an agreement is not reached. The new policy can then be to distribute resources based on the contribution of an individual to the community. That policy might not trigger such a loud negative reaction in the community and therefore this policy would be selected for application. However, even if that policy is accepted, the collective should be aware of, and remember, the compromises made by some individuals (who were not that satisfied by that policy), which would be translated to some form of social capital [177] that those individuals can use for future negotiation on policies.

5.3.2 Political Sciences Insights

This Section borrows concepts from political sciences to design HyRL systems with the capability to reason and learn in order to form collective agreements in the context of founding a self-governing institution. Hence, this presents some insights from political sciences concerning institutional foundation, consensus, and dissent, which were also analysed in Section 2.2.

According to Ober’s thought experiment, called *Demopolis*, which was developed to explore Basic Democracy theory [42], inter alia, outlines a three-stage process for establishing a self-governing institution. This involves agreeing on the collective’s goals, establishing core rules for participation, legislation, and protection of these rules, and finally, operationalising these foundations through rule modification, selection, and enforcement. Additionally, Ostrom’s theory [38] is concerned with the sustainable self-governance in the context of Common Pool Resource (CPR) management.

In both studies, there is a need for having a mechanism for reaching agreements on matters of self-governance. According to Canevaro[20], deliberative assemblies in Classical Athens defined a structured process for building consensus through meaningful deliberation, known as *thorybos* [20]. During *thorybos*, *speakers* presented policy ideas and rules to the *demos* (the citizen assembly), while the *proedroi* (meeting leaders) encouraged the audience to express their approval or disapproval vocally. This process allowed *proedroi* and *speakers* to gauge whether consensus had been reached, helping to guide the assembly’s next steps.

5.3.3 Problem Specification

The study combines ideas from consensus formation and RL to address the challenge of establishing agreements for founding and self-governing an institution. In particular, the previous Section discussed the need for forming collective agreements with respect to institutional goals, and proposed *thorybos* as an effective mechanism for achieving that. Here, we combine the concept of *thorybos* with MARL, within an HyRL framework, to *provide a general mechanism for reaching agreements and deliberating on the process of forming agreements itself*.

Note that the goal of this research is not to identify the best approach or specific policies. In our experiments, we simply refer to “the policy” without specifying its content. We conceptualise different policies as points in a space that may change over time and design a mechanism that operates independently of the specific policy, making it adaptable across various applications and contexts. Ultimately, this work aims to propose a mechanism for learning how to reach and

maintain agreements on policies over time, irrespective of the policies' specific content.

5.3.4 Θ -learning Algorithm for Sustainable Consensus Formation

We introduce the Θ -learning algorithm, which combines two key components: *thorybos* and Q-learning. This corresponds to another socially-inspired agent-based model of a hybrid system comprising a machine reasoning (*thorybos*) and a ML (Q-Learning) component. Specifically, in each epoch of the Θ -learning algorithm, the group engages in *thorybos* (a process for consensus formation and collective decision-making) and a learning phase (to deliberate on the consensus-building process itself). As such, next we provide the formal model of *thorybos* (represented as “ Θ ” in Θ -learning) and multi-agent Q-learning (corresponding to “Learning” in Θ -learning), along with the pseudocode and formal specification of the algorithm.

5.3.4.1 Thorybos

Inspired by Demopolis and decision-making practices in Classical Athens [39], we present a formal specification of *thorybos*. In this context, *thorybos* represents the vocal reactions (such as laughter, cheers, and heckling) of citizens to an idea or policy, intended to make the *speaker(s)* (those presenting the idea) and *proedroi* aware of the community's stance. Based on this response, corresponding to the generated noise, an agreement may be reached.

The first component of the Θ -learning algorithm centres on this process of *thorybos*. In each epoch, t , every agent proposes a policy v from a defined specification space M , with a dimension of N_m . The agents' proposals are then aggregated, and the policy with the highest support (most votes) becomes the collective decision. Consequently, a *thorybos* Θ is produced in each epoch t , reflecting the index of the agreed policy $\mathcal{M}_t$ and the number of agents backing it $\mathcal{V}_t$. The value of $\mathcal{V}_t$ determines whether an agreement is achieved.

Additionally, agents hold specific policy preferences, so each agent i has a preference $pr_{i,v}$ for policy v . These preferences can either remain constant throughout the process or vary every $\frac{T}{c_{pr}}$ epochs, where c_{pr} is a tunable parameter. Preferences are sampled randomly from the interval $[0, N_{pr}]$, with 0 indicating no preference and N_{pr} representing the maximum possible preference value, which is externally defined.

It is worth noting that in contrast to the problems investigated in the previous Chapter, in which the policy was formed by aggregating multiple rules with different weights, here the policy

corresponds to a single rule, and agents have preferences on those rules. For instance, if we investigate a problem of common pool resource distribution and we consider two rules, in the first case, multiple rules are being used to decide on the distribution, but they affect in different amount (depending on their weight - e.g. if a rule has weight 0.2 and another 0.8, the latter affects the policy more). In that case the rule weights have to sum to one. In the second case that we use a single rule for the policy, and agents have preferences on those, the preferences to rules can be any number from 0 to 1 and an agent can even have a value of preference equal to 1 to all the rules, which means that the agent is equally satisfied with the application of any of these rules. Those preferences affect the compromises that agents do and therefore a policy is determined considering the level of compromise of each agent.

5.3.4.2 Q-Learning

The learning component of this framework aims to help agents learn how to respond to proposed policies based on their preferences and the *thorybos* generated in the prior epoch. This environment is modelled as a multi-agent Markov Decision Process (MA-MDP) with N agents participating in the Markov Game. Here, all agents observe the *thorybos* from the previous epoch at each new epoch t , so the observation for each agent $i \in A$ is $o_t \equiv s_t \equiv s_t^i \equiv s_t^j \equiv \Theta_{t-1}$. Agents then choose from the N_m possible policies in the action space M . The agents use an ε -greedy policy, so in each epoch t , agent i selects an action $a_t^i = m$, with $m \in M$, choosing the action that maximises Q with probability $1 - \varepsilon$, or a random action with probability ε :

$$a_t^i = \begin{cases} \operatorname{argmax}_{a_t^i} Q(\Theta_{t-1}, a_t^i), & \text{w.p. } 1 - \varepsilon \\ a_t^i \sim \operatorname{Uniform}(M), & \text{w.p. } \varepsilon \end{cases} \quad (5.7)$$

where $a_t^i \sim \operatorname{Uniform}(M)$ denotes a randomly sampled action from M .

Based on the action chosen, preferences, and an experimental parameter $\mathcal{L}$ (which reflects the learning goal in the reward function $\mathcal{R}$), agent i receives a reward r_t^i and updates its Q -table using Q-learning.

5.3.4.3 Θ -learning Algorithm

The proposed model, which follows a hybrid reasoning and learning approach, is illustrated in Figure 5.8. The left side of the figure represents the Θ -learning process, with the grey part

indicating the algorithm's machine reasoning component (*thorybos*) and the red part representing the machine learning component (Q-learning). The right side of Figure 5.8 provides an abstraction of Θ -learning, depicting it as a sequential process of collective decision-making and deliberation. With the specifications of *thorybos* and the formal reinforcement learning model outlined, we can

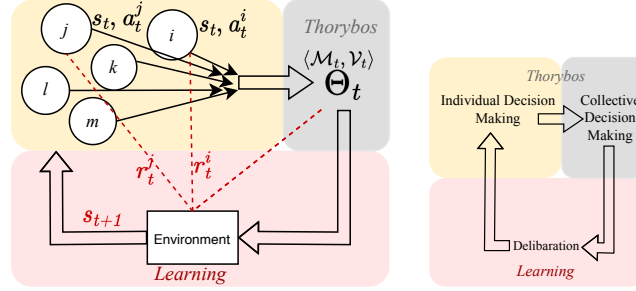


Figure 5.8: Θ -learning algorithm (left) and abstraction of the process (right)

now examine the reward function. This reward function is designed to incentivise actions that either support the common good (promoting consensus), minimise the compromise required by the agent (preserving well-being and diversity), or both. It is composed of two terms: the first term reflects the agent's contribution to the collective, while the second ensures diversity by discouraging suppression of dissent.

The first term is based on the absolute difference between the number of agents supporting the policy m that agent i chose (i.e. those selecting action $a_t^i = m$) and the votes for policy v , where v is the policy with the highest support, $\mathcal{M}_t$, having received $\mathcal{V}_t$ votes. The second term measures the difference between agent i 's preference for policy m and their highest preference, $\max_v pr_{i,v}$.

To analyse the effects of collective contribution and individual compromise, each part of the reward function is weighted by coefficients c_{col} and c_{ind} , which vary across experiments but always sum to one. In experiments reinforcing only individual rationality (Individual Setting), the collective term is deactivated by setting c_{col} to 0. For the Collective Setting, the individual term is deactivated with c_{ind} set to 0, while the Mixed Setting activates both terms with the coefficients normalised to 1. Thus, the reward for agent i at time t is given by:

$$r_t^i = c_{ind} * |pr_{i,m} - \max_v pr_{i,v}| + c_{col} * |Votes_m - \mathcal{V}| \quad (5.8)$$

The Θ -learning algorithm is detailed in Algorithm 8. In summary, Θ -learning integrates a collective decision-making process (through the *thorybos* mechanism) with a learning process (via Q-learning) to support agreement-reaching in institutional foundations while balancing individual and collective rationality [178]. To examine how different core values, established by the

institution's participants in its foundational stage, influence decision outcomes, we define three reward function settings as experimental parameters, namely *Individual*, *Collective*, and *Mixed*. We conduct experiments under the *Individual*, where agents prioritise individual rationality and minimising compromises; the *Collective*, aimed at fostering collective rationality and consensus; and the *Mixed*, which combines both goals.

Algorithm 8: Θ -learning Algorithm

```

 $t = 0$ 
Set  $T = \text{max epochs}$ 
Specify  $M, N_{pr}$ 
Initialise agents  $N: \forall i \in N. \forall v \in M. \text{initialise}(pr_{i,v} \sim \text{Uniform}(N_{pr}))$ 
Specify experimental parameters  $\gamma, \varepsilon, c_{col}, c_{ind}$ 
Initialise  $Q$  table:  $\forall Q_{a,b} \in \{\text{state}, \text{action}\}. \text{set } Q_{a,b} = 0$ 
Set initial states  $s_0$  of agents
Specify initial actions  $a_0$ 
 $\Theta_0 = \langle 0, \text{Uniform}(M) \rangle$ 
while  $t < T$  do
  for  $i \in N$  do
    if  $t \bmod \frac{T}{c_{pr}} == 0$  // Modify agents' preferences
      then
        for  $v \in M$  do
           $pr_{i,v} \sim \text{Uniform}(N_{pr})$ 
     $a_t^i =$  // Define agents' actions
     $\begin{cases} \text{argmax}_a Q(\Theta_{t-1}, a_t^i), & \text{w.p. } 1 - \varepsilon \\ a_t^i \sim \text{Uniform}(M), & \text{w.p. } \varepsilon \end{cases}$ 
  for  $v \in M$  // aggregate votes on policy
  do
     $Votes_v = \sum_{i \in A} (a_t^i = v)$ 
   $\mathcal{M}_t = \text{argmax}_v Votes_v$  // Identify agreed policy
   $\mathcal{V}_t = \max_v Votes_v$  // Measure amount of individuals agreeing on that policy
   $\Theta_t = \langle \mathcal{M}_t, \mathcal{V}_t \rangle$  // Specify thorybos
   $t = t + 1$ 

```

5.3.5 Experimental Design

Our objective is to determine, through these experiments, whether Θ -learning can effectively address challenges in collective decision-making, institutional foundation, and sustainable self-governance. To analyse system behaviour across different scenarios, such as varying learning objectives and population types, the socially-inspired Θ -learning algorithm is implemented in a multi-agent simulator. The proposed system corresponds to an HyRL-based system that encompasses MR (to approximate the process of *thorybos*) and ML (to enable learning from past decisions and monitoring compromises) capabilities, and specifically MARL.

The implementation is performed using Python. In this setup, a population of $N = 100$ agents,

each functioning as an RL agent, participates in each epoch. Each agent observes the environment, updates its state based on the prior epoch's *thorybos*, and selects an action. The agent then receives a reward and updates its Q-table based on its chosen action, the environment's current *thorybos*, and the learning objective (defined by the agreed-upon core values, represented by the experimental parameter $\mathcal{L}$). The agents' goal is to adapt to *thorybos* and choose actions that maintain it at a low level, promoting collective agreement while minimising compromises.

In more detail, agents at each time t observe the true environment state, where $o_t^i \equiv s_t \equiv s_t^i \equiv s_t^j = \Theta_{t-1}$, and Θ_{t-1} belongs to $\mathbb{R}^{N_m \times N} = \mathbb{R}^{10 \times 100}$. Here, Θ combines $\mathcal{V}$ (vote count for the winning policy) and $\mathcal{M}$ (winning policy index). Agents can choose from N_m policies, each with preference values between 0 and $N_{pr} = 10$.

Actions taken by agent i in epoch t , denoted as $a_t^i \in \mathbb{R}^{N_m} = \mathbb{R}^{10}$, represent policy indices. During training, agents follow an ϵ -greedy policy with a discount factor γ set to 0, focusing solely on current rewards, and a learning rate α set to 0.2 for faster learning. To ensure adaptability in a dynamic environment (as agents' preferences shift) and to prevent stagnation, Θ -learning is evaluated continuously throughout training. Table 5.1 summarises the simulator's independent variables. In these experiments, the term 'consensus' (represented by the red line) refers to the

Variable	Symbol	Value
<i>Environmental Parameters</i>		
Epochs:	T	150000, 250000, 500000
Agents	N	100
Available Policies:	N_m	[0,10]
Preferences Interval:	N_{pr}	[0,10]
<i>RL Parameters</i>		
State Space:	$o_t^i \equiv s_t \equiv s_t^i \equiv s_t^j = \Theta_{t-1}$	$\mathbb{R}^{N_m \times N} = \mathbb{R}^{10 \times 100}$
Action Space:	$a_t^i \equiv N_m$	$\mathbb{R}^{N_m} = \mathbb{R}^{10}$
Discount Factor:	γ	0
Epsilon:	ϵ	0.2
Learning Rate:	α	0.2

Table 5.1: Independent Variable

proportion of the population proposing the winning policy at epoch t , denoted as $\mathcal{V}_t$. 'Diversity' (black line) reflects the degree of dissent or disagreement, defined as the average number of agents proposing different policies, calculated by:

$$\frac{\sum_{i=1}^N \sum_{j=1}^N 1(v_i \neq v_j)}{N^2 - N}$$

where v_i and v_j denote the policies proposed by agents i and j (i.e. $a_t^i = v_i$ and $a_t^j = v_j$).

‘Agreement’ (yellow line) represents the average number of agents sharing the same opinion, essentially the complement of diversity, and together they add up to 1.

Lastly, ‘compromise’ measures the average of the absolute difference between each agent’s preference for their favoured policy and the policy ultimately selected, computed as:

$$\mathcal{C}_i = |pr_{i,m} - \max_v pr_{i,v}|$$

The experiments are organised as follows:

- Experiment 1: Examines the behaviour of static populations with distinct learning objectives, reflecting varied core values.
- Experiment 2: Analyses the nature of agreement and the level of compromise within dynamic populations with different learning objectives.
- Experiment 3: Studies how varying rates of preference change among agents impact collective agreement, with core values fixed to the *Mixed* setting (a combination of collective and individual objectives).

The first two experiments assess three types of learning objectives, while the third experiment evaluates six different rates of preference change across the population. The values for the independent experimental parameters used are detailed in Table 5.2.

We execute these experiments in our multi-agent simulator in order to test whether our hypotheses are satisfied. Those correspond to:

- Hypothesis 1: Based on the results from the experimental findings of the experiments in Section 4.6 (concerning the need for diversity for self-improvement), we expect that the static population (i.e. no change of preferences) would stagnate.
- Hypothesis 2: We expect that the learning objectives of the agents (reflecting whether they prioritise individual or collective good) would affect the process of opinion formation, and specifically when the population comprises purely individualist agents (*Individual*) no agreement would be reached, while when it comprises collectivists a consensus would be reached but there would be an increased level of compromises.

Experiment	Type of Population	c_{pr}	Objective	(c_{ind}, c_{col})	Epochs
1	Static	0	<i>Individual</i>	(1,0)	250000
1	Static	0	<i>Collective</i>	(0,1)	250000
1	Static	0	<i>Mixed</i>	(0.5,0.5)	250000
2	Dynamic	5	<i>Individual</i>	(1,0)	250000
2	Dynamic	5	<i>Collective</i>	(0,1)	250000
2	Dynamic	5	<i>Mixed</i>	(0.5,0.5)	250000
3	Dynamic	5	<i>Mixed</i>	(0.5,0.5)	500000
3	Dynamic	10	<i>Mixed</i>	(0.5,0.5)	500000
3	Dynamic	20	<i>Mixed</i>	(0.5,0.5)	500000
3	Dynamic	32	<i>Mixed</i>	(0.5,0.5)	500000
3	Dynamic	50	<i>Mixed</i>	(0.5,0.5)	500000
3	Dynamic	100	<i>Mixed</i>	(0.5,0.5)	500000

Table 5.2: Experimental Parameters

- Hypothesis 3: Since the collective state takes more time to change, we expect that when changes of preferences are very quick, the group would prioritise following the collective and reaching consensus.

5.3.6 Experimental Results

This Section outlines the experimental findings from three experiments aimed at assessing the effectiveness of Θ -learning in establishing and maintaining self-governing institutions. Specifically, these experiments investigate whether agents can reach collective agreement, preserve diversity, and thus minimise compromise across different population types (defined by objectives, such as core values, and the frequency of change).

5.3.6.1 Experiment 1: Static Population

This experiment investigates the behaviour of static populations, where agents do not alter their preferences. Figure 5.9 illustrates how different learning objectives—namely *Individual*, *Collective*, and *Mixed* (shown in separate rows)—influence the number of agents achieving a general agreement (first column) and the compromises they are required to make to attain that agreement (second column). As previously described, the first row reflects the values of ‘consensus’, ‘diversity’, and ‘agreement’, while the second row depicts the concept of ‘compromise’.

Examining the different columns, when the goal is to adhere to individual rationality (row

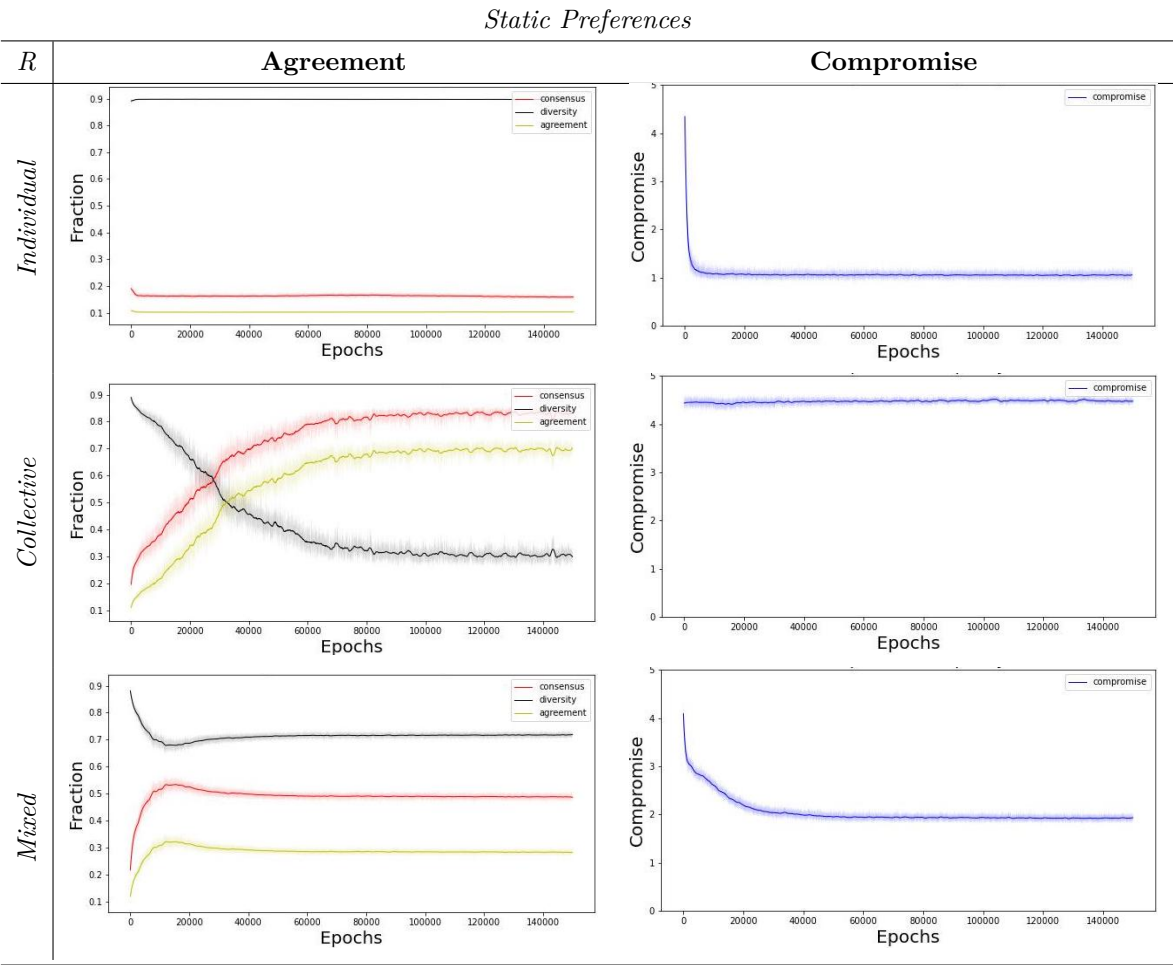


Figure 5.9: Experiment 1: Fraction of population reaching an agreement and compromises made by individuals for different conditions showing a strong relation between learning objective and level of agreement, which is persistent in static populations

1), the decisions (policy proposals) made by the agents diverge, resulting in maximal dissent and an inability for the group to reach an agreement, with only 20% of the agents supporting the maximum consensus achievable. Moreover, agents make minimal compromises, as indicated in column 2, since they select policies solely based on their personal preferences. In contrast, when the agents' motivation is to achieve a general agreement regardless of compromises (row 2), they establish a persistent consensus, as over 80% of agents concur on the same policy (red line in column 1 row 2), although they must make significantly more compromises, which could lead to diminished diversity and instability in the long term (row 2 column 2). When the objective is to balance contributing to a collective agreement while preserving diversity and limiting individual compromises (row 3), a majority achieves agreement (i.e. 50% of agents support the same policy) (red line in column 1 row 3), while making fewer compromises (blue line in column 2 row 3).

Thus, we conclude that *the agents' objectives*, determined by their reward function, *are strongly linked to their ability to achieve a general agreement and the compromises they make*. Nevertheless, regardless of the learning objective, *static populations tend to stagnate*, as they settle into a particular state (e.g., a state of disagreement, general agreement, or majority agreement) after a few iterations due to a lack of diversity over time.

5.3.6.2 Experiment 2: Dynamic Population

This experiment examines the behaviour of dynamic populations regarding agreement (i.e. general agreement, majority agreement, disagreement) and compromises under various learning objectives. Specifically, the nature of agreement can be analysed from two perspectives: an external viewpoint, which assesses the proportion of the population that agrees on the same policy (denoted by the term 'consensus' introduced in the previous Section), and an internal viewpoint, which reflects the agents' perspectives regarding the index of the chosen policy. We present graphs illustrating both viewpoints to provide a comprehensive understanding of the outcomes.

Accordingly, Figure 5.10 displays the proportion of the population achieving consensus on the policy (column 1) and the compromises required by agents to reach that consensus (column 2), similar to Figure 5.9. To support the findings in Figure 5.10, Figure 5.11 presents the index of the policy selected by each agent across different epochs, specifically at the start of the process (column 1) and as it nears completion (column 2) for the different objectives (shown in separate rows). The various colours in the graphs correspond to different policy indices, with dark blue representing the policy with index 0, orange for index 1, and so forth. It is noteworthy that the

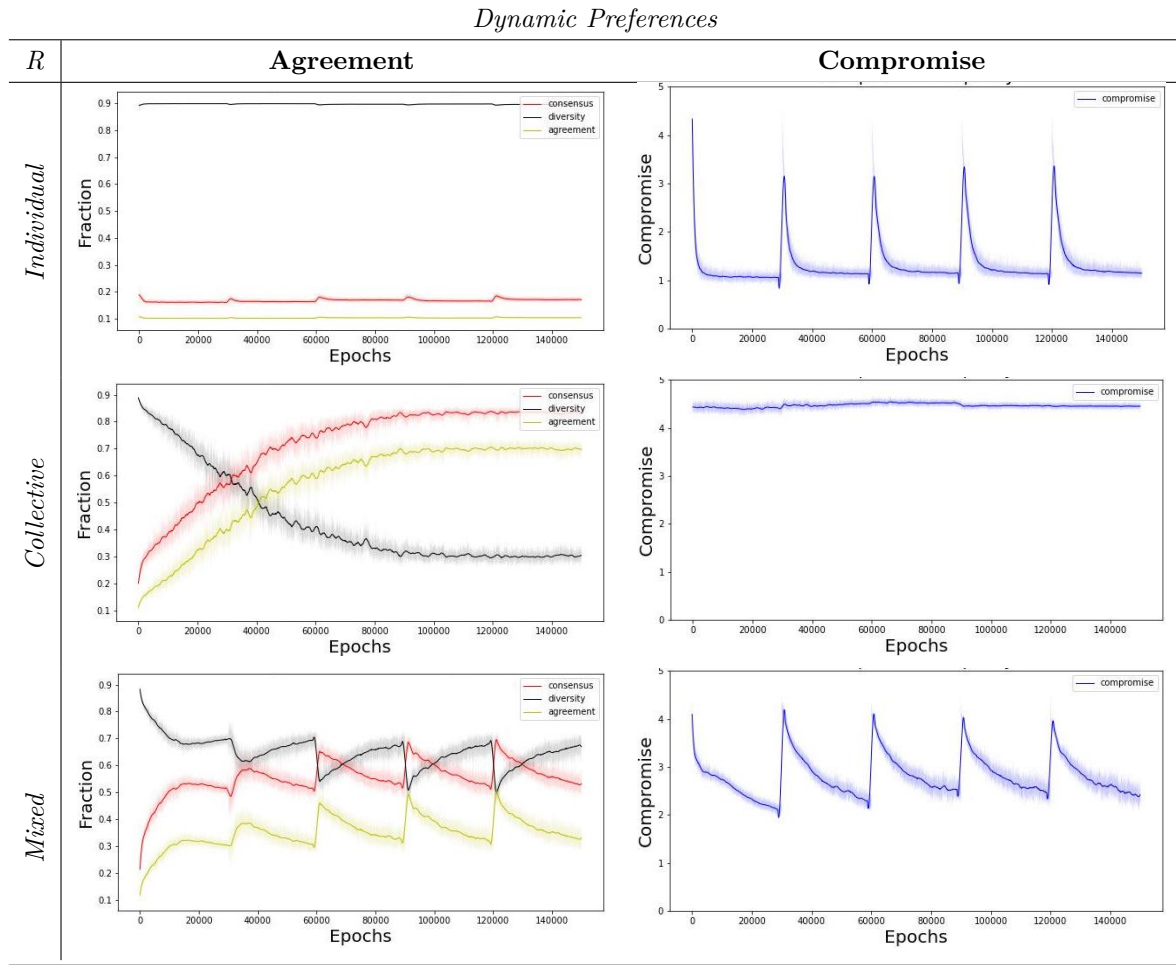


Figure 5.10: Experiment 2: Fraction of population reaching an agreement and compromises made by individuals for different conditions showing that a collective agreement can be reached even if the preferences of the population change, as long as agents do not act individualistic

observation window in column 2 is considerably wider than that in column 1, providing a long-term perspective of the system.

In this experiment, we observe that when agents aim to adhere to their individual rationality (i.e. minimise compromises), a general agreement cannot be achieved in either dynamic (row 1 column 1 in Figure 5.10) or static (row 1 column 1 in Figure 5.9) populations. Consequently, when the *Individual* objective is selected, the population remains fragmented. This fragmentation is further illustrated in row 1 column 2 of Figure 5.11: each agent chooses the policy that minimises their individual compromises; with each change in preferences, the agent alters their policy selection (horizontal lines), resulting in different policies being proposed (various colours in the vertical lines).

