The Emergence of Complex Behaviours in Agent-Based Models using Reinforcement Learning

Sedar Olmez

School of Geography

University of Leeds

A thesis submitted for the degree of

Doctor of Philosophy

October 2023

Intellectual Property and Publication Statements

The candidate confirms that the work submitted is his own, except where work which has formed part of jointly authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others. The peer-reviewed and published articles, co-authored, are found in the Appendix as A.1 and A.2:

Olmez, S.; Douglas-Mann, L.; Manley, E.; Suchak, K.; Heppenstall, A.; Birks, D.; Whipp, A. Exploring the Impact of Driver Adherence to Speed Limits and the Interdependence of Roadside Collisions in an Urban Environment: An Agent-Based Modelling Approach. Appl. Sci. 2021, 11, 5336. https://doi.org/10.3390/app11125336

Olmez, S.; Thompson, J.; Marfleet, E.; Suchak, K.; Heppenstall, A.; Manley, E.; Whipp, A.; Vidanaarachchi, R. An Agent-Based Model of Heterogeneous Driver Behaviour and Its Impact on Energy Consumption and Costs in Urban Space. Energies 2022, 15, 4031. https://doi.org/10.3390/en15114031

I declare that, the research articles were born out of my personal interest and ideas from readings into traffic modelling. All software code was written by myself and hosted on my open-source repositories. The main body of the articles were written by myself, however, co-authors have contributed to various sections, these include the literature review and discussion and conclusion. Furthermore, co-authors have recommended and made changes to the research article and supported the planning and execution of model experimentation and analysis of results.

Acknowledgements

The work you are about to embark on deciphering could not have been achieved without the tutelage and brilliance of my supervisors Alison Heppenstall, Dan Birks and Jiaqi Ge. For the past four years, I have felt like a nomad, living in a yurt somewhere in Mongolia with nothing but a laptop and a free subscription to Indian cricket. Speaking of cricket, I should take this opportunity to thank my colleague and friend Akhil Ahmed for not boring me to death during our research at Accenture, that corporate fuelled six-month experience will probably be published in my memoir when I am on death's door. Given that this thesis will live through the ages, while our bones turned to dust, I should also congratulate the highly knowledgeable and charismatic Jason Thompson from the University of Melbourne for trusting me in managing a research project which, to my surprise, was published. One summer's morning, while fannying around in my pyjama bottoms like the rest of the world, I was graced by a random video call by Nigel Gilbert an idol of mine. To my surprise, Nigel was interested in the housing model I had built in Python, thanks to his work, therefore, I would like to thank him and his colleagues at Surrey University. Lastly, I would like to thank my friend Annabel Whipp, with whom I share a four-legged child with called Rupert, for loving me at my worst and deserving me at my best. Before I sign off forever, I must also thank those that shall not be named, the viva examiners, thank you for taking the time out of your busy schedules to read, digest and not throw up.

Abstract

Over the last two decades, agent-based modelling (ABM) has become an invaluable tool across various disciplines for simulating the intricacies of complex systems. At the core of these models are the decision-making frameworks that guide agent behaviour. Traditionally, these have been constrained to predetermined rules, providing deterministic insights yet often omitting the dynamic process of learning and adaptation inherent in real-world behaviour. This research examines the integration of Reinforcement Learning (RL) within a set of ABMs—designated as 'hybrid-ABMs'. These models are characterised by their ability to merge traditional ABMs with the adaptive, experience-driven learning processes facilitated by RL. By doing so, this thesis explores to what extent such integration enables agents to develop and refine intelligent behaviours in response to environmental stimuli and changing scenarios. Focusing on three distinct domains—predator-prey ecosystems, burglary and criminal behaviour, and the economic behaviours of housing markets—this thesis assesses how RL can empower ABMs to simulate complex phenomena with greater detail. Through these hybrid-ABMs, we scrutinise whether RL-enhanced agents can autonomously acquire behaviours that not only resonate with theoretical and empirical findings but also adapt to unforeseen circumstances with a degree of experience and reliability. The findings are: agents within these hybrid-ABMs successfully learn from their environment, the emergent behaviours align with established literature in the respective fields, and the agents exhibit a capacity to adapt to novel situations effectively. This thesis hypothesises that neurologically inspired algorithms like RL can enhance sociological ABMs by introducing an element of learning and adaptability, thereby equipping these models to better mirror the complexity of real-world systems and decision-making processes.

Abbreviations

ABM	Agent-Based Model/Modelling
RL	Reinforcement Learning
PPO	Proximal Policy Optimisation
RCP	Rational Choice Perspective
RAT	Routine Activity Theory
CPT	Crime Pattern Theory
TRPO	Trust Region Policy Optimisation
A2C	Actor 2 Critic
ACT-R/PM	Adaptive Control of Thought - Rational / Procedural Memory
AI	Artificial Intelligence
ANN	Artificial Neural Network
ANOVA	Analysis of Variance
BDI	Beliefs, Desires and Intentions
BOID	Beliefs, Desires, Obligations, and Intentions
BRIDGE	Beliefs, Rules, Intentions, Desires, Goals and Emotions
CLARION	Connectionist learning with adaptive rule induction ON-line
CONSUMAT	Consumer Satisfaction and Utility Maximisation
CPI	Conservative Policy Iteration
DBQ	Driver Behaviour Questionnaire
DQN	Deep-Q Learning
EAV	Electric Autonomous Vehicles
EEG	Electroencephalography
EMIL-A	Ethical Minded Intelligent Learning Agent
EV	Electric Vehicle
GA	Genetic Algorithm
GAE	Generalised Advantage Estimate

GB	Great Britain
GBP	Great British Pound
GWP	Global Warming Potentials
MHP/RT	Model Human Processor with Real-Time Constraints
ICEV	Internal Combustion Engine Vehicle
JTC	Journey to Crime Curve
KDE	Kernel Density Estimation
MAS	Multi-agent System
MDP	Markov Decision Process
MHP	Model Human Processor
ML	Machine Learning
NDS	Naturalistic Driving Study
ODD	Overview, Design and Details
PCO	Per-kilometre Cost of Ownership
PECS	Physical conditions, Emotional states, Cognitive capabilities, Social status
PHEV	Plug-in Hybrid Electric Vehicles
PRS	Production Rule System
RAN	Routine Activity Node
SCPI	Situational Crime Prevention Intervention
SOAR	State, Operator, and Result
TCR	Target Cumulative Reward
TD	Temporal Difference
UTS	Urban Traffic Simulator
WTW	Well-to-wheel
eBDI	Extended Beliefs, Desires and Intentions
DNA	Deliberative Normative Agents
NoA	Normative Agent

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction to the research

In recent decades, researchers have adopted agent-based modelling (ABM) (Epstein, 1999; Epstein & Axtell, 1997) as a tool to explore domain-specific phenomena, such as offender behaviours and interventions, stress testing economic policy, housing market trends and cooperative decision-making, to name but a few (Birks *et al.*, 2012; Crooks, 2015; Hamill & Gilbert, 2015; Malleson *et al.*, 2010; Secchi, 2015; Squazzoni, 2012; Zhuge & Shao, 2018). ABM offers several key strengths for understanding complex systems. ABM focuses on modelling individual agents and their interactions, allowing for a detailed representation of heterogeneity and individual behaviours. This bottom-up approach captures emergent phenomena and system-level patterns that arise from micro-level interactions. ABM is flexible and serves as a virtual laboratory for conducting experiments and exploring system dynamics. It provides insights into nonlinear dynamics and feedback loops. ABM can inform policy design, and support evidence-based strategies. For instance, ABMs have been pivotal in urban planning to evaluate the impacts of boundary changes on city growth and traffic patterns (Gulden *et al.*, 2011; Orsi, 2019). In public health, they have been used to simulate the spread of infectious diseases and the potential effects of interventions such as social distancing policies (Kerr *et al.*, 2021; Silva *et al.*, 2020). In environmental policy, ABMs have helped in understanding the potential outcomes of water resource management strategies, allowing for the assessment of the sustainability of different usage policies (Berglund, 2015; Darbandsari *et al.*, 2020). Such applications emphasise the utility of ABMs in developing

policies. Integration with empirical data enhances model realism and predictive power through calibration and validation, enabling the exploration of real-world "what-if" scenarios.

To ensure the reliability and accuracy of ABM, two key processes are essential: verification and validation. Verification is the process of confirming that the computational model's implementation is consistent with the conceptual model, and it is correctly executed within the simulation environment (Anderson & Titler, 2014; Curreli *et al.*, 2021). Validation, however, is the process of determining that the model is an accurate representation of the real-world system, with a focus on the model's output and its ability to predict future states (Guerini & Moneta, 2017; Napoletano *et al.*, 2018; Tieleman, 2022).

Like all analytical approaches, ABM has its own set of limitations, perhaps one of the most predominant is embedding accurate behavioural rules, representing these behaviours and simulating them, ultimately, these issues arise from the absence of decision-making frameworks that better reflect the modelled agent's behaviours (Balke & Gilbert, 2014; Francès *et al.*, 2015). Researchers have argued that incorporating rules-based or state-transition based frameworks lead to "simplistic" behaviours, which in turn impact the macroscopic patterns that emerge from the interactions of these agents negatively (Francès *et al.*, 2015; Ramchandani *et al.*, 2017). Additionally, others have argued that these behavioural rules do not represent how "real individuals" decide on actions, which they argue oversimplifies behaviours (DeAngelis & Diaz, 2019). Kennedy (2012) argues that simulating complex behaviours such as human behaviours is not obvious; these behaviours are not random and are heterogeneous. Due to the complexity of simulating human decision-making, Groeneveld *et al.* (2017) found that most models reviewed resulted in ad-hoc representations of human decisions in agent-based land-use models, and most often, these representations are not grounded in theory. Researchers in health and disease control found that a critical challenge in utilising ABMs to inform public policy was the conceptualisation process of decision rules. Identifying representative condition-action rules to model the agent's behaviours is resource-intensive, and as model complexity increases, the margin for error exponentially increases (Badham *et al.*, 2018; Cocea *et al.*, 2012). As computational social scientists (and other disciplines studying agent-agent and agent-environment interactions through computational methods) investigate research questions involving behaviours of individual people in an

environment (Conte *et al.*, 2012; Torrens, 2010), the methods must better reflect the decision-making processes of these individuals (Ramchandani *et al.*, 2017).

Complex systems are characterised by individual components, or agents, whose interactions lead to emergent properties that cannot be predicted by analysing the components in isolation (Simon, 1962). These systems exhibit features such as nonlinearity, self-organisation, and adaptability. Herbert Simon's seminal work on complexity outlines that a complex system exhibits a hierarchical structure, where components are nested within each other, each level providing a different lens of understanding (Simon, 1962). In contrast, 'less' complex systems, while still exhibiting interactions between their components, do not necessarily show a high degree of emergent behaviours or hierarchical organisation. These systems are often easier to predict and manage due to their reduced number of interactive elements and simpler organisational structure.

The term 'complicated,' on the other hand, refers to systems or problems that, although they may have many parts, the parts have predictable and well-defined interactions. Miller and Page differentiate complicated systems from complex ones by emphasising that in complicated systems, the interactions do not give rise to new emergent properties (Miller & Yu, 2008; Page, 2015). Complex real-world decision-making, therefore, involves navigating through systems that are inherently complex, where the decision-maker must account for a multitude of interacting variables and emergent phenomena that are a result of these interactions. This contrasts with decision-making in 'less' complex or complicated scenarios, where variables and outcomes are more predictable and controllable.

A key challenge in modelling complex adaptive decision-making in ABMs is the ability to integrate decision-making algorithms with the model architecture whereby agents can learn behaviours naturally from individual-level perceptions without any pre-determined behaviour rules (Bryant, 2004; Midgley *et al.*, 2007). An area of research that has gained momentum in the past decade and can be used to overcome the challenge mentioned above is Reinforcement Learning (RL) (Sutton & Barto, 2018a). There have and continue to be many types of computational RL algorithms, from Deep-Q Learning (DQN), Proximal Policy Optimisation (PPO) to Actor-Critic Style (A2C) (Deepanshu Mehta, 2020; Schulman *et al.*, 2015). However, the principle processes all algorithms are founded on are outlined by (Sutton & Barto, 2018a, p. 1): "reinforcement learning is learning what to do, how to map situations to actions to

maximise a numerical reward signal. The learner is not told which actions to take but instead must discover which actions yield the most reward by trying them. In the most interesting and challenging cases, actions may affect not only the immediate reward but also the following situation and, through that, all subsequent rewards. These two characteristics, trial-and-error search and delayed reward, are the two most important distinguishing features of reinforcement learning." RL is grounded in neurological and psychological theory inspired by animal learning (Mnih *et al.*, 2015; Subramanian *et al.*, 2022; Thorndike, 1911). Its earliest conception in literature dates to 1911 - 1927, where Thorndike (1927) established the Law of Effect. The Law of Effect states that behavioural responses that are most proximal to a satisfying result are more likely to become patterns. Rewards for appropriate behaviour continuously strengthen associations. Similarly, behavioural scientists have shown that RL signal processes occur within parts of the human brain, such as the striatum, parietal and frontal cortices (Niv, 2009; O'Doherty *et al.*, 2015) making it a suitable framework to model behaviour.