Moreover, when the objective is *Collective*, the population achieves a general agreement (red line row 2 column 1 in Figure 5.10), but depending on preferences, the system may stabilise in a

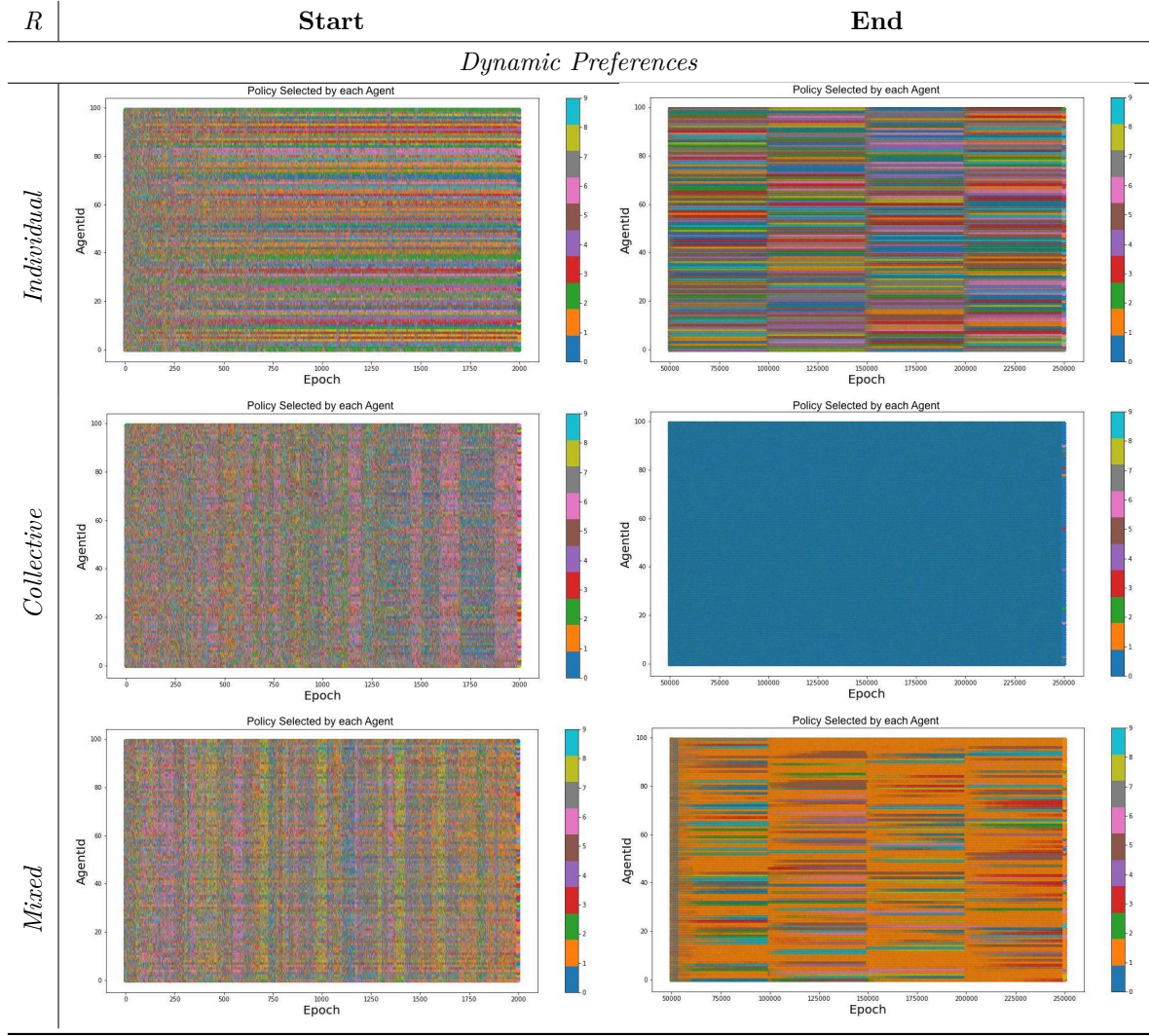


Figure 5.11: Experiment 2: Different learning objectives allow different forms of agreement to emerge when the preferences of the population change over time

state where agents must make significantly more compromises. This is evident in row 2 column 2 of Figure 5.10, where the blue line fluctuates as preferences change. Additionally, the graph in row 2 column 2 of Figure 5.11 shows that consensus is achieved since nearly all agents agree on the same policy (corresponding to the policy with index 0, represented by the blue colour).

Conversely, when the objective is to blend individual with collective rationality, referred to as *Mixed*, each time the population changes, agents quickly reach an agreement (red line in row 3 column 1 in Figure 5.10), though they must sacrifice some individual preferences (blue line in row 3 column 1 in Figure 5.10). However, after several epochs, they adapt to these changes and learn to balance the tensions between individual rationality and collective contribution, managing to achieve majority agreement while making fewer compromises. The results in row 3 of Figure 5.11 support this observation. It shows that agents start in a state of high diversity, as seen in column

1. After a few epochs, they reach an agreement, indicated by the prevalence of orange in column
2. During this process, they make fewer compromises while still maintaining diversity, which is illustrated by the various horizontal colours in the same graph.

Furthermore, column 2 of Figure 5.11 demonstrates that, depending on the learning objective, not only do various forms of agreement and levels of compromise arise, but this also results in different conditions in terms of stability. More specifically, when the objective is *Individual*, the system becomes unstable as collective agreement cannot be established. When the objective is *Collective*, the system remains stable but risks stagnation and potential instability, as a general agreement is reached while dissent is suppressed. Lastly, when the objective is *Mixed*, the system is quasi-stable, transitioning from unstable states (indicated by the colourful vertical lines in row 3 column 2) to more stable conditions (with almost homogeneous orange vertical lines).

Additionally, Figures 5.10 and 5.11 reveal that, although the system starts from a similar state (characterised by high diversity, as reflected in the various colours in column 1), the end state varies considerably depending on the objective (whether it be consensus, majority rule, or fragmentation). This exemplifies a typical property of a *cybernetic system*, in which the outcome is not dictated by the initial state, but by the developments that occur within it.

Finally, from Figures 5.10 and 5.11, we conclude that *Θ -learning enables the population to achieve the desired state, whether that be general agreement, majority agreement, or minimal compromise, as long as the participants share the same objectives*. This could manifest as a situation where the population agrees on a common voting mechanism (*Collective*), a scenario where they minimise compromises and can freely express dissent (*Individual*), or a state where they balance legitimate consent with legitimate dissent (*Mixed*). Furthermore, this illustrates a strong correlation between the form of agreement and the learning objective in both dynamic and static environments.

Additionally, we note that if the aim is to reach collective agreements while allowing for the maintenance of differing opinions and not suppressing dissent, agents must consider both individual preferences and the need for collective agreement (*Mixed*). However, if the goal is to form a consensus, agents should prioritise achieving a general agreement (*Collective*). Thus, *Θ -learning provides a mechanism for sustainable self-governance by balancing the tensions between consent and compromise when agents aim for a Mixed objective, while also providing a structure for forming consensus when the objective is Collective*.

5.3.6.3 Experiment 3: Speed of Change

Having examined the dynamics of both static and dynamic populations with various objectives, and recognising that the *Mixed* is more effective in balancing tensions between consent and diversity, this experiment centres on the *Mixed*. It aims to determine whether Θ -learning can facilitate a group's ability to achieve collective agreement while minimising compromises, even when preferences change at various rates.

Figure 5.12 displays the proportion of agents reaching an agreement and the diversity within the population for different values of the parameter c_{pr} (which indicates the rate of preference change). The first column illustrates the agreement and diversity, while the second shows the compromises made.

The graphs reveal that as the speed of change increases, the population tends to sacrifice individual preferences to achieve a collective agreement. Specifically, the results from the first 3 rows indicate that, as the population adapts, the group prioritises establishing an agreement before focusing on minimising compromises. However, when there is limited *time to react* due to frequent changes in preferences (rows from 4 to 6), the group sacrifices diversity for the sake of coherence.

Notably, the red line, representing ‘consensus’, surpasses the black line, which denotes ‘diversity’, when $c_{pr} > 32$. Furthermore, while the blue line in row 4 indicates a relatively low level of compromise, it rises significantly after row 4 as the speed of change increases, reflecting higher levels of compromise.

This relationship between collective agreement and the compromises made provides insight into cognitive biases, as highlighted by Kahneman [152]. It suggests that when agents must respond quickly to changes, they prioritise urgent tasks—such as reaching an agreement and maintaining short-term stability—while deferring less urgent but potentially more challenging tasks, like minimising compromises, which can have long-term implications for well-being, sustainability, and innovation.

In conclusion, we find that *the speed of change in the population correlates with the nature of collective agreement and the degree of compromise*. Moreover, *slow change is essential for achieving systemic sustainability*.

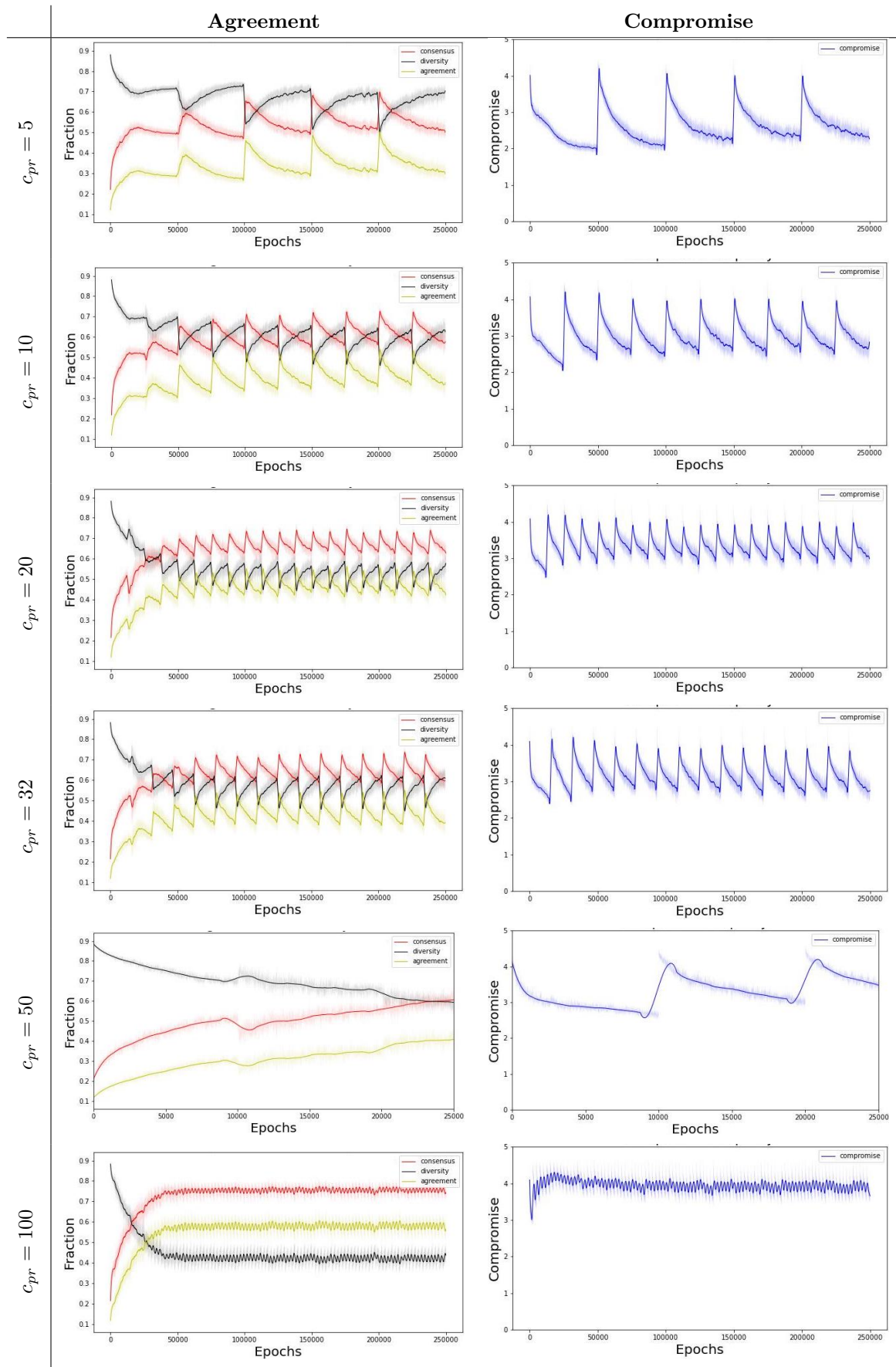


Figure 5.12: Experiment 3: Fraction of population reaching an agreement and corresponding level of compromise for different speed of change of population preferences showing that the quick response to change is to reach an agreement (in order to maintain stability) even if this requires a lot of compromise, however sustainable collective agreements require slow change

5.3.7 Conclusions on Sustainable Consensus Formation

The formation of consensus for collective self-governance encounters two main dilemmas. On one hand, there is the issue of *legitimate consent*: how can a collective achieve a general agreement on an issue, as well as maintain a consistent agreement on multiple issues, when achieving unanimity is both inefficient and improbable, yet majority voting carries the risk of a potential tyranny? On the other hand, there is the challenge of *legitimate dissent*: how can the collective honour a “valid” disagreement when the parameters for dissent are not established and continually redefined by existing authorities?

In this context, this interdisciplinary study merges a concept from historical political science, namely *thorybos*, with a domain of computer science, specifically RL, to create an algorithm for consensus formation and collective decision-making in self-governing MASs.

The technical contributions of this study include:

- Building on the democracy-by-design framework introduced in [62] by offering a more in-depth exploration of consensus during the foundational and operational phases of Demopolis;
- Introducing a formal specification for the Θ -learning algorithm, a two-step hybrid algorithm combining vocal expression (*thorybos*) with RL; and
- Proposing a framework for effective sequential decision-making that minimises compromise (see Figure 5.11 row 1 and Figure 5.9 columns 2&3).

The empirical contributions are:

- Demonstrating that agents’ objectives influence the form of collective agreement (unanimous, majority, disagreement) and lead to varying outcomes in terms of stability (instability, stability, quasi-stability) (see Figures 5.9 and 5.9 rows 2&3, and Figure 5.11 rows 2&3);
- Highlighting that both legitimate consent and legitimate dissent are essential for achieving sustainable self-governance (see Figure 5.11 row 3 column 2); and
- Indicating that the pace of policy change should correspond with the pace of preference change to balance consent and dissent (see Figure 5.12).

These experimental results reveal that depending on the learning objective, various forms of agreement and levels of compromise can arise, resulting in different stability outcomes. They

also illustrate that when agents act independently, aiming to minimise compromises and preserve diversity, the group fails to reach an agreement, leading to instability. Conversely, when agents strive to make compromises to foster a general agreement, the group may consent to a decision without expressing dissent, which can result in stagnation due to a lack of diversity or long-term instability from excessive compromise. By integrating individual and collective rationality, the group can achieve an agreement while minimising compromises, leading to sustainable (i.e. quasi-stable) self-governance.

Ultimately, the findings underscore that the immediate response to a change is to favour processes that ensure short-term stability, such as reaching an agreement regardless of the compromises involved, while deferring processes with long-term implications that may be necessary for sustainability (long-term stability). Additionally, it is noted that gradual and incremental change is a prerequisite for viable *constitutional choice rules* in sustainable systems (cf. [38], [152]).

In conclusion, we assert that Θ -learning is a fit-for-purpose algorithm for achieving consensus through meaningful democratic participation. This algorithm also offers insights on how to deal with multiple issues of self-governance, including sustainability, legitimate consent and legitimate dissent, iron law of oligarchy, unrestricted self-modification, voting paradoxes, and dignity, which would be analysed in turn. Initially, it leverages compromises to provide a sustainable mechanism for forming agreements. Also, Θ -learning incorporates an RL component to enables agents to learn when to consent and when to express dissent, while through the expression of dissent it mitigates oligarchic tendencies.

Additionally, the vocal expressions are used to inform the policy makers whether the policy is satisfactory or should be modified, fact that allows for unrestricted modification of rules (i.e. policies). The process of *thorybos*, generally, aims for mitigating issues related to the ineffectiveness of consensus and paradoxes related to voting mechanisms. Finally, the fact that individuals become aware of the policy option, and can affect it through their vocal expressions, constitutes Θ -learning a mechanism that promotes civic dignity.

5.4 Managing Sustainability in Self-Governing Socio-Technical Systems

5.4.1 Motivation

The complexity of societies has been increasing due to the integration of new *technologies* which affect the *social system* as well as the *environment* in which this social system is situated. Under these conditions, the sustainability of self-governance becomes more challenging. This is because it involves communities agreeing on policies while considering their implications to themselves and their environment, as well as the ways that their society and the environment would change.

As described by Ahlborg et al.[179], technology can act as an intermediary between humans and the environment. In particular, technology can be an essential element of the ecosystem comprising humans and nature [180] (i.e. plants, animals and other organisms, as well as weather, and landscape), not only serving as a mediator but also creating an ambivalent relationship between humans and the environment, reshaping human actions and enabling new forms of interaction. Consequently, we contend that today, technology is not merely another component of the ecosystem but a foundational factor transforming socio-technical systems and fostering stronger interconnections between them.

This requires developing systems that are capable of foreseeing the changes, adapting to them, and dynamically switching between different social arrangements. HyRL methodology is fit-for-purpose for addressing problems concerning *hybrid* (i.e. human-AI) systems. Specifically, hybrid mechanisms can support the integration of new tools to adapt to the ever-changing socio-technical environment. Accordingly, first, we analyse sustainability in this context, and we extend the existing definition to identify the categories of metrics required for measuring it. Also, we discuss how hybrid approaches can address issues related to uncertainty and sustainability while acknowledging their limitations.

5.4.2 Evaluation & Control of Sustainability

Chapter 3 explored the Hybrid Reasoning and Learning (HyRL) methodology, which can aid the adaptive management of the next generation of socio-technical systems. However, to determine whether the system effectively sustains itself, both the system designers (external agents) and members of the social (eco-)system (internal agents) must understand its state and have insight

into its performance across various evaluation criteria—such as accuracy, coherence, inclusivity, specialisation, flexibility, stability, economy, and diversity—as outlined in [134]. This evaluative information can be provided through a range of metrics.

Before examining the specific metrics needed, it is essential to consider what we mean by sustainability. Professionals from diverse disciplines, including science, engineering, and economics, as gathered by the U.S. EPA’s National Risk Management Research Laboratory, agree that sustainability represents a shift towards a trajectory where both human health and the state of the environment are preserved or enhanced.

Thus, sustainability entails not only maintaining or optimising environmental ecosystems but also introducing an additional focus on social welfare and equity, as proposed by [181]. According to Sikdar, sustainability is achieved when the economic, environmental, and social dimensions of the system intersect and operate in a controlled, favourable manner, as illustrated in the upper part of Figure 5.13.

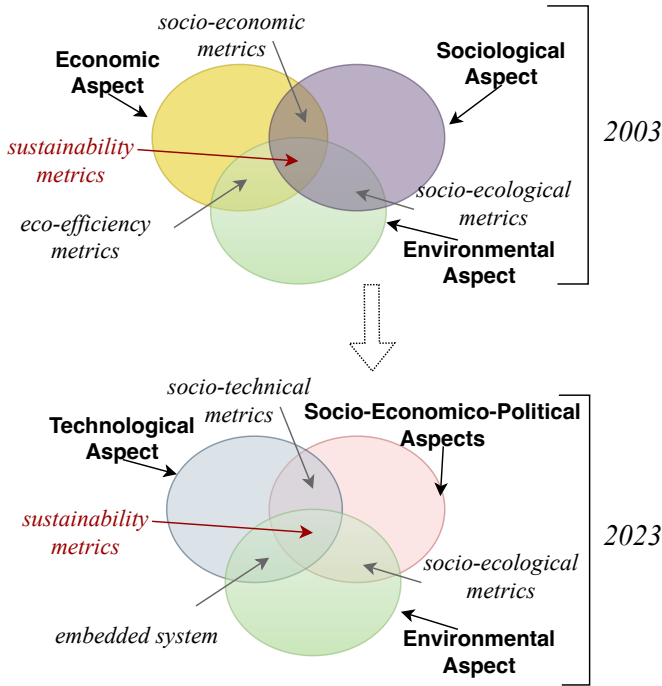


Figure 5.13: Extended version of socio-environmental system including technology as a factor affecting sustainability

Since technology now forms an essential component of the system and economic activities are increasingly seen as intrinsic to the social sphere, we propose that sustainability resides at the intersection of technological, social, and environmental dimensions, as shown in the lower part of Figure 5.13. Additionally, we argue that the social dimension encompasses societal, political, and

economic aspects. Thus, to ensure the system transitions to a sustainable state, all aspects must be monitored simultaneously.

While defining measurement approaches for these aspects can be challenging, given their specific relevance to each social-ecosystem, we propose that metrics should provide feedback on both the system's stability and flexibility (concerning the three discussed dimensions), as the combination of these two factors promotes sustainability. Furthermore, evaluating the system's sustainability would allow insights into its efficiency and economic viability. Lastly, metrics should also include data on societal well-being, covering aspects such as the legitimacy of decisions and the inclusivity of processes.

5

5.4.3 HyRL Systems For Uncertainty & Sustainability

There are two promising avenues for this HyRL approach to sustainable self-organisation, as it combines scientific self-modelling (self-determination phase) with artificial intelligence (learning phase). These opportunities are, firstly, constructing a model of socio-techno-ecological systems, and secondly, developing algorithms from this model to assist in managing socio-technical (eco-) systems. However, there are associated risks and limitations with each of these.

First, modelling cyber-physical systems involves managing multiple sources and types of uncertainty, with various methods suggested, such as self-healing [182] and Model-Driven Engineering [183]. Adding the social component further complicates this, as efforts to ground its interactions with the other two components in psychological, political, and anthropological theories to predict organisms' behaviours and responses only increase uncertainty in the modelling process. Additionally, the persistent challenge of “unknown unknowns” arises, supporting the use of conservative RL approaches [184], or methods aimed at accurately estimating uncertainty, such as Epistemic Neural Networks [185] and Prior Networks [186].

Second, as with any cross-disciplinary effort, it is very challenging to create a comprehensive model of socio-technical systems, derive AI algorithms, and obtain accurate, reliable measurements to gather the data essential for these algorithms to enable ‘correct’ management decisions. Also, if we assume that we can succeed in all the previous, still the notion of ‘correctness’ is relative to the scientific discipline and the stakeholder, with different stakeholders prioritising different aspects of sustainability. For instance, some focus on preventing negative global impacts [187], while others view sustainability as avoiding tipping points and unintended consequences [188]. In this setting, risk-averse exploration becomes important [189] since it allows for early detection of

dead-end states and taking proactive actions to avoid them. This also connects to the ideas of Graeber and Wengrow discussed earlier in this thesis, where the critical question may not be the origin of inequality but rather how certain social and political structures became ‘stuck’ [90].

5.4.4 Conclusions on Metrics for Sustainability

This Section addressed issues related to sustainable self-governance of next generation socio-technical systems and proposed the following:

- to extend sustainability Sikdar’s framework for evaluating sustainability [181], by elevating the technological aspect and including political, economic and societal elements of the “sociological” aspect;
- to use HyRL methodology for addressing issues of sustainability while being aware of their limitations due to uncertainty.
- to provide a better understanding on how to maintain sustainable self-organisation, which corresponds to one of the main challenges faced by SGSs.

5.5 Summary & Conclusions

In conclusion, Chapter 5 demonstrated that HyRL methodology offers valuable frameworks for fostering sustainability and self-governance within socio-technical systems. The approaches discussed present distinct methodologies for combining RL, psychological insights, and democratic processes to address complex self-regulation and decision-making challenges.

The first approach incorporated a psychologically-inspired *4voices* algorithm alongside Q-learning in order to specify a model where units within a system can selectively prioritise various “voices” to achieve a shared goal, i.e. the effective self-regulation of the system. The experimental results of this HyRL-based framework demonstrated that it could support continuous monitoring and ethical regulation through quasi-stability and requisite social influence. The second approach drawn inspiration from classical Athenian deliberation, and introduced Θ -learning uses the mechanism of *thorybos* together with MARL to reach sustainable consensus across diverse institutional conditions. Finally, the third approach highlighted the significance of the HyRL methodology in achieving sustainability by aligning technological, environmental, and socio-political elements of next generation socio-technical systems.

While each of these approaches offers promising contributions to self-governance and sustainability, socio-technical systems are evolving at an unprecedented pace, posing the risk that control may be increasingly dominated by AI. This creates an urgent need for tools that empower humans to deliberate on their Social Arrangements (SAs) and systematically innovate them to respond to new and unforeseen challenges. Chapter 6 addresses those issues by using higher order cybernetics patterns to support human empowerment, and by proposing an exemplar of an innovation support system that uses the Socially-Guided Reasoning and Learning (SGRL) methodology, which was introduced in Section 3.4, to leverage the human-AI co-production to enable humans envision collectively deliberate upon their Social Arrangements (SAs) and maintain an equitable distribution of power during hybridisation.

Chapter 6

Innovation & Empowerment

Contents

6.1 Introduction	135
6.2 Empowerment	137
6.3 Definition of Empowerment	137
6.3.1 Higher-Order Cybernetics Patterns & Empowerment	139
6.3.2 Case Studies of Cybernetics & Empowerment	142
6.3.3 Conclusions on Empowerment	147
6.4 From Empowerment to Innovation	147
6.5 Innovation	148
6.5.1 Conceptual Analysis of Innovation	148
6.5.2 Process-Oriented Definition of Innovation	149
6.6 Innovation for Empowerment	150
6.6.1 ISSD Detailed Overview	151
6.6.2 ISSD Implementation	155
6.6.3 Evaluation as a Tool for Empowerment	165
6.6.4 Evaluation as a Cybernetic System	170
6.6.5 Social Implications of SGRL Methodology	174
6.6.6 Conclusions on Innovation for Empowerment through ISSD	176
6.7 Evaluation & Metrication	177
6.8 Empowerment and Innovation: A Chicken-Egg Problem	180
6.9 Summary & Conclusions	181

6.1 Introduction

As discussed in previous Sections, the rapid advances of Artificial Intelligence (AI) technologies have resulted in an increasing hybridisation of our societies [63]. Hence, next-generation socio-technical systems will be characterised by the convergence of human intelligence, i.e. Natural

Life (NLife), with artificial intelligence, i.e. Artificial Life (ALife), in shared existence, in an NA-Life system [23]. The sustainability of those hybrid (human-AI) systems would rely on the effective interactions between the naturally and artificially intelligent components [190].

The advent of such hybrid systems would also lead to profound changes in how human and artificial agents operate, learn, and adapt within a shared environment. NA-Life systems, by fostering these interdependencies, present unique challenges and opportunities, calling for an examination of the roles that both human and ALife play in influencing socio-technical outcomes. On the one hand, ALife can improve NLife's living and increase human empowerment. On the other hand, the integration of ALife technologies can diminish human empowerment, either by intensifying existing imbalances in power dynamics or through the unintentional transfer of power from NLife to ALife [23].

Human empowerment within these hybrid Self-Governing Systems (SGSs) necessitates individuals' ability to autonomously structure and deliberate upon their Social Arrangements (SAs). Therefore, a critical question arises: how can AI tools and methodologies contribute to or enable such empowerment? We argue that higher-order cybernetics and meta-cybernetics, due to their expressiveness, are effective in reducing the uncertainty by uncovering hidden material (corresponding to interdependencies and power asymmetries) or regulatory relationships [191]. This way they can enable humans develop a better understanding of the dynamics of their environment and therefore support their empowerment by helping them to shape their social arrangements.

Accordingly, Section 6.2 of this Chapter addresses this question by employing higher-order cybernetic patterns, which offer a framework for understanding and designing interactions that support human agency in NA-Life systems. Through these cybernetic patterns, humans may gain increased capacity to shape their environments and interactions with ALife, thus asserting meaningful participation in the governance of these complex systems.

This Chapter also considers the imperative of innovation within Self-Governing Systems (SGSs), specifically in relation to SAs that can foster human empowerment. To this end, after discussing innovation in the context of self governance in Section 6.5, Section 6.6 introduces an Socially-Guided Reasoning and Learning (SGRL)-based tool, referred as Innovation Support System through Deliberation (ISSD), that supports innovation conducive to empowerment. In Section 6.7, the discussion extends to issues related with the evaluation of the level of empowerment and innovation, underscoring the need for measurable outcomes within this evolving landscape. Furthermore, it highlights the duality of empowerment and innovation in these systems. It suggests that as self-governing

socio-technical systems evolve, the mechanisms that gauge and enhance their capacity to empower have to evolve too. Together, the Chapter concludes in Section 6.8 with a reflection on innovation and empowerment by highlighting the interconnection between those terms in the context of SGS.