While other individual-based modelling methods such as cellular automata (CA) and microsimulation (MS) have been used to simulate complex systems (Gómez-Marín *et al.*, 2018; Huynh *et al.*, 2019; Kamo *et al.*, 2022; Khan & Habib, 2021; Lin & Yao, 2018; Ruan *et al.*, 2020). The selection of ABM for this research is underpinned by its compatibility with RL and the practicalities of the computational resources available. ABM offers detailed training features for RL due to its inherent capacity to model individual agents within spatially detailed environments. These agents are capable of complex interactions and exhibit heterogeneity in their behaviours, which is a crucial aspect when simulating real-world scenarios. This individual-centric approach is essential for RL, which thrives on the ability to train agents based on their unique experiences within the environment. It is this combination of agent complexity, detailed environmental interaction, and the support of mature programming frameworks such as Unity, Tensorflow and Keras (A Gulli, 2012; Juliani *et al.*, 2018; Martinez, 2016) that positions ABM as the most appropriate method for the integration of RL in this research. In contrast, MS typically simulates more agents than ABMs, however, with no interaction between the individuals or heterogeneous behaviour. Given the hardware at hand, training this many agents is not computationally feasible (Birkin, 2021; Brearcliffe & Crooks, 2021; Peichl, 2016). CA, on the other hand, are less computationally expensive. While it is acknowledged that programming frameworks facilitating the

integration of Cellular Automata (CA) and Microsimulation (MS) with Reinforcement Learning (RL) are in the stages of development, it is also important to recognise the strides RL has made within the domain of CA, especially concerning land-use modelling. A growing body of literature illustrates the successful application of RL in enhancing CA models, enabling them to navigate complex spatial decisions and simulate dynamic land-use changes effectively (He *et al.*, 2018a; Sajan *et al.*, 2022; Xing *et al.*, 2020a). Nevertheless this advancement, when considering the breadth and depth of integration as a whole, especially in terms of accessible and robust frameworks, RL's application in CA and MS models does not yet match the established prevalence seen within ABM. This is evident from a Web of Science search, which shows a considerably lower number of publications combining "cellular automata" or "microsimulation" with "reinforcement learning" (46 and 115 respectively since 1997) compared to those of "agent-based model" with "reinforcement learning" (419 since 2013). The disparity in these figures may reflect the more mature development and broader adoption of RL in ABMs, signifying a greater level of integration in terms of the tools and frameworks available for researchers within this particular scope of computational modelling.

RL as a decision-making framework paved the way for human-computer interaction where algorithms were able to beat the world's greatest players in Chess (Deep Blue by IBM in 1996) and Go (Alpha Go by Deep Mind in 2015) (Silver *et al.*, 2017). During this period, the focus was on RL's ability to learn, adapt and achieve goals. Some research disciplines such as Computer Science, Robotics and Engineering preempted the development of these algorithms for their relative disciplines (Doya, 2010). However, RL has gained momentum in the soft sciences. An example application was to investigate social dilemmas and strategy (Izquierdo *et al.*, 2008) scholars demonstrate how RL can identify some mistakes made by the decision-making of actors in social interactions. Schelling's segregation model (Schelling, 1969) was revisited by RL researchers to re-build the ABM, where the properties of social systems are investigated through the interplay and integration of actions by RL agents (Sert *et al.*, 2020). Modelling the dynamics of street robbery in South Africa using an ABM with RL decision-making enabled researchers to simulate realistic offender behaviours, which led to synthetic street robbery patterns similar to those observed in the real-world (Joubert *et al.*, 2022). RL has also been used in Ecology, where researchers modelled the behaviours of predator-prey interactions (Niv *et al.*, 2002; Olsen & Fraczkowski, 2015). Economic

studies where stock trading RL agents are used to optimise the trading process (Charpentier *et al.*, 2021; Dang, 2020). Health research, where models identify the optimum dose of medicine for heterogeneous patient needs (Jalalimanesh *et al.*, 2017; Liu *et al.*, 2020). With these advantages in mind, this thesis will explore how RL embedded within agents in three domain specific ABMs can organically learn behaviours from their environment and adapt to changes while trying to achieve goals without any predetermined decision-rules. A research domain that makes for a good use-case to test this approach in the first instance is predator-prey interactions. These models tend to be more simplistic in nature with only two-types of agents, a simple environment and several simplistic agent rules. Traditionally, these models use rule-based heuristics (Grimm & Railsback, 2012; Heppenstall *et al.*, 2012), therefore, RL can potentially lead to new insights and more realistic models of predator-prey interactions.

A discipline that can benefit from RL is environmental criminology. Some researchers in environmental criminology have recognised that most ABMs developed to investigate criminological phenomena often rely on simplistic condition-action decision-making (Groff *et al.*, 2019; Johnson & Groff, 2014; Johnson *et al.*, 2014; Park & Buckley, 2016). Many examples in the literature discuss the lack of complex decision-making frameworks to simulate realistic decision-making characteristics of offenders, victims and crime preventers, such as adaptive behaviours and learning inspired by their environment (Bosse & Gerritsen, 2008; Bosse *et al.*, 2011; Cornelius *et al.*, 2017; Groff, 2007), which psychological studies of offender decision-making have pointed out as crucial mechanisms utilised by offenders during target selection (Gialopsos & Carter, 2014; Sigurdsson *et al.*, 2008; Topalli, 2005).

Another discipline in which RL has been less explored, yet can be of benefit is housing market research. Some articles have investigated the emergence of housing bubbles (Axtell, 2014; Erlingsson *et al.*, 2014; Ge, 2014, 2017) using ABM, and some have looked at how dynamics in the market lead to urban regeneration (Jordan *et al.*, 2011, 2012; Picascia, 2014) or by embedding survey data to a housing resale market model to simulate search behaviour (Zhang & Li, 2014). Some have looked at how ABMs can help investigate the causes and implications of real-world shocks such as the 2008 financial crash to the housing market (Gilbert *et al.*, 2009; Hamill & Gilbert, 2015). A less-explored topic where an opportunity for further research arises is the application of RL algorithms to support decision-making in alleviating market shocks

in real-time. Central banks can enhance intervention policies when managing real-world economic shocks through the utility of this work. The previously mentioned models have examined how, when, and why shocks occur. However, developing real-time techniques to counteract these shocks is vital to reduce the knock-on effect on the overall economy and people's health and well-being (Oguibenine, 2011).

This research contributes to the field of geography by harnessing the intrinsically detailed nature of Agent-Based Modelling (ABM) to explain the spatial mechanisms that underlie policy-making and criminal behaviour. Integrating geospatial packages for analysis with ABM, this study not only sheds light on the intricate spatial interactions at the individual level that cumulate into broader spatial patterns but also offers a robust methodology for simulating and inspecting the spatial implications of policy interventions. This spatially-attuned methodology is essential for dissecting the intricacies of housing market fluctuations, where location intertwines with socio-economic variables and policy choices. The spatial insights provided by these models contribute significantly to geographical science by offering an innovative avenue to not only observe but also to actively engage with and comprehend spatial processes within a controlled computational simulation, thereby enhancing our understanding of spatially detailed systems and informing geographically nuanced policy formulation.

In the process of verification for this research, I have executed an ensemble of simulations for the same experiment to ensure outliers did not exist and to show that the distribution of runs corresponded to the stylised facts about the real world. This multi-run approach helps in identifying any discrepancies between the model's design and its operational counterpart. Furthermore, in the instance where a model was replicated, statistical tests such as descriptive statistics and Pearson's correlations were conducted to ensure that there were correlations between the same parameters for both models and statistical figures such as Q-Q plots supplemented findings. Additionally, I have conducted hypothesis tests such as one-way ANOVA, Cohen's D, and t-tests to ensure the distribution means were different across the various experiments. These steps were taken to demonstrate that the models behaved as intended when parameters were altered, such as changing the target rewards from 0 to 1 to a uniform distribution between [0, 1] inclusive (Badham *et al.*, 2018; Cocea *et al.*, 2012).

To summarise, through the application of several domain-specific models, this thesis will investigate the utility of RL over pre-existing decision-making frameworks through

the following:

- Learning - agents should be able to learn from individual perceptions and use past knowledge to support future decisions (empirical example: where offenders learn that heightened security measures increase their chances of being apprehended) (Lorscheid, 2014).
- Adapting - agents should be able to adapt to stochastic environment configurations and continue to perform sub-optimally (empirical example: when a person enters a new elevator, they adapt to the button configuration) (Maqbool *et al.*, 2011; Ramchandani *et al.*, 2017).
- Achieving Goals - agents should be able to achieve long or short-term goals through the heterogeneous decisions they make (empirical example: the goal is to reduce COVID-19 infection rates by 40% through policy changes) (Prystawski *et al.*, 2021).

1.2 Research aim and objectives

This research aims to explore how reinforcement learning algorithms can be incorporated into agent-based models to generate complex agent behaviours across several domains. To fulfil this aim, several objectives have been established:

1. Review the existing literature describing methodological applications of behavioural frameworks in ABMs to simulate emergent complex behaviours.
2. Develop a proof-of-concept ABM using reinforcement learning where complex behaviours organically learnt by agents are quantified and analysed, assessing the extent by which emergent complex behaviours are observed.
3. Assess whether reinforcement learning can be used to investigate theoretical perspectives in social science using criminology as an example.
4. Assess the capabilities of reinforcement learning agents to emulate the behaviours of policymakers who are tasked to make informed decisions about the dynamics of a complex system.

1.3 Thesis structure

This thesis is presented in the traditional format (more information can be found here) as described by the University of Leeds. The six chapters of the thesis are outlined below, and Table 1.1 highlights the structure of the thesis in relation to the research objectives.

Objective	**Chapter**
1. Review the existing literature describing theoretical and methodological applications of behavioural frameworks in ABMs to simulate emergent complex behaviours.	Chapter 2
2. Develop a proof-of-concept ABM using reinforcement learning where complex behaviours organically learnt by agents are quantified and analysed, assessing the extent by which emergent complex behaviours are observed.	Chapter 3
3. Assess whether reinforcement learning can be used to investigate theoretical perspectives in social science using criminology as an example.	Chapter 4
4. Assess the capabilities of reinforcement learning agents to emulate the behaviours of policymakers who are tasked to make informed decisions about the dynamics of a complex system.	Chapter 5

Table 1.1: The thesis structure in relation to the research objectives.

Chapter 2 delivers a critical review of literature relevant to emergent complex behaviours and the methodological applications of common decision-making frameworks used to simulate these behaviours within ABMs. This chapter will define essential terms such as 'emergent complex behaviours', emphasising studies employing these terms. Then presents a discussion of the adopted decision-making frameworks' strengths and weaknesses, stressing gaps that RL can overcome. The chapter then explores studies within environmental criminology and housing markets that have utilised ABM to assess domain-specific behavioural gaps RL can fill. The work in this chapter emphasises the need to develop ABMs that better represent the decision-making processes of entities modelled as agents, which subsequently aims to produce models that better reflect the real world.

Chapter 3 aims to test an observational example of RL to assess its ability to demonstrate adaptability, intelligence and emergent behaviour as a foundation for the following two chapters. The work begins with a critical discussion on behavioural simulation using RL and how it can contribute to agent-based modelling research. It then describes the anatomy of the predator-prey ABM from its conceptual design to the parameters utilised for RL training. Lastly, the work quantifies and visualises experimental results demonstrating how agents learn to adapt their behaviours to unbeknown situations, closing with a discussion and conclusion of the learnt behaviours and the overarching ramifications of the methodology for broader ABM research.

Chapter 4 develops an ABM of burglary events using RL for offender behaviours. The model aims to demonstrate how RL agents can learn behaviours that agree with empirical and theoretical assertions of environmental criminology literature. The work starts with a critique of ABMs utilised within environmental criminology, emphasising the behavioural frameworks used. It then describes the conceptualisations of predominant theories from environmental criminology adopted in the ABM. Several experiments are designed whereby spatial interventions are introduced run-time, and the adaptive qualities of offender agents are assessed. Lastly, the results are visualised, demonstrating crime patterns in agreement with empirical and theoretical literature in environmental criminology.

In the preceding chapters, RL has been instrumental in enabling agents within ABMs to exhibit behaviours that align more closely with empirical observations of decision-making processes. Chapter 5 builds on the findings from the previous two chapters and explores if RL can be used to train an agent that learns to respond to patterns observed in a complex system such as the housing market to produce a specific outcome. To that end, a well-known housing market ABM of the UK is replicated in the Python programming stack. A central-bank RL agent is developed to learn trends from the model and adapt the market to potential exogenous shock events. The work explores ABMs that simulate housing market dynamics, critiquing the decision-making frameworks utilised. It then defines various economic shocks and alleviating measures adopted by central bank intervention policies. The anatomy of the housing market model is described, and the replicated model is compared to the original to ensure validity. Several shock events are simulated, and the decisions made by the central bank agent are quantified at various spatial-temporal resolutions. The study finds that

RL can learn global patterns of the housing market and, by utilising goal criteria, adjust a housing market dynamically to overcome exogenous shocks. The work ends by discussing results showing each step of the learning phase of the RL agent and the subsequent implications on the housing market.

The thesis is concluded in Chapter 6. The chapter begins by outlining the novelty of the thesis. The chapter then proceeds to provide a summary of the thesis and demonstrates the extent to which the research aim and objectives have been met. The limitations of the research are noted and recommendations for future work are made. An outlook on producing and utilising ABMs using RL as a decision-making framework and concluding remarks are also documented.