6.2 Empowerment

Given that the next-generation of socio-technical systems comprise NLife and ALife components [63], we aim to design systems that benefit from the capabilities of latter to support the empowerment of the former. Thus, we need to consider how the existing AI technologies can increase human empowerment. This Section defines empowerment in the context of SGS, analyses conceptually how patterns of higher order cybernetics can support individual and collective empowerment, and also grounds it to four cases studies [23].

6.3 Definition of Empowerment

Reflecting on the definitions of power and empowerment that were presented in Chapter 2 and grounding it in the context of self-governance, we argue that empowerment relates to the deliberation on social arrangements, which correspond to the set of rules, roles, structures, procedures, policies, norms, conventions, contracts or laws with which individuals in a group voluntarily agree to comply, in order to hold each other accountable [23]. Hence, it refers to the cognitive capabilities of ‘intelligent’ and ‘reflective’ components within a socio-technical system to understand and reason about five cognitive dimensions of empowerment [192]:

- self-determination: the degree to which individuals in the group take part in creating, modifying, and enforcing the social arrangements that affect them or that they have mutually agreed upon to regulate their behaviour;
- competence: the group’s capacity to meaningfully contribute to decision-making and control, and its awareness of this contribution;
- influence: the extent to which an individual can impact processes like deliberation, decision-making, truth-tracking, adherence to norms, and maintaining community coherence;
- knowledge: the extent to which the group can gather and access distributed knowledge, align and curate common knowledge, and involve domain-specific expertise; and

- meaning: the value placed on direct problem-solving and the social bonds that emerge from it.

Therefore, the level of empowerment within a group can be measured by its ability to assess the socially constructed rules, structures, and networks in relation to the five dimensions, and to use that assessment to make informed decisions and exert control over them. The next Section introduces higher order cybernetics as a way to enable social actors in a socio-technical system to determine the social arrangements that affect their lives.

Relating the five dimensions of empowerment in SGSs with the eight features for the sustainability of SGSs, each of the latter can be ‘linked’ with one or more dimensions. Specifically, those dimensions of empowerment support those features in order to enable sustainable self-governance.

In more detail, the first feature (i.e. articles of (voluntary) association, for inclusion and exclusion of entities, and expectations of entities in the system (e.g. adherence to the rules, participation in the application of the rules) is correlated with the first dimension of empowerment (i.e. self-determination). This is because individuals having the capability to participate in the processes of development, adaptation, and application of social arrangements can act on the specification of articles related to association, and processes of inclusion, exclusion. The second dimension of empowerment (i.e. competence) enables individuals to develop structures and procedures for facilitating the intended functions of the system, as well as for run-time self-modification of the over-arching set of social arrangements (corresponding to the second and third features). The reason is that individuals that are capable of contributing to processes of decision-making are also capable of developing those structures.

Knowledge and influence facilitate knowledge management (i.e. fourth feature) and identify and appoint trustworthy gatekeepers of knowledge (i.e. fifth feature), because they have the knowledge and channels to support that. Then, compliance (i.e. sixth feature) is assured through influence and self-determination, since individuals have both the channels and the willingness to apply the rules. The seventh feature that concerns the limits of and rights to self-determination is relevant with the dimensions of self-determination and meaning. This is because on the one hand it involves individuals’ participation in the selection, modification, application and enforcement of the rules but also requires them to consider matters of rights, core values, and morals. Finally, polity (i.e. eighth feature) requires the cognitive capacity of self-determination but also involves influence, which is essential for making decisions that involve or affect external actors and authorities.

However, there is a second aspect of empowerment that corresponds to an objective measure-

ment of collective empowerment, referred to herein as community health. Accordingly, the empowerment can be evaluated in a unit level with respect to each individual's sense of self-determination itself, and an awareness of their competence in, influence on, knowledge of, and meaning of this process. At a collective level, this corresponds to the health of a community and can be assessed by the following six metrics: inclusivity, transparency, diversity, equality, accountability and satisfaction (as a form of interactional justice [72], i.e. how well individuals feel they have been treated by the social arrangements). The metrics of individual empowerment and collective empowerment (i.e. community health) are illustrated in Figure 6.1.

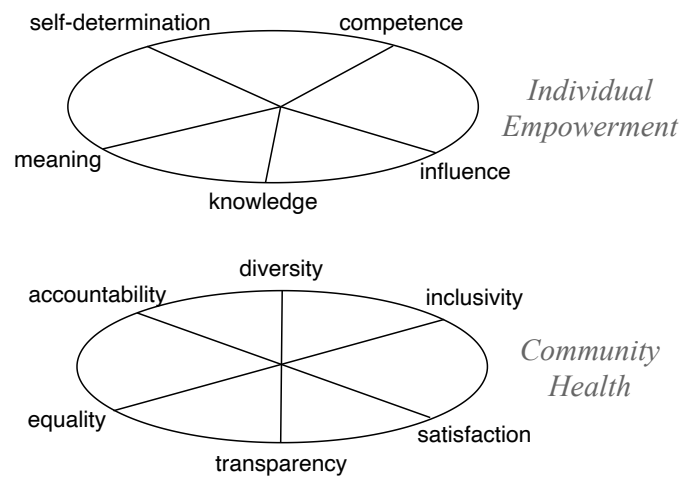


Figure 6.1: Measurements of Individual and Collective Empowerment

6.3.1 Higher-Order Cybernetics Patterns & Empowerment

As discussed in Chapter 3, Power Sensitive Design (PSD) constitutes a conceptual model for designing human-AI systems empowering humans. However, we need to consider how human capabilities and artificial intelligent technologies can be brought together in a system to meet the PSD design requirements, leading to Power Sensitive Operationalisation (PSO). We contend that one way to achieve that is by using higher-order cybernetics patterns.

This is because cybernetics systems possess feedback loops, which are essential for control and organisation since they allow to those systems to modify their behaviour in order to meet a goal. Therefore, by adding feedback loops at different levels of the system, we can design SGS with different capabilities. Therefore, a cybernetics analysis enables us, as well as the components of the system, to develop a better understanding of it. At the same time, those feedback mechanisms give leverage to human participants of the cybernetic system to get involved, engage in, and have

control over it.

Accordingly, in this Section, we consider the different orders of cybernetics, which were introduced in Section 2.5.2, and specifically the feedback loops they possess and the leverage they give to the human users and participants. Reflecting on those, we explain how different orders can enable different levels of human involvement, participation, and engagement in human-AI systems.

Various levels of cybernetics offer distinct forms of empowerment for NLife. In *first-order cybernetics* [16], NA-Life systems are designed by NLife to perform specific tasks efficiently. These systems are self-regulating and intended to manage delegated functions while conserving NLife's physical and cognitive resources. Their capabilities are limited, as they are built to execute relatively simple tasks, like distributing electricity across devices in a smart home. Inadequate design, however, can lead to wasted resources and inefficiency.

Second-order cybernetics involves a regulated system alongside an observer [193], where humans function as observers who can influence the system's operations but are unaffected by it. This model is useful for designing NA-Life systems where regulation requires an external assessor to evaluate and shape the system's behaviour. Such systems use ALife under NLife's oversight to address more complex problems. Nonetheless, an inadequate design might result in ineffective regulation or unnecessary resource expenditure, whether in time, physical, or cognitive domains.

Second-order models can also be intentionally structured to undermine collective empowerment. Here, an NA-Life system might be configured so that the 'self-regulating' system includes NLife (and ALife), while the observer, also an NLife entity, monitors and influences it in ways that benefit the observer rather than the system itself.

Third-order cybernetics introduces an interactive element[92], allowing for the co-evolution of ALife and NLife components with humans involved in the loop. This makes humans both insiders and outsiders within the NA-Life system, recognising that they influence the system and, in turn, are influenced by it. The role of the observer can alter the system's dynamics.

The awareness of the NLife components that they not only affect the system they observe, but they are also observed and affected by it can trigger different behaviours. Hence, designing systems with third-order patterns enhances NLife's understanding of the hybrid NA-Life system, and enables humans address issues of influence, social coordination, and social cohesion.

Third-order cybernetics also allows for multi-stakeholder engagement, where numerous observers, each with different objectives, contribute to self-developing, multi-subject systems. This

Order	Type of Empowerment	Type of Disempowerment
First Order	Saves Physical and Cognitive Resources	Resource Loss and Ineffective Regulation
Second Order	Solves Complex Problems	Ineffective Regulation or Manipulation
Third Order	Deals with Social Constructions	Power Misuse and Reaching Detrimental States
Fourth Order	Mitigate Existential Crises	Misinformation and Manipulation

Table 6.1: Relation between Higher Order Cybernetics and Empowerment

reduces manipulation risks, but NLife components must acknowledge that their influence on the system is reciprocal. Recognising this can prevent unintended consequences or negative impacts from the choices they make.

Fourth-order cybernetics centres humans within the design, making them integral to the system. Here, NLife components can view themselves as both inside and outside the system [93], comprehending it holistically while situating it within a larger ecosystem. This enables NLife to address complex, ‘wicked’ issues, such as sustainability, resilience, and antifragility, and manage existential crises.

However, while fourth-order cybernetics can significantly empower NLife, it also has the potential for heightened disempowerment if complexity is used to the detriment of the collective. For instance, an NLife authority or ALife could mislead the system, fostering false beliefs, biased understanding, or misinformation. In such cases, NLife may be completely disempowered, as decision-making power is seized by others while NLife components perceive themselves as influential.

In summary, increasing a system’s complexity through higher-order cybernetic patterns progressively enhances empowerment, as shown in Table 6.1 and illustrated in Figure 6.2. Specifically, moving from simple *self-regulation* (first order) to *regulation with external feedback and responsive awareness* (second order), then to *co-evolving system-observer patterns* (third order), and finally to *contextualised self-aware and self-regenerating systems* (fourth order) allows NLife to tackle increasingly complex issues, from optimising smart energy grids to addressing sustainability. This is because greater complexity can foster deeper understanding and control over the system, as also described in the Laws of Requisite Variety & Complexity [174],[171].

Thus, the empowerment and disempowerment of humans within socio-technical systems designed around complex cybernetic patterns develop non-linearly. Consequently, it is vital to design systems that not only assign humans an active and interactive role but also strive to maintain their awareness of power dynamics, drawing attention to any inequitable power distributions.

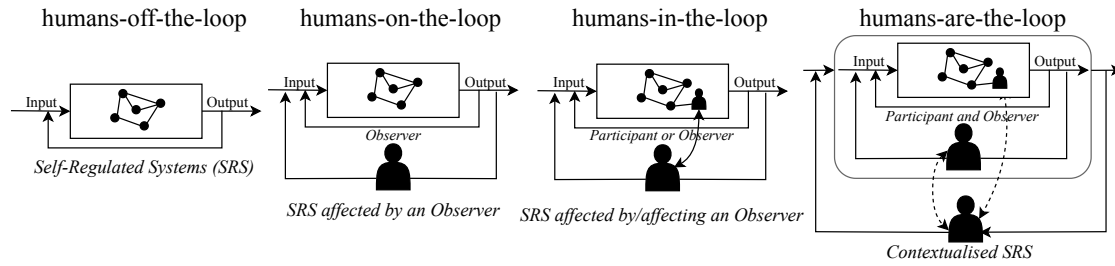


Figure 6.2: Higher Order Cybernetics Learning to Empowerment

6.3.2 Case Studies of Cybernetics & Empowerment

This Section grounds the theory of higher order cybernetics and its relation to empowerment and disempowerment in the context of SGSs to four case studies. The case studies include: varying levels of coordination in community energy system distribution and consumption [194]; a moral aspect of empowerment requiring diverse justice qualifiers within a system of systems [40], [41], [72], [89], [195]; a specific (business) model for the spread of misinformation on social media [196], [197]; and a broader understanding of power in human-machine integration within cyborgs [198].

6.3.2.1 Community Energy Systems

A decentralised community energy system (CES) functions as a type of isolated SmartGrid, where each home (or “SmartHouse”) houses several people (NLife components) and is equipped with multiple ALife components. Among these is an interactive SmartMeter that monitors the energy consumption of various NLife components, such as household appliances like refrigerators, washing machines, dishwashers, and ovens. SmartHouses may also include renewable energy sources (e.g., roof-mounted solar panels or small wind turbines) and some form of energy storage.

However, issues of self-governance and empowerment arise across multiple scales. At the basic level, synchronising refrigerators’ compressor cycles is essential to maintain voltage and frequency within set limits; otherwise, connected devices may face unexpected operation conditions. Although this synchronisation is crucial for the grid’s functionality, NLife owners are typically concerned only with their refrigerators keeping food at the right temperature.

At the next level, appliances such as dishwashers and washing machines offer two modes: *A*, meaning “run now”, and *B*, allowing the user to specify cost and timing preferences. The machines can then arrange an optimised schedule based on the community’s collective constraints and resources. Nonetheless, NLife residents generally remain disengaged from decisions handled

by ALife components.

The SmartMeter(s) may, however, detect when current consumption levels risk exceeding available energy from generation and storage, suggesting that certain devices should be switched off temporarily to prevent service interruptions [194]. Choosing which appliances to turn off involves human judgement and benefits from community attention, supported by interactions and visualisations through SmartMeters.

At the broader scale, the community must make strategic investment choices, establish agreements with other communities, assess carbon emissions, and address other qualitative, value-driven goals. In these cases, direct human involvement is essential, while ALife components can offer innovation support.

Accordingly, the CES design aligns with four engagement and empowerment levels, reflecting the four cybernetic patterns, as shown in Figure 6.3:

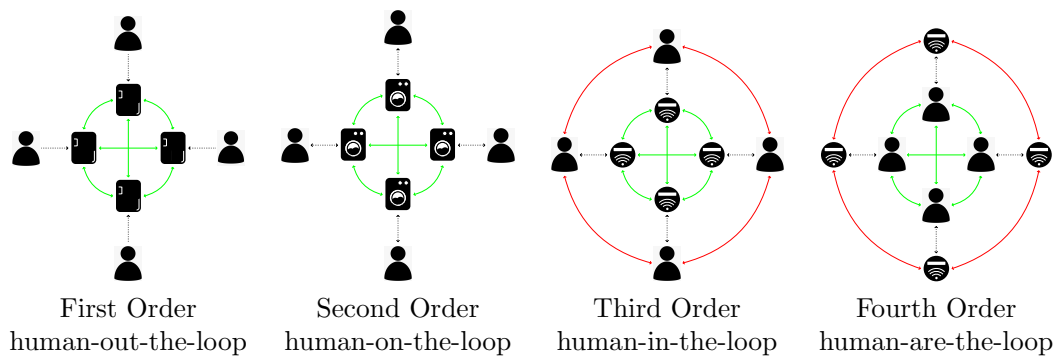


Figure 6.3: n -th Order Cybernetics for Empowerment in Community Energy Systems

- *First order (humans off the loop)*: Strengthening NLife competence through indirect voltage and frequency control, foundational for the islanded SmartGrid.
- *Second order (humans on the loop)*: Increasing collective deliberation through optimised scheduling informed by human feedback on criteria like cost and sustainability.
- *Third order (humans in the loop)*: Boosting influence for proactive, socially beneficial outcomes by fostering collective awareness, where humans are both participants and observers.
- *Fourth order (humans are the loop)*: Expanding knowledge to support innovation and enhance quality of life, enabling the community to act as a model for others through diverse adaptations.

6.3.2.2 System of Systems

Expanding on the provision and appropriation scenario from the previous case study, let us consider a collective action setting involving access to a shared resource through a common infrastructure. This scenario presents two interconnected decision-making contexts: a short-term appropriation context for extracting common-pool resources and a long-term provision context for contributing to infrastructure maintenance.

To illustrate this with NLife components, consider a user-managed irrigation system within a rural farming community. This system requires short-term operational decisions to manage water distribution to specific farms within the community (appropriation), which may need to account for the interaction between water scarcity and pest dynamics [199]. Additionally, it demands long-term operational management, involving contributions towards infrastructure construction, maintenance, and monitoring (provision).

In these contexts, [200] identifies three core elements of *communality*: *the commons*, which refers to mutually agreed local practices voluntarily upheld to sustain the resource (highlighted in the previous case study); *community*, which fosters a shared purpose while respecting individual identities and local belonging [90]; and *polity*, the collective capacity to engage with external actors.

This framework introduces *moral* considerations within empowerment dimensions [41]. Deliberation now requires the ability to propose, negotiate, and act on justice principles that are accepted by those affected, with compromise as an essential component [21].

Within NA-Life systems, different forms of justice need to be operationalised, including: *procedural justice* [89], which ensures that institutional rules meet specific standards; *distributive justice* [40], which addresses fair allocation of rewards and penalties within the group; *restorative justice* [195], allowing disputes to be resolved locally without external legal intervention; and *interactional justice* [72], assessing the quality of polity through its treatment of community members, translating subjective assessments into a unified, objective evaluation.

These justice qualifiers contribute to empowerment by enhancing communality as follows:

- first order: Procedural justice fosters collective deliberation by supporting commons that align with environmental conditions [38].
- second order: Distributive justice enhances competence by employing an observer to assess decision impacts, achieving qualitative objectives such as fairness.

- third order: Restorative justice strengthens influence, allowing components to act as both observers of breaches and participants in alternative dispute resolution.
- fourth order: Interactional justice leverages collective knowledge to deepen the community's embodiment of communality.

6.3.2.3 Social Media

Social media platforms enable NLife to use ALife to engage with other NLife entities, exchange information, and leverage social influence to enhance collective knowledge. Initially, these platforms might be seen as first-order systems where users perform specific tasks and sign out when done. However, in contemporary social media, user engagement goes beyond mere participation. By choosing to be part of these systems, individuals share their personal data and content with the platform and, indirectly, with those who own or control that ALife component (i.e. the social media platform). This means that those in control have access to users' personal information, enabling them to shape opinions through the spread of misleading information, endorsement of false authorities, and dissemination of misinformation. This aligns with second-order cybernetics, where those in power act as observers, influencing the system without participants fully realising they are being observed, monetised, or manipulated [196], as shown in Figure 6.4.

This influence can be even more subtle, leading users to perceive an illusion of control based on existing (often inaccurate) beliefs rather than their active choices [197]. Those in power can circulate false narratives, shaping collective beliefs so that users experience a constructed reality. Here, users may believe they are both participants and observers within a fourth-order cybernetic system, with control over and within the system, though they are in fact influenced without genuine agency.

A discerning user, however, might recognise flaws in the system and advocate for social media to adopt patterns of third-order systems. This would allow each participant to be aware of the influence the system has on them, providing both control and responsibility over decisions that will impact them directly.

Upon closer examination, social media should ideally follow fourth-order cybernetic patterns. Today, these platforms influence us, explicitly or implicitly, regardless of whether we actively participate. To empower communities and combat disinformation (intentional spreading of falsehoods), misinformation (unintentional spread of false beliefs), and manipulation, users need to be both

active participants and critical observers of these platforms. They should also remain aware of the range of perspectives within the digital environment.

In this framework, users could not only redistribute power within the network but also understand the system's intent (e.g., information sharing, entertainment, communication) and have tools to address issues (like disinformation and hate speech) or redesign systems to better serve their purpose. Research by [201] suggests that decentralisation can help counteract deception and misinformation, acting as a corrective mechanism against corrupt governance.

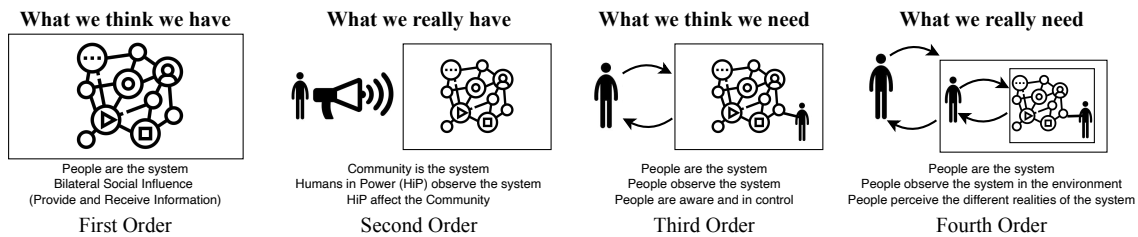


Figure 6.4: n -th Order Cybernetics for Empowerment and Disempowerment in Social Media

6.3.2.4 Cyborgs

Another area of human-computer interaction, or more aptly, human-computer integration, involves merging NLife and ALife components, defining the realm of *cyborgs* [198]. A cyborg is typically characterised by the integration or co-existence of human (NLife) and machine (ALife), with a particular subset of cyborgs being assistive technology prosthetics. These range from wearable health monitors on wrists or fingers, to robotic wheelchairs that aid (but do not control) movement, to brain-controlled prosthetics, and implantable devices such as insulin pumps.

This progression represents a spectrum of cyborg innovations: aesthetic devices, which are external to the body; assistive devices, worn on the body; bionic devices, integrated or implanted; and full cyborg systems, where the technology becomes part of the body itself.

This “cyborg transition” aims to enhance NLife performance and quality of life, though it raises questions around the levels of empowerment and dependency created by different cybernetic patterns as assistive technologies acquire more ALife intelligence.

For instance, some children have been observed to rely on voice-activated virtual assistants for convenience, despite having the capacity to solve the problem independently. Additionally, people may “dumb down” in response to an ALife perceived as superior [140], a tendency seen in joint tasks. In focal point games, for example, people often defer to an ALife they have a socially

constructed relationship with, mistakenly attributing it more capability than it possesses [202]. Furthermore, intentional digital design aesthetics can foster a form of “digital dependence”, which can be exploited for commercial gain or to undermine autonomy [203].

In summary, future socio-technical systems with cyborg capabilities will feature ALife components acting on behalf of, or following instructions from, NLife components. Yet, it remains essential to clarify precisely *which* NLife component holds the authority: the user, the programmer, or the manufacturer/distributor.

6.3.3 Conclusions on Empowerment

In summary, combined PSD methodology, presented in Section 3.4.3, with higher-order cybernetics, presented in Section 2.5.2, to provide a conceptual framework for dealing with asymmetries in the distribution of power in hybrid systems, which were described in Section 2.5.1. Also, through four case studies it demonstrated how these design patterns can be applied to extend empowerment in line with the defined dimensions, i.e. self-determination, competence, knowledge, influence, and meaning.

The significance of this was to provide a conceptual mechanism for maintaining civic dignity by enabling human components of the hybrid systems to understand their empowerment and act to increase it. Also, by giving the power to individuals to determine the level of their empowerment and control the distribution of power amongst themselves and AI, it took a step towards dealing with issues of citizenship and rights which are threatened due to the increase automation and rapid adoption of AI technologies. Overall, cybernetics-empowered empowerment establishes the basis for addressing issues of dignity, citizenship, and rights in SGSs.

6.4 From Empowerment to Innovation

In the previous sections, we explored the importance of human empowerment during the transition to a digital society, one in which human and computational agents coalesce within socio-technical frameworks. As we approach this integration, the inevitable shifts and complexities it introduces impose a critical requirement: a robust framework to navigate and manage these changes effectively. This necessity, in turn, drives a pressing need for innovation, not only in the technological realm but, more importantly, in how we design and operationalise social arrangements.

In the context of SGSs, human empowerment extends beyond traditional capacities and involves the ability to deliberate on the social structures that shape one's life. Here, empowerment embodies the capability to influence and structure one's environment and social interactions autonomously, without undue external control. As such, innovation in this context is fundamentally linked to the capacity for individuals and communities to generate, apply, and modify social arrangements in ways that are meaningful to their shared objectives and values. In other words, the advancement we seek in SGSs must facilitate the evolution of social relationships and the distribution of agency across both human and AI components to uphold human empowerment (in an individual and collective level).

The following discussion delves into innovation within the domain of social arrangements. This exploration seeks to address the challenges and opportunities associated with implementing empowering structures in digital societies, examining what is innovation and how it can foster collective deliberation and equitable agency within socio-technical systems.

6

6.5 Innovation

Section 2.5 introduced related terms, concepts, and types of innovation with a focus on political innovation and related instances and attempts to achieve it. This section builds up on those issues by providing a conceptual analysis of innovation together with a definition of it in the context of innovation for empowerment in self-governing socio-technical systems.

6.5.1 Conceptual Analysis of Innovation

6.5.1.1 Critical Factors of Innovation

In this subsection, we examine the three essential elements in innovation: reflection, generation, and application. The first factor, *generation*, relates to a shared foundation across these terms: the concept of novelty and establishing a mechanism to achieve it. Such novelty, depending on the context (e.g., definition or perspective), may emerge from a distinct approach to handling knowledge. In any case, it is the generation of a novel idea that forms one of the core pillars of innovation.

The second factor, *reflection*, involves engaging with the surrounding environment to identify both current and emerging challenges and assess the available information that can help mitigate

these effectively.

The third factor emphasises that innovation seeks not only to introduce novelty but to ground and contextualise it within a specific setting. Thus, it is as much about devising something new as it is about implementing it effectively. This third element, *application*, entails developing a methodology for addressing a present or anticipated issue, adapting the new idea to the environment, and recognising its distinctive attributes, strengths, and weaknesses.

In summary, innovation involves a process of *reflection* on the environment, *generation* of novelty through envisioning new combinations of factual and conceptual elements, and *application* to transform an abstract idea into a practical solution. An underlying factor is that environmental change often triggers this need for novelty, with the pursuit of new solutions arising in response to shifts in the environment and achieved by integrating pre-existing elements and processes therein.

6.5.1.2 Reflection on the Types of Innovation

Different forms of innovation may be more suitable in various conditions to drive the required or desired changes within a socio-technical system. Analysing innovation types based on their origin highlights that, when innovation (or growth, as termed in economic theory) is endogenous, the range of new ideas that can emerge is inherently constrained.

While economic theory lacks consensus on which type of innovation—endogenous or exogenous—leads to greater productivity, we argue that for continuous innovation within socio-technical systems without external contributions, there must be intricate mechanisms to enable the effective recombination of existing knowledge. Alternatively, to simplify this complexity, incorporating external knowledge can stimulate the generation of novel ideas, processes, and structures.

Additionally, when classifying innovation types by speed, we suggest that it is crucial to adaptively shift among incremental, additive, complementary, and radical innovations according to the system's state. The optimal innovation type depends on environmental conditions, with one type potentially being more advantageous than another in differing contexts.

6.5.2 Process-Oriented Definition of Innovation

Innovation can take many forms, depending on its context and purpose. However, this thesis is concerned with structural and procedural innovation, and specifically innovation of social ar-

rangements in hybrid socio-technical systems [24]. This is because the fundamental problem that individuals in an SGS have to address is the selection of effective social arrangements that would improve their collective standards of living. Therefore, innovation, in this context, corresponds to a tool supporting individual and collective empowerment, by equipping individuals with choice and control over the rules, roles, structures, and procedures that directly affect their lives. It is a form of adopted disruption that aims to enable individuals to address existing societal challenges (reactive innovation) or to initiate proactive changes that prevent unintended consequences of current practices (curative innovation), thereby ensuring long-term sustainability.

Combining the above with the three features of innovation (i.e. reflection, generation, and application) identified in Section 6.5.1.1, we provide here an informal definition of innovation. As such, innovation can be understood as the synthesis of three key processes: *reflection* on the environment, the *generation* of new ideas (in the form of social arrangements), and their *application* to the environment, as depicted in the ‘equation’ of innovation in Figure 6.5.

$$\text{Innovation} = \text{Reflection} + \text{Generation} + \text{Application}$$


Figure 6.5: Illustration of the Innovation ‘Equation’

Reflection involves recognising the problem or identifying processes that can be improved. *Generation* entails imagining potential solutions by creating diversity. *Application* refers to the ability to implement these solutions in a way that effectively addresses the issue within its specific context.

6.6 Innovation for Empowerment

Section 3.4.3 introduced PSD as a methodology for designing hybrid socio-technical systems in such a way that we empower NLife components, while Section 3.4 introduced the SGRL as a methodology for achieving innovation in self-governing socio-technical systems through the generation and evaluation of SAs. In this Section, we present Innovation Support System through Deliberation (ISSD) which is a framework that uses SGRL methodology for designing and implementing an AI-powered tool for social deliberation concerning issues of innovation of SAs leading to community empowerment. This is a tool that achieves power-sensitive operationalisation, by not only designing a power-sensitive system but also giving to the its participants a mechanism for

maintaining an equitable distribution of power.

Specifically, this Section describes ISSD as a modified model-view-controller framework, in which two controllers (i.e. human and Generative Artificial Intelligence (GenAI)), a Multi-Agent System (MAS) model, and a visualisation module are employed. In this framework, both controllers can suggest alternative SAs, which are then simulated in the MAS to assess their impact on the collective, shown visually for the human to interpret. This co-production of social arrangements leverages GenAI's strengths to enhance human creativity while mitigating issues like bias and hallucinations.