Chapter 2

Understanding Emergent Complex Behaviours

2.1 Introduction

This thesis aims to explore the emergence of complex behaviours across several distinct agent-based models (ABMs) using reinforcement learning (RL) algorithms. To fulfil this aim, a thorough understanding of emergent complex behaviours is required. Section 2.2 defines the term 'emergent complex behaviours' highlighting early studies on the topic. Section 2.3 provides a comprehensive description of RL and deep-RL with their relative strengths and weaknesses. Section 2.4 explores the applications of complex adaptive behaviours. Section 2.5 follows with a critical discussion of computational simulations of emergent complex behaviours which may be of value during the development life cycle of the models presented in this thesis. Section 2.6 compares the chosen method and packages Unity and ml-agents against other approaches. Section 2.7 builds on this by discussing the core strengths and weaknesses of computational decision-making frameworks used to simulate emergent complex behaviours. Section 2.8 presents studies within environmental criminology and the housing market domain that have utilised ABMs to simulate emergent complex behaviours.

2.1.1 Definitions and descriptions of key terms

Complex behaviours

Complex behaviours emerge from interactions between organisms and their environments. These behaviours are patterns observed at the aggregate level, such as social structures or ecosystem dynamics, that are not explicitly programmed into the individuals but arise from collective interactions. They encompass both instinctual and learned responses to environmental stimuli.

Adaptive behaviour

Adaptive behaviour refers to how agents meet their personal goals while responding to environmental and social challenges. This concept is central to modelling communities and understanding how agents adjust their actions to maximise their success within a given context, such as finding food or avoiding predators.

Learning

Learning is the process by which organisms acquire new behaviours or modify existing ones in response to changes in their environment. It involves a cognitive process where information is processed and utilised for making decisions that enhance an organism's ability to thrive.

Emergent phenomena

Emergent phenomena are patterns or properties that arise from the collective interactions of simpler elements of a system. These are not properties of the individual elements themselves but result from the interplay between them.

Emergence and real emergence

Emergence is the process through which complex behaviours arise from simple rules or interactions. Real emergence might refer to the manifestation of these behaviours in tangible, observable ways in the real world, as opposed to simulated environments.

Emerging interactions

Emerging interactions are the unexpected and novel behaviours or patterns that come about from the interaction of agents within a system. These interactions can lead to the development of new properties or behaviours at the system level that are not predictable from the properties of the individual agents.

Complex vs complicated behaviours

Complex behaviours are dynamic, often unpredictable, and arise from non-linear interactions within a system. Complicated behaviours, while possibly intricate, are predictable and result from linear cause-and-effect relationships.

2.2 Emergent complex behaviours, from theory to practice

The term 'complex behaviours' is vast within literature and multi-faceted across research areas. Complex behaviours, in their purest form, emerge from interactions between organisms and their environments at the individual level. Usually, these interactions lead to patterns at the aggregate level (i.e., from individual ants to an ant colony) that are emergent and are not hardcoded in the individual (Hemelrijk, 2013). The emergence of these behaviours is partly genetic make-up (such as reproducing) but mostly a process of learning information and cognitively processing responses to situations. For example, learning among mammals starts from an early stage. A critical behaviour that emerges is survival as a response to hostile situations (Schakner & Blumstein, 2016). According to Percy (1989), the process of learning among humans, which ultimately leads to behaviours, involves studying, instruction, practice and experience. To learn these complex behaviours and navigate complex environments, it is agreed among researchers that some stimulus is triggered within the brain to aid the decision-making process. Dopamine is considered the stimulus which affects the basal ganglia (a part of the brain that deals with motor responses) (Lockwood & Klein-Flügge, 2021; Niv, 2009; Sutton & Barto, 2018b; Walsh & Anderson, 2014). Cochet & Byrne (2015) suggests that motor precision, coordination and anticipatory planning all lead to behavioural complexity. The aforementioned demonstrates that the emergence

of complex behaviours is a complex process with various physiological and neurological facets making it challenging to computationally replicate (Kennedy, 2012).

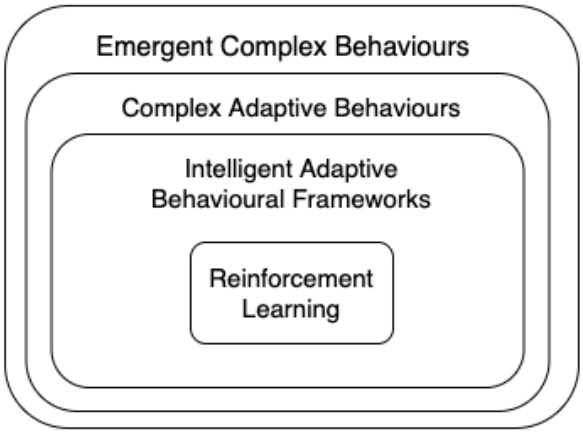


Figure 2.1: A set diagram depicting the relationship between key terms.

2.2.1 Adaptive nature of complex behaviours in modelling

Adaptive behaviours—how agents meet personal goals and navigate environmental demands—are pivotal in modelling communities (Oakland & Harrison, 2008). This includes both natural settings and social constructs, where the modelling of such behaviours has led to insights into the formation of higher-quality communities (Sun & Wang, 2008) and enriched knowledge bases (Cocea *et al.*, 2012). This understanding of community dynamics sets the stage for addressing the broader challenges of system management.

2.2.2 The challenge of managing complex systems

Despite advancements, managing complex systems remains a significant challenge. The introduction of tools like the resource/agent map by Kazakov *et al.* (2021) helps in appreciating these systems' multifaceted nature. Moreover, such complexities prompt the development of new theories, as novel modelling techniques shed light on previously unexplored behaviours within these systems (Baumann, 2015). As we recognise

these modelling advancements, we must also consider their implications for theoretical development within the field of complex adaptive systems.

2.2.3 Complex adaptive behaviours and theoretical development

Through complex adaptive behaviours, ABMs hold the potential to investigate new theories by simulating a wide array of behaviours and understanding the brain's information processing steps (Joyce *et al.*, 2012). Supply chain management serves as a testament to the efficacy of these models, where autonomous agents have led to policies aligning with empirical market observations (Brintrup, 2010). The success of these applications brings to light the core strengths of ABMs, especially in terms of their ability to capture decision-making processes in fluctuating environments.

2.2.4 Understanding decision-making in fluctuating environments

One profound strength of modelling complex adaptive behaviours is the facilitation of understanding organism's decision-making in variable environments, crucial for both ecological studies and broader applications (Houston *et al.*, 1999). This understanding is further clarified when we examine the evolution of key terms and concepts over time, as depicted in the subsequent visual representation.

2.2.5 Key terms and their evolution

The evolution of key terms from early theoretical foundations to practical computational applications can be observed in Figure 2.1, providing clarity on the trajectory from "emergent complex behaviours" to their embodiment in RL algorithms. In the following paragraph, we explore weaknesses or drawbacks of complex adaptive behavioural research in context of this thesis.

2.2.6 Predictive challenges and verification in complex adaptive systems

The inherent unpredictability of complex adaptive systems can sometimes lead to undesirable outcomes, a limitation acknowledged by recent research efforts aiming to develop model-free formal verification approaches (Hachicha *et al.*, 2017; Janssen *et al.*, 2015).

This unpredictability is less problematic in simpler systems, where environmental constraints can lead to more deterministic learning outcomes (Olsen & Fraczkowski, 2015; Yamada *et al.*, 2011; Zhdankin & Sprott, 2010). Such approaches highlight the crucial importance of data fidelity in computational models aiming to simulate emergent complex behaviours.

2.2.7 Data-driven modelling and validation

Data-driven approaches are vital in ensuring that simulated environments reflect real-world observations, with many researchers advocating for robust validation and calibration processes (Capasso *et al.*, 2020; Troitzsch, 2017; Windrum *et al.*, 2007). This ensures that computational models are not only theoretically sound but also empirically grounded, enhancing their applicability to real-world problems. The domain of vehicle behaviour amplifies the challenges faced in modelling individualistic and context-specific actions, such as driver responses at traffic lights.

2.2.8 Modelling vehicle driver behaviour

Modelling driver behaviour has exposed limitations in capturing the nuances of how individuals react in varied driving conditions, suggesting that some aspects of complex adaptive behaviour may elude current modelling techniques (Toledo, 2007). These models often struggle to incorporate every detail accurately, highlighting the indirect relationship between the properties of interacting elements and emergent system results (Funes *et al.*, 2003; Nolfi, 2004). The difficulty in designing these behavioural systems underscores the necessity for formal verification tools to ensure model integrity during the design phase.

2.2.9 The need for formal verification tools and data-driven design

The challenges in designing behavioural systems emphasise the importance of formal verification tools that can aid in the design of models intended to simulate complex adaptive behaviours (Funes *et al.*, 2003; Nolfi, 2004). Ensuring these models are data-driven could help ground them in empirical reality, thereby increasing their utility and relevance. Still, the complexity science community continues to grapple with how to measure and manage the complexity of problems, seeking to develop scales for complexity that might streamline the process of model selection and hardware utilisation.

2.2.10 Measuring complexity and framework selection

The quest for a scale to measure complexity aims to differentiate levels of complexity and align them with appropriate computational resources (Kitto, 2006). The absence of a universally accepted framework adds a layer of complexity to the selection of the most suitable modelling approach for a given problem (Hawryszkiewycz, 2009; Kazakov *et al.*, 2021). These challenges, however, are set against a backdrop of rapid advancement in computational modelling and an ongoing debate about the feasibility of simulating real-world emergence within complex adaptive systems.

2.2.11 Discussing the possibility of simulating emergence

While some scholars have expressed scepticism regarding the ability to simulate genuine emergence due to computational limitations, this thesis hypothesises a more optimistic view, aiming to demonstrate emergence through complex adaptive behavioural frameworks across three unique domains (Tolk, 2019). The methodology to be explored next includes a variety of algorithmic solutions that cater to the behavioural decision-making of agents, ultimately leading to the emergence of complex behaviours.

2.2.12 Algorithmic solutions to behavioural decision-making

Algorithmic approaches, such as genetic algorithms and reinforcement learning, have been instrumental in exploring learning and behaviour across various fields since the late '90s (Back, 1996; Lockwood & Klein-Flügge, 2021; Sutton & Barto, 2018b; Yu *et al.*, 2008). These algorithms have been particularly effective in domains such as robotics, optimising control parameters to navigate efficiently through environments (Beer & Gallagher, 1992; da Silva Assis *et al.*, 2016; Zhu & Zhang, 2021).

2.3 Comprehensive understanding of reinforcement learning

2.3.1 Introduction to reinforcement learning

Reinforcement Learning (RL) is a computational approach in which an agent learns to make decisions by performing actions and receiving feedback in the form of rewards

or penalties. This learning paradigm is structured around the agent-environment interaction, where the agent aims to maximise a cumulative reward over time. The fundamental components of an RL system include the environment, states, actions, policy, reward signal, and value function (Sutton & Barto, 2018b).

In 1944, a book titled "The Theory of Games and Economic Behaviour" by John von Neumann and Oskar Morgenstern was published by Princeton University Press. This book introduced key concepts of maximising expected utility as a decision-making framework. This work laid the foundation for developing RL algorithms (Von Neumann & Morgenstern, 1944). In 1959, Samuel (1959) presented the first machine learning algorithm that was able to learn from experience, paving the way for the development of RL algorithms. In 1982, Thompson (1982) wrote the paper "The Single-Sample Problem of Detection, Learning and Prediction", which introduced the concept of temporal difference (TD) learning, which forms the basis of many modern RL algorithms. The single-sample problem refers to the challenge of learning optimal policies in an environment based on a single sample of data. This problem usually arises when an RL agent has to make decisions based on incomplete or noisy information about the environment, making it challenging to learn an optimal policy. The temporal difference approach was the cure to this problem, which used a single sample of data to update the estimates of expected values of states and actions.

2.3.2 Deep reinforcement learning techniques

Deep Reinforcement Learning (Deep RL) combines neural networks with RL, enabling agents to process complex, high-dimensional inputs and learn directly from raw sensory data (Lecun *et al.*, 2015). Below are some important Deep RL techniques:

- Deep Q-Networks (DQN): Extends traditional Q-learning by using deep neural networks to approximate the Q-value function, which evaluates the quality of particular actions in given states (Fan *et al.*, 2020).
- Proximal Policy Optimization (PPO): A policy gradient method that aims to take the largest possible improvement step on a policy without causing performance collapse, ensuring stable and reliable training (Schulman *et al.*, 2017).
- Actor-Critic Methods (A2C/A3C): These methods use two neural networks: the actor, which decides the action to take, and the critic, which evaluates the action.

A2C (Advantage Actor-Critic) improves stability by considering the advantage function, whereas A3C (Asynchronous Advantage Actor-Critic) introduces parallel training to further stabilise and speed up learning (Lewis *et al.*, 2020; Wang *et al.*, 2016).

- Trust Region Policy Optimisation (TRPO): A more advanced policy gradient method that ensures the new policy is not too far from the old, maintaining stability in the learning updates (Schulman *et al.*, 2015).
- Soft Actor-Critic (SAC): An off-policy algorithm that seeks to maximise both the expected return and entropy (entropy is a concept borrowed from information theory, which measures the randomness or unpredictability in the agent's policy), encouraging the policy to explore more widely (Lewis *et al.*, 2020; Wang *et al.*, 2016).

2.3.3 Reinforcement learning training and neural networks

In RL training, the agent is exposed to the environment and learns through episodes of interaction. Neural networks serve as function approximators in this process. For example, in value-based methods like DQN, a deep neural network can approximate the Q-value function, while in policy-based methods, it can represent the policy itself.

2.3.4 Neural network architecture in reinforcement learning

The design of the neural network in RL is critical and typically involves several layers:

- Input Layer: Represents the state of the environment.
- Hidden Layers: May vary in number and size, designed to capture the complexities of the environment and the required policy representation.
- Output Layer: Provides the action selections or value predictions.

2.3.5 Critique of reinforcement learning

Despite its successes, Deep RL and RL in general are critiqued for their sample inefficiency, opaqueness, and overfitting to specific environments. Moreover, many Deep RL algorithms require significant tuning and are sensitive to hyperparameter changes,

which can impede their practical deployment (Asadi, 2015; Choshen *et al.*, 2019). Below we describe in detail the specific drawbacks of RL:

- Sample inefficiency: Deep RL algorithms often require a large number of samples to learn effectively. This is because they need to experience a wide variety of states and outcomes to formulate a reliable policy. In real-world applications, where each sample can represent a costly or time-consuming interaction with the environment, this inefficiency becomes a critical drawback. The trial-and-error nature of RL can lead to slow convergence rates, particularly in environments with sparse or deceptive rewards.

- Opaqueness and lack of interpretability: Deep RL models, especially those involving deep neural networks, are often criticised for being black boxes. The complex, non-linear interactions within deep networks make it difficult to interpret how the system is making decisions. This opaqueness becomes a significant issue in domains where explainability is crucial, such as healthcare or finance.

- Overfitting and generalisation: Deep RL agents are prone to overfitting their policies to the specific nuances of the training environment. This can result in a lack of generalisation to new environments, even when they share similarities with the training scenario. The models might fail to perform well when faced with variations not encountered during training, which is a significant hurdle for deploying RL in dynamic, real-world settings.

- Hyperparameter sensitivity: The performance of Deep RL algorithms is often highly sensitive to the choice of hyperparameters. Finding the right set of hyperparameters can be a resource-intensive process involving a lot of trial and error. This sensitivity also means that a successfully trained model in one environment might perform poorly in another without significant re-tuning.

- Credit assignment problem: In many RL scenarios, rewards are sparse or delayed, making it challenging for the agent to determine which actions are responsible for obtaining the reward. This delayed reward structure can significantly complicate the learning process, as the agent struggles to associate the correct behaviours with the outcomes observed.

- Exploration vs exploitation: Effective RL requires a balance between exploring the environment to find new strategies and exploiting known strategies to gain rewards. Striking this balance is non-trivial and often requires complex heuristics or additional mechanisms, like curiosity-driven exploration, which adds to the model's complexity and computational requirements.

- Dependency on reward function: The quality of the learned policy is heavily dependent on the design of the reward function. If the rewards do not accurately reflect the desired outcome, the agent may learn non-realistic or even harmful behaviours. Designing a reward function that encapsulates the objectives correctly is a challenging task.

- Scalability and computation Requirements: Deep RL requires significant computational power, particularly for problems with high-dimensional state and action spaces. Training can take an impractically long time or may not be feasible with limited computational resources.

- Robustness and safety: Ensuring the robustness and safety of RL agents in operation is a significant concern. Agents may learn to exploit bugs or unintended features in the simulation, leading to unsafe behaviours when deployed in the real world.

- Environmental dynamics: RL agents are often trained in simulated environments, which may not capture all the dynamics of the real-world setting they are intended for. This discrepancy can lead to policies that perform well in simulation but fail when transferred to the real world.

Reinforcement learning is one approach to developing complex adaptive behaviours, in the following section, we compare RL against other approaches to simulating complex adaptive behaviours.

2.4 Comparing complex adaptive decision-making algorithms

RL is a type of machine learning that involves learning through trial and error, in which an agent learns to take actions in an environment to maximise a reward signal. RL

algorithms are typically used to solve optimisation problems in which the agent has to make a sequence of decisions over time, such as deciding which actions to take in a game or which controls to use to operate a machine. RL algorithms learn by interacting with their environment and receiving feedback as a reward or penalty (Lapan, 2018; Sutton & Barto, 2018b). Similarly, GAs are a type of optimisation algorithm that is inspired by the process of natural evolution. GAs represent solutions to a problem as a set of parameters, or "genes," and iteratively improve these solutions through selection and reproduction. GAs are typically used to solve optimisation problems that involve finding the best set of parameters to achieve a given objective, such as finding the optimal shape of an aeroplane wing or the optimal combination of ingredients in a recipe (De Jong, 2012; Katoch *et al.*, 2021).

One key difference between RL and GAs is how they represent solutions to problems. RL algorithms represent solutions as a set of actions or policies that an agent can take to maximise a reward signal. An RL solution involves mapping the states of the environment to actions the agent should take in those states. This mapping is often represented as a function, called a policy, that takes the current state of the environment as input and outputs an action to take (Buşoniu *et al.*, 2010; Zhan-quan, 2006). GAs, on the other hand, represent solutions as a set of parameters or genes (Katoch *et al.*, 2021). Another difference is that RL algorithms learn through interaction with their environment, while GAs learns through selection and reproduction. Lastly, RL is generally more suitable for problems involving sequential decision-making, complex environments and continuous action spaces. GAs, on the other hand, are better suited for population-based search, and combinatorial problems (Goldberg, 1989; Sutton & Barto, 2018b).

Both RL and GAs have their respective strengths and weaknesses, particularly when it comes to computational tractability. For example, RL allows agents to learn complex behaviours through direct interaction with their environment at an individual level, without the need for explicit supervision or labelled training data (Kaelbling *et al.*, 1996). On the other hand, GAs are capable at searching vast solution spaces, which is advantageous for optimisation problems featuring numerous variables or complex constraints (De Jong, 2012). One notable challenge with RL is the difficulty in precisely defining the reward function, which guides the agent on the actions to take or behaviours to learn to maximise its reward (Buşoniu *et al.*, 2010; Lapan, 2018; Sutton & Barto,

2018b). As the complexity within the modelled domain escalates, the likelihood that the reward function accurately encapsulates desired behaviours diminishes, potentially leading to sub-optimal or incorrect behaviours. To mitigate this, researchers should formulate objective functions aligned with theoretical constructs and empirical data. Conversely, GAs' performance is contingent on hyperparameters such as population size and mutation rate, influencing the optimisation process's speed and stability (Back, 1996; Katoch *et al.*, 2021). GAs also tend to be less efficient in smaller search spaces but excel when dealing with long-term goals or objectives (De Jong, 2012).

With RL, achieving "good results" encompasses a variety of benchmarks, such as maximising cumulative rewards, completing tasks, converging on stable and reliable policies, demonstrating generalisation across different scenarios, and meeting established domain-specific benchmarks. Thus, RL requires numerous trials and errors to attain such results, often demanding significant computational resources and time. This iterative process of learning and improvement is essential in reaching the desired level of performance that meets these criteria for success.

RL and GAs have been applied to tasks in natural language processing, such as language generation and machine translation. For example, in the article (Luong *et al.*, 2015), RL was used to optimise machine translation models. In contrast, in the article (Rios & Springer, 2017), GAs were used to optimise the architecture of recurrent neural networks for language generation. RL and GAs have been applied to game-playing tasks, such as Chess and Go. For example, in the article (Silver *et al.*, 2017), RL was used to develop a Go-playing agent that achieved superhuman performance. Conversely, in the article (Tong *et al.*, 2011), GAs were used to evolve neural network controllers for real-time strategy games.

A question that might need unpicking at this stage is why use RL over other approaches such as GAs? The main reason is the ability of RL algorithms to adapt an agent's behaviours to changing environments and learn new behaviours. Furthermore, RL allows agents to discover various behaviours, including decision-making, learning and problem-solving (Sutton & Barto, 2018b). In contrast, GAs can model individual behaviours by evolving a population of individuals with different behaviours and selecting individuals with the most desirable behaviours for reproduction. However, this makes the approach computationally expensive and may not be as effective at modelling more complex behaviours (Back, 1996).

In the following sub-section, an analysis of the limits and implications of agent-based models that simulate emergent complex behaviours using RL is undertaken.

2.5 Exploring the limits and implications of computational simulations of emergent complex behaviours

ABMs that utilise RL have been applied in various research domains to simulate complex behaviours and study multiple social, economic, and environmental systems. Some notable examples include traffic control system optimisation (Chen *et al.*, 2017b; Wang *et al.*, 2019a), group behaviour such as group hunting, foraging and decision-making (Daw *et al.*, 2006; Hwang *et al.*, 2011), optimisation of environmental management systems, such as water resource management and wildlife conservation (Leibo *et al.*, 2017b; Tang *et al.*, 2017; Wang *et al.*, 2019b), investigate and understand social dilemmas such as prisoner's dilemma (Koster & De Jong, 2009; Leibo *et al.*, 2017a)

There are numerous examples of ABMs that utilise RL to model complex behaviours, where these methods have demonstrated promising results, particularly in domains yet in their infancy but are promising research avenues, such as in the modelling of autonomous vehicle behaviour (Bellemare *et al.*, 2013; Cai *et al.*, 2019; Koutn'k *et al.*, 2013; Wang *et al.*, 2018). As evidenced in this thesis, RL has been widely used in developing complex decision-making agents and applied to various problem domains. Many academic articles have explored the use of RL in ABMs, and have demonstrated its effectiveness in solving complex decision-making tasks. For example, Wang & de Silva (2008) used RL to develop a multi-agent system for coordinating the behaviour of a group of robots. They found that the RL-based approaches outperformed other decision-making algorithms, such as evolutionary algorithms, in terms of both convergence speed and final performance. A similar claim was made earlier in sub-section 2.2. Conversely, many academic articles do not use RL in their ABMs. These articles utilise other decision-making algorithms, including but not limited to decision trees, Markov decision processes (MDPs) or condition-action rules, which can be effective in certain situations.

It is essential to define the limitations of employing RL as a decision-making framework within ABMs. This investigation must extend across different disciplines to pinpoint issues that might arise within the scope of this thesis. A recurring theme in the

literature is the application of RL to social dilemmas. For instance, Leibo *et al.* (2017a) observed that while agents could eventually co-operate through RL, the learning process was notably demanded extensive interactions. A significant limitation identified was the absence of inter-agent communication, which reduced the range of complex behaviours that could be represented. Similarly, Lowe *et al.* (2017) and Foerster *et al.* (2017) reported that RL could induce sub-optimal behaviours, especially in environments where agent goals clashed or where the environment was excessively unpredictable. The lack of communication capabilities among agents was also noted here as a constraint that restricted the depth of simulated behaviours.

This limitation, however, may depend on the specific applications being examined. With comprehensive planning and a detailed conceptualisation phase, the variability of agent behaviours can be more effectively managed. There are instances where RL has yielded encouraging outcomes, particularly when agents have communication abilities. For example, the study by Foerster *et al.* (2016) represented a considerable advancement in the field of RL and multi-agent systems by demonstrating that deep multi-agent reinforcement learning (MARL) can facilitate emergent communication among agents. This allowed agents to autonomously tackle complex cooperative tasks that would be difficult without communication. Similarly, Su *et al.* (2017) identified that communication could lead to enhanced coordination amongst agents, thereby adjusting their collective performance on collaborative endeavours. Furthermore, Wang *et al.* (2020) found that inter-agent communication in a traffic domain resulted in improved navigational awareness and collision avoidance. These studies demonstrate that communicative agents can achieve coordination, problem-solving, and operational efficiency, which are recognised as "encouraging outcomes" within the RL literature.

A further limitation is encountered when considering the integration of social and cultural factors within the model, which hold pivotal importance in domains such as social networking and policy-making. The following study Zhang *et al.* (2019), observed that RL can facilitate agents' evolution in decision-making, allowing them to respond more appropriately to dynamic changes in the environment over time. However, the lack of social and cultural factors in the model could potentially constrain their findings to a subset of domains. To mitigate this limitation, it is vital to incorporate the RL objective function with a robust theoretical underpinning and empirical data. Such an approach could ensure that decisions made by the agents not only reflect optimal

strategies within the simulated environment but also echo the complexity of societal norms and cultural dynamics, thereby enhancing the validity and applicability of the RL approach. This denotes accurate decisions as those which are not only optimal within the defined parameters of the RL environment but also coincide with real-world social and cultural behavioural patterns.

In the field of autonomous driving, Bojarski *et al.* (2016) identified that RL could result in less-than-ideal behaviours in highly dynamic settings or when agents had limited environmental information. Furthermore, their model did not account for the actions of other drivers, which could diminish the realism of the simulated driving behaviour.