The operation of ISSD approach is based on the iterative application of two phases: first, running the MAS and visualising the system's state, and second, having the user evaluate and propose changes (optionally consulting GenAI). The system's non-deterministic nature means that community outcomes rely more on iterative interactions than initial conditions. Resetting the system enables exploration of multiple social arrangement iterations, allowing comparison of long-term impacts across different human-GenAI combinations, and fostering continual co-production of new social arrangements.

This Power-Sensitive Designed ISSD framework could be applied in the co-housing community as well, in two different ways. The first way is to enable humans to self-determine and innovate their social arrangements for distributing the power to humans and machines. Here, the power corresponds to the control of the decisions on the policy for distributing the resources between individual and collective devices. The second way is to facilitate innovation of social arrangements relevant to the policy of resource distribution. In both ways, the MAS would be designed to simulate the shared-housing community, and the social arrangements proposed by GenAI would be relevant to the distribution of the decision-making processes, in the first case, or with the distribution of resources in the second. Accordingly, humans would co-produce with GenAI novel fit-for-purpose social arrangements, and evaluate those based on the emerging behaviours on MAS.

6.6.1 ISSD Detailed Overview

This Section presents the features of the current implementation of the proof-of-concept for the ISSD. The system currently runs locally, and the user inputs information in natural language in the system's interactive interface and it outputs responses in natural language as well as visualisations, such as the ones presented below. Hence, the rest of this Section presents the interface of the current

implementation of the system, an illustrative walk-through, as well as the system's modules and their operation.

6.6.1.1 System Interface

The system's interface is designed to facilitate seamless interaction between the user, GenAI, and MAS. It supports the following activities for users:

- Inputting social arrangements they wish to test within the MAS.
- Drawing inspiration for new social arrangements by querying GenAI.
- Accessing visual representations (e.g., graphs) of the outcomes from applying a social arrangement to the MAS.

An example visualisation, displayed in Figure 6.6, illustrates the impact of iterative social arrangement applications on the MAS. These spider plots depict the extent to which each cognitive dimensions of empowerment and community health metric is met, ranging from 0% to 100%. This interface is designed for web-based access following a client-server model, although for the purpose of this thesis the system was implemented and hosted locally.

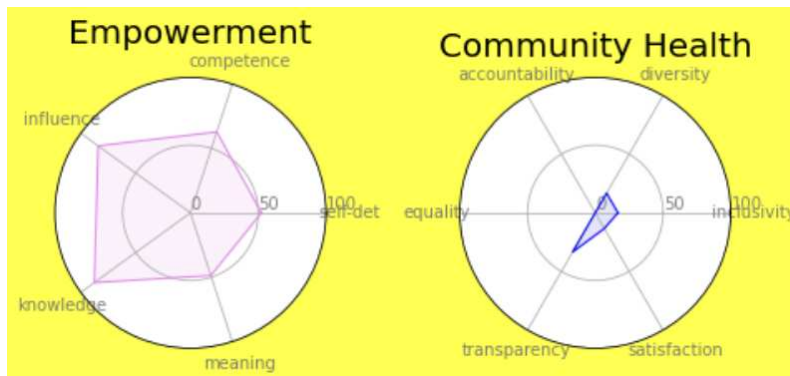


Figure 6.6: Example Visualisation of SGRL

6.6.1.2 Operational Cycle

The system's interaction and control flow are described in Algorithm 9 and illustrated in Figure 6.8, with each module discussed in the following Section. User input is highlighted in green, displayed output in blue, and module calls in red.

The operational cycle begins by initialising the MAS, setting up interaction with the GenAI API, and running an initial MAS simulation for a predefined number of rounds, m , without any proposed or implemented SAs. Upon completion of this first phase, the visualiser displays the dimensions of empowerment and community health.

The *Program* then prompts the user to decide if they wish to change the SA. If they opt for a change, they must also choose whether the new SA will be user-generated or generated by GenAI. The system identifies which parameters are underperforming and either requests a new SA from the user or queries GenAI for a suggestion. The user can either accept GenAI's suggestion or modify it, resulting in a new SA; if no change is desired, the current SA is retained for the next cycle.

This new SA is then applied in a new MAS cycle, with each round's effects saved for reference. Once the cycle completes (after m rounds), the visualiser presents the SA's impact across each cognitive dimensions of empowerment and community health metric for user review.

Finally, the user is prompted to decide whether to reset, or “bootstrap”, the system, which would restart the cycle from the initialisation stage. If the user opts not to reset, the next epoch begins, looping back to the original prompt about changing the SA.

By resetting to its starting conditions and re-running, both the user and the system can observe how different SAs—and their sequence—impact outcomes in various scenarios, enhancing learning and insights on SA effectiveness in distinct situations.

6.6.1.3 Illustrative Walkthrough

Figure 6.7 provides a step-by-step illustration of the system's operation: on the left, empowerment and community health visualisations; on the right, the user-system interaction dialogue.

In the first row, the user opts to input their own SA; the system suggests an SA to enhance the *knowledge* dimension of empowerment, and the user submits a new SA. In the second row, the user chooses to consult GenAI; the system then proposes an SA to improve the *influence* dimension, querying GenAI for a response, which the user accepts without modification. In the third row, the user decides to retain the current SA.

In each instance, the impact of the SA on the MAS is calculated, with the resulting level of empowerment—across cognitive dimensions of empowerment and community health properties—reflected by changes in the spider plot shapes.

Algorithm 9: Control Flow

```

1:  $t = 0, epochs = 0, rounds = 0, reset = false, m = \text{max rounds}$ 
2: System Initialiser. (Section 6.6.2.2)
3: for  $rounds = m$  do
4:   Baseline Scenario Executor. (Section 6.6.2.4)
5:   Parameter Collector.
6: end for
7: Visualisation  $\leftarrow$  Visualiser (Section 6.6.2.11)
8:  $rounds = 0, epochs = epochs + 1$ 
9: while  $t < T$  and not( $reset$ ) do
10:   $rounds = rounds + 1, t = t + 1.$ 
11:  Input1  $\leftarrow$  Ask if need for change of SA.
12:  if  $Input1 \neq No$  then
13:    Need Detector. (Section 6.6.2.5)
14:    Output1  $\leftarrow$  Need from Need Detector.
15:    if  $Input1 == Human$  then
16:      Input2  $\leftarrow$  Input of SA from User.
17:    else if  $Input1 == GenAI$  then
18:      GenAI Messenger. (Section 6.6.2.6)
19:      Output2  $\leftarrow$  SA from GenAI Messenger.
20:      Input2  $\leftarrow$  Ask for validation or modification.
21:    end if
22:    SA Analyser. (Section 6.6.2.7)
23:     $t_{sa} = t.$ 
24:  end if
25:  for  $rounds = m$  do
26:    Baseline Scenario Executor. (Section 6.6.2.4)
27:    SA Effect Calculator for SA provided in Input2. (Section 6.6.2.8)
28:    DoE Calculator. (Section 6.6.2.9)
29:    Opinion Formation Executor. (Section 6.6.2.10)
30:     $t = t + 1, rounds = rounds + 1.$ 
31:  end for
32:  Output3  $\leftarrow$  Informative Message from SA Effect Calculator.
33:  Visualisation  $\leftarrow$  Visualiser (Section 6.6.2.11)
34:  Input3  $\leftarrow reset = consult(user, bootstrap)$ 
35: end while

```

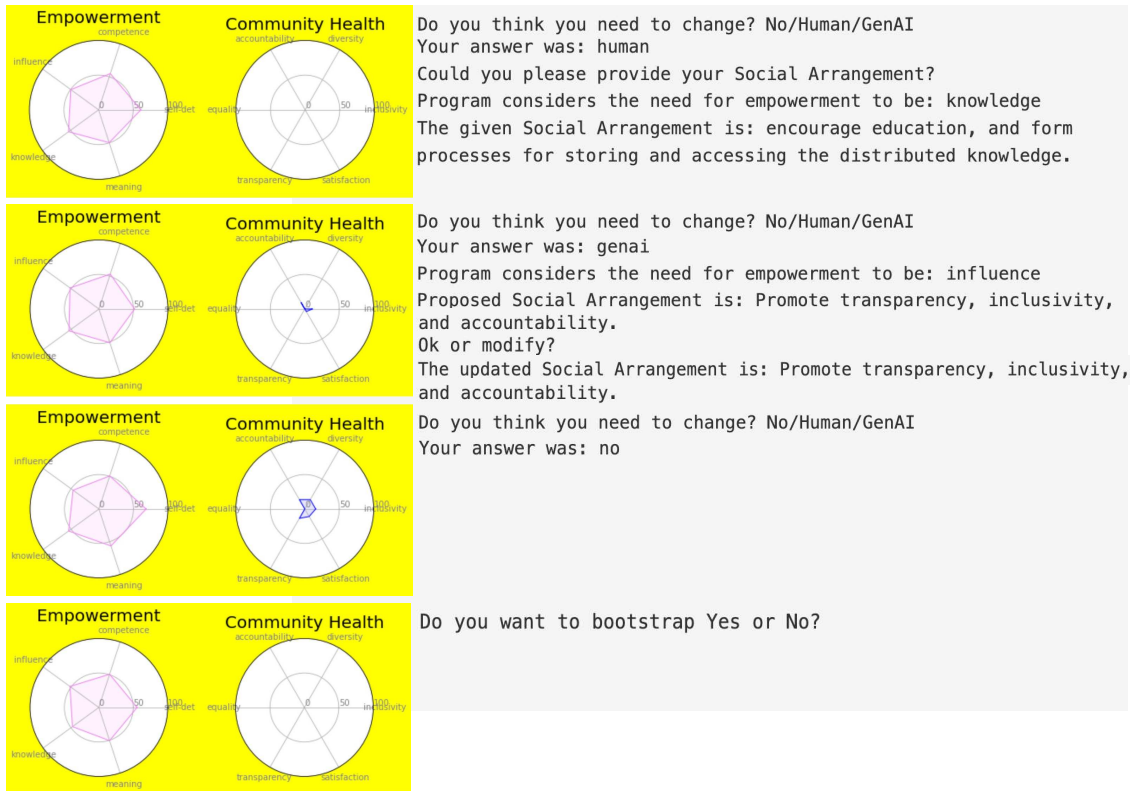


Figure 6.7: Illustrative Visualisation and Dialogue

6.6.2 ISSD Implementation

This subsection outlines the system's implementation, detailing the information processing within each module and providing a thorough specification of each component, as described in [204].

6.6.2.1 System Modules and Information Processing

The *Program* is composed of the following modules which operate in turn in each iteration:

- the *System Initialiser*, which instantiates the MAS and the independent parameters.
- the *MAS Simulator*, which generates the MAS;
- the *Baseline Scenario Executor*, which runs the MAS for an epoch (m rounds) without any SA;
- the *Need Detector*, which evaluates the state of the MAS and determines the need for empowerment;
- the *GenAI Messenger*, which interacts with GenAI API
- the *SA Analyser*, which maps SAs to dimensions of empowerment;

- the *SA Effect Calculator*, which grounds an SA to the current round;
- the *DoE Calculator* (where Dimensions of Empowerment (DoE) stands for ‘dimension of empowerment’), which calculates the cognitive dimensions of empowerment and community health properties for the current round;
- the *Opinion Formation Executor*, in which agents interact and compute the updated cognitive dimensions of empowerment and community health properties; and
- the *Visualiser*, which shows the empowerment after the iterative application of the new SA.

The modules process information as follows. The *System Initialiser* assigns parameter values based on user inputs, such as setting the number of agents to x . The *MAS simulator* then generates the MAS using these inputs, resulting in an instance of the MAS (e.g., a system with x agents). Next, the *Baseline Scenario Executor* takes the MAS instance and calculates agents’ cognitive dimensions of empowerment over m rounds without any SA. For example, for each agent a in round r , it calculates each cognitive dimension of empowerment i , producing $cd_{a,i,own,base}^r = y$, which it then aggregates as $CD_{i,own,base}^r = Y$.

The *Need Detector* receives $CD_{i,base}^r$ and determines the community’s current need, such as *influence*. This, along with existing SAs, serves as input for the *GenAI Messenger*, which communicates with the GenAI API to suggest a new SA (e.g., “Foster a culture of open communication”). This proposed SA is sent to the *SA Analyser*, which maps the SA to specific dimensions of empowerment and produces a *bonus* vector (e.g., $bonus = [\text{self-determination:1, competence:0, influence:2, ...}]$). The *SA Effect Calculator* also sets the effectiveness distribution of an SA over time, outputting the current effectiveness, such as ef_i^r for SA i in round r .

The outputs from the *Baseline Scenario Executor*, *SA Analyser*, and *SA Effect Calculator* are inputs for the *DoE Calculator*, which computes the agents’ updated dimensions of empowerment. For instance, this results in values like $cd_{i,a,own}^r = Z$ and $CH_j^r = W$. The agents’ cognitive dimensions of empowerment are passed to the *Opinion Formation Calculator*, where agents interact and choose between their own $cd_{i,a,own}^r$ or neighbours’ $cd_{i,a,nei}^r$ dimensions of empowerment, forming their final $cd_{i,a,f}^r$. The aggregated output then becomes $CD_{i,a}^r$, which, along with CH_j^r , is sent to the *Visualiser*, producing a graphical output similar to Figure 6.6.

For simplicity, references to agents’ dimensions of empowerment omit a and i . A summary of modules is provided in Table 6.2, and the system’s architecture is shown in Figure 6.8.

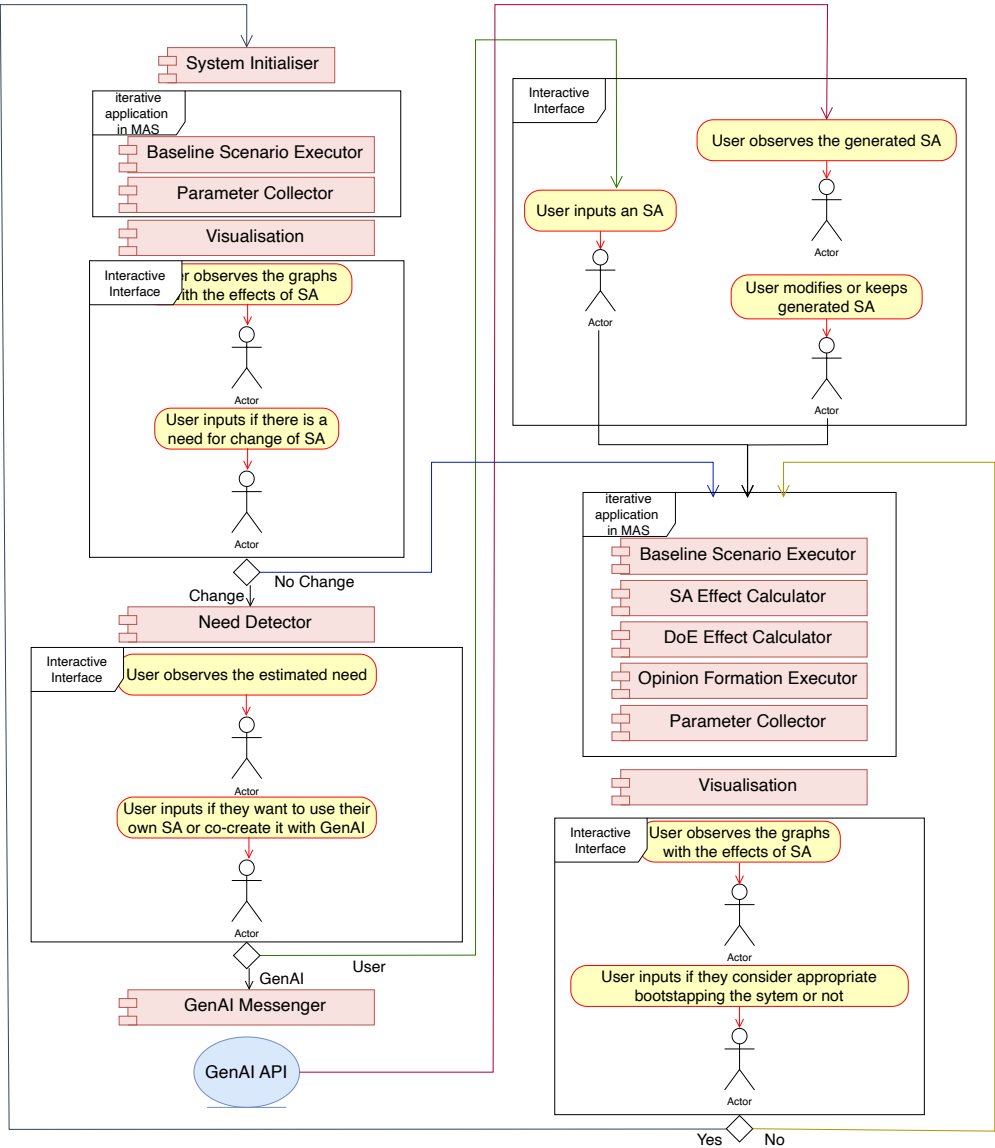


Figure 6.8: Architecture of ISSD comprising (a) human user(s), a MAS, and a GenAI agent

Modules		
Name	Input	Output
System Initialiser	Independent parameters	Instantiated independent parameters
MAS Simulator	Instantiated independent parameters	MAS instance
Baseline Scenario Executor	MAS instance	$cd_{a,base,own}^t, CD_{own,base}^t$
Need Detector	$CD_{own,base}^t$	$need_t$
GenAI Messenger	$need^t, KB$	new_SA
SA Analyser	new_SA	$bonus$
SA Effect Calculator	Gamma params., t_{sa}, t	ef_i^t
DoE Calculator	$cd_{own,base}^t, bonus, ef_i^t$	cd_{own}^t, CH^t
Opinion Formation Executor	cd_{own}^t, SN, cr, sc	CD^t
Visualiser	CD^t, CH^t	Graphs (e.g. Fig6.6)

Independent and Exp. Determined parameters			
Symbol	Description	In/ExD	Value
$\mathcal{N}$	number of agents	In	100
T, m	maximum duration in epochs and rounds in an epoch	In	50,25
KB_{init}	initial knowledge base of SAs	In	$\emptyset$
c_{gmax}, dev	max fixed agents' cognitive dimension of empowerment and allowed deviation	ExD	10,-
$-d, d$	agents' cognitive dimension of empowerment allowed deviation interval	ExD	$[-1, 1]$
k, θ, loc	shape, scale and time shift of Gamma distr.	ExD	5, 10, -
loc_f, loc_r	time shift fixed and random part	ExD	1,-
loc_{min}, loc_{max}	lower/upper bounds of time shift random part	ExD	$[-5, 5]$
mul	PDF multiplier of Gamma	ExD	-
mul_f, mul_r	mult. fixed and random parts	ExD	50,-
mul_{min}, mul_{max}	lower/upper bounds of mult. random part	ExD	$[-25, 25]$
$bonus_{max}$	max. reinforcement to a dimension of empowerment	ExD	10
τ	critical health message threshold	ExD	45
c	reinforcement of credence/confidence in a round	ExD	0.01
$cr_{a,n,init}, sc_{a,init}$	initial credence of agent a in n /self-confidence of a	ExD	0.5,0.5

Table 6.2: Modules and Experimental Parameters

6.6.2.2 System Initialiser

The initialisation of the system includes the instantiation of the MAS and the definition of the experimental parameters. Table 6.2 gives an overview of the independent and experimentally determined parameters and presents the values assigned to them in the experiments below.

6.6.2.3 MAS Definition

The MAS corresponds to a self-organising institution $\mathcal{I}$, given by:

$$\mathcal{I} = \langle \mathcal{A}, SN, KB, CD, CH \rangle$$

where $\mathcal{A}$ is the set of agents, SN is the social network, KB is the knowledge base describing known SAs for each cognitive dimension of empowerment and is initialised to be empty, CD is the list with the five cognitive dimensions of empowerment, and CD_i^t corresponds to the collective value of them in time t , and CH is the list with the collective dimension of empowerment (corresponding to community health properties), and CH_i^t is the collective value of them in time t .

Each agent a is initialised with a fixed individual value for each cognitive dimension of empowerment $cd_{a,own}$ (randomly elected from the interval $[0, cgmax]$, where $cgmax$ is an experimentally determined parameter) and in each round this value can deviate from the fixed by a random value dev from the interval $[-d, d]$, $d > 0$, e.g. $cd_{a,own}^r = cd_{a,own} + dev$. This is to generate some agents being ‘experts’, i.e. having higher cognitive dimensions of empowerment and consequently being more empowered. Also, each agent a has a social network, and has credence $cr_{a,n}$ to each of their neighbours n , but also confidence to their own cognitive dimensions of empowerment sc_a , and agents are initialised to have equal self-confidence and credence to all neighbours. Also, the upper part of Table 6.3 provides an example of KB .

KB	
Cognitive DoE	SAs
Self-determination	Encourage grassroots involvement in community governance.
	Promote active engagement.
Competence	Empowerment through education and support.
	Promoting self-advocacy and decision-making skills.
Influence	Foster a culture of open communication.
Knowledge	-
Meaning	Encourage civic education and run several assessments
	Conduct surveys to gather input and address any discrepancies.

Table 6.3: Example of Knowledge Base

6.6.2.4 Baseline Scenario Executor

In the baseline scenario, no SAs are present. During each round, agents generate their $cd_{own,base}$, and the total value of each cognitive dimension of empowerment i for the population is calculated as follows: the sum of the agents’ values for the corresponding cognitive dimension of empowerment, divided by the maximum possible value for each individual cognitive dimension of empowerment, as represented by:

$$CD_{i,base}^t = \frac{\sum_{a \in N} cd_{a,i,own}^r}{cgmax * \mathcal{N}} * 100$$

where $\mathcal{N}$ is the number of agents, and $cd_{a,i,own}^r$ is the value of each cognitive dimension of empowerment i for agent a in round r . However, the collective dimensions of empowerment are zero in this baseline scenario, i.e. $CH_{base} = 0$, reflecting the absence of community health when

there is no knowledge of SAs.

6.6.2.5 Need Detector

The *Need Detector* monitors the MAS to identify the current need for SAs that promote empowerment. It functions as an internal mechanism of interoceptive awareness [205] (the first level of interoceptive awareness) established within the MAS, comparing cognitive dimensions of empowerment and identifying any 'threats to the body politic'. Specifically, it processes the previous cognitive dimensions of empowerment, calculates the rate of change for each $CD_{i,rate}$, and produces the $need^t$, which corresponds to the cognitive dimension of empowerment i exhibiting the most significant negative rate of change in the current epoch t , represented by:

$$need^e = \operatorname{argmax}_i \frac{\sum_{e=1}^t (CD_i^e - CD_i^{e-1})}{t}$$

6.6.2.6 GenAI Messenger

The *GenAI Messenger* takes the current need $need^t$ and the knowledge base KB as inputs and constructs a message outlining the identified need along with the known SAs related to that need (i.e. the corresponding row in the KB table). For example, if the need is 'competence' and the knowledge base is as provided in Table 6.3, the message would be formatted as follows:

“The population requires a social arrangement to enhance individuals’ cognitive dimension of competence, and is already aware of the following: ‘empowerment through education and support’, ‘promoting self-advocacy and decision-making skills’. Can you suggest another one?”

This message is then sent to the GPT-4o mini API [206], and the response received corresponds to the new_SA , which is presented to the user for their approval or modification.

6.6.2.7 SA Analyser

The analysis of an SA includes the following steps:

1. the *bonus* vector initially assigns a zero to each dimension of empowerment
2. the punctuation marks of the new SA are removed and the text is converted to lowercase

3. the SA is parsed and the analyser is looking for key-terms that are included in the vocabulary
4. if a keyword is detected, then the analyser looks for 'Y' in the vocabulary and increases the value of that dimension of empowerment in the *bonus*.

The analysis of an SA includes the following steps:

1. the *bonus* vector initially assigns a zero to each dimension of empowerment
2. the punctuation marks of the new SA are removed and the text is converted to lowercase
3. the SA is parsed and the analyser is looking for key-terms that are included in the vocabulary
4. if a keyword is detected, then the analyser looks for 'Y' in the vocabulary and increases the value of that dimension of empowerment in the *bonus*.

For instance, if the *new_SA* suggests to 'Foster a culture of open communication, collaboration, critical thinking, and ethical behaviour to engage individuals in meaningful contributions to collective knowledge, and decision-making', then the corresponding *bonus* might be:

bonus = [self-determination: 3, competence: 2, influence: 3, knowledge: 3, meaning: 4, inclusivity: 3, transparency: 2, diversity: 0, equality: 1, accountability: 1]

6.6.2.8 SA Effect Calculator

Drawing an analogy to a community where individuals need time to process and adapt to changes, the system is structured so that each time a new SA is implemented in the MAS, the population requires a certain period to comprehend it, engage with it, and reap its benefits, as noted in [152]. Additionally, over time, the population or environment may evolve, rendering the SA potentially obsolete. To accommodate this, we assume that the effectiveness of an SA follows a Gamma distribution. Furthermore, since different SAs necessitate varying amounts of time for acceptance and can have differing levels of effectiveness, a new Gamma distribution is generated for each SA, characterised by specific properties.

The shape k and scale θ parameters of the Gamma distribution for each SA are determined experimentally. Additionally, the distribution is adjusted over time by a parameter called *loc*, which consists of a fixed component loc_f and a random component loc_r drawn from a uniform distribution within the $[loc_{min}, loc_{max}]$ interval. This results in $loc = loc_f + loc_r$, where $loc_r \sim Uniform(loc_{min}, loc_{max})$, and loc_f , loc_{min} , and loc_{max} are experimental parameters. To

Term	S-D	C	Infl	K	M	Incl	D	A	Eq	T
Active Participation	Y	N	N	N	N	N	N	N	N	N
Ethical Behaviour	N	N	N	N	Y	N	N	Y	Y	N
Core Values	N	N	N	N	Y	N	N	N	N	N
Open Communication	N	N	Y	N	N	Y	N	N	N	Y
Address Inconsistencies	N	N	N	N	Y	N	N	N	N	N
Distributed Knowledge	N	N	Y	Y	N	N	N	N	N	N
Knowledge Sharing	N	N	Y	Y	N	Y	Y	N	N	N
Gather Input	N	N	Y	Y	Y	Y	Y	N	Y	N
Civic Engagement	Y	N	N	N	N	N	N	Y	N	N
Civic Participation	Y	N	N	N	N	Y	N	N	N	N
Civic Education	N	Y	N	Y	N	Y	N	Y	Y	N
Critical-thinking	Y	Y	N	N	Y	N	N	N	N	N
Decision Making	Y	N	N	N	Y	N	N	N	N	N
Collective Knowledge	N	N	Y	Y	N	Y	N	N	N	N
Diverse Perspectives	N	N	Y	N	N	N	Y	N	N	N
Social Network	N	N	Y	N	N	N	Y	N	N	Y
Education	Y	N	N	N	N	N	N	N	N	N
Skill-building	Y	N	N	N	N	N	N	N	N	N
Participation	N	Y	N	Y	Y	N	N	Y	Y	N
Collaborate	Y	N	Y	N	N	N	N	N	N	Y
Feedback	N	N	Y	N	N	N	N	N	N	N
Assess	N	N	N	N	Y	Y	Y	N	Y	N
Values or Principles	N	N	N	N	Y	N	N	N	N	N
Ethics	N	N	N	N	Y	N	N	N	N	N
Safe or Support	N	N	N	N	Y	Y	N	N	N	N
Ownership	Y	N	N	N	Y	N	Y	Y	N	N
Accountability	Y	N	N	N	Y	N	Y	Y	N	N
Memory	N	N	N	Y	N	N	N	N	N	N
Document or Record	N	N	N	Y	N	N	N	N	N	N
Sources or Evidence	N	N	Y	Y	N	N	N	N	N	N
Dissent or Concerns	N	N	N	N	Y	N	Y	N	N	N
Community	Y	N	N	N	N	N	N	N	N	N
Inclusive	Y	N	N	N	Y	N	N	N	N	N
Dialogue or Voice	Y	N	N	N	N	N	N	N	N	N
Diversity or Identity	Y	N	N	N	N	N	Y	N	N	N
Culture	N	N	N	Y	N	N	N	N	Y	N
Structure	N	Y	N	Y	N	N	N	N	N	N
Transparent or Open	N	N	N	N	N	N	N	N	Y	Y
Act	Y	N	N	N	N	N	N	N	N	N
Advocate or Negotiate	Y	N	Y	N	N	N	N	N	N	N
Attend or Listen	Y	N	Y	N	N	N	N	N	Y	N
Opinion or Viewpoint	N	N	Y	N	N	Y	Y	N	Y	N
Consistent	N	N	N	N	Y	N	N	N	N	N
Democracy	Y	N	N	N	N	Y	N	N	N	N
Initiatives	Y	N	N	N	N	N	N	N	N	N
Awareness	N	Y	N	N	Y	N	N	N	N	N
Skill	N	Y	N	Y	N	N	N	N	N	N
Shape	Y	N	Y	N	N	N	N	N	N	N
Distributing	N	N	Y	Y	N	N	Y	N	N	N
Critical	Y	Y	N	Y	Y	N	N	N	N	N
Problem-solving	N	Y	N	Y	N	N	N	N	N	N
Equality	N	N	N	N	N	N	N	N	Y	N
Quality	N	Y	N	N	Y	N	N	N	N	N
Share	Y	N	Y	Y	N	Y	Y	N	Y	Y
Autonomy	Y	Y	N	N	N	Y	Y	Y	Y	N
Management	N	Y	N	Y	N	N	N	N	N	Y
Reasoning	Y	Y	N	Y	Y	N	N	N	N	N
Align	N	N	N	N	Y	N	N	Y	N	N
Meaningful	N	Y	N	N	Y	N	N	N	N	N

Figure 6.9: Vocabulary used for the testing of the system

ensure that the probability density function (PDF) of the Gamma distribution yields values ranging from -1 to 1 within the specified interval, its output is multiplied by mul , which also consists of a fixed mul_f and a random mul_r drawn similarly to the parameters of the shift.