Now that the limitations of RL applied to ABMs as decision-making frameworks have been described, it is worth exploring the strengths of applied examples that have enabled researchers to test new theories and make discoveries. An area of research that RL has been widely adopted over other decision-making frameworks is cooperative and competitive problem-solving. Lowes *et al.* (2017) presented a multi-agent deep RL approach for complex cooperative and competitive environments. They demonstrate that deep RL approaches can effectively learn complex behaviours in various settings, including resource gathering, predator-prey, and traffic control. They also show that their approach can learn to adapt to changing environments and exhibit emergent behaviours that are not explicitly programmed. Similar findings were also documented in (Mataric, 2002; Singh *et al.*, 2004). Every article read and analysed for this review makes similar points regarding the limitations and strengths of RL applied to ABMs. Based on the literature reviewed thus far, the strengths of RL can be summarized as follows (Buşoniu *et al.*, 2010; Lapan, 2018; Lowes *et al.*, 2017; Sutton & Barto, 2018b):

- RL can learn complex behaviours that vary across environments.
- RL can adapt an agent's behaviours to deal with new environments and challenges.
- RL does not require explicit behaviours hard-coded.
- RL agents can learn how to achieve long-term goals.
- RL agents can learn from one another by interacting and adjusting their behaviours accordingly.

The limitations can also be summarised as follows (Kober & Peters, 2013; Lewis *et al.*, 2020; Sutton & Barto, 2018b; Taylor *et al.*, 2009):

- RL algorithms often require many interactions with the environment to learn optimal behaviours.
- RL algorithms are computationally expensive to run.
- RL algorithms can have difficulty determining which actions are responsible for specific outcomes, especially when there are long delays between the actions and outcomes. This is also known as the "credit assignment problem".
- RL algorithms may not be able to generalise their behaviour, making them less suitable in more abstract settings.
- RL algorithms perform sub-optimally if the model stochasticity or complexity is great.

Overall, both RL and non-RL approaches have their relative strengths and weaknesses, and the choice of the algorithm may depend on the specific problem at hand. In the following sub-section, the commonly used decision-making frameworks in ABMs are described with their relative strengths and weaknesses compared to the RL approach.

2.6 Unity and ml-agents for Reinforcement Learning Applications

The selection of Unity in conjunction with ml-agents for the development of reinforcement learning (RL) based agent-based models (ABMs) in this thesis is predicated on several strengths that this pairing provides over traditional RL libraries such as Keras (Ketkar & Ketkar, 2017), TensorFlow Agents (Developers, 2023), or PyTorch's (Paszke *et al.*, 2019) RL implementations. Unity offers a highly detailed development environment with an emphasis on 3D simulation, which is useful given the spatial and visual demands of ABMs, particularly when complex interactions and environmental nuances are essential for accurate model representation.

Unity's strength lies in its robust graphical engine, which allows for the rendering of agent environments, offering an intuitive visual depiction that is important for the

observation and analysis of emergent complex behaviours in ABMs. This level of visual and environmental details is not as readily achievable with other RL libraries that typically do not possess such advanced graphical capabilities.

Integrating Unity with ml-agents, which is a project by Unity Technologies dedicated to applying machine learning to Unity, enhances the platform's suitability for computational social science research. ml-agents provides a framework that allows agents within Unity to learn using deep RL. This facilitates the modelling of a wide range of behaviours, from simple locomotion to complex decision-making processes, which are fundamental for the simulation of social, economic, and ecological systems to name but a few. The integration of ml-agents into Unity transforms the game engine into a powerful research tool for RL applications, enabling the simulation of rich, multi-agent ecosystems with dynamic environments. Furthermore, the ml-agents toolkit is explicitly designed to harness Unity's engine, simplifying the process of creating learning scenarios. The toolkit offers a suite of learning algorithms, including Proximal Policy Optimisation (PPO), which has been employed in this research. The use of PPO, a policy gradient method for RL, is advantageous due to its stability and robustness, especially in the context of environments with high-dimensional observation spaces, as often found in ABMs, this is also described further in Chapters 3, 4 and 5.

Compared to other RL libraries, ml-agents in Unity provides a user-friendly interface for defining agent behaviour and training configurations. This not only streamlines the model development process but also makes it more accessible to researchers who may not have extensive backgrounds in machine learning or computer programming.

In contrast, while libraries like Keras (Ketkar & Ketkar, 2017) and TensorFlow Agents (Developers, 2023) are powerful and flexible, they often require more elaborate setup and integration to produce a comparable level of environmental sophistication compared to ml-agents and Unity. Moreover, these libraries are more focused on the algorithmic aspects of RL rather than the comprehensive development of simulations. This makes Unity and ml-agents a more efficient choice for research that benefits from a visually rich and interactive modelling approach, aligning with the thesis's objective to introduce a novel modelling method to the computational social science community.

In conclusion, the functionality of Unity and ml-agents is not just a convenience of compatibility, it represents a strategic choice to leverage a platform suited to the demands of spatially and visually driven ABMs. This compatibility, coupled with the

advanced learning capabilities provided by ml-agents, justifies their selection over other RL libraries for the research presented in this thesis.

2.7 Comparative investigation of traditional decision-making frameworks in agent-based models

ABMs consist of several interlinked components, such as the agent and environment. However, one of the critical elements of ABMs is the decision-making framework that guides the agents' actions. One such framework is the beliefs, desires, and intentions (BDI) model (Rao & Georgeff, 1995), which proposes that agents make decisions based on their beliefs about the world, their desires or goals, and their intentions or plans to achieve those goals. The BDI model has been widely used in ABMs and applied to various domains, including organisational behaviour, artificial intelligence, and game theory (Wooldridge, 2002).

However, as depicted earlier in the review, the BDI model is not the only decision-making framework used in ABMs. Many other frameworks have been proposed and applied in this field, including the rational choice model, the behavioural economics model (Kahneman & Tversky, 2013; Tesfatsion, 2006), and the evolutionary game theory (Yu *et al.*, 2008) model. In this research, we aim to investigate and compare traditional decision-making frameworks used in ABMs, focusing on the BDI model and other prominent frameworks. We will review the literature on these frameworks and explore their strengths and limitations.

A broad search across multiple disciplines was undertaken using academic databases such as the Web of Science to identify literature investigating these decision-making frameworks. The search criteria: "(TOPIC) agent based model" AND "(TOPIC) decision making framework" returned over 1,971 results. Most of these results were articles that did not describe a decision-making framework per se. However, they described the entire agent-based model as a framework to achieve a particular task such as enterprise management (Miao & Xu, 2007), or the article described a review of agent-based models in a particular research area such as land use models (Matthews *et al.*, 2007). If instead, we search these terms in the title rather than the topic category, we only receive one result. One reason for the difficulty of identifying papers that propose approaches to modelling decision-making in agent-based models could be that the multi-disciplinary

use of agent-based models makes it challenging to search specific components of models. Therefore, the widely used and accepted decision-making frameworks are identified through articles that have surveyed these approaches, such as (Balke & Gilbert, 2014; DeAngelis & Diaz, 2019; Francès *et al.*, 2015).

In Tables 2.1 to 2.5 the commonly used decision-making frameworks across various disciplines that utilise ABM have been reviewed with references to some articles that describe these approaches. These decision-making frameworks are split into categories as described by (Balke & Gilbert, 2014), these are Normative, Cognitive and Psychologically and Neurologically inspired frameworks.

Normative decision-making frameworks (Mahmoud *et al.*, 2012; von Neumann & Morgenstern, 1947) are based on the idea that agents should make choices that are in line with certain values or principles, such as maximising utility or minimising risk and effort. These frameworks often involve concepts from economics, such as expected utility theory (where agents weigh the potential outcomes of different options according to their probabilities and the utility of those outcomes) (Harrison, 1994; Mongin, 1998), to analyse and predict how agents will make decisions. Under this framework, an agent's decision is determined by the option that maximises their expected utility, which is calculated as the sum of the utilities of each possible outcome multiplied by the probability of that outcome occurring. The Schelling segregation model and the Prisoner's Dilemma fit within this category, as they both utilise rational choice theory to predict agent behaviour. The Schelling model, for instance, demonstrates how individual agents follow simple, utility-maximising rules that can lead to emergent patterns such as segregation, despite no explicit intention to segregate (Schelling, 1971). Similarly, the Prisoner's Dilemma explores the outcomes of utility-maximising decisions in a game-theoretic context (Axelrod, 1980). An example framework described in Table 2.2 is the Deliberative Normative framework, which sometimes utilises decision trees to analyse the trade-offs between different options. It is worth exploring the broad limitations of the normative approach. One example is that they may not accurately reflect the decision-making process of real agents, who may be influenced by various factors that are not accounted for in the models—for example, behaviours influenced by emotions and biases (Kahneman, 2003). Similarly, the oversimplification of decisions may not adequately capture the complexities of the real world. One common assumption is that the agents are perfectly rational or have complete information (Simon, 1955).

Framework	Description	Strengths	Weaknesses	Example	References
Production Rule System (PRS)	The production rule system (PRS) is a framework for implementing agent-based models (ABMs) that are based on the concept of production rules. Production rules are a type of artificial intelligence rule that specifies a condition and an action to be taken if that condition is met. In a PRS-based ABM, each agent is equipped with production rules that define its behaviour.	It allows for the specification of complex and flexible behaviour for the agents in the model. Production rules can encode a wide range of decision-making processes and behaviours, and the rules can be easily modified or added to as the model is developed.	It can be computationally intensive to evaluate the production rules for each agent at each time step of the simulation. This can be particularly challenging for models with large numbers of agents or models with long simulation runs.	Rule 1: If the agent's energy level is low and there is food available in the environment, then eat the food. Rule 2: If the agent is not hungry and there is a predator nearby, then flee from the predator.	(Durfee, 1993; Maes, 1987; Wooldridge & Jennings, 1995)
Beliefs, Desires and Intentions (BDI)	The BDI framework for implementing agent-based models (ABMs) is based on mental states or attitudes. According to this framework, agents make decisions based on their beliefs about the world, desires or goals, and intentions or plans to achieve those goals.	It provides a natural and intuitive way of representing agents' decision-making process. The BDI model is based on common-sense notions of human psychology and is relatively easy to understand and implement.	It may not accurately capture real-world agents' complex and often irrational decision-making processes. The BDI model assumes that agents are rational and have a clear set of goals and beliefs. However, human behaviour can be influenced by a wide range of factors, such as emotions, biases, and social influences.	Group of agents competing for limited resources in an economic system. Each agent in the model might have beliefs about the availability and value of different resources, desires to acquire certain resources, and intentions to take specific actions (such as negotiating with other agents or competing with them) to acquire the resources.	(der Hoek & Wooldridge, 2003; Wooldridge, 2002; Wooldridge & Jennings, 1995; Yen & Lu, 2008)
Extended Beliefs, Desires and Intentions (eBDI)	The extended beliefs, desires, and intentions (eBDI) framework are a variant of the beliefs, desires, and intentions (BDI) framework. Like the BDI framework, eBDI represents agents' decision-making process regarding their beliefs, desires, and intentions. However, eBDI adds additional constructs to the BDI model, such as emotions, social norms, and personality traits, to provide a more realistic and nuanced representation of agent behaviour.	It allows for the inclusion of a wide range of factors that can influence agents' decision-making processes. By incorporating emotions, social norms, and personality traits into the model, eBDI can better capture the complexity and diversity of human behaviour.	It can be more difficult to implement and analyze than the simpler BDI framework. The additional constructs of the eBDI framework can make the ABM more complex and require more computational resources to simulate.	A group of agents interacting in a social network. Each agent in the model might have beliefs about the attitudes and behaviours of other agents, desires to achieve certain goals, and intentions to take specific actions (such as forming relationships or competing with others) to achieve those goals. The agents' emotions, social norms, and personality traits might influence their decision-making processes and behaviours.	(Robins *et al.*, 2001; Son *et al.*, 2013; Yaich *et al.*, 2011).

Table 2.1: Agent-based decision-making frameworks (part one)

Framework	Description	Strengths	Weaknesses	Example	References
Beliefs, Desires, Obligations, and Intentions (BOID)	The BOID framework is a variant of BDI that includes the concept of obligations as an additional factor that can influence agents' decision-making process. The BOID framework is often used in multi-agent systems, where agents are required to adhere to specific rules or norms of behaviour to achieve their goals or maintain their relationships with other agents.	It allows for the representation of social norms and rules of behaviour in agent-based models (ABMs). The inclusion of obligations in the BOID ABM can help to capture the complex social dynamics that can arise in multi-agent systems and how agents might balance their own goals and desires with the expectations of others.	It can be challenging to define and model the concept of obligations precisely and rigorously. Different agents might have different interpretations of their obligations and prioritise them differently, making it challenging to predict their behaviour.	Simulation of a group of agents engaged in a negotiation process. Each agent in the model might have beliefs about the preferences and motivations of the other agents, desires to achieve specific outcomes from the negotiation, and obligations to adhere to specific rules or protocols. The agents' intentions to take specific actions (such as making offers or concessions) might be influenced by all these factors.	(Broersen *et al.*, 2001, 2002).
Beliefs, Rules, Intentions, Desires, Goals and Emotions (BRIDGE)	BRIDGE is a multi-layered agent-based modelling (ABM) framework that aims to represent agents' decision-making processes in a realistic and nuanced way. The BRIDGE framework is based on the belief-desire-intention (BDI) model of agency. However, it adds additional layers to the model to better capture human behaviour's complexity and diversity.	It allows for the representation of a wide range of factors that can influence agents' decision-making processes, including beliefs, rules, desires, goals, and emotions. This makes the BRIDGE framework well-suited for modelling complex social systems where various factors influence agents.	It can be complex and computationally intensive to implement, especially when compared to simpler ABM frameworks such as the BDI model. Given the additional layers of complexity in the BRIDGE model, it may be more challenging to design and simulate a phenomenon.	Simulation of a group of agents interacting in an organisational setting. Each agent in the model might have beliefs about the goals and objectives of the organisation, rules and procedures to follow, desires to achieve certain outcomes, and goals to achieve both within and outside the organisation.	(Dignum *et al.*, 2006, 2009; Langton *et al.*, 2005, 2007).
Deliberative Normative Agents (DNA)	Deliberative Normative Agents (DNA) are a type of agent-based modelling (ABM) framework that represents agents as having the ability to reason about and evaluate the normative constraints and rules that govern their behaviour. The DNA framework is based on the BDI framework of agency. However, it adds an additional layer to the model to represent the normative constraints and rules that influence agents' decision-making processes.	It allows for the representation of a wide range of normative factors that can influence agents' behaviour, such as moral values, social norms, and legal regulations. This makes the DNA framework well-suited for modelling complex social systems where various normative factors influence agents.	Limited representation of emotions: The DNA framework primarily focuses on representing normative constraints and rules and may not adequately capture the influence of emotions on agent behaviour. If a model requires a more detailed representation of emotions, a different ABM framework may be more appropriate.	Simulation of a group of agents interacting in a social setting, where each agent is influenced by a set of moral values and social norms that govern their behaviour. The agents might use their reasoning and evaluation skills to determine how to act following these normative constraints and rules.	(Castelfranchi *et al.*, 2000; Dignum & Dignum, 2015)

Table 2.2: Agent-based decision-making frameworks (part two)

Framework	Description	Strengths	Weaknesses	Example	References
Ethical Minded Intelligent Learning Agent (EMIL-A)	EMIL-A is a type of ABM decision-making framework representing agents as having the ability to reason about and evaluate ethical considerations in their decision-making processes. The EMIL-A framework is based on the BDI framework of agency but adds an additional layer to represent the ethical considerations that influence agents' behaviour.	It allows for the representation of a wide range of ethical factors that can influence agents' behaviour, such as moral values, ethical principles, and social norms. This makes the EMIL-A framework well-suited for modelling complex social systems where various ethical factors influence agents.	Limitations of the BDI model: As the EMIL-A framework is based on the BDI model, it is subject to the same limitations as the BDI model. Some limitations of the BDI model include its assumption of rational agents and its limited representation of the influence of emotions on behaviour.	Simulation of a group of agents interacting in a social setting, where each agent is influenced by a set of moral values and ethical principles that govern their behaviour. The agents might use their reasoning and evaluation skills to determine how to act following these ethical considerations.	(Andrighetto *et al.*, 2007).
Normative Agent (NoA)	The Normative Agent (NoA) is a framework that simulates agents' behaviour in complex systems. The NoA framework is based on the idea that agents in a system follow certain norms or rules of behaviour, which the modeller defines. These norms can be based on the agents' goals, preferences, or constraints, and they guide the agents' decision-making and behaviour in the model.	It allows the modeller to explicitly define the norms that govern the agents' behaviour, making the model more transparent and easier to understand. This can also make testing and comparing different normative systems easier to see how they affect the agents' behaviour and the overall system.	It relies on the modeller to define the norms that govern the agents' behaviour, which means that the model may not capture all real-world systems' complexity. In addition, the NoA framework may not be well-suited to modelling systems where agents have more complex or nuanced decision-making processes or where the agents' behaviour is influenced by factors other than the norms defined in the model.	Simulating the behaviour of farmers who grow and sell crops. In this model, each farmer would be represented as an agent with certain goals (e.g., maximizing their profits), preferences (e.g., preferring to grow certain types of crops), and constraints (e.g., limited land and resources).	(Mahmoud *et al.*, 2012; Savarimuthu & Cranefield, 2011).
Physical conditions, emotional states, cognitive capabilities, social status (PECS)	The PECS framework aim to capture the complexity of agents' behaviour by considering multiple factors that can influence their actions and interactions. These factors include physical conditions, such as the agents' physical abilities and limitations; emotional states, such as their mood or arousal level; cognitive capabilities, such as their ability to perceive and process information; and social status, such as their position within a social hierarchy or group.	It provides a comprehensive approach to modelling agent behaviour, considering various factors influencing an agent's actions. This allows for a more realistic and nuanced representation of agent behaviour, as it can account for the complex interplay between these different factors.	It can be challenging to accurately model and quantify the various factors influencing agent behaviour. For example, measuring and incorporating emotional states or cognitive capabilities into a model can be difficult.	Simulation of the behaviour of consumers in a virtual shopping environment. The model might consider factors such as the physical conditions of the consumers (e.g., their mobility or ability to carry items) and their emotional states (e.g., whether they are in a good mood or feeling rushed). Their cognitive capabilities (e.g., their ability to process information about different products) and their social status (e.g., whether they are influenced by peer pressure or the opinions of others).	(Kvassay *et al.*, 2017; Malleson *et al.*, 2012; Schmidt & Schneider, 2004).

Table 2.3: Agent-based decision-making frameworks (part three)

Framework	Description	Strengths	Weaknesses	Example	References
Consumer Satisfaction and Utility Maximisation (CONSUMAT)	CONSUMAT is an ABM framework for simulating consumer behaviour in a market environment. The model aims to capture the complex decision-making processes of consumers as they evaluate and choose among different products or services to maximise their satisfaction and utility.	It can provide insight into the factors that influence consumer behaviour and how they interact with one another. For example, the framework can help researchers understand how consumers weigh the costs and benefits of different options or how their preferences change over time.	It may rely on assumptions about consumer behaviour that may not always hold true in the real world. For example, the framework may assume that consumers are rational actors who always seek to maximise their utility, which may not always be true in practice. Additionally, the framework may be limited by the data and information used to parameterise it, which may not always represent real-world consumer behaviour.	Simulation of consumer behaviour in a virtual grocery store, in which consumers are modelled as agents who evaluate and choose among different products based on their preferences and budget constraints. The model might consider factors such as the prices and qualities of different products and the consumers' characteristics (e.g., their income, tastes, or health concerns).	(Forero *et al.*, 2019; Moglia *et al.*, 2018).
Model Human Processor (MHP)	The model human processor (MHP) is a framework for understanding and modelling human cognitive processes. The MHP framework assumes that human cognition can be understood as a set of interacting processes that work together to produce complex behaviours and mental states. These processes include perception, attention, memory, decision-making, and language.	A strength of the MHP framework is its flexibility. It can model various cognitive processes and behaviours, from simple perception and attention to more complex decision-making and language use. It is also well-suited for modelling dynamic, interactive systems, such as social networks and multi-agent systems.	It can be complex and computationally intensive to implement, especially for large-scale models. Additionally, it can be challenging to validate the results of MHP models, as it can be difficult to observe many of the processes that the model simulates directly.	An example of an applied MHP model is the "Virtual Humans" model, developed by researchers at the University of California, Berkeley. This model used the MHP framework to simulate the cognitive processes involved in natural language communication between humans and virtual agents.	(Card, 1981; Kitajima & Toyota, 2012).
Connectionist learning with adaptive rule induction ON-line (CLARION)	CLARION is a framework for understanding and modelling human cognitive processes. It is based on the idea that human cognition is mediated by a network of interconnected neurons or "units," which communicate with each other through weighted connections called "links." The weights of these connections are adjusted through learning, allowing the network to adapt and change over time.	It is well-suited for modelling complex, dynamic systems, such as those involved in decision-making and problem-solving. It can also simulate the interaction between different cognitive processes, such as perception, attention, and memory.	CLARION has mainly been applied to modelling decision-making and problem-solving in financial decision-making and artificial intelligence. It may not be as well-suited for modelling other cognitive processes or behaviours. Furthermore, the models are based on a fixed set of assumptions about the underlying cognitive processes and how they interact. This can limit the range of phenomena that the model can capture.	An example of an applied CLARION model is the "Adaptive Decision Maker" model, developed by researchers at Carnegie Mellon University. This model used the CLARION framework to simulate the decision-making processes involved in financial decision-making.	(Licato *et al.*, 2014; Lynch *et al.*, 2011; Sun, 2007).

Table 2.4: Agent-based decision-making frameworks (part four)

Framework	Description	Strengths	Weaknesses	Example	References
Adaptive Control of Thought - Rational / Procedural Memory (ACT-R/PM)	ACT-R/PM framework is a cognitive architecture used to understand and model human cognitive processes. The ACT-R/PM framework is implemented as a computer simulation in which the various cognitive processes are modelled as interacting agents or modules. These modules communicate with each other through a set of predefined rules, which specify how the different processes should interact and coordinate their activity.	The ACT-R/PM framework is based on a set of explicit assumptions and rules, making it possible to test the model's predictions against experimental data. This allows researchers to validate the model and refine it over time.	The ACT-R/PM framework can be computationally intensive to implement, especially for large-scale models. This can make it difficult to run and test the model in a reasonable amount of time.	Researchers used the ACT-R/PM framework to build a computer model of the cognitive processes involved in reading and comprehending text. They used the model to simulate the effects of interruptions on reading performance and found that it could accurately predict the experimental data.	(Byrne, 2000; Cao & Liu, 2011; Fleetwood & Byrne, 2002).
State, Operator, and Result (SOAR)	SOAR is designed to model complex behaviour as the actions of an autonomous agent in an environment. It is based on the idea that complex behaviour can be represented as a set of productions, which are rules that specify what actions an agent should take in a given situation. The "State" component of SOAR refers to the current state of the agent and the environment, the "Operator" component refers to the actions that the agent can take, and the "Result" component refers to the consequences of those actions.	It can model various cognitive tasks and processes, including problem-solving, decision-making, planning, and learning. SOAR is also designed to be efficient and scalable, allowing it to handle complex environments and tasks.	SOAR models may not always be robust and may not generalize well to new situations or environments.	An example of an applied model that has used SOAR is the "Mars Rover" model, which was developed to study how a rover might explore the surface of Mars and make decisions about where to go and what to do. The Mars Rover model used SOAR to simulate the rover's decision-making process and evaluate different exploration strategies.	(Laird *et al.*, 1987; Newell, 1994).

Table 2.5: Agent-based decision-making frameworks (part five)

Cognitive decision-making frameworks focus on how agents process information and make choices based on their mental representations and processes. These frameworks often draw on theories and models from psychology and cognitive sciences to understand how agents perceive, store and retrieve information and how they use it to make decisions. The PECS and CONSUMAT frameworks in Tables 2.3 and 2.4, respectively, fall under the cognitive category. According to (Balke & Gilbert, 2014), cognitive models are split into two main categories: those based on the assumption of rationality and those based on the assumption of bounded rationality. Rationality-based models (Luce & Raiffa, 1989; Simon, 1955) assume agents make decisions by carefully considering all the available options and choosing the one that maximises their utility or satisfaction. These models are based on the assumption that agents have complete information about the options available and are perfectly rational (similar to normative frameworks discussed earlier).

Conversely, bounded rationality (Simon, 1955, 1957) considers the limitations of agents and the complexity of real-world decision-making situations. These models propose that agents use mental shortcuts or heuristics to make decisions when faced with complex or uncertain situations and may also be influenced by cognitive and emotional biases. The El Farol Bar model by Brian Arthur exemplifies this approach, as it models agents using heuristics to make decisions under bounded rationality (Arthur, 1994). Unlike the purely rational agents in normative models, agents in the El Farol Bar model have limited information and cognitive capabilities, leading them to use simplifying strategies to navigate decision-making scenarios. Some limitations are worth describing, such as the limited scope of cognitive approaches (Johnson & Goldstein, 2003), as they only focus on specific aspects of the decision-making process. They may not adequately capture the full range of factors influencing agents' decisions, such as communication with others (Payne *et al.*, 1988). Moreover, the most common limitation of most decision-making frameworks (including cognitive ones) is computational complexity, where the simulation requires a substantial amount of data or experience (training), which requires a significant amount of computational resources to complete successfully.

Psychologically and neurologically inspired agent-based decision-making frameworks build on the cognitive approach by incorporating insights from neuroscience and psychology into the modelling of decision-making. These frameworks may use neuroimag-

ing techniques, such as functional magnetic response imaging (fMRI), to study how the brain processes information and makes decisions and may also incorporate psychological theories of motivation, emotion and decision-making biases. The frameworks MHP (Card *et al.*, 1983), CLARION (Sun, 2006) and ACT-R/PM (Anderson, 2013) in Tables 2.4 and 2.5 are all cognitive but also neurologically and psychologically inspired. According to (Balke & Gilbert, 2014), these models of decision-making seek to understand the neural basis of decision-making by studying the brain activity of agents as they make decisions. These models often incorporate fMRI or electroencephalography (EEG) techniques to measure brain activity and can provide insights into the neural processes underlying decision-making that scientists try to incorporate into computational agents. The psychological aspect incorporates emotions, motivations and biases. Modelling the full complexity of human psychology and neuroscience can be challenging, and it may not be feasible to include all relevant factors in an ABM. This can lead to a simplified, incomplete representation of decision-making processes (Gilbert & Troitzsch, 2005).