Thus, this module takes as input the experimental parameters k , θ , loc , loc_{min} , loc_{max} , mul , mul_{min} , and mul_{max} , along with the current time t and the number of rounds since the SA was applied t_{sa} . It calculates the effectiveness $ef_i^{t-t_{sa}}$ of SA i at time t . This is represented as the value of the probability density function (PDF) of the Gamma distribution at the current time t minus the time when the new SA was first applied t_{sa} , multiplied by mul and adjusted by loc .

$$\begin{aligned} ef_i^{t-t_{sa}} &= mul * f(t - t_{sa}; k, \theta) + loc \\ &= mul * \frac{(t - t_{sa})^{k-1} e^{-\theta(t-t_{sa})} \theta^k}{\Gamma(k)} + loc \end{aligned} \quad (6.1)$$

In addition, this module incorporates a secondary control mechanism that establishes a second layer of interoceptive awareness (second-level interoceptive awareness mechanism developed in human-computer interaction), building upon the initial layer developed in the *Need Detector* (within the MAS). This mechanism calculates the gradient of the effectiveness distribution at the current time point following the application of the SA ($t - t_{sa}$) and informs the user whether the SA has attained its maximum effectiveness (indicated by a negative gradient) or not (indicated by a positive gradient).

Based on the sign of $\nabla ef_i^{t-t_{sa}}$, the system presents the user with messages such as ‘*Maximum effectiveness has been reached*’ or ‘*Wait more for maximum effectiveness*’. Moreover, if the SA has achieved maximum effectiveness and any of the dimensions of empowerment falls below a specified threshold τ , a warning message stating ‘*Community is disempowered, consider a change!*’ will be displayed.

6.6.2.9 DoE Calculator

The *DoE Calculator* processes the baseline cognitive dimensions of empowerment, the *bonus* associated with a specific SA i , and its effectiveness ef_i^t to determine the values of the agents’ dimensions of empowerment. Specifically, the cognitive dimension of empowerment for agent a is denoted as $cd_{a,i,own}^t$, while the population’s cognitive dimension of empowerment $CD_{i,own}^t$ is calculated as the average of $cd_{a,i,own}^t$ for all agents a in $\mathcal{A}$. The collective dimensions of empowerment are

represented as CH_i^t , defined by the following equations:

$$cd_{a,i,own}^t = \frac{cd_{a,i,base}^t \cdot bonus \cdot ef_i^t}{bonus_{max}},$$

$$CH_i^t = \frac{bonus \cdot ef_i^t}{bonus_{max}}$$

In these equations, *bonus* refers to the *bonus* associated with the current SA, and $bonus_{max}$ is an experimental parameter that indicates the maximum bonus that an SA can provide to a dimension of empowerment.

6

6.6.2.10 Opinion Formation Executor

The *Opinion Formation Executor* facilitates agent interactions and the development of their final opinions regarding cognitive dimensions of empowerment. Agents engage with their social networks and can choose to rely on their own cognitive dimensions of empowerment or seek input from a neighbour. This interaction mimics a community where agents can either use their own judgements or consult a trusted source to enhance their decision-making, as outlined in [12].

Each agent a possesses their own cognitive dimension of empowerment $cd_{a,own}^t$ and selects one of their most credible neighbours n to obtain the neighbour's cognitive dimension of empowerment $cd_{n,own}^t$ (denoted as $cd_{a,nei}^t$). The decision on whether to use their own dimensions of empowerment or those of the neighbour depends on the credence $cr_{a,n}$ given to the neighbour and the agent's self-confidence sc_a . This mechanism allows expertise to emerge and optimises the decision-making process within the MAS.

The final cognitive dimension of empowerment for agent a is expressed as follows:

$$cd_{a,f} = \begin{cases} cd_{a,nei}^t, & \text{if } cr_{a,n} \geq sc_a \\ cd_{a,own}^t, & \text{if } cr_{a,n} < sc_a \end{cases} \quad (6.2)$$

Following this decision, agent a updates their credence for neighbour n and their self-confidence based on whether the average of n 's cognitive dimensions of empowerment exceeds that of agent a . The degree of reinforcement per round is governed by an experimentally determined parameter denoted as c . The adjustments to credence are calculated with the following equation, while the

opposite adjustment applies to self-confidence:

$$cr_{a,n} = \begin{cases} cr_{a,n} * (1 + c), & \text{if } \sum_{i \in CD} cd_{i,a,nei} \geq \sum_{i \in CD} cd_{i,a,own} \\ cr_{a,n} * (1 - c), & \text{otherwise} \end{cases} \quad (6.3)$$

It is important to note that the opinion formation influences only the cognitive dimensions of empowerment of the agents, not the community health properties. While agents can individually enhance their cognitive capabilities, community health is an emergent quality that results from cognitive empowerment and overall well-being. Therefore, it cannot be improved merely by consulting an external source.

6.6.2.11 Visualiser

After executing the MAS for m rounds, the system must communicate the effects of a specific SA to the user. To achieve this, the *Visualiser* processes the dimensions of empowerment and produces informative graphs. An illustrative example of this is shown in Figure 6.6.

6.6.3 Evaluation as a Tool for Empowerment

The system developed is non-deterministic, meaning the sequence of events holds more importance than the initial conditions, so varying the input order can lead to different outcomes. Nonetheless, this Section provides proof-of-concept examples of experimenting with the implemented system that illustrate the emergence of interoceptive awareness, the diverse effects of SAs, the user's role, and the emergence of expertise through bootstrapping. This section also concludes with a discussion of future improvements of the current implementation of the system.

6.6.3.1 Interoceptive Awareness

To support community empowerment through deliberation on SAs, the system includes mechanisms that inform users and help them develop interoceptive awareness regarding the system's overall health (a second-level interoceptive awareness mechanism). Figure 6.10 illustrates this feature: initially, the user recognises a need for change and engages GenAI (first grey box). The suggested SA is “establish clear communication channels, utilise collaborative tools, encourage knowledge sharing and collaboration, and regularly review and update knowledge repositories”. This proposal

is approved, progressively increasing empowerment, especially in knowledge, influence, diversity, inclusivity, transparency, and satisfaction.

Over time, however, the SA loses effectiveness, resulting in a decline in the dimensions of empowerment, as reflected in the spider plots. The system identifies this downturn and issues a warning (highlighted in red in the figure).

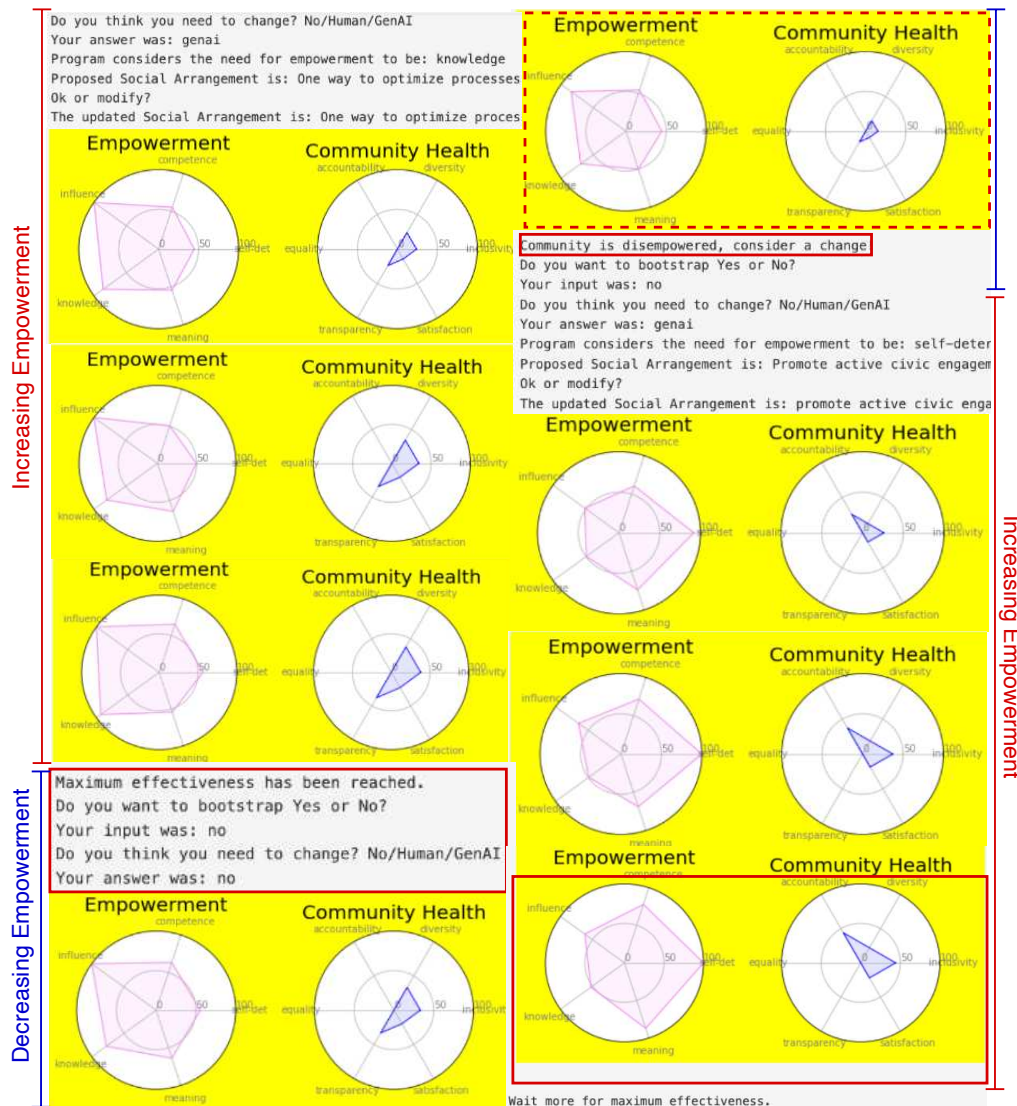


Figure 6.10: Snapshots from the interaction with the systems showing how it enables the users to develop interoceptive awareness.

The user then initiates a new GenAI query, proposing the SA, ‘Promote active civic engagement and participation, providing equal opportunities for meaningful involvement in decision-making’. This new SA is approved and results in an increase in the dimensions of empowerment.

Thus, the system exhibits a form of self-healing enabled by effective user-system interaction.

Specifically, the system alerts the user to a need for intervention, the user responds by initiating a change, and this change is implemented, allowing the system to recover from a state of reduced empowerment.

6.6.3.2 Variety of SA Effects

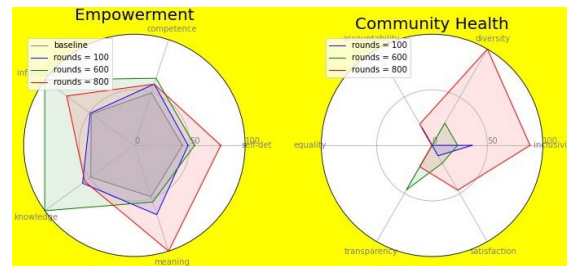


Figure 6.11: Comparison of Effects of SAs and Baseline

As previously discussed, various SAs produce distinct outcomes across the dimensions of empowerment. Figure 6.11 compares the dimensions of empowerment in the baseline condition to those resulting from the application of two different SAs: “Establish clear communication channels, utilise collaborative tools, and regularly update and share information among team members” introduced at round 600, and “Conduct regular surveys or feedback sessions to gather input and address any discrepancies” applied at round 800.

Additionally, Figure 6.12 shows the effect of the following SAs:

- Regularly review and assess practices against core values, and provide a platform for dissent to be voiced.
- Promote inclusive and transparent decision-making to empower individuals to actively participate in shaping their communities.
- Foster a culture of inclusivity, transparency, critical thinking, and accountability.
- Establish clear communication channels, utilise collaborative tools, and encourage knowledge sharing.
- Regularly review practices against core values and encourage open communication for dissent.

Each SA addresses a unique need, resulting in distinct dimension of empowerment, with all outperforming the baseline scenario. For instance, the first SA focusing on accountability primarily enhances meaning, diversity, and inclusivity, while the fourth SA strengthens influence, knowledge,

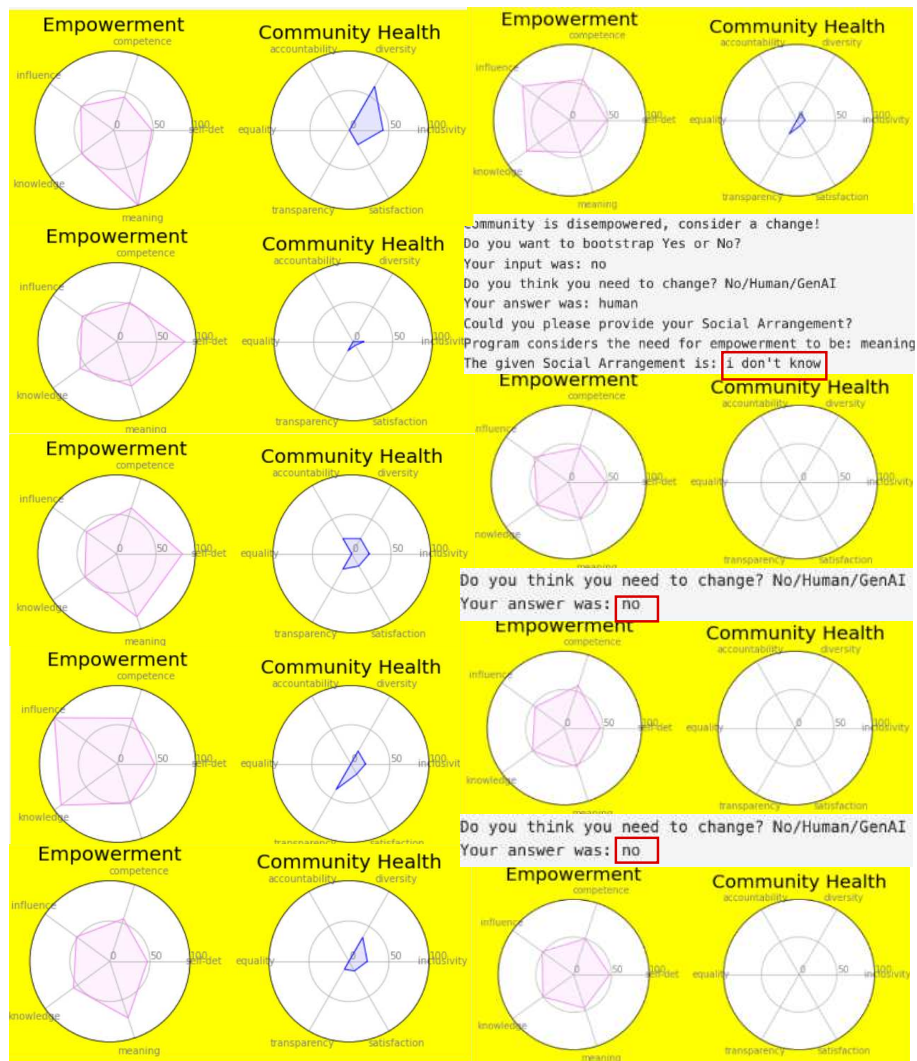


Figure 6.12: Snapshots of the systems showing how different SAs result in increasing different dimensions of empowerment and that human interest is required to maintain empowerment

and transparency. This underscores the importance of tailoring SAs to system needs—a principle guiding the system’s need detection to optimise decision-making in the context of SAs.

6.6.3.3 Importance of Human Interest and Patience

Emphasising the user’s role, the right side of the second row in Figure 6.12 illustrates an interaction where the user appears indifferent and, as a result, dimension of empowerment not benefit from the system’s guidance. Specifically, although the user initially chooses to make a change, their input is irrelevant (e.g., “I don’t know”). In subsequent epochs, even as the dimension of empowerment decline, the user fails to intervene—for example, they do not initiate a change, as highlighted by red boxes in the figure (e.g., “no”). This indifference and disregard for system warnings ultimately lead

to both individual and collective disempowerment, evidenced by low dimension of empowerment.

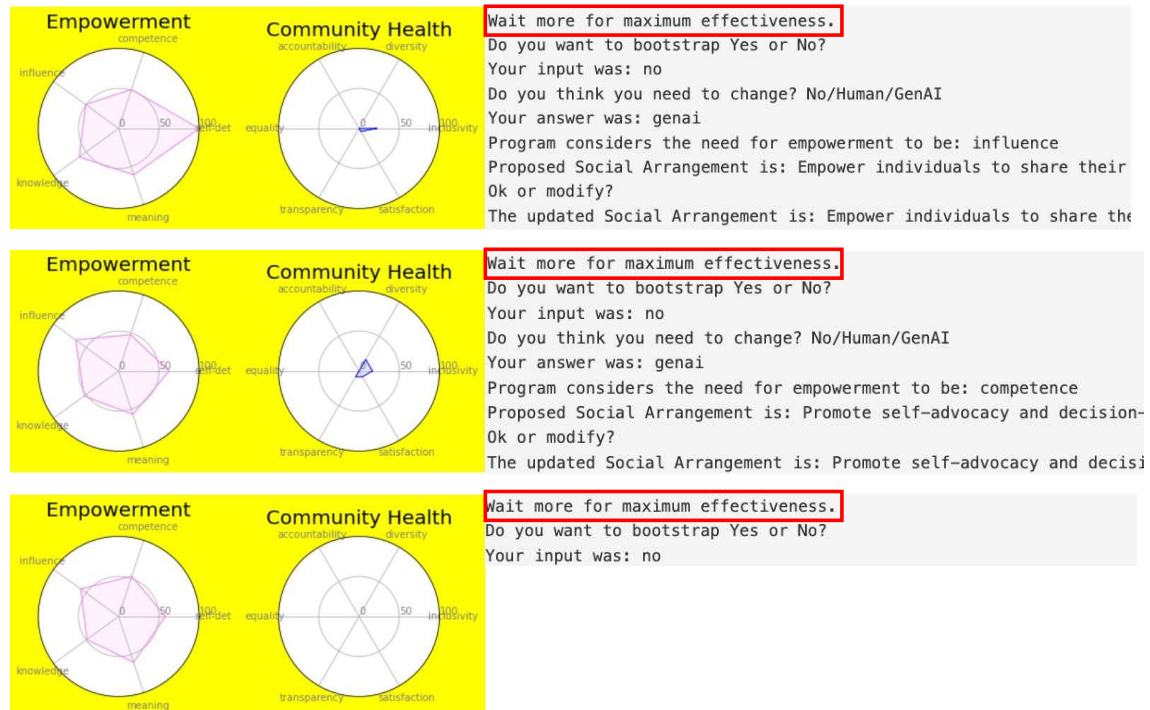


Figure 6.13: Snapshots of the system showing that humans need to be patient in order to see the impact of the different SAs

An essential factor for the system's effective operation is the user's patience. Figure 6.13 presents a scenario where the user repeatedly asks GenAI to suggest a new SA without allowing sufficient time for each one to take effect. Specifically, the AI-generated SAs in this example are as follows:

- Encourage individuals to engage in shaping their communities through inclusive decision-making processes;
- Support people to share perspectives, listen actively, seek evidence-based solutions, and uphold ethical principles;
- Promote self-advocacy and decision-making.

Although the SAs have the potential to benefit the system, their impact remains unseen because the user impatiently switches from one SA to the next. This constant change prevents the MAS population from fully engaging with each SA, which would otherwise lead to an increase in dimension of empowerment. Thus, it is crucial for the user to follow the system's guidance (i.e. the message "Wait more for maximum effectiveness") and allow time for the effects of each change to manifest.

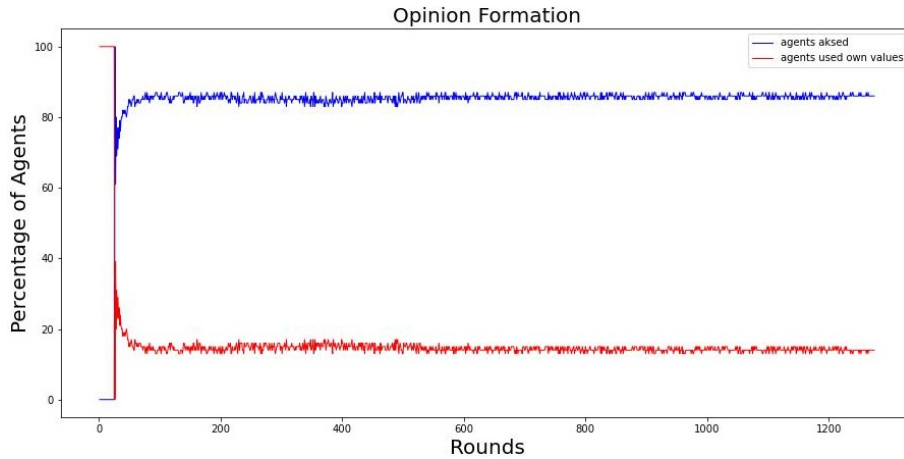


Figure 6.14: Selected Cognitive dimension of empowerment showing the emergence of expertise, since agents tend to use the opinion of the “experts”

6

6.6.3.4 Opinion Formation in MAS

In the *Opinion Formation Executor*, agents have the option to consult a source (a neighbouring agent from the social network with higher cognitive dimension of empowerment). This functionality demonstrates the emergence of expertise within the system, as agents learn to identify “experts” and actively seek their input based on cognitive dimension of empowerment.

This emergent property is illustrated in Figures 6.14 and 6.15. In the first column, the number of agents using either their own (in red) or their neighbours’ (in blue) cognitive dimension of empowerment to form their final cognitive dimension of empowerment is shown. The second row displays the population’s cognitive dimension of empowerment when they rely on their own (in red) versus those they ultimately select from neighbouring sources (in blue). The first graph indicates that agents in the MAS learn to draw on the cognitive dimension of empowerment of trusted sources, leading to the higher cognitive dimension of empowerment evident in the second graph. This process highlights that agents not only consult their neighbours but also discern experts within the network, effectively enhancing the community’s collective cognitive capacity and fostering empowerment.

6.6.4 Evaluation as a Cybernetic System

Previous Sections discussed how patterns of higher-order cybernetics foster human empowerment. Accordingly, here we evaluate ISSD as a cybernetics system.

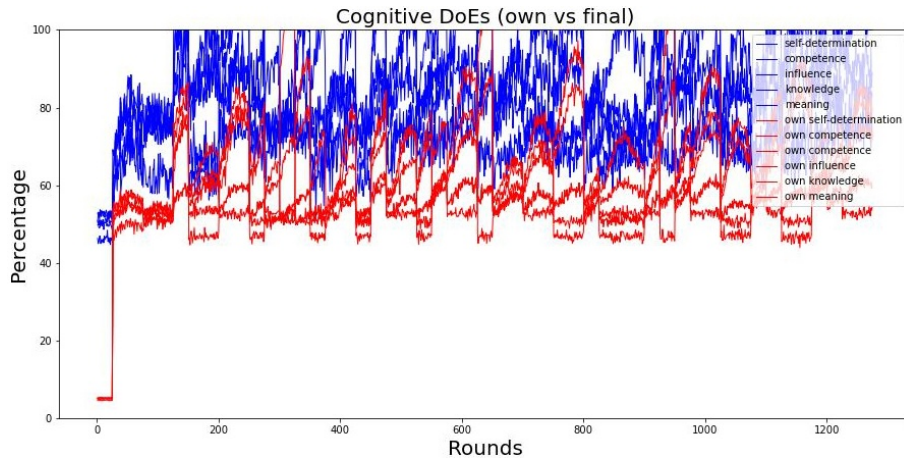


Figure 6.15: Own and Selected cognitive dimension of empowerment showing the emergence of expertise, since agents tend to use the opinion of the “experts”

6.6.4.1 First Order Cybernetics

First Order Cybernetics examines self-regulated systems that operate in a closed manner, featuring an input, an output, and negative feedback mechanisms [16]. In this type of system, the MAS component functions as a system that accepts agent properties as input, performs internal measurements, and produces values for dimensions of empowerment as output. These output values serve as negative feedback, which subsequently identifies a new need, indicating that the MAS component aligns with a first-order cybernetic system.

Thus, when analysing our system through the MAS, it can be observed as a first-order system that receives SA as input, processes it, and outputs the values of the dimensions of empowerment. Based on this output, the system can determine needs, as depicted in Figures 6.10 and 6.13 above. In this process, the system utilises an internal, first-level mechanism of interoceptive awareness (such as the *Need Detector*) to organise itself effectively.

6.6.4.2 Second Order Cybernetics

Second Order Cybernetics explores systems influenced by an observer. When viewing the MAS as a black-box system and considering the human user as an observer capable of affecting it, our system corresponds to a second-order system. For instance, in Figure 6.12, the user can view the values of the dimensions of empowerment (output of the black-box system) and input a new SA, which would alter the system and yield different results.

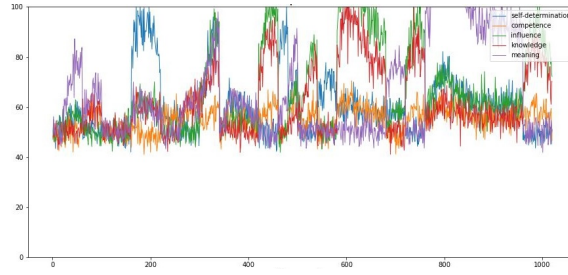


Figure 6.16: Systems with an active observer showing the development of interoceptive awareness leading to increased cognitive dimensions of empowerment

6.6.4.3 Third Order Cybernetics

Third Order Cybernetics focuses on systems involving a regulated system and an active observer who simultaneously holds an outsider and insider role within that system. In such a scenario, the system and observer co-evolve, with each capable of affecting and being affected by the other [91].

When applying the model-view-controller pattern in our system and regarding the user as not merely an external observer or controller but as an internal component of the model, the system achieves the characteristics of a third-order system. The user thereby assumes both an outsider (observer) and an insider (integral part of the model and MAS) role, using the system's first level of interoceptive awareness (in the need detector) to identify empowerment needs (i.e. cognitive dimensions of empowerment) and to influence SA changes. These changes are applied to the MAS, impacting both the MAS and the user as an integral part of the system.

The system (comprising the user and MAS) thus employs its interoceptive awareness to adjust in ways that foster increased levels of individual empowerment (i.e. enhanced cognitive dimensions of empowerment values). This dynamic is illustrated by the fluctuations in Figure 6.16, where the x -axis denotes time and the y -axis reflects the percentage of cognitive dimension of empowerment.

6.6.4.4 Fourth Order Cybernetics

Fourth Order Cybernetics [207] addresses systems embedded in an environment, characterised by contextualisation and self-awareness, and containing mechanisms for self-healing and regeneration. These capabilities arise through embodied knowledge, which equates to knowledge acquired through experience. If we “zoom out” from the isolated model-view-controller system to view it as embedded within a community that determines SA for equitable power distribution, it becomes evident that collective empowerment and community health are paramount. In this scenario, a user capable of developing interoceptive awareness of community health, as described in Section

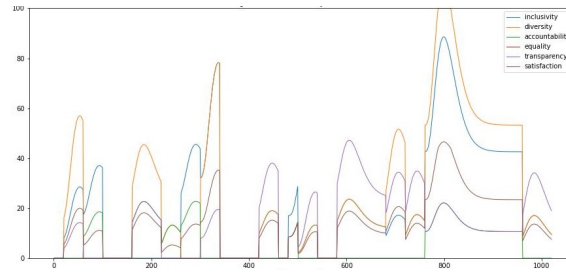


Figure 6.17: Systems with re-generating mechanisms showing the capacity of the system to heal itself and recover from situations in which the community health is low

6.6.3.1, can instigate self-healing and regeneration within the system. This dynamic is highlighted in Figure 6.17, where the system alerts the user to the importance of community health, prompting action that ultimately leads to increased empowerment. Note that the graph's x -axis represents system time, while the y -axis denotes the values of the collective dimensions of empowerment.

6.6.4.5 Further Work on ISSD

Although substantive results have been demonstrated as a proof of concept for empowerment and with respect to orders of cybernetics, this system is at present a work in progress, and there is a number of limitations to overcome and improvements to be made in further research and development. These include:

- in the *SA Analyser*, we need a substantive user survey to establish a stronger correlation between keywords and the dimension of empowerment;
- in the *MAS*, a method for mapping the features and characteristics of a community to the MAS is required;
- in the *Need Detector*, we need to develop a more complex mechanism for interoceptive awareness that takes into consideration multiple parameters;
- in the *Visualiser*, although some text entry (of SA) is required, we should upgrade the text-based dialogue to improve the UX;
- generally, considering groups of users or experts could improve its effectiveness of the system; and
- overall, we aim to package the system as a plug-in for PlatformOcean [208] and evaluate performance in field trials.

However, in the next Section we discuss the potential of the SGRL methodology to support humans mitigate ‘wicked’ problems.

6.6.5 Social Implications of SGRL Methodology

This Section outlines the enhanced value of employing a tool to support innovation leading to community empowerment within self-governing socio-technical systems. It examines the implications for addressing complex societal challenges and reducing existing asymmetries in the distribution of power.

6

6.6.5.1 Managing Wicked Problems

Wicked problems represent a category of social challenges whose inherent characteristics (as previously outlined) prevent them from being ‘fixed’ in a conventional manner. A socio-technical system designed and implemented to address such problems must be correspondingly open-ended to manage incomplete, inconsistent, and conflicting information and requirements. Consequently, the system itself must be sustainable while simultaneously meeting the collective goals of the community’s individuals.

These attributes indicate that effectively addressing wicked problems necessitates reflective and progressive solutions, drawn from a potentially extensive ‘solution space.’ This may involve continuous monitoring and modification of solution approaches (which correspond to SAs) while avoiding the selection of those that could be harmful in the long term. In the absence of opportunities for trial and error, this necessitates making ‘safe’ choices (Popperian evolution [209]). Moreover, the imperative for innovation—both reactive to unforeseen changes and proactive in anticipation of expected developments—poses a potential limitation on conventional decision support systems.

However, as evidenced by AlphaGo, where the interaction between human and artificial intelligence appeared to inspire remarkably ‘creative’ and ‘imaginative’ moves, an advanced innovation support system could effectively combine the strengths of both, facilitating the integration of knowledge and expertise to foster more adaptive institutions. Thus, while we do not assert that innovation support systems provide a definitive solution to global challenges, we contend that the synergy between human and artificial intelligence aimed at co-producing innovation can yield local benefits and serve as a valuable contribution to the agenda of public interest technology (PIT) research.

For instance, by employing the SGRL methodology for innovation support, users can engage with the system to collaboratively generate new SAs using GenAI. Visualising the outcomes of the iterative application of these new arrangements to the MAS allows for the assessment of both short-term and long-term consequences. In this manner, users and AI can mutually support each other in co-creating more suitable solutions for addressing problems and adapting to evolving circumstances. Therefore, the proposed innovation system presents an opportunity for more adaptive, creative, responsible, and fit-for-purpose solutions.

This mutual support hinges on participatory design: innovation support systems must provide affordances that assist users lacking proficient technological knowledge. Furthermore, the community must be instructed on how to utilise the system, engage with it effectively, and interpret its results. This process extends beyond mere design to encompass participatory consultation as much as it dimension of empowerment design itself. One avenue is the creation of self-explanatory systems that provide clear explanations and justifications for their proposals.

Another essential step is to enhance user experience (UX) design to facilitate and improve user interactions. A more challenging endeavour is to educate individuals not only on how to utilise these tools but also on the significance of participating in the processes of deliberation regarding social arrangements. Additionally, UX design must address the reduction of transaction costs for participation in social arrangements, akin to the way classical Athenian democracy incentivised involvement in knowledge aggregation processes, while also ensuring equitable participation in these processes.

6.6.5.2 Equitable Distribution of Power

As previously discussed, intelligent systems are transforming the global economy and are anticipated to reshape the (self-)organisation of socio-technical systems. While humans can benefit from the co-production of new content with AI, they are simultaneously exposed to systemic risks that threaten sustainability, such as algorithmic bias, unequal access and benefits, and cascading failures.

Moreover, the growing adoption of AI systems by government agencies can positively influence social dynamics; however, it also poses a risk of exacerbating social inequality. Specifically, the adoption of AI may increase disparities in the distribution of power in three ways: by transferring existing inequalities from the social sphere to the socio-technical system, by utilising AI as a proxy to intensify existing inequalities, and by inadvertently conceding power to AI. This presents an

opportunity to leverage AI in a constructive manner, specifically to alter or alleviate inequities in the distribution of power. Such an objective can be realised through the development of tools aimed at enhancing public awareness of power asymmetries. For instance, existing literature describes a meta-platform designed to facilitate multi-stakeholder participation in addressing problems that necessitate collective action.

Consequently, we contend that the SGRL exemplifies a system capable of both democratising technology and enabling individuals to derive benefits from AI technology. Specifically, individuals can utilise the system to determine their own SAs, particularly those pertaining to power distribution, envision the effects of power dynamics on the social system through visualisation of the results derived from application within the MAS, modify these arrangements either independently or based on suggestions from GenAI, and ultimately identify those that would mitigate asymmetries in the distribution of power.

Overall, this system employs the SGRL methodology to democratise individuals' understanding of existing asymmetries and the relationship between SAs and power dynamics. This democratisation of understanding has the potential to enhance public awareness, which may serve as a significant step toward mitigating inequalities in the distribution of power.

6.6.6 Conclusions on Innovation for Empowerment through ISSD

Section 6.5 performed a conceptual analysis of innovation which resulted in the establishment of a process-oriented definition of innovation, in the context of self-governance. Based on this definition, this Section introduced a tool for innovation support, called ISSD, that follows the SGRL methodology, which was introduced in Section 3.4, to support human empowerment in hybrid societies. This tool was also evaluated as a proof-of-concept and as a cybernetics system.

The experimental results showed that ISSD helps people design their communities better, using a model of equitable society as a guide. This is because it enabled users to visualise the extent of their empowerment in cognitive and collective dimensions, and, through a synthesis of GenAI, MAS, and human feedback, to explore and evaluate the impact of new SAs on their empowerment. Finally, this Section explored the social implications of the SGRL methodology in innovation support, examining issues related to wicked problems and a fair distribution of power. This demonstrated that human-AI collaborative production of new SAs empowers individuals to use AI technology not merely for decision support but to achieve innovation support.

Overall, study of innovation and SGRL for developing a tool for innovation support provides insights on sustainable self-governance, unrestricted self-modification of rules (i.e. SAs), and protection of dignity. The former is achieved through the continuous and informed modification of SAs leading to maintaining a high level of individual and collective empowerment. The second is because humans can exercise control over their SAs, generate new, evaluate them, and modify then without restrictions. Finally, proposed tool, i.e. ISSD, fosters dignity by enabling humans making informed decisions with respect to issues of governance.

6.7 Evaluation & Metrication

This Section outlines other potential metrics that could be used to inform the humans about the effectiveness of SAs, especially in cases that groups of users corresponding to different stakeholders are taking part into the processes. A series of indicative metrics is presented in the form of queries in Table 6.4.

Metric	Query
<i>Innovation-Specific</i>	
Novelty	Is it bringing something new in?
Alignment	Is it aligned with the core values?
Effectiveness	Is it solving an existing or upcoming problem?
Expense	Is it economic?
Flexibility	Is it applicable in both short- and long-term?
Holism	Is it fixing a problem without creating another?
<i>Population-Specific</i>	
Coherence	dimension of empowerment it fragment the population?
Engagement	Do people engage with it?
Satisfaction	dimension of empowerment it increase the satisfaction of the population?
Equality	dimension of empowerment it treat people as equals?
Participation	dimension of empowerment it encourage public participation?

Table 6.4: Metrics for evaluating innovation for empowerment

Initially, we identify two broad categories of metrics: population-specific and innovation-specific (or policy-specific). Specifically, innovation-specific metrics aim to evaluate the proposed policy (i.e. social arrangement) and include novelty, alignment, effectiveness, expense, flexibility, and holism. Novelty assesses whether the chosen policy introduces a new concept within the system that produces different outcomes, rather than merely representing a ‘rebranding’ of an existing social arrangement. Alignment pertains to the conformity of the innovation with the core values

of the system, which correspond to the foundational principles, ethics, and characteristics of the population.

Furthermore, effectiveness examines whether the policy addresses an existing or impending issue or enhances the well-being of the population. Should this novelty fail to achieve any of the aforementioned objectives, it becomes undesirable, as it heightens the likelihood of instability and incoherence without providing any benefits. Another metric related to effectiveness is expense, which evaluates whether the introduced novelty is feasible and capable of solving a problem without depleting the (physical or conceptual) resources of the system. Additionally, flexibility integrates effectiveness with time, assessing whether the new idea is applicable and useful in the short term or constitutes a sustainable solution that dimension of empowerment not necessitate changes over an extended time horizon. The final innovation-specific metric, holism, examines whether the proposed change optimises one dimension while causing deterioration in multiple others.

In terms of population-specific metrics, coherence evaluates the extent to which the new policy is publicly accepted and engenders a consensual agreement or, conversely, fragments public opinion. Closely related to coherence is engagement, which considers whether the population is actively involved in and willingly adopting the change. Moreover, satisfaction measures the overall utility experienced by the population, reflecting the well-being of individuals. Equality addresses whether the chosen social arrangement treats individuals equitably, while participation assesses whether the selected policy encourages individuals to engage actively in the commons, propose ideas, and voice concerns.

It is essential to incorporate both innovation-specific and population-specific metrics in human-centred innovation support systems. Evaluating new social arrangements solely concerning their effectiveness in optimising certain dimensions of the system is inadequate; it is equally necessary to capture the reactions and emotions of individuals who are part of those systems, as discussed in the literature.

Moreover, measurements should be conducted repeatedly throughout the application and testing phase for two primary reasons: first, to adhere to the Nyquist Theorem and avoid evaluating metrics that may yield incorrect assessments due to their evolving nature; and second, to facilitate analysis of both the immediate effects arising from their implementation and their short-term and long-term implications. Additionally, the manner in which these properties (described as metrics) are measured and presented should evolve over time to prevent situations where a metric becomes a target, in accordance with Goodhart's Law.

Furthermore, within multi-stakeholder socio-technical systems, various systemic drivers may conflict with one another, necessitating the definition of multiple metrics as previously outlined, alongside their continual monitoring to account for the system's performance in relation to all these drivers. Additionally, an inherent contradiction exists between the willingness to change (which aims for long-term sustainability) and reluctance to change (prompted by identity and entropy). Specifically, while a critical mass must agree to implement change, it is also beneficial for a minority within the population to resist change to maintain community identity and prevent the enhancement of new trends and ideas without deliberating upon them.

The presence of multiple drivers not only necessitates the management of various parameters but also requires evaluating the trajectory of a plane in an N-dimensional space. In detail, the N systemic drivers define an N-dimensional space, and observing how they change over time (in response to environmental shifts, population dynamics within the MASs, and the selected social arrangements) necessitates evaluating the trajectory of a plane within this N-dimensional space. Consequently, assessing that trajectory in relation to all drivers necessitates ensuring that none of the drivers is out of control and that all remain balanced. Thus, differentiating between various social arrangements becomes quite challenging, as a human user may struggle to evaluate elements corresponding to a high-dimensional space that they cannot readily perceive.

To facilitate user engagement and enable them to exert requisite influence over the system, the proposed innovation support framework should provide multiple types of metrics, including quantitative, qualitative, and narrative metrics. In particular, quantitative metrics deliver information regarding the value of a measurement and are useful for swiftly understanding measurement evolution; however, they do not suffice to convey the truthfulness or quality of a decision. This type of information can be supplemented by qualitative metrics, which, though beneficial, do not sufficiently capture the meaning of the selected policies. For the latter, the system should incorporate narrative metrics, which correspond to stories that elucidate the actual effects of the selected decisions on the population. By incorporating these types of metrics, user experience would be enhanced, enabling individuals not only to comprehend and empathise better but also to provide comprehensive feedback. Overall, the three types of metrics provide diverse information and should all be included in innovation support systems to accommodate environmental complexity and facilitate effective adaptation, as posited by the law of variety and the law of requisite complexity.

To conclude, the analysis above highlighted issues related with metrics, especially when those metrics aim to inform human users and not computational components. Hence, if those metrics

are properly designed, and are integrated in a system aiming to humans empowerment, such as the ISSD, they can enable humans make inform decisions on their SAs. This way humans have the information and the tools to actually control “states of affairs” that directly affect them, which supports dignity.

6.8 Empowerment and Innovation: A Chicken-Egg Problem

Having discussed issues of empowerment and innovation, here we reflect on previous Sections and discuss the interconnection between those two terms in the context of SGSs designed to support human decision making during digital transformation. In more detail, as discussed above, the integration of AI technologies into social systems necessitates a robust framework for managing the significant shifts and complexities this transition introduces. This requirement, in turn, emphasises the need for innovation—not only in advancing technology itself but, crucially, in how we design, operationalise, and sustain social arrangements. In SGSs, empowerment becomes essential, defined as the capacity for individuals and communities to autonomously shape their social structures. Empowerment here involves the ability to influence and adapt one’s environment and interactions without undue external control, thus positioning innovation as a critical tool in fostering meaningful collective deliberation.

Innovation and empowerment in socio-technical systems are intrinsically connected. As society undergoes digital transformation, individuals and communities need empowerment to adapt their social arrangements in ways that reflect shared values and goals. Floridi [210] describes empowerment through two complementary dimensions: “equal opportunities”, which ensures human inclusion in decision-making processes, and “more opportunities”, which involves expanding the quality and diversity of choices. In SGSs, the first dimension aligns with empowerment as self-determination, as it prioritises human participation in the democratic structuring of social arrangements. The second dimension reinforces the inherently self-referential nature of innovation and empowerment, where expanding choices actively cultivates empowerment.

Additionally, as AI’s integration into human systems often heightens existing asymmetries of power [23], [24], innovative adaptation of social arrangements becomes essential to counterbalance these imbalances. By revisiting and modifying structures, roles, and rules, individuals can mitigate disparities, thus empowering themselves to innovate upon their empowerment. This dynamic is particularly vital in SGSs, where the evolution of social relationships and distribution of agency

across both human and AI components can reinforce empowerment. In summary, power enables innovation, which in turn reinforces and expands empowerment, fostering an adaptive and equitable framework for a digitally transformed society.

6.9 Summary & Conclusions

In conclusion, Chapter 6 introduced concepts of innovation and community empowerment within self-governing socio-technical systems, focusing particularly on the role of AI in facilitating human agency. It examined AI methodologies rooted in higher-order cybernetics patterns to demonstrate how AI can enable empowerment in systems where humans and AI agents co-exist and co-evolve. By framing innovation as a three-step process involving reflection, generation, and application, this Chapter presented a process-oriented definition of innovation pertinent to self-governing contexts. Furthermore, it introduced ISSD, an AI-powered framework designed according to the SGRL methodology, which was evaluated both as an innovative tool for human and community empowerment and as a cybernetic system. In addition, the chapter explored how the SGRL methodology can aid in addressing complex, “wicked problems” and counteracting inequities in power distribution within human-AI systems. Challenges related to the measurement and evaluation of innovation within SGSs were also discussed, with a closing analysis on the self-referential nature of both empowerment and innovation.

Firstly, in advancing the discourse on “ethical” or “responsible” AI, this Chapter foregrounds power dynamics and empowerment within the design, development, and operationalisation of AI. By positioning replication as a method for community empowerment, the chapter advocates for an approach that embodies communality, uniting commons, community, and polity. This integration offers communities immediate access to deep social knowledge, enabling them to actively influence and manage issues that directly impact them.

Secondly, the chapter argued that the SGRL methodology has the potential to elevate AI technology from a decision-support tool to an innovation-support mechanism. This shift holds a dual significance. The first is that it enriches the “human collective mind” by fostering the imagination to envisage impactful outcomes that span societal and generational timescales, facilitating a “social micro-macro-scope” that illuminates the effects of policy decisions over short and long-term horizons. The second is that, by embedding human involvement and feedback at each stage of the decision-making process, this framework paves the way for safer, more democratised design ap-

proaches, reducing the potential for technological exploitation and enhancing community control over AI applications.

Finally, this SGRL approach exemplifies a multi-component system that combines human social knowledge and expertise with the creative potential of GenAI, which draws from diverse knowledge connections, and MAS simulations that allow the observation of emergent behaviours. In enabling humans to explore alternative solutions and foresee possible future trajectories, the system supports the deliberation on SAs and promotes an equitable distribution of power within socio-technical ecosystems.

Chapter 7

Conclusions and Further Research

7.1 Summary

7

Next-generation socio-technical systems combine human (natural) and computational (artificial) intelligence, but this raises unique challenges for self-governance and Self-Governing Systems (SGSs). Specifically, a SGS comprises independent components that, while pursuing personal and collective interests and possessing different values or goals, operate within an overarching, mutually agreed-upon, and amendable set of conventional Social Arrangements (SAs). The system operates without external authority, with entities responsible for selecting, modifying, and enforcing the SAs themselves. As such, to maintain sustainable self-governance, the system requires mechanisms for inclusion, rule modification, and dispute resolution, as well as processes for collective decision-making and knowledge management.

Consequently, the main challenges of self-governance involve ensuring fairness, justice, and dignity, preventing oligarchy, learning when to consent and when to express dissent, and establishing effective processes for knowledge management and collective decision-making. This thesis has brought together theories of human psychology and political science with novel algorithms for Machine Reasoning (MR), Machine Learning (ML) and Human-Machine Innovation (MI) to mitigate those challenges and maintain sustainable self-governance in socio-technical systems.

Initially, we combined a theory of social influence [12] with Machine Reasoning (MR) and followed a Socially-Inspired Computing (SIC) methodology, which was implemented using the Agent-Based Modelling and Simulation (ABMS), approach to investigate under which initial conditions a

group of autonomous interconnected agents, acting as a Distributed Information Processing (DIP), can develop explanatory adequacy, and what is the epistemic knowledge of the population on that state. The results of this research offer some further insight into the need for epistemic diversity for self-improvement through exogenous innovation.

We combined insights from psychoacoustics, first-order cybernetics theory [16],[19] and Reinforcement Learning (RL) to achieve effective self-regulation. Specifically, following an Hybrid Reasoning and Learning (HyRL) methodology, we designed a system to examine whether a group of agents (i.e. components of a regulated system) can learn to distinguish between a set of inner voices. The voice selected by each agent is used to form its individual expression, and the aggregation of those results in the collective expression. The latter is used to inform the regulator regarding the state of the regulated unit. This way, the regulator can adjust the policy in such a way that the systemic stability is maintained (i.e. agents are satisfied). This work therefore enables us to develop a better understanding of issues of requisite (social) influence and ethical regulation.

Using a HyRL methodology, we incorporated concepts from political sciences, and specifically from classic Athenian Democracy [20], [42], with Multi-Agent Reinforcement Learning (MARL) to designed a fit-for-purpose algorithm for sustainable self-governance through meaningful democratic deliberation. In this study, we showed how sustainable self-governance requires making compromises while also expressing dissent, and we illustrated that a majority decision is acceptable if the underlying principle is democratic deliberation.

Moreover, we analysed issues of empowerment and innovation in the context of hybrid (human-AI) SGS, we leveraged higher-order cybernetic patterns to enable Power Sensitive Design (PSD) in such systems, and we followed the Socially-Guided Reasoning and Learning (SGRL) methodology to introduce a novel tool for innovation support and Power Sensitive Operationalisation (PSO) leading to human and community empowerment during Digital Transformation.

This thesis also offered insights into ways of evaluating the systems' performance in terms of sustainability, performance, and human empowerment. We demonstrated that the performance of a socio-technical system, which operates as a DIP, requires reflecting on the values of a set of four pairs of conflicting systemic drivers, namely accuracy vs. economy, stability vs. flexibility, inclusivity vs. expertise, and coherence vs. diversity. Additionally, we discussed issues of sustainability in complex socio-technical systems, specifically highlighting that, in the next generation socio-technical systems, the goal of sustainability can be satisfied when the intersection of technology, environmental and socio-politico-economic aspects of the system operate in a controlled

and desired way. In addition, we examined issues of evaluating the effectiveness of SAs in innovation and empowerment by proposing metrics for optimising with respect to objective functions (corresponding to population specific metrics) as well as human perception (corresponding to innovation specific metrics). These enable us to evaluate the system with respect to a trajectory rather than an index, to account for conflicting drivers of multiple stakeholders, but also to enhance the user experience by providing not only value measurements but also qualitative assessments and explanatory narratives.

Overall, the contributions of this thesis were five-fold:

- *theoretical*: a characterisation of Self-Governing Systems (SGSs) together with an analysis of issues of self-governance, innovation, and empowerment in next-generation socio-technical systems;
- *methodological*: an incremental refinement and integration of three methodologies: SIC, HyRL, and SGML, to produce a new methodology called SGRL for designing and implementing SGSs;
- *technical*: three novel algorithms for explanatory adequacy and epistemic coherence through social influence (i.e. *RTSI+*), for effective self-regulation via requisite influence (i.e. *4voices*), and for reaching sustainable collective agreements through consensus and dissent (i.e. *Θ-learning*), of which the experimental results provide insights into three key aspects: first, how initial conditions influence the collective's epistemic coherence and its explanatory adequacy; second, how they affect systemic stability and the nature of individual 'truth'; and third, how foundational institutional goals shape collective decisions, with population dynamics influencing both agreements and the compromises required to achieve them.
- *practical*: a novel tool, called Innovation Support System through Deliberation (ISSD), for innovation support and PSO, combining MAS, GenAI, and human users to promote human and community empowerment.
- *metrical*: an advancing analysis of issues of metrics for assessing self-governing socio-technical systems by: presenting a set of four conflicting pairs of systemic drivers for performance, highlighting the importance of aligning technology, environment, and socio-political factors for sustainability, and introducing metrics that optimise both system functions and human perception, allowing for dynamic, multi-stakeholder evaluations and enhanced user experience leading to human empowerment.

Additionally, those contributions offered mechanisms of dealing with some of the main challenges of self-governance which were introduced in Section 2.2.4. The relevance of each of the contributions (technical, practical, or metrical) of this thesis was analysed in the corresponding version, and Table 7.1 summarises it by presenting the challenge and the contributions that offer insights into it.

SGS Challenge	Contribution
sustainability	<i>RTSI+</i> , Θ -learning, metrics for sustainability (Section 5.4), ISSD
fairness and justice	<i>RTSI+</i>
knowledge management	<i>RTSI+</i> , <i>4voices</i> , metrics for performance (Section 4.10)
legitimate consent	Θ -learning
legitimate dissent	Θ -learning
iron law of oligarchy	metrics for performance (Section 4.10), Θ -learning
unrestricted self-modification	<i>4voices</i> , Θ -learning, ISSD
voting paradoxes	<i>4voices</i> , Θ -learning
citizenship and rights	cybernetics for PSD (Section 6.2)
dignity	metrics for performance (Section 4.10), <i>4voices</i> , Θ -learning
	cybernetics for PSD (Section 6.2), ISSD, metrics for empowerment (Section 6.7)

Table 7.1: Contribution to Issues of SGSs

Finally, this research offers some further insights on the importance of designing SGSs that not only *adapt* to ever-changing conditions, but also *proactively learn* and *self-improve* over time, ensuring their sustainability and human empowerment in increasingly complex socio-technical landscapes.

7.2 List of Publications

This thesis produced the following scientific publications:

- A. Mertzani and J. Pitt, “Socially-Guided Machine Learning for Self-Organised Community Empowerment”, Accepted for publication in the 17th International Conference on Agents and Artificial Intelligence (ICAART), 2025.
- A. Mertzani and J. Pitt, “Social Implications of Socially-Guided Machine Learning for Innovation Support”, in Proceedings of the IEEE International Symposium on Technology and Society (ISTAS), Puebla, Mexico, 2024, pp. 1-8.
- A. Mertzani and J. Pitt, “Power-sensitive design using higher-order cybernetics patterns”, in Proceedings of the Artificial Life Conference (ALIFE2024), 2024, p. 126.

- 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.
- A. Mertzani, J. Ober, and J. Pitt, “Theta learning: An algorithm for the self-organisation of collective self-governance”, in Proceedings of 4th IEEE International Conference on Autonomic Computing and Self-Organizing Systems (ACSOS2023), Toronto, ON, Canada, 2023, pp. 97-106.
- A. Mertzani and J. Pitt, “Sustainable self-organisation of socio-techno-ecological systems”, in 4th IEEE International Conference on Autonomic Computing and Self-Organizing Systems Companion (ACSOS-C), Toronto, ON, Canada, 2023, pp. 92–97.
- A. Mertzani, J. Pitt, A. Nowak, and T. Michalak, “Epistemic diversity and explanatory adequacy in distributed information processing”, in Coordination, Organizations, Institutions, Norms, and Ethics for Governance of Multi-Agent Systems XV: International Workshop (COINE 2022), Virtual Event, Revised Selected Papers. Springer-Verlag, Berlin, Heidelberg, 19–37.
- A. Mertzani and J. Pitt, “Metrics for reflection in distributed information processing”, in 14th International Workshop Agent-Based Modelling Human Behaviour (ABMHuB2022), 2022. Available at: http://abmhub.cs.ucl.ac.uk/2022/camera_ready/Mertzani_Pitt.pdf

Also, content of this thesis contributed to the research of our group and resulted in the following scientific outputs:

- Pitt, J., Mertzani, A., Scott, M. and Smit, C., 2024. “The Architecture of Re-Empowerment”, Accepted for publication in the Technology and Society Magazine 2024.
- M. Scott, A. Mertzani, C. Smit, S. Sarkadi, and J. Pitt, “Social deliberation vs. social contracts in self-governing voluntary organisations”, arXiv preprint arXiv:2403.16329, 2024.
- A. Mertzani, J. Pitt, S. Sarkadi, M. Sas, M. Scott, C. Smit, “Cohesion and the Explanation of Constitutional Choice in Self-Governing Systems”, in 6th International Workshop Agent-Based Modelling Human Behaviour (ABMHuB), 2024.
Available at: http://abmhub.cs.ucl.ac.uk/2024/camera_ready/Mertzani_etal.pdf
- Smit, C., M. Scott, A. Mertzani, and J. Pitt, “SALab: Computer-Supported Social Arrangements Laboratory”, in Frontiers of Artificial Intelligence, Ethics, and Multidisciplinary

Applications: 1st International Conference on Frontiers of AI, Ethics, and Multidisciplinary Applications (FAIEMA), Greece, 2023. Springer Nature Singapore. pp. 299-312.

- B. Blackledge, A. Papaoikonomou, M. Scott, A. Mertzani, N. Le Renard, H. Masoud, and J. Pitt, “Incentivising Participation with Exclusionary Sanctions”, in International Workshop on Coordination, Organizations, Institutions, Norms, and Ethics for Governance of Multi-Agent Systems (COINE). Lecture Notes in Computer Science, Springer Nature Switzerland. 2023, vol. 14002, pp. 37-54.
- A. Mertzani, M. Scott, C. Smit and J. Pitt, “Engage-Envision-Enact: Self-Organised Governance for Self-Improving Socio-Technical Systems”, in 4th IEEE International Conference on Autonomic Computing and Self-Organizing Systems Companion (ACSOS-C), Toronto, ON, Canada, 2023, pp. 53-58.
- A. Mertzani and J. Pitt, “Social Influence and the Normalization of Surveillance Capitalism: Legislation for the Next Generation”, in IEEE Technology and Society Magazine, 2022, vol. 41, no. 2, pp. 57-63.

7.3 Limitations and Further Research

This thesis offered several contributions, though there are some ways in which it could be further expanded. Those include improvements in the studies of *RTSI for explanatory adequacy* [10], *metrics for reflection* [25], *Effective Self-Regulation* [18], *Sustainable Consensus Formation* [21], and *Innovation Support for Community Empowerment* [23], [24].