The strengths and weaknesses of RL have been discussed previously in sub-section 2.3. However, comparing these strengths and weaknesses in the context of other decision-making frameworks commonly used by researchers can set the scene for the models described later in the thesis. The BDI decision-making framework models an agent's beliefs, desires and intentions as separate mental states and assumes agents make decisions by reasoning about these mental states. RL, however, models decision-making as a learning process to maximise a reward signal through trial and error. One critical difference between the two is the way they model an agent's mental state. RL does not explicitly model mental states but instead learns a policy that maps states to actions through local experience, given this 'black box' configuration, it can be hard to interpret behaviours, compared to BDI. Another difference is how RL and BDI handle uncertainty. BDI models decision-making under uncertainty using probabilistic reasoning, while RL handles uncertainty through exploration and exploitation. RL algorithms can learn directly from raw sensory input, allowing them to learn complex tasks without requiring explicit decision rules. BDI models may require explicit modelling of beliefs, desires and intentions, which can sometimes be complex depending on domain and more likely to contain irregularities. RL algorithms can make decisions in real-time, making them well-suited for tasks that require fast decision-making. BDI models may

take longer to make decisions as they rely on probabilistic reasoning and may need to consider multiple mental states. Most importantly, RL algorithms can learn efficiently from a few examples, making them well-suited for tasks where data is scarce. BDI models may require more data to make accurate decisions (Russel & Norvig, 2012; Wooldridge, 2009).

Normative agent frameworks like BDI also differ from RL in how it makes decisions. Here, agents make decisions based on constrained moral or ethical principles. The study by (Asai & Ikeda, 2008) compared RL to a normative decision-making framework called multi-attribute decision making (MADM) in the context of resource allocation in a manufacturing system. The study found that RL outperformed MADM in terms of efficiency and accuracy, as RL could more effectively balance the trade-offs between different objectives. On the other hand, the Model Human Processor (MHP) framework is a good comparator framework to RL as both are cognitive and have been used to simulate human-like behaviours and intuition. The MHP framework is often used to simulate and study human-like decision-making in various contexts, including human-computer interaction. It has been applied in various domains, including psychology, computer science, and engineering, and has been influential in developing AI and decision-support systems (Card *et al.*, 1983). One main strength of RL over MHP is that RL allows agents to learn from their own experiences and interactions with the environment rather than relying on predefined rules or cognitive models. This allows RL to adapt to changing environments. Additionally, RL is more efficient than MHP models in terms of computational complexity. The issue with MHP is that it has many interlinked layers that pass information, such as the perceptual, cognitive and motor processors (Balke & Gilbert, 2014; Card *et al.*, 1983). Reinforcement learning has been applied to a wide range of problems in agent-based and multi-agent systems, leading to new insights and advances in fields such as economics, computer science and robotics, to name a few. Its roots date back to the 1940s, meaning it has been founded and researched long before some decision-making frameworks described earlier. At this stage in the review, there may be value in unpicking some of the critical milestones of RL, which can form the basis for the following section, where we describe RL within the context of the papers presented in this thesis.

Traditional algorithms, such as Q-learning and SARSA (Sutton & Barto, 2018b), are effective at learning optimal policies even when the data is incomplete or noisy. In

1989, Watkins (1989) introduced the Q-learning algorithm, one of the most commonly used RL algorithms for learning optimal policies in Markov Decision Processes (MDPs). The Q-learning approach is an early example of model-free algorithms where the RL agent does not require a model of the environment at initialisation. It can purely learn directly from experience without needing prior knowledge of the transition probabilities or rewards in the environment.

Moreover, Q-learning is an off-policy algorithm which means it can learn from data generated by policies that are different from the optimal policy. It can learn from past experiences (an analogy of this could be exploring several branches of a decision tree, not only the branch with the optimal policy). An MDP is a framework used to describe a reinforcement learning problem. It consists of a set of states, actions, a transition function (probability of transitioning from one state to another based on the action taken), a reward function (the utility the agent receives for taking a particular action in a given state) and a discount factor (trade-off the importance of immediate rewards versus future rewards) (Sutton & Barto, 2018b). Reinforcement learning research leapt into the video game domain, where Justesen *et al.* (2017) presented a comprehensive review of RL applications in the following classic titles: Breakout, Zelda, Vizdoom, TORICS, Minecraft and Starcraft to name but a few. The video game Roller Champions by Ubisoft was also played by an RL agent (Iskander *et al.*, 2020).

Furthermore, they demonstrated the power of this collaboration through the learning of complex Atari video game environments where deep-RL agents outperformed human players. Neural networks are used as function approximators to learn value functions or policies that map states or state-action pairs to expected returns or probabilities of taking a particular action. The Neural network is trained by presenting it with examples of state-action pairs and their corresponding returns or probabilities and adjusting the network's parameters to minimise the prediction error. The function approximators can take the form of linear models, decision trees, and neural networks. Neural networks are a particularly powerful function approximator because they can learn complex, non-linear relationships between inputs and outputs. Lastly, the prediction error is the difference between the predicted output of a model and the true output. Also in 2013, Mnih *et al.* (2013) presented the first application of deep RL to complex control tasks, demonstrating the ability of RL algorithms to learn complex behaviours from raw sensory inputs.

Overall, the main difference between these approaches is the level of abstraction and the focus of the analysis. Normative frameworks are more abstract and focus on the idealised behaviour of agents. In contrast, cognitive, psychologically, and neurologically inspired frameworks are more grounded in empirical observations and seek to understand decision-making mechanisms. The following sub-section analyses articles that have utilised ABM to investigate housing markets and environmental criminology. It is worth noting that these topics will be reviewed in more detail in Chapters 4 and 5 for each article presented in this thesis.

2.8 Agent-based models in environmental criminology and housing market research

This sub-section discusses the application of agent-based models (ABMs) in two distinct research areas: environmental criminology and housing markets. The review will focus on the decision-making approaches researchers adopt in their ABMs. Subsequent chapters will delve into the benefits and innovative aspects of using reinforcement learning (RL) in these domains and its advantages over pre-existing frameworks employed by researchers.

2.8.1 Agent-based models in environmental criminology

Agent-based modelling is a relatively new approach in the field of criminology, but it has already been used to study a wide range of topics related to offender behaviour. These models have been used to investigate a wide range of questions related to the relationship between the built environment and crime. Some examples of research questions that have been addressed using agent-based models in environmental criminology include: how do individual behaviours and decision-making processes contribute to patterns of crime and disorder in a particular neighbourhood or community? (Bosse & Gerritsen, 2008; Bosse *et al.*, 2007; Caskey *et al.*, 2018; Groff, 2007). How do different environmental features, such as the layout of streets, the presence of surveillance, or the availability of lighting, affect crime and disorder? (Birks & Davies, 2017; Malleson & Andresen, 2016; Malleson *et al.*, 2010). How do the actions of individual agents, such as offenders or law enforcement officers, influence the overall level of crime and disorder in a particular area? (Brantingham & Brantingham, 1995; Malleson & Andresen, 2016).

2.8 Agent-based models in environmental criminology and housing market research

An early example of agent-based modelling in crime was developed to understand and empirically test social phenomena that manifest across both space and time (Groff, 2006). The paper proposes a methodological innovation, the use of simulation models as a tool for theory evaluation. In this approach, theoretical assumptions are transformed into operational components within a simulation model. The model is then used to conduct a variety of experiments that mimic possible real-world scenarios. The resulting data from these simulations are then analysed to assess whether the outcomes are consistent with the predictions made by the underlying theory.

The model presented in (Bosse & Gerritsen, 2008) investigated the emergence of "hot spots" or areas with high crime levels. The model incorporated the concept of reputation in the analysis of crime and displacement and utilised predicate logic languages such as TTL and LEADSTO to combine qualitative and quantitative concepts into the model. The results show a repeating pattern of displacement, in which passersby move away from offenders, offenders follow passersby and guardians follow offenders. According to the authors, this pattern is consistent with displacement trends described in criminological literature.

While the article by (Caskey *et al.*, 2018) also simulates similar dynamics to examine the effects of offender behaviour, targets, and guardians on the displacement of crime hot spots. However, (Caskey *et al.*, 2018) utilised a more advanced decision-making framework called belief learning, which allowed agents to learn and adapt to the actions of other agents by observing them rather than by receiving rewards which are common in RL. The article mentions the need to consider more realistic parameter settings in future works and the scaling issues that may arise when studying large populations.

(Birks *et al.*, 2012) and (Groff, 2007) both developed ABMs of crime (residential burglary and street robbery, respectively) that employed condition-action rules (production-rule system) and included both targets and offender agents in the environment. The ABM in (Birks *et al.*, 2012) simulates the behaviour of both offenders and victims, with the offender behaviour being specified by micro-level mechanisms drawn from theories of environmental criminology, including the routine activity approach (Cohen & Felson, 1979), the crime pattern theory (Brantingham & Brantingham, 2019), and the rational choice perspective (Clarke, 1980). The study's results suggest that the three mechanisms have differential impacts on the macroscopic regularities studied. The routine activity and awareness space mechanisms have the most

significant impacts, followed by rational choice. The paper concludes that all three mechanisms interact to varying degrees to produce the observed crime patterns.

In comparison, the (Groff, 2007) model conceptualises the routine activity theory and rational choice perspective to simulate the complex and dynamic interactions of individuals that produce observed crime patterns. Both models used a decision-making framework that assumed equal weighting for perceived utility and local knowledge. However, (Birks *et al.*, 2012, p. 244) acknowledged that this assumption was "unlikely to reflect real-world offending". In the case of (Groff, 2007), implementing activity spaces for civilians to enhance agents' behaviours and awareness levels is crucial.

(Malleson *et al.*, 2012) developed an ABM of residential burglary to introduce a novel framework for enhancing both human and environmental factors in criminological ABMs. They employed the Physical conditions, Emotional states, Cognitive capabilities and Social status (PECS) framework (Urban & Schmidt, 2001). The paper argues that current approaches to agent-based crime modelling either include only simple behavioural frameworks that allow for limited dynamic behaviour or no framework at all. The results show that the model can replicate the spatial and temporal patterns of residential burglaries in Leeds, United Kingdom and produce results consistent with criminological theories. However, the authors acknowledged the need to increase the complexity of agents to allow them to correctly perceive the environment and the importance of including the decision not to offend as a viable option to improve the model's accuracy in replicating offender behaviour.

In contrast, (Zhu & Wang, 2021) proposes a hybrid model combining cellular automata and ABM to more accurately represent the complex dynamics of crime events in Baton Rouge, United States. The model is validated by comparing predicted and reported crime hot spots. The authors find that the model performs well and supports the least-effort principle of criminal behaviour, in which offenders choose suitable targets within a limited time and with less intervention at places already known for high crime levels. However, the model performs poorly when motivated offenders do not consider the reputation of neighbourhoods. The authors suggest that the model could be improved by including more realistic offender decision-making processes and incorporating more detailed spatial data.

In summary, the articles used various ABM frameworks and decision-making approaches to study the effects of various criminological theories on crime dynamics and

the impact of space on crime. However, there is an appetite to improve the complexity and realism of the models in order to replicate offender behaviour more accurately.

2.8.2 Agent-based models of housing markets

Agent-based models (ABMs) provide a lens through which the interplay of individual decisions and market dynamics in housing can be examined. These models encapsulate the decision-making processes of various market participants such as buyers, sellers, landlords, and tenants, allowing for an exploration of the emergent phenomena such as pricing patterns and urban land use changes (Axtell, 2014; Filatova *et al.*, 2009; Gilbert & Troitzsch, 2005).

A seminal work in this area by Filatova *et al.* (2009) pioneered the study of agents' pricing behaviour within urban land markets. Their ABM was instrumental in demonstrating how micro-level interactions could lead to macro-level patterns in land prices and urban development, thus providing critical insights into the spatial-temporal dynamics of urban land use change. This work laid the groundwork for subsequent studies that delve into the complexity of housing markets and how they are shaped by the behaviours and interactions of individual agents.

ABMs offer the ability to scrutinise the impact of market forces, like supply and demand, on housing prices and the volume of transactions. Researchers employ these models to simulate scenarios such as the effect of new housing construction on market equilibrium or the influence of varying buyer or seller archetypes on market dynamics (Ge, 2014, 2017; Zhuge & Shao, 2018). Bagheri-Jebelli *et al.* (2020) extended the application of ABMs to the phenomenon of urban gentrification, capturing the intricate decision-making processes of diverse household types and the resultant socioeconomic transformations within urban neighbourhoods.

The flexibility of ABMs allows for the exploration of the implications of behavioural assumptions. For instance, He *et al.* (2018b) investigated the Chinese housing market through an ABM, uncovering the model's sensitivity to variations in agent decision-making assumptions. Such sensitivity analysis highlights the criticality of the foundational hypotheses in ABMs and their influence on model outcomes.

However, ABMs are not without their challenges in representing housing markets. The accuracy of these models is contingent upon the behavioural assumptions they embed, which might not always align with real-world behaviours. Moreover, the computa-

tional demands of ABMs can be substantial, often restricting the scope of simulations that can be feasibly executed (Tian & Qiao, 2014).

Despite these challenges, ABMs have consistently demonstrated their value in elucidating the complexities of housing markets and other intricate systems. Axtell (2014) provided evidence of this by simulating the precursors to housing bubbles in the US market, while Haase *et al.* (2010) successfully captured the peculiarities of the housing market in Leipzig, Germany. Collectively, these studies underscore the potential of ABMs to serve as powerful tools in understanding and forecasting the dynamics of housing markets.