Firstly, the study of *RTSI for explanatory adequacy* [10] assumed that agents were willing to share their knowledge with their peers, and were well-intended. Future research could explore different types of populations, such as agents possessing distinct personalities, and agents deliberately sharing selective knowledge to influence opinions and thoughts, or even agents aiming to deceive others [196], [201]. This way we would explore negative aspects of social influence and the way they affect the system’s dynamics. More sophisticated methods for developing self-confidence and credence within the network could be investigated, such as costly signalling models or block-based approaches. Future studies could also extend network communication to include not only the exchange of processing rules but also the rationale behind selecting those rules. Finally, deeper questions concerning epistemic coherence as well as collective empowerment could be examined.

Secondly, *Metrics for Reflection* [25] highlighted the limitations of relying solely on metrics-based evaluations. It noted that some things can be measured but may not be important, while others, particularly human values, may matter greatly but are not easily measurable. Two key values that are difficult to quantify but critically important are *quality* and *dignity*. Future research could further explore the relationship between quality and dignity, or even could shed light on issues of rights, particularly in the context of human empowerment during digital transformation.

Thirdly, *Effective Self-Regulation* [18] focused on inner voice comprising four voices. Future research could extend the set of voices as well as consider individuals being biased towards a or some voices due to personal experiences. Moreover, the regulated units lacked an intelligent mechanism for choosing between voices, and the regulator was restricted to a particular learning algorithm. Further investigation could focus on scenarios where agents have more advanced methods for selecting between voices, or where regulators utilise different learning algorithms to adapt their strategies based on agent behaviour.

Furthermore, the current approach assumed full environmental observability, with all agents having access to their neighbours' opinions. Future research could explore systems where agents have limited environmental visibility and restricted communication with other agents, as seen in studies like Aguzzi et al.[211] that use of Deep Q-Networks combined with Graph Neural Networks to address information limitations. Also, this work focuses on the beneficial aspects of social influence, although in real life social influence can also have negative effects. Therefore, future work could explore the system dynamics when agents have also bad intentions and try to manipulate others to support own interests and whether it is possible to protect the system for such 'malignant' agents and achieve effective self-regulation regardless of their presence.

Additionally, the framework proposed in *Sustainable Consensus Formation* [21] could be extended to address scalability issues, specifically by investigating the impact of different population sizes on the stability and effectiveness of the approach. The current algorithm uses a reward function with two components: one promotes agreement (ensuring stability and coherence), while the other minimises compromises (preserving diversity). Related research has explored how to balance conflicting drivers such as diversity and coherence or stability and flexibility in multi-agent environments [9]. Therefore, future work could focus on refining the reward function to account for additional drivers and maintaining balance through deliberation and learning.

Moreover, this work lays the foundation for further investigation of both the *civic bargain*, which refers to the mutual commitment of citizens to guarantee civic rights and fulfil basic duties,

and *civic bargaining*, which frames democratic self-governance as a means of reconciling competing priorities. Citizens must be able to compromise without creating enmity, instead fostering externalities (or “conceptual resources”) that allow reciprocal concessions in future negotiations [212]. Future studies could combine the proposed algorithm with formal models of these externalities and reasoning with such conceptual resources.

On *Innovation Support for Community Empowerment* [24],[23], as noted in Section 6.6, the existing configuration of the system serves as a proof of concept and possesses several limitations. The primary and most notable limitation is the lack of testing on human participants. Therefore, future research should assess the system’s effectiveness with human users who possess varying levels of knowledge, including both experts and novices. Furthermore, while we present a specific implementation of the system, there are alternative configurations that could enhance human-AI co-production through the utilisation of an SGRL methodology to foster innovation, warranting further investigation. Additionally, it is important to note that the current implementation does not incorporate any RL within MAS, which represents another potential area for exploration.

Furthermore, to foster innovation in SGS we introduced an LLM-agent in the population of the proof-of-concept implementation. In those experiments, the use of the LLM was controlled by the user and because of the limitations of the existing implementation, the LLM could not have a significant impact in the system. However, future research aims to extend the functionalities of the system and use it for ‘outside the lab’ experiments. In that case, prior to testing the model in real populations, we need to make sure that the LLM responses are consistent and hallucinations are limited if not eliminated, or that the system possesses mechanisms that protect the user from such issues.

Generally, the experiments that were carried out in this thesis have a large parameter space. Although there were multiple combinations of parameter sets that could be explored, we set most of the parameters such that they produce a stable behaviour. We varied several parameters to investigate the conditions under which our hypotheses are satisfied or the behaviours of interest emerge within a reasonable time-frame. Future research could explore different combinations of fixed and varying parameters. Also, in most of the experiments the population of the multi-agent system comprised 100 agents. It would be particularly interesting to explore different scales of populations and specifically whether the system ‘collapses’ above a certain size of population.

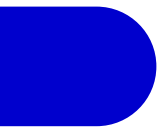
Finally, further research could combine the algorithms and the concepts described in a single end-to-end system. Such an approach would bring together multiple issues of self-governance,

and leverage the different methodologies and underlying theories from social sciences, to provide a test-bed for a thorough and complete solution for addressing issues of self-governing socio-technical systems. It would be interesting to apply this to real communities facing different issues. Examples of such communities include football teams, a group of people sharing an open plan office [213], [214], co-housing projects, and startups. In that case, it would be particularly interesting to explore the integration of the system and its interaction with other systems used by or affecting that community, such as legal systems, educational systems, or ecological systems. Last but not least, if we are to apply such a tool to a real community, we need to consider issues related with the change of the human behaviour due to the interaction with the machine, and how the system can adapt dynamically to match the behavioural change of the human users.

7.4 Final Remarks

Apart from the scientific contributions, this thesis contributed to the community in multiple ways:

- It has laid the foundation for the final year projects of undergraduate and postgraduate students.
- It supported the research of our group on SGSs and resulted in numerous scientific contributions on issues concerning constitutional choice and social cohesion [215], computer-supported governance [26], [216], social deliberation [217], and incentivisation of participation through sanctions and social capital [218].
- Chapter 6 on empowerment, innovation, and self-improvement constituted the introduction to a panel at Self-Improving System Integration (SISSY) workshop in an international scientific conference (ACSOS 2024).
- The scientific contributions of this thesis formed the content for the half of a course on SGSs which was taught in the European Summer School of Artificial Intelligence (ESSAI2024).
- It has been one of the cornerstones for research project submissions in ‘Democracy, Deception, and Dignity in Human-AI SGSs’, and ‘Supporting Human-AI Co-Production for Empowering Communities’.
- Parts of this thesis will be used for the development of the syllabus of an inter-disciplinary module on AI, Ethics, Democracy and User-Centred Systems Design which will be presented in collaboration with Stanford University and University of Oxford.



Bibliography

- [1] S. Rawas, “AI: The future of humanity,” *Discover Artificial Intelligence*, vol. 4, p. 25, Mar. 2024. DOI: 10.1007/s44163-024-00118-3.
- [2] R. Salvi and D. R. Singh, “Artificial intelligence and human society,” *International Journal of Social Science and Human Research*, vol. 6, Sep. 2023. DOI: 10.47191/ijsshr/v6-i9-13.
- [3] M. Shafiq, Z. Gu, O. Cheikhrouhou, W. Alhakami, and H. Hamam, “The rise of “internet of things”: Review and open research issues related to detection and prevention of iot-based security attacks,” *Wireless Communications and Mobile Computing*, vol. 2022, no. 1, p. 8 669 348, 2022.
- [4] R. Beck, “Beyond bitcoin: The rise of blockchain world,” *Computer*, vol. 51, no. 2, pp. 54–58, 2018.
- [5] E. K. Oermann and D. Kondziolka, “On chatbots and generative artificial intelligence,” *Neurosurgery*, vol. 92, no. 4, pp. 665–666, 2023.
- [6] H. Hamann, Y. Khaluf, J. Botev, *et al.*, “Hybrid societies: Challenges and perspectives in the design of collective behavior in self-organizing systems,” *Frontiers in Robotics and AI*, vol. 3, 2016, ISSN: 2296-9144. DOI: 10.3389/frobt.2016.00014. [Online]. Available: <https://www.frontiersin.org/journals/robotics-and-ai/articles/10.3389/frobt.2016.00014>.
- [7] F. Cloete, “Governing artificial intelligence (AI) and other technologies in the digital era,” *Administratio Publica*, vol. 32, no. 1, pp. 1–30, 2024.
- [8] A. Sigfrids, J. Leikas, H. Salo-Pöntinen, and E. Koskimies, “Human-centricity in AI governance: A systemic approach,” *Frontiers in artificial intelligence*, vol. 6, p. 976 887, 2023.
- [9] A. Mertzani, J. Pitt, A. Nowak, and T. Michalak, “Expertise, social influence, and knowledge aggregation in distributed information processing,” *Artificial Life*, vol. 29, no. 1, pp. 37–65, 2022.

-
- [10] A. Mertzani, J. Pitt, A. Nowak, and T. Michalak, “Epistemic diversity and explanatory adequacy in distributed information processing,” in *Coordination, Organizations, Institutions, Norms, and Ethics for Governance of Multi-Agent Systems XV: International Workshop, COINE 2022, Virtual Event, May 9, 2022, Revised Selected Papers*, Berlin, Heidelberg: Springer-Verlag, 2022, pp. 19–37, ISBN: 978-3-031-20844-7.
 - [11] J. Pitt, *Self-Organising Multi-Agent Systems: Algorithmic Foundations of Cyber-Anarcho-Socialism*. Singapore: World Scientific, 2021. DOI: 10.1142/q0307. [Online]. Available: <https://www.worldscientific.com/doi/abs/10.1142/q0307>.
 - [12] A. Nowak, R. Vallacher, A. Rychwalska, *et al.*, *Target in Control: Social Influence as Distributed Information Processing*. Cham, CH: Springer, Jan. 2019, ISBN: 978-3-030-30621-2. DOI: 10.1007/978-3-030-30622-9.
 - [13] L. Bottou, “From machine learning to machine reasoning: An essay,” *Machine learning*, vol. 94, pp. 133–149, 2014.
 - [14] S. Bandini, S. Manzoni, and G. Vizzari, “Agent based modeling and simulation: An informatics perspective,” *Journal of Artificial Societies and Social Simulation*, vol. 12, no. 4, p. 4, 2009.
 - [15] C. M. Macal and M. J. North, “Agent-based modeling and simulation,” in *Proceedings of the 2009 winter simulation conference (WSC)*, IEEE, 2009, pp. 86–98.
 - [16] W. Ashby, *Design for a Brain*. London, UK: Chapman-Hall, 1952.
 - [17] C. J. C. H. Watkins, “Learning from delayed rewards,” *Dissertation, King’s College, Cambridge*, 1989.
 - [18] 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.
 - [19] 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>.
 - [20] M. Canevaro, “Majority rule vs. consensus: The practice of democratic deliberation in the Greek poleis,” in *Ancient Greek History and Contemporary Social Science*, M. Canevaro, A. Erskine, and B. Gray, Eds. Edinburgh, UK: Edinburgh University Press, 2018, pp. 101–156.
 - [21] A. Mertzani, J. Ober, and J. Pitt, “Theta-learning: An algorithm for the self-organisation of collective self-governance,” in *IEEE International Conference on Autonomic Computing and Self-Organizing Systems (ACSOS)*, 2023, pp. 97–106.

-
- [22] N. Wiener, *Cybernetics Or Control and Communication in the Animal and the Machine* (Actualités Scientifiques et Industrielles). M.I.T. Press, 1961, ISBN: 9780262730099. [Online]. Available: <https://books.google.gr/books?id=NnM-uISyywAC>.
- [23] A. Mertzani and J. Pitt, "Power-sensitive design using higher-order cybernetics patterns," ser. Artificial Life Conference Proceedings, vol. ALIFE 2024: Proceedings of the 2024 Artificial Life Conference, 2024, p. 126.
- [24] A. Mertzani and J. Pitt, "Social implications of socially-guided machine learning for innovation support," in *2024 IEEE International Symposium on Technology and Society (ISTAS)*, IEEE, 2024, pp. 1–8.
- [25] A. Mertzani and J. Pitt, "Metrics for reflection in distributed information processing," in *4th International Workshop on Agent-Based Modelling of Human Behaviour (ABMHuB'22)*, 2022, pp. 52–59.
- [26] A. Mertzani, M. Scott, C. Smit, and J. Pitt, "Engage-envision-enact: Self-organised governance for self-improving socio-technical systems," in *2023 IEEE International Conference on Autonomic Computing and Self-Organizing Systems Companion (ACSOS-C)*, IEEE, 2023, pp. 53–58.
- [27] C. Fernyhough, *The Voices Within*. Wellcome Collection, 2017.
- [28] F. Heylighen, "The science of self-organization and adaptivity," English, in *Knowledge Management, Organizational Intelligence and Learning, and Complexity: v. 3*, L. Kiel, Ed. EOLSS Publishers Co Ltd, 2002, ISBN: 978-1-84826-913-2.
- [29] J. Gjaltema, R. Biesbroek, and K. Termeer, "From government to governance...to meta-governance: A systematic literature review," *Public Management Review*, vol. 22, pp. 1–21, Aug. 2019. DOI: 10.1080/14719037.2019.1648697.
- [30] B. Damgaard and J. Torfing, "Network governance of active employment policy: The danish experience," *Journal of European Social Policy*, vol. 20, no. 3, pp. 248–262, 2010. DOI: 10.1177/0958928710364435.
- [31] P. Glasbergen, "Mechanisms of private meta-governance: An analysis of global private governance for sustainable development," *International Journal of Strategic Business Alliances*, vol. 2, pp. 189–206, Jan. 2011. DOI: 10.1504/IJSBA.2011.040886.
- [32] E. Sørensen and J. Torfing, "Network governance and post-liberal democracy," *Administrative Theory and Praxis*, vol. 27, no. 2, pp. 197–237, May 2005, ISSN: 1949-0461. DOI:

- 10.1080/10841806.2005.11029489. [Online]. Available: <http://dx.doi.org/10.1080/10841806.2005.11029489>.
- [33] M. D. Molin and C. Masella, "From fragmentation to comprehensiveness in network governance," *Public Organization Review*, vol. 16, no. 4, pp. 493–508, Jul. 2015, ISSN: 1573-7098. DOI: 10.1007/s11115-015-0320-4. [Online]. Available: <http://dx.doi.org/10.1007/s11115-015-0320-4>.
- [34] C. Ansell and A. Gash, "Collaborative governance in theory and practice," *Journal of Public Administration Research and Theory*, vol. 18, no. 4, pp. 543–571, Nov. 2007, ISSN: 1053-1858. DOI: 10.1093/jopart/mum032. [Online]. Available: <http://dx.doi.org/10.1093/jopart/mum032>.
- [35] M. P. Hansen, P. Triantafillou, and S. H. Christensen, "Two logics of democracy in collaborative governance: A mapping of clashes and compromises," *Public Management Review*, vol. 26, no. 3, pp. 635–656, Aug. 2022, ISSN: 1471-9045. DOI: 10.1080/14719037.2022.2107696. [Online]. Available: <http://dx.doi.org/10.1080/14719037.2022.2107696>.
- [36] P. D. Aligiga and V. Tarko, "Polycentricity: From Polanyi to Ostrom, and beyond," *Governance*, vol. 25, no. 2, pp. 237–262, Sep. 2011, ISSN: 1468-0491. DOI: 10.1111/j.1468-0491.2011.01550.x. [Online]. Available: <http://dx.doi.org/10.1111/j.1468-0491.2011.01550.x>.
- [37] C. Smit and J. Pitt, "Digital polycentricity and value-sensitive operationalization," *IEEE Technology and Society Magazine*, vol. 42, no. 4, pp. 107–118, Dec. 2023, ISSN: 1937-416X. DOI: 10.1109/mts.2023.3341464. [Online]. Available: <http://dx.doi.org/10.1109/mts.2023.3341464>.
- [38] E. Ostrom, *Governing the commons: The evolution of institutions for collective action*. Cambridge University Press, 1990.
- [39] J. Ober, *Democracy and Knowledge: Innovation and Learning in Classical Athens*. Princeton University Press, 2008, ISBN: 9780691146249.
- [40] N. Rescher, *Distributive Justice: a Constructive Critique of the Utilitarian Theory of Distribution*. Indianapolis, IN: Bobbs-Merrill Company, Inc., 1966.
- [41] J. Rawls, *A Theory of Justice*. Harvard MA: Harvard University Press, 1971.
- [42] J. Ober, *Demopolis: Democracy before Liberalism in Theory and Practice* (The Seeley Lectures). Cambridge University Press, 2017. DOI: 10.1017/9781108226790.

-
- [43] D. Burth Kurka, J. Pitt, P. Lewis, A. Patelli, and A. Ekárt, “Disobedience as a mechanism of change,” in *12th IEEE International Conference SASO*, 2018, pp. 1–10.
- [44] R. Michels, *Political Parties: A Sociological Study of the Oligarchical Tendencies of Modern Democracy*. New York, NY: The Free Press, 1962 [1911].
- [45] P. Suber, *The Paradox of Self-Amendment: a study of law, logic, omnipotence, and change*. Peter Lang International Academic Publishers, 1990.
- [46] C. List and R. Goodin, “Epistemic democracy: Generalizing the condorcet jury theorem,” *Journal of Political Philosophy*, vol. 9, no. 3, pp. 277–306, 2001.
- [47] K. Arrow, *Social Choice and Individual Values*. New York, NY: Wiley, 1951.
- [48] J. Tasioulas, *Artificial Intelligence, Ethics, and a Right to a Human Decision*. Harold T. Shapiro Lecture on Ethics, Science, and Technology, Princeton University, 2023.
- [49] J. Pitt, J. Dryzek, and J. Ober, “Algorithmic reflexive governance for socio-techno-ecological systems,” *IEEE Technology and Society Magazine*, vol. 39, no. 2, pp. 52–59, 2020. DOI: 10.1109/MTS.2020.2991500.
- [50] S. J. Russell and P. Norvig, *Artificial intelligence: a modern approach*. Pearson, 2016.
- [51] P. G. Balaji and D. Srinivasan, “An introduction to multi-agent systems,” in *Innovations in Multi-Agent Systems and Applications - 1*, D. Srinivasan and L. C. Jain, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 1–27, ISBN: 978-3-642-14435-6. DOI: 10.1007/978-3-642-14435-6_1. [Online]. Available: https://doi.org/10.1007/978-3-642-14435-6_1.
- [52] V. Julian and V. Botti, “Multi-agent systems,” *Applied Sciences*, vol. 9, no. 7, 2019, ISSN: 2076-3417. DOI: 10.3390/app9071402. [Online]. Available: <https://www.mdpi.com/2076-3417/9/7/1402>.
- [53] T. Kelsall, N. Schulz, W. D. Ferguson, M. vom Hau, S. Hickey, and B. Levy, “46 Collective-action Problems and Development: A Typological Theory of Political Settlements,” in *Political Settlements and Development: Theory, Evidence, Implications*, Oxford University Press, May 2022, ISBN: 9780192848932. DOI: 10.1093/oso/9780192848932.003.0003. eprint: <https://academic.oup.com/book/0/chapter/361970132/chapter-pdf/53447566/oso-9780192848932-chapter-3.pdf>. [Online]. Available: <https://doi.org/10.1093/oso/9780192848932.003.0003>.

-
- [54] C. V. de Graaf, “Procedural fairness: Between human rights law and social psychology,” *Netherlands Quarterly of Human Rights*, vol. 39, no. 1, pp. 11–29, 2021. DOI: 10.1177/0924051921992749. eprint: <https://doi.org/10.1177/0924051921992749>. [Online]. Available: <https://doi.org/10.1177/0924051921992749>.
- [55] J. M. Peiró, V. Martínez-Tur, and C. Moliner, “Perceived fairness,” in *Encyclopedia of Quality of Life and Well-Being Research*, A. C. Michalos, Ed. Dordrecht: Springer Netherlands, 2014, pp. 4693–4696, ISBN: 978-94-007-0753-5. DOI: 10.1007/978-94-007-0753-5_2125. [Online]. Available: https://doi.org/10.1007/978-94-007-0753-5_2125.
- [56] J. Pitt, D. Busquets, and S. Macbeth, “Distributive justice for self-organised common-pool resource management,” *ACM Trans. Auton. Adapt. Syst.*, vol. 9, 14:1–14:39, 2014.
- [57] C. List, “Social choice theory,” in *Stanford Encyclopedia of Philosophy*, E. Zalta, Ed., 2013.
- [58] J. Surowiecki, “The wisdom of crowds: Why the many are smarter than the few,” *Abacus*, p. 295, Jan. 2004.
- [59] A. Ishii and N. Okano, “Politicophysics: A new model of opinion dynamics and its possibility of an application to politics/political science,” in *ABMHuB’2021: 3rd International Workshop on Agent-Based Modelling of Human Behaviour*, Jul. 2021.
- [60] G. R. Moreno, J. Padilla, and M. Brede, “The effects of party competition on consensus formation,” in *ABMHuB’2021: 3rd International Workshop on Agent-Based Modelling of Human Behaviour*, Jul. 2021.
- [61] J. Pitt, R. Cardoso, E. Hart, and J. Ober, “Relevant expertise aggregation for policy selection in collective adaptive systems,” in *2018 IEEE 3rd International Workshops on Foundations and Applications of Self* Systems (FAS*W)*, 2018, pp. 136–141.
- [62] J. Pitt and J. Ober, “Democracy by design: Basic democracy and the self-organisation of collective governance,” in *12th IEEE International Conference SASO*, 2018, pp. 20–29. DOI: 10.1109/SASO.2018.00013.
- [63] S. Sarkadi, “Self-governing hybrid societies and deception,” *ACM Trans. Auton. Adapt. Syst.*, vol. 19, no. 2, p. 9, 2024. DOI: 10.1145/3638549. [Online]. Available: <https://doi.org/10.1145/3638549>.
- [64] R. H. Gass, “Social influence, sociology of,” in *International Encyclopedia of the Social and Behavioral Sciences (Second Edition)*, J. D. Wright, Ed., Second Edition, Oxford: Elsevier, 2015, pp. 348–354, ISBN: 978-0-08-097087-5.

-
- [65] K. Izuma, "Chapter 16 - the neural bases of social influence on valuation and behavior," in *Decision Neuroscience*, J.-C. Dreher and L. Tremblay, Eds., San Diego: Academic Press, 2017, pp. 199–209, ISBN: 978-0-12-805308-9.
- [66] S. E. Asch, "Studies of independence and conformity: I. a minority of one against a unanimous majority.," in *Psychological Monograph: General and Applied*, vol. 70, 1956, Whole No. 416.
- [67] S. Milgram, "Behavioral study of obedience.," *The Journal of abnormal and social psychology*, vol. 67, no. 4, p. 371, 1963.
- [68] R. E. Petty and J. T. Cacioppo, "The elaboration likelihood model of persuasion," in *Communication and persuasion*, Springer, 1986, pp. 1–24.
- [69] R. Cialdini, *Pre-suasion: A revolutionary way to influence and persuade*. Simon and Schuster, 2016.
- [70] J. R. Baldwin and S. K. Hunt, "Information-seeking behavior in intercultural and intergroup communication," eng, *Human communication research*, vol. 28, no. 2, pp. 272–286, 2002, ISSN: 0360-3989.
- [71] J. Pitt, A. Nowak, T. Michalak, W. Borkowski, and R. Vallacher, "Knowing what the bits know: Social influence as the source of collective knowledge," in *Second International Workshop on Agent-Based Modelling of Human Behaviour*, 2020. [Online]. Available: <http://abmhub.cs.ucl.ac.uk/2020/papers/Pitt.pdf>.
- [72] J. Pitt, "Interactional justice and self-governance of open self-organising systems," in *11th IEEE International Conference SASO*, 2017, pp. 31–40.
- [73] K. Klemm and V. M. Eguíluz, "Growing scale-free networks with small-world behavior," *Phys. Rev. E*, vol. 65, p. 057 102, 5 May 2002.
- [74] B. Prettejohn, M. Berryman, and M. McDonnell, "Methods for generating complex networks with selected structural properties for simulations: A review and tutorial for neuroscientists," *Frontiers in computational neuroscience*, vol. 5, p. 11, Mar. 2011.
- [75] P. T. Nicholls, "Price's square root law: Empirical validity and relation to Lotka's law," *Information Processing Management*, vol. 24, no. 4, pp. 469–477, 1988, ISSN: 0306-4573. DOI: [https://doi.org/10.1016/0306-4573\(88\)90049-0](https://doi.org/10.1016/0306-4573(88)90049-0). [Online]. Available: <https://www.sciencedirect.com/science/article/pii/0306457388900490>.
- [76] C. Landauer and K. Bellman, "Reflective systems need models at run time," in *Proceedings of the 11th International Workshop on Models@run.time*, 2016, pp. 52–59.

-
- [77] S. Patil and H. Shankar, “Transforming healthcare: Harnessing the power of AI in the modern era,” *International Journal of Multidisciplinary Sciences and Arts*, vol. 2, no. 1, pp. 60–70, 2023.
 - [78] Z. Akhtar, “Artificial intelligence (AI) within manufacturing: An investigative exploration for opportunities, challenges, future directions,” *Metaverse*, vol. 5, p. 2731, Jul. 2024. DOI: 10.54517/m.v5i2.2731.
 - [79] W. Holmes, M. Bialik, and C. Fadel, *Artificial Intelligence in Education. Promise and Implications for Teaching and Learning*. Center for Curriculum Redesign, Mar. 2019, ISBN: 978-1794293700.
 - [80] C. Watkins and P. Dayan, “Q-learning,” *Machine Learning*, vol. 8, pp. 279–292, 1992. [Online]. Available: <https://api.semanticscholar.org/CorpusID:208910339>.
 - [81] R. Bellman, “On the theory of dynamic programming,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 38, no. 8, pp. 716–719, 1952, ISSN: 00278424, 10916490. [Online]. Available: <http://www.jstor.org/stable/88493> (visited on 10/28/2024).
 - [82] Y. Li, “Deep reinforcement learning: An overview,” *arXiv preprint arXiv:1701.07274*, 2017.
 - [83] V. Mnih, K. Kavukcuoglu, D. Silver, *et al.*, “Human-level control through deep reinforcement learning,” *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.
 - [84] H. van Hasselt, A. Guez, and D. Silver, *Deep reinforcement learning with double q-learning*, 2015. arXiv: 1509.06461 [cs.LG].
 - [85] Z. Wang, T. Schaul, M. Hessel, H. van Hasselt, M. Lanctot, and N. de Freitas, *Dueling network architectures for deep reinforcement learning*, 2016. arXiv: 1511.06581 [cs.LG].
 - [86] S. R. Rick, G. Giacomelli, H. Wen, *et al.*, *Supermind ideator: Exploring generative AI to support creative problem-solving*, 2023. arXiv: 2311.01937 [cs.AI]. [Online]. Available: <https://arxiv.org/abs/2311.01937>.
 - [87] R. Adams, *Empowerment, participation and social work*. New York, NY: Palgrave Macmillan, 2008.
 - [88] A. Jones and M. Sergot, “A formal characterisation of institutionalised power,” *Journal of the IGPL*, vol. 4, no. 3, pp. 427–443, 1996.
 - [89] J. Pitt, D. Busquets, and R. Riveret, “Procedural justice and ‘fitness for purpose’ of self-organising electronic institutions,” in *Proc. PRIMA*, ser. LNCS, vol. 8291, Springer, 2013, pp. 260–275.

-
- [90] D. Graeber and D. Wengrow, *The Dawn of Everything: A New History of Humanity*. London, U.K.: Allen Lane, 2021.
- [91] E. Chepin, "Robotics: From first-order cybernetics to third-order cybernetics," *Procedia Computer Science*, vol. 190, pp. 130–136, 2021.
- [92] S. Umpleby, T. Medvedeva, and V. Lepskiy, "Recent developments in cybernetics, from cognition to social systems," *Cybernetics and Systems*, vol. 50, pp. 1–16, Feb. 2019. DOI: 10.1080/01969722.2019.1574326.
- [93] B. Kumar, *IoT and fourth order cybernetics gisfi - 5g "networked society"*, <https://slideplayer.com/slide/3056538/>, 2014.
- [94] J. Breen, "Enterprise, entrepreneurship and small business: Where are the boundaries?" *International Journal of Entrepreneurship and Small Business*, vol. 1, no. 1-2, pp. 21–34, 2004.
- [95] S. Szczelkun, *SENSE THINK ACT: a collection of exercises to experience total human ability*. Stefan Szczelkun, 2018, ISBN: 9781870736107. [Online]. Available: <https://books.google.co.uk/books?id=ZIxODwAAQBAJ>.
- [96] V. Shklovsky, "Art, as device," *Poetics Today*, vol. 36, pp. 151–174, Sep. 2015. DOI: 10.1215/03335372-3160709.
- [97] H. A. Love, "On aerial perspective, socio-technical systems, and interdisciplinarity: Reading modernism alongside cybernetics," *IEEE Technol. Soc. Mag.*, vol. 42, no. 4, pp. 35–41, 2023.
- [98] M. Weitzman, "Recombinant growth," *The Quarterly Journal of Economics*, vol. 113, no. 2, pp. 331–360, 1998.
- [99] I. Douven, "Abduction," in *The Stanford Encyclopedia of Philosophy*, E. N. Zalta, Ed., Summer 2021, Metaphysics Research Lab, Stanford University, 2021.
- [100] K. Kahn, "Understanding innovation," *Business Horizons*, vol. 61, no. 3, pp. 453–460, 2018.
- [101] G. Grossman and E. Helpman, "Endogenous innovation in the theory of growth," *Journal of Economic Perspectives*, vol. 8, no. 1, pp. 23–44, 1994.
- [102] P. Drucker, "The rise of the knowledge society," *The Wilson Quarterly*, vol. 17, no. 2, pp. 52–72, 1993.
- [103] J. E. Tardy, "Chapter 18: Generative AI and consciousness," in *The Creation of a Conscious Machine: The Quest for Artificial Intelligence*. Berlin, Boston: Mercury Learning and Information, 2023, pp. 177–188, ISBN: 9781501518317. DOI: doi:10.1515/9781501518317-019. [Online]. Available: <https://doi.org/10.1515/9781501518317-019>.

-
- [104] T. Ahmed, “Collective wisdom in language models: Unveiling the power of LLM-swarm,” *Accepted for publication in NeurIPS 2024 Workshop on Open-World Agents*, Apr. 2024.
 - [105] A. Rychwalska, M. Roszczyńska-Kurasińska, K. Ziembowicz, and J. Pitt, “Fitness for purpose in online communities: Community complexity framework for diagnosis and design of socio-technical systems,” *Front. Psychol.*, 12:739415, 2021.
 - [106] M. Roszczyńska-Kurasińska, A. Domaradzka, M. O’Grady, *et al.*, “Beyond data: Recognizing the democratic potential of citizen science,” *IEEE Technology and Society Magazine*, vol. 42, no. 4, pp. 47–56, 2023. DOI: 10.1109/MTS.2023.3344904.
 - [107] G. Zarkadakis and D. Tapscott, *Cyber Republic: Reinventing Democracy in the Age of Intelligent Machines*. MIT Press, 2021.
 - [108] J. Lála, O. O’Donoghue, A. Shtedritski, S. Cox, S. Rodriques, and A. White, *Paperqa: Retrieval-augmented generative agent for scientific research*, arXiv:2312.07559, 2023. arXiv: 2312.07559 [cs.CL].
 - [109] J. Thiagarajan, V. Narayanaswamy, D. Rajan, J. Liang, A. Chaudhari, and A. Spanias, *Designing counterfactual generators using deep model inversion*, arXiv:2109.14274, 2021. arXiv: 2109.14274 [cs.LG].
 - [110] S. Rick, G. Giacomelli, H. Wen, *et al.*, *Supermind ideator: Exploring generative AI to support creative problem-solving*, arXiv:2311.01937, 2023. arXiv: 2311.01937 [cs.AI].
 - [111] A. Thomaz and C. Breazeal, “Reinforcement learning with human teachers: Evidence of feedback and guidance with implications for learning performance,” in *AAAI Conference on Artificial Intelligence*, 2006.
 - [112] A. Shahmansoori, *Concurrent brainstorming and hypothesis satisfying: An iterative framework for enhanced retrieval-augmented generation (r2cbr3h-sr)*, arXiv:2401.01835, 2024. arXiv: 2401.01835 [cs.IT].
 - [113] E. Oliveira, K. Fischer, and O. Stepankova, “Multi-agent systems: Which research for which applications,” *Robotics and Autonomous Systems*, vol. 27, no. 1, pp. 91–106, 1999, Multi-Agent Systems Applications, ISSN: 0921-8890. DOI: [https://doi.org/10.1016/S0921-8890\(98\)00085-2](https://doi.org/10.1016/S0921-8890(98)00085-2). [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0921889098000852>.
 - [114] R. E. Marks, “Analysis and synthesis: Multi-agent systems in the social sciences,” *The Knowledge Engineering Review*, vol. 27, no. 2, pp. 123–136, 2012. DOI: 10.1017/S0269888912000094.

-
- [115] P. Kollock, "Social dilemmas: The anatomy of cooperation," *Review of Sociology*, vol. 24, pp. 183–214, 1998. [Online]. Available: <https://api.semanticscholar.org/CorpusID:21021101>.
- [116] Z.-H. Zhou, *Machine learning*, S. Liu, Ed. Singapore: Springer Singapore, 2021, ISBN: 978-981-15-1967-3. DOI: <https://doi.org/10.1007/978-981-15-1967-3>.
- [117] F. Fui-Hoon Nah, R. Zheng, J. Cai, K. Siau, and L. Chen, "Generative AI and chatgpt: Applications, challenges, and AI-human collaboration," *Journal of Information Technology Case and Application Research*, vol. 25, no. 3, pp. 277–304, 2023.
- [118] A. L. Thomaz, C. Breazeal, A. G. Barto, and R. Picard, "Socially guided machine learning," Ph.D. dissertation, Cambridge, MA, USA, 2006.
- [119] A. Sheth, K. Roy, and M. Gaur, *Neurosymbolic AI – why, what, and how*, 2023. arXiv: 2305.00813 [cs.AI]. [Online]. Available: <https://arxiv.org/abs/2305.00813>.
- [120] J. Conklin, *Dialogue mapping: building shared understanding of wicked problems*. Chichester, UK: Wiley, 2006.
- [121] R. Axelrod, "Advancing the art of simulation in the social sciences," in *Simulating Social Phenomena*, R. Conte, R. Hegselmann, and P. Terna, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 1997, pp. 21–40, ISBN: 978-3-662-03366-1.
- [122] J. Bongard, "Biologically inspired computing," *IEEE Computer*, vol. 42, pp. 95–98, Apr. 2009. DOI: 10.1109/MC.2009.104.
- [123] J. P. Garbiso, A. Diaconescu, M. Coupechoux, J. Pitt, and B. Leroy, "Distributive Justice for Fair Auto-Adaptive Clusters of Connected Vehicles," in *2017 IEEE 2nd International Workshops on Foundations and Applications of Self* Systems (FAS*W)*, Los Alamitos, CA, USA: IEEE Computer Society, Sep. 2017, pp. 79–84. DOI: 10.1109/FAS-W.2017.124. [Online]. Available: <https://doi.ieeecomputersociety.org/10.1109/FAS-W.2017.124>.
- [124] Y. Li, H. Xu, Q. Cao, Z. Li, and S. Shen, "Evolutionary game-based trust strategy adjustment among nodes in wireless sensor networks," *International Journal of Distributed Sensor Networks*, vol. 11, no. 2, p. 818903, 2015. DOI: 10.1155/2015/818903. eprint: <https://doi.org/10.1155/2015/818903>. [Online]. Available: <https://doi.org/10.1155/2015/818903>.
- [125] J. Pitt, J. Schaumeier, and A. Artikis, "Coordination, conventions and the self-organisation of sustainable institutions," in *Agents in Principle, Agents in Practice: 14th International*

- Conference, PRIMA 2011, Wollongong, Australia, November 16-18, 2011. Proceedings 14*, Springer, 2011, pp. 202–217.
- [126] M. Shanahan, “An abductive Event Calculus planner,” *The Journal of Logic Programming*, vol. 44, no. 1-3, pp. 207–240, 2000.
 - [127] N. Duan, D. Tang, and M. Zhou, “Machine reasoning: Technology, dilemma and future,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, A. Villavicencio and B. Van Durme, Eds., Online: Association for Computational Linguistics, Nov. 2020, pp. 1–6. DOI: 10.18653/v1/2020.emnlp-tutorials.1. [Online]. Available: <https://aclanthology.org/2020.emnlp-tutorials.1>.
 - [128] R. Solso, O. Maclin, and M. MacLin, *Cognitive Psychology: Pearson New International Edition*. Pearson Education, 2013, ISBN: 9781292055497. [Online]. Available: <https://books.google.co.uk/books?id=RtCpBwAAQBAJ>.
 - [129] N. Minar, R. Burkhart, C. Langton, and M. Askenazi, “The swarm simulation system: A toolkit for building multi-agent simulations,” *Santa Fe Institute Working Paper*, vol. 96-06-042, Jul. 1996.
 - [130] M. North, N. Collier, and J. Vos, “Experiences creating three implementations of the repast agent modeling toolkit,” *ACM Trans. Model. Comput. Simul.*, vol. 16, pp. 1–25, Jan. 2006. DOI: 10.1145/1122012.1122013.
 - [131] R. Tobias and C. Hofmann, “Evaluation of free java-libraries for social-scientific agent based simulation,” *Journal of Artificial Societies and Social Simulation*, vol. 7, pp. 1–6, Mar. 2004.
 - [132] J. Kazil, D. Masad, and A. Crooks, “Utilizing Python for agent-based modeling: The mesa framework,” in *Social, Cultural, and Behavioral Modeling*, R. Thomson, H. Bisgin, C. Dancy, A. Hyder, and M. Hussain, Eds., Cham: Springer International Publishing, 2020, pp. 308–317, ISBN: 978-3-030-61255-9.
 - [133] W. Geary, M. Bode, T. Doherty, *et al.*, “A guide to ecosystem models and their environmental applications,” *Nature ecology and evolution*, vol. 4, pp. 1459–1471, Sep. 2020. DOI: 10.1038/s41559-020-01298-8.
 - [134] A. Mertzani and J. Pitt, “Sustainable self-organisation of socio-techno-ecological systems,” in *IEEE International Conference on Autonomic Computing and Self-Organizing Systems, ACSOS Proceedings Companion*, IEEE, 2023, pp. 92–97.
 - [135] M. Tsvetkova, T. Yasseri, N. Pescetelli, and T. Werner, “Human-machine social systems,” *arXiv preprint arXiv:2402.14410*, 2024.

-
- [136] P. F. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, and D. Amodei, “Deep reinforcement learning from human preferences,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS’17, Long Beach, California, USA: Curran Associates Inc., 2017, pp. 4302–4310, ISBN: 9781510860964.
- [137] J. MacGlashan, M. K. Ho, R. Loftin, *et al.*, “Interactive learning from policy-dependent human feedback,” in *International conference on machine learning*, PMLR, 2017, pp. 2285–2294.
- [138] A. L. Thomaz, G. Hoffman, and C. Breazeal, “Experiments in socially guided machine learning: Understanding how humans teach,” in *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction*, ser. HRI ’06, Salt Lake City, Utah, USA: Association for Computing Machinery, 2006, pp. 359–360, ISBN: 1595932941. DOI: 10.1145/1121241.1121315. [Online]. Available: <https://doi.org/10.1145/1121241.1121315>.
- [139] P. R. Lewis, S. Marsh, and J. Pitt, “AI vs “AI”: Synthetic minds or speech acts,” *IEEE Technology and Society Magazine*, vol. 40, no. 2, pp. 6–13, 2021. DOI: 10.1109/MTS.2021.3077052.
- [140] J. Robbins, “The intelligence factor: Technology and the missing link,” *IEEE Technol. Soc. Mag.*, vol. 41, no. 1, pp. 82–93, 2022. DOI: 10.1109/MTS.2022.3147528. [Online]. Available: <https://doi.org/10.1109/MTS.2022.3147528>.
- [141] J. Robbins, “If technology is a parasite masquerading as a symbiont?are we the host?” *IEEE Technol. Soc. Mag.*, vol. 38, no. 3, pp. 24–33, 2019. DOI: 10.1109/MTS.2019.2930267. [Online]. Available: <https://doi.org/10.1109/MTS.2019.2930267>.
- [142] B. Friedman, P. Kahn, and A. Borning, “Value sensitive design and information systems,” in *The Handbook of Information and Computer Ethics*, K. Himma and H. Tavani, Eds., Wiley-Interscience, 2008, pp. 69–101.
- [143] M. A. Boden, “Creativity and Artificial Intelligence,” *Artificial Intelligence*, vol. 103, no. 1-2, pp. 347–356, 1998.
- [144] G. Harman, “Practical reasoning,” *The Review of Metaphysics*, pp. 431–463, 1976.
- [145] B. Chen and X. Zhu, “Integrating generative AI in knowledge building,” *Computers and Education: Artificial Intelligence*, vol. 5, p. 100184, Nov. 2023. DOI: 10.1016/j.caeai.2023.100184.

-
- [146] W. Chen, Y. Su, J. Zuo, *et al.*, “Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors in agents,” *arXiv preprint arXiv:2308.10848*, vol. 2, no. 4, p. 6, 2023.
 - [147] M. Afshar and M. Asadpour, “Opinion Formation by Informed Agents,” *Journal of Artificial Societies and Social Simulation*, vol. 13, no. 4, pp. 1–5, 2010.
 - [148] D. J. Watts and S. H. Strogatz, “Collective dynamics of ‘small-world’ networks,” *Nature*, vol. 393, no. 6684, pp. 440–442, Jun. 1998, ISSN: 1476-4687.
 - [149] A.-L. Barabási, R. Albert, and H. Jeong, “Mean-field theory for scale-free random networks,” *Physica A: Statistical Mechanics and its Applications*, vol. 272, no. 1, pp. 173–187, 1999, ISSN: 0378-4371.
 - [150] M. Foucault, *The Order of Things: An Archaeology of the Human Sciences* (Routledge classics). Routledge, 2002, ISBN: 9780415267366. [Online]. Available: <https://books.google.co.uk/books?id=xVeRwX-AvLAC>.
 - [151] P. Mellodge, A. Diaconescu, and L. J. D. Felice, *Timing configurations affect the macro-properties of multi-scale feedback systems*, 2021. arXiv: 2106.01651 [eess.SY]. [Online]. Available: <https://arxiv.org/abs/2106.01651>.
 - [152] D. Kahneman, *Thinking, fast and slow*. Farrar, Straus and Giroux, 2011.
 - [153] C. Landauer and K. Bellman, “Reflective systems need models at run time,” in *Proceedings of the 11th International Workshop on Models@run.time*, 2016, pp. 52–59.
 - [154] J. Dryzek and J. Pickering, “Deliberation as a catalyst for reflexive environmental governance,” *Ecological Economics*, vol. 131, pp. 353–360, 2017.
 - [155] S. Emmons, S. Kobourov, M. Gallant, and K. Börner, “Analysis of network clustering algorithms and cluster quality metrics at scale,” *PloS one*, vol. 11, no. 7, e0159161–e0159161, 2016, ISSN: 1932-6203.
 - [156] D. Welburn, “Rawls, the well-ordered society and intergenerational justice,” *Politics*, vol. 33, no. 1, pp. 56–65, 2013.
 - [157] R. Collier and D. Collier, *Shaping The Political Arena*. Princeton, NJ: Princeton University Press, 1991.
 - [158] J. Hauser and G. Katz, “Metrics: You are what you measure!” *European Management Journal*, vol. 16, no. 5, pp. 517–528, 1998, ISSN: 0263-2373. DOI: [https://doi.org/10.1016/S0263-2373\(98\)00029-2](https://doi.org/10.1016/S0263-2373(98)00029-2). [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0263237398000292>.

-
- [159] E. Patterson and J. Miller, *Macrocognition Metrics and Scenarios: Design and Evaluation for Real-World Teams*. London, UK: Ashgate Publishing Limited, 2012, ISBN: 9781409486367. [Online]. Available: <https://books.google.gr/books?id=fDx7BgAAQBAJ>.
- [160] J. Muller, *The Tyranny of Metrics*. Princeton, NJ: Princeton University Press, 2019.
- [161] C. Goodhart, "Problems of monetary management: The u.k. experience," *Papers in monetary economics 1975*, vol. 1, no. 1, pp. 1–20, 1975.
- [162] E. Ries, *The Lean Startup*. London, UK: Portfolio Penguin, 2011.
- [163] E. Sarigöl, R. Pfitzner, I. Scholtes, A. Garas, and F. Schweitzer, "Predicting scientific success based on coauthorship networks," *EPJ Data Science*, vol. 3, article number 9, 2014.
- [164] M. Sandel, *The Tyranny of Merit*. London, UK: Penguin, 2021.
- [165] C. J. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, pp. 279–292, 1992.
- [166] R. Conant and W. Ashby, "Every good regulator of a system must be a model of that system," *International Journal of Systems Science*, vol. 1, pp. 89–97, Oct. 1970. DOI: 10.1080/00207727008920220.
- [167] J. Driver, "A selective review of selective attention research from the past century," *British Journal of Psychology*, vol. 92, no. 1, pp. 53–78, 2001.
- [168] D. Osborne and T. heat, "The role of social space requirements in ergonomics," *Applied Ergonomics*, vol. 10, no. 2, pp. 99–103, 1979.
- [169] B. Horne, D. Nevo, J. Freitas, H. Ji, and S. Adali, "Expertise in social networks: How do experts differ from other users?" In *Proceedings Tenth International AAAI Conference on Web and Social Media*, 2016, pp. 583–586.
- [170] M. Deutsch and H. Gerard, "A study of normative and informational social influences upon individual judgment," *The Journal of Abnormal and Social Psychology*, vol. 51, no. 3, pp. 629–636, 1955.
- [171] M. Boisot and B. McKelvey, "Complexity and organization–environment relations: Revisiting Ashby's law of requisite variety," in *The SAGE handbook of complexity and management*, P. Allen, S. Maguire, and B. McKelvey, Eds., SAGE, Jan. 2011, pp. 279–298. DOI: 10.4135/9781446201084.n16.
- [172] V. Mnih, K. Kavukcuoglu, D. Silver, *et al.*, "Playing Atari with deep reinforcement learning," *CoRR*, vol. abs/1312.5602, 2013. arXiv: 1312.5602. [Online]. Available: <http://arxiv.org/abs/1312.5602>.

-
- [173] 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. [Online]. Available: <http://jmlr.org/papers/v22/20-1364.html>.
- [174] W. R. Ashby, *An Introduction to Cybernetics*. London: Chapman and Hall, 1956.
- [175] M. Hardt, E. Mazumdar, C. Mendler-Dünner, and T. Zrnic, “Algorithmic collective action in machine learning,” ser. ICML’23, Honolulu, Hawaii, USA: JMLR.org, 2023.
- [176] C. List, “Social Choice Theory,” in *The Stanford Encyclopedia of Philosophy*, E. N. Zalta and U. Nodelman, Eds., Winter 2022, Metaphysics Research Lab, Stanford University, 2022.
- [177] P. E. Petruzzi, J. Pitt, and D. Busquets, “Electronic social capital for self-organising multi-agent systems,” *ACM Transactions on Autonomous and Adaptive Systems (TAAS)*, vol. 12, no. 3, pp. 1–25, 2017.
- [178] A. Rapoport, Ed., *Game theory as a theory of conflict resolution*. Dordrecht: Reidel, 1974.
- [179] H. Ahlborg, I. Ruiz-Mercado, S. Molander, and O. Masera, “Bringing technology into social-ecological systems research—motivations for a socio-technical-ecological systems approach,” *Sustainability*, vol. 11, no. 7, 2019, ISSN: 2071-1050.
- [180] M. Soga and K. Gaston, “The ecology of human–nature interactions,” *Proceedings of the Royal Society B: Biological Sciences*, vol. 287, p. 20 191 882, Jan. 2020. DOI: 10.1098/rspb.2019.1882.
- [181] S. K. Sikdar, “Sustainable development and sustainability metrics,” *AIChE journal*, vol. 49, no. 8, pp. 1928–1932, 2003.
- [182] B. H. C. Cheng, A. Ramirez, and P. McKinley, “Harnessing evolutionary computation to enable dynamically adaptive systems to manage uncertainty,” in *Proc. 1st International Workshop on Combining Modelling and Search-Based Software Engineering*, 2013, pp. 1–6.
- [183] T. Vogel, S. Neumann, S. Hildebrandt, H. Giese, and B. Becker, “Model-driven architectural monitoring and adaptation for autonomic systems,” in *Proceedings of the 6th International Conference on Autonomic Computing*, ser. ICAC ’09, Barcelona, Spain: Association for Computing Machinery, 2009, pp. 67–68, ISBN: 9781605585642. DOI: 10.1145/1555228.1555249. [Online]. Available: <https://doi.org/10.1145/1555228.1555249>.
- [184] A. Kumar, A. Zhou, G. Tucker, and S. Levine, *Conservative q-learning for offline reinforcement learning*, 2020. arXiv: 2006.04779 [cs.LG].

-
- [185] I. Osband, Z. Wen, S. M. Asghari, *et al.*, *Epistemic neural networks*, 2023. arXiv: 2107.08924 [cs.LG].
- [186] A. Malinin and M. Gales, *Predictive uncertainty estimation via prior networks*, 2018. arXiv: 1802.10501 [stat.ML].
- [187] U. Secretary-General, *Report of the World Commission on Environment and Development: Our Common Future*. Oxford University Press, 1987.
- [188] E. A. Moallemi, E. Bertone, S. Eker, *et al.*, “A review of systems modelling for local sustainability,” *Environmental Research Letters*, vol. 16, no. 11, p. 113004, Nov. 2021. DOI: 10.1088/1748-9326/ac2f62. [Online]. Available: <https://dx.doi.org/10.1088/1748-9326/ac2f62>.
- [189] T. W. Killian, S. Parbhoo, and M. Ghassemi, “Risk sensitive dead-end identification in safety-critical offline reinforcement learning,” *arXiv preprint arXiv:2301.05664*, 2023.
- [190] A. Balanzó-Guzmán and M. Ramos-Mejía, “Towards epistemic diversity in sustainability transitions: An exploration of hybrid socio-technical systems,” *Sustainability Science*, vol. 18, no. 6, pp. 2511–2531, 2023.
- [191] M. Yolles, “Metacybernetics: Towards a general theory of higher order cybernetics,” *Systems*, vol. 9, no. 2, 2021, ISSN: 2079-8954. [Online]. Available: <https://www.mdpi.com/2079-8954/9/2/34>.
- [192] F. Wach, J. Karbach, S. Ruffing, R. Brünken, and F. Spinath., “University students’ satisfaction with their academic studies: Personality and motivation matter,” *Frontiers in Psychology*, vol. 7, pp. 1–55, 2016.
- [193] E. Chepin, “Robotics: From first-order cybernetics to third-order cybernetics,” *Procedia Computer Science*, vol. 190, pp. 130–136, 2021.
- [194] A. Bourazeri and J. Pitt, “Collective attention and active consumer participation in community energy systems,” *Int. J. Hum. Comput. Stud.*, vol. 119, pp. 1–11, 2018. DOI: 10.1016/J.IJHCS.2018.06.001. [Online]. Available: <https://doi.org/10.1016/j.ijhcs.2018.06.001>.
- [195] K. Stockdale, “Police understandings of restorative justice — the impact of rank and role,” *Restorative Justice*, vol. 3, no. 2, pp. 212–232, 2015.
- [196] X. Zhan, Y. Xu, and S. Sarkadi, “Deceptive AI ecosystems: The case of chatgpt,” in *Proceedings of the 5th International Conference on Conversational User Interfaces*, 2023, pp. 1–6.

-
- [197] J. Klusowski, D. A. Small, and J. P. Simmons, “Does choice cause an illusion of control?” *Psychological Science*, vol. 32, no. 2, pp. 159–172, 2021.
- [198] K. Michael, J. Pitt, R. Abbas, C. Perakslis, and M. Michael, “Cyborgs and human–machine communication configurations,” in *The SAGE Handbook of Human-Machine Communication*, SAGE, 2023, p. 32.
- [199] J. Lansing and J. Kremer, “Emergent properties of Balinese water temple network,” *American Anthropologist*, vol. 95, pp. 97–114, 1993.
- [200] J. Hoogesteger, A. Bolding, C. Sanchis-Ibor, *et al.*, “Communality in farmer managed irrigation systems: Insights from spain, ecuador, cambodia and mozambique,” *Agricultural Systems*, vol. 204, p. 103 552, 2023, ISSN: 0308-521X. DOI: <https://doi.org/10.1016/j.agry.2022.103552>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0308521X22001883>.
- [201] Ş. Sarkadi, A. Rutherford, P. McBurney, S. Parsons, and I. Rahwan, “The evolution of deception,” *Royal Society open science*, vol. 8, no. 9, p. 201 032, 2021.
- [202] K. Milanovic and J. Pitt, “Misattribution of error origination: The impact of preconceived expectations in co-operative online games,” in *Designing Interactive Systems (DIS)*, W. Ju, L. Oehlberg, S. Follmer, S. E. Fox, and S. Kuznetsov, Eds., 2021, pp. 707–717.
- [203] K. Michael, “High-tech lust,” *IEEE Technol. Soc. Mag.*, vol. 32, no. 2, pp. 4–5, 2013.
- [204] A. Mertzani and J. Pitt, “Socially-guided machine learning for self-organised community empowerment,” in *Accepted for publication in 17th International Conference on Agents and Artificial Intelligence (ICAART)*, 2025.
- [205] J. Pitt and A. Nowak, “Collective awareness and the new institution science,” in *The Computer After Me*, J. Pitt, Ed., London: ICPress, 2014, ch. 12, pp. 207–218.
- [206] OpenAI, *ChatGPT API*, <https://chat.openai.com>, 2024.
- [207] R. Mancilla, “Introduction to sociocybernetics (part 3): Fourth order cybernetics,” *Journal of Sociocybernetics*, vol. 11, pp. 47–73, 2013.
- [208] S. Pitt, M. van Meelis Lacey, E. Scaife, and J. Pitt, “No app is an island: Collective action and sustainable development goal-sensitive design,” *Int. J. Interact. Multim. Artif. Intell.*, vol. 6, no. 5, pp. 24–33, 2021. [Online]. Available: <https://www.ijimai.org/journal/bibcite/reference/2890>.

-
- [209] N. I. Platnick and E. S. Gaffney, "Evolutionary biology: A popperian perspective," *Systematic Zoology*, vol. 27, no. 1, pp. 138–141, 1978, ISSN: 00397989. [Online]. Available: <http://www.jstor.org/stable/2412832> (visited on 10/13/2024).
- [210] L. Floridi, *The Fourth Revolution: How the Infosphere is Reshaping Human Reality*. Oxford University Press, USA, 2014.
- [211] G. Aguzzi, M. Viroli, and L. Esterle, "Field-informed Reinforcement Learning of Collective Tasks with Graph Neural Networks," in *2023 IEEE International Conference on Autonomic Computing and Self-Organizing Systems (ACSOS)*, Los Alamitos, CA, USA: IEEE Computer Society, Sep. 2023, pp. 37–46. DOI: 10.1109/ACSOS58161.2023.00021. [Online]. Available: <https://doi.ieeecomputersociety.org/10.1109/ACSOS58161.2023.00021>.
- [212] B. Manville and J. Ober, *The Civic Bargain: How Democracy Survives*. Princeton University Press, 2023.
- [213] M. S. Santos and J. Pitt, "Macs - affective conditioning system for shared working environments," in *2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, vol. 3, 2011, pp. 57–60. DOI: 10.1109/WI-IAT.2011.35.
- [214] M. S. Santos and J. Pitt, "Emotions and norms in shared spaces," in *Coordination, Organizations, Institutions, and Norms in Agent Systems IX*, T. Balke, F. Dignum, M. B. van Riemsdijk, and A. K. Chopra, Eds., Cham: Springer International Publishing, 2014, pp. 157–176, ISBN: 978-3-319-07314-9.
- [215] A. Mertzani, J. Pitt, S. Sarkadi, M. Sas, M. Scott, and C. Smit, "Cohesion and the explanation of constitutional choice in self-governing systems," *ALife Workshop Agent-Based Modelling of Human Behaviour (ABMHuB)*, 2024.
- [216] C. Smit, M. Scott, A. Mertzani, and J. Pitt, "Salab: Computer-supported social arrangements laboratory," in *Frontiers of Artificial Intelligence, Ethics, and Multidisciplinary Applications*, M. Farmanbar, M. Tzamtzi, A. Verma, and A. Chakravorty, Eds., Singapore: Springer Nature Singapore, 2024, pp. 299–312.
- [217] M. Scott, A. Mertzani, C. Smit, S. Sarkadi, and J. Pitt, "Social deliberation vs. social contracts in self-governing voluntary organisations," *arXiv preprint arXiv:2403.16329*, 2024.
- [218] B. Blackledge, A. Papaoikonomou, M. Scott, *et al.*, "Incentivising participation with exclusionary sanctions (full)," in *International Workshop on Coordination, Organizations, Institutions, Norms, and Ethics for Governance of Multi-Agent Systems*, Springer, 2023, pp. 37–54.