In Chapter 5 a more comprehensive review of housing market applications will be disseminated.

2.9 Summary

This chapter examines emergent complex behaviours and their applications, which arise from interactions between organisms and their environments, with genetic factors and learning processes playing a role. Genetic Algorithm (GA) and Reinforcement Learning (RL) are proposed as methodological solutions for behavioural decision-making and the emergence of complex adaptive behaviours. The chapter compares RL and GAs and how they are used in optimisation problems, analyses the limitations and implications of agent-based RL models for simulating complex behaviours, explores the use of RL in Agent-Based Models (ABMs) to simulate complex behaviours, describes traditional decision-making frameworks in ABMs, and finally, discusses the application of ABMs in two distinct research areas: environmental criminology and housing markets. The chapter concludes with a focus on the benefits and innovative aspects of using RL in these domains and its advantages over pre-existing frameworks employed by researchers.

The following three chapters 3, 4 and 5 describe the results of three different studies conducted using the primary methodologies, namely RL and ABM. Due to the variation in the domains by which the models explore, i.e., predator-prey interactions (ecology), crime and housing markets respectively, the structure of the chapters captures this variation. However, commonalities between the chapters can be found, for example, each chapter includes an introduction, literature review, model description, results and discussion and conclusion. All three chapters were written in a self-contained format, with references to the main thesis aim and objectives, this way, readers can pick and

choose the order in which they read the chapters.

The investigation undertaken in chapter 3 had several objectives, the first was to develop an effective development pipeline whereby RL and ABM can be integrated in an efficient and effective way, once this was achieved, the following subsequent applications in chapters 4 and 5 would be better placed to be investigated efficiently. The second objective was to identify a suitable RL algorithm, as these algorithms develop rapidly, with a large open-source community, identifying the perfect algorithm was not possible. Instead, the most prominent algorithms (highly-cited publications) were weighed against their strengths and weaknesses and suitability to modelling the problem(s). This and a mixture of trial and error led to the identification of the most suitable algorithm at the time of writing this thesis. The third objective was to prove that RL can learn emergent complex behaviours and adapt to novel situations (learning about and reacting to unbeknown stimuli), to achieve this, a simple predator-prey ABM is developed and the RL-powered prey agents are trained on several environmental configurations. The chapter conveys the results using both qualitative and quantitative methods, once these were assessed and identified as promising, the subsequent case studies in chapters 4 and 5 are ready to be conducted and investigated.

Chapter 3

Learning Complex Spatial Behaviours in ABM: An Experimental Observational Study

Capturing and simulating complex adaptive behaviours within spatially explicit individual-based models remains an ongoing challenge for researchers. While an ever-increasing abundance of real-world behavioural data are collected, few approaches exist that can quantify and formalise key individual behaviours and how they change over space and time. Consequently, commonly used agent decision-making frameworks, such as event-condition-action rules, are often required to focus only on a narrow range of behaviours. We argue that these behavioural frameworks often do not reflect real-world scenarios and fail to capture how behaviours can develop in response to stimuli. There has been an increased interest in Machine Learning methods and their potential to simulate complex adaptive behaviours in recent years. One method that is beginning to gain traction in this area is Reinforcement Learning (RL). This chapter explores how RL can be applied to create emergent agent behaviours using a simple predator-prey Agent-Based Model (ABM). Running a series of simulations, we demonstrate that agents trained using the Proximal Policy Optimisation (PPO) algorithm behave in ways that exhibit properties of real-world complex adaptive behaviours, such as hiding, evading and foraging.

3.1 Introduction

Prior to delving into a formal definition of "complex adaptive behaviour", it is crucial to establish a foundation upon which such a concept can be understood. Chapter 2 provides a comprehensive literature review that lays the groundwork for this, examining the theories that describe how agents interact with and adapt to their environments. In understanding the dynamics of complex systems, particularly social systems, it becomes apparent that they are not static. They evolve over time as a result of the interactions and decisions made by individuals within them. The literature hypothesises that these systems are characterised by a network of individuals whose behaviours are interlinked, leading to emergent properties that define the system as a whole.

O'sullivan *et al.* (2012) offers valuable insights into this phenomenon, particularly discussing how neighbourhood-level changes are indicative of underlying social processes shaped by many individual decisions. Building on this, Batty (2013) further explains that these social systems are networks of individuals, with information continuously transmitted between them, thereby driving the evolution of the system. These perspectives underscore the necessity for any simulation of individual decision-making to capture the nuances of these social processes at the individual level. It is within this context that we can now define "complex adaptive behaviour" as the capacity of agents to utilise their knowledge about their environment and their interactions to make decisions that enable them to adapt to new, evolving situations. This definition is pivotal to our understanding of the processes that shape complex systems, providing a lens through which we can examine the adaptive phenomena that occur as a result of individual actions and interactions.

In this research, we apply this definition to our agent-based models, investigating how agents adapt their decision-making processes over time and space, and how these adaptations contribute to the larger dynamics observed within the system. Understanding these dynamics is particularly applicable in an era where data availability is increasing.

The recent increase in data from varied sources such as footfall cameras, pollution monitors, and mobile navigation systems has enhanced our insights into the behaviours and movements of individuals (Benenson *et al.*, 2008; Dawson *et al.*, 2011; Luo *et al.*, 2008). Despite this wealth of data, embedding individual-level behaviours within Agent-Based Models (ABMs) presents a considerable challenge. To address this, scholars

have sought to develop frameworks to better capture behaviour within ABMs (Balke & Gilbert, 2014; DeAngelis & Diaz, 2019; Groeneveld *et al.*, 2017).

ABMs incorporate decision-making processes (Crooks & Hailegiorgis, 2014; Epstein, 1999; Kangur *et al.*, 2017; Malleson *et al.*, 2013; Olmez *et al.*, 2021a), yet the frameworks underpinning these models vary significantly in their purpose and intricacy. A frequent criticism of traditional ABMs is the reliance on predetermined decision rules, often rooted in historical data analysis, which inevitably limits the behavioural range of the agents to specific contexts and time frames. DeAngelis & Diaz (2019) notes that decision-making is complex, involving numerous factors such as environmental interactions, rewards and penalties, and the agent's internal state and knowledge.

Reinforcement Learning (RL) represents an evolution in modelling complex decision-making, offering multiple levels of sophistication including cognitive and social aspects (Balke & Gilbert, 2014). In contrast to existing decision-making frameworks in ABMs, this research explores how agents can learn complex adaptive behaviours that evolve across space and time through RL, notably Proximal Policy Optimisation (PPO)—an advanced RL algorithm (Schulman *et al.*, 2017).

The aim of RL algorithms is to map situations to actions effectively, learning through trial-and-error reinforced by a reward function (Sutton & Barto, 2018b; Wooldridge, 2020). RL's potential for exploring complex behaviours is well-documented, with applications ranging from adaptive drone behaviour (Lopes *et al.*, 2018) to the evolution of segregation dynamics (Sert *et al.*, 2020) and air traffic control management (Spatharis *et al.*, 2019).

The objective of this research is to demonstrate the usability of RL in developing complex adaptive behaviours in an illustrative agent-based model and subsequently, to interpret these emerging behaviours qualitatively and quantitatively. To achieve this objective, two experiments are devised, where (1) looks at the impact training length has on task efficiency, answering the question: does learning for a longer duration lead to behaviours with better outcomes than those trained for a shorter period? In experiment (2), three model scenarios are devised, these are: prey agents are trained without the predator, and the predator is not present post-training (Scenario 1), prey agents are trained with the predator, and predator is present post-training (Scenario 2); lastly, the prey agents are trained without the predator, while the predator is introduced post-training (Scenario 3). The question to be answered is: how do agents adapt to the

presence of an unknown stimulus? Does this have an adverse effect on task efficiency? Given the two experiments, we aim to compile outputs from these experiment scenarios and (1) assess the quantitative outputs by comparing task efficiency across the different experiment scenarios. (2) analyse the individual behaviours from recorded simulations to interpret complex adaptive behaviours.

The research objective is achieved by: developing a simple ABM containing two types of agents (prey and predator) in the Unity software platform using the ml-agents software package (Juliani *et al.*, 2018). Training this model under several experimental conditions using PPO (Schulman *et al.*, 2017), and subsequently examining the outcomes of the trained models both quantitatively and qualitatively. Finally, a framework is devised to record, analyse and interpret behaviours agents portray during the simulation runs.

In summary, this research seeks to contribute to the field of ABMs by applying RL to enhance the modelling of complex adaptive behaviours, offering insights that could be pivotal for future, more realistic behavioural simulations.

In section 3.2, the PPO RL algorithm is described, including how it works, its past uses, its features and relative strengths and weaknesses. Section 3.3 outlines the ABM, the types of agents, environment, rewards and penalties that agents yield. Section 3.4 provides a comprehensive description of the steps taken to verify the behaviours of the experiments. Section 3.5 defines the steps taken during training of the RL model; this includes the parameters used for training, formal definition of these parameters and subsequently, the results from the training process. Section 3.6 details the simulation results; these results are quantitatively analysed using various data science techniques. The individual behaviours from these experiments are qualitatively described; a systematic approach was developed to ascribe visually inspected behaviours to determine complex adaptive behaviours. Section 3.7 section, discusses the research outcome, what was learnt from the research and how the research would be useful in future applications conducted by ABM researchers.

3.2 Literature review

This section describes the PPO RL algorithm (Juliani *et al.*, 2018; Schulman *et al.*, 2017), including previous studies that have utilised and compared the algorithm. The algorithm is illustrated by outlining the main components, including the underlying

formulae and pseudocode. It then reviews the Unity platform and ml-agents applications.

3.2.1 Proximal policy optimisation (PPO)

PPO is a policy gradient method; these are a type of RL methods that depend on optimising parameterised policies concerning the expected return (long-term cumulative reward) by gradient descent (Sutton & Barto, 2018b).

The PPO algorithm used in this research was developed by Open AI researchers (Schulman *et al.*, 2017). These algorithms alternate between sampling data by interacting with the environment and optimising a proxy objective function using the stochastic gradient descent algorithm (Ruder, 2016). The developers of PPO argue that given the recent advancements made in RL algorithms that adopt neural network function approximators. There are still areas that could be improved, such as making these algorithms scalable to larger models (even more so given the COVID-19 pandemic and the need for large-scale models that simulate population behaviours), parallel computation applications and solving multiple problems without the need for hyper-parameter tuning (Schulman *et al.*, 2017).

During testing, Schulman *et al.* (2017) compared the PPO algorithm against algorithms that are known to perform well at solving continuous control problems. The algorithms compared were: Trust Region Policy Optimisation (TRPO) (Schulman *et al.*, 2015), Cross-Entropy Method (CEM) (Szita & Lorincz, 2006), Advantage Actor-Critic (A2C) (Mnih *et al.*, 2016), A2C with Trust Region (Wang *et al.*, 2016). During these tests, PPO outperformed the algorithms mentioned above in almost all instances of continuous control scenarios.

$$L^{CPI}(\theta) = \hat{\mathbb{E}}_t \left[\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta old}(a_t|s_t)} \hat{A}_t \right] = \hat{\mathbb{E}}_t \left[r_t(\theta) \hat{A}_t \right] \tag{3.1}$$

In the above Formula 3.1, CPI stands for "conservative policy iteration" (Kakade *et al.*, 2002). Without a constraint, maximisation of $\mathrm{L}^{\mathrm{CPI}}$ would lead to an extremely large policy update; therefore, the objective function needs to be modified to penalise changes to the policy that shift $r_{\mathrm{t}}(\theta)$ away from 1. Subsequently, the following Formula 3.2 was developed (Schulman *et al.*, 2017).

The previous variant of the PPO algorithm detailed in (Schulman *et al.*, 2017) used

an adaptive Kullback-Leibler divergence (a measure of how one probability distribution is different from a second, reference probability distribution) (Kullback & Leibler, 1951) penalty to control the change of policy at each iteration. The newly updated variant of the PPO algorithm adopts a different objective function (a method to measure the quality of any solution to a problem) proposed by (Schulman *et al.*, 2017) which can be found below (Formula 3.2).

$$L^{CLIP}(\theta) = \hat{\mathbb{E}}_t[min(r_t(\theta)\hat{A}_t, clip(r_t(\theta), 1-\varepsilon, 1+\varepsilon)\hat{A}_t)] \tag{3.2}$$

Where θ is the policy parameter, π is the policy, a and s are action and state respectively, $\hat{\mathbb{E}}_t$ is the empirical expectations over time steps. $r_t$is the ratio of the probability under the new and old policies, respectively. $\hat{A}_t$ is the estimated advantage at time t. ε is a hyperparameter, usually between 0 and 1; the hyperparameter value is used to control the learning process. As described by (Schulman *et al.*, 2017), the first term inside the min is $\mathrm{L}^{\mathrm{CPI}}$ (Formula 3.1). The second term, $clip(r_t(\theta), 1-\varepsilon, 1+\varepsilon)\hat{A}_t)$, adjusts the surrogate objective by clipping the probability ratio, which eliminates the incentive for moving r_t outside of the period $[1-\varepsilon, 1+\varepsilon]$. Finally, the minimum of the clipped and unclipped objective is taken, so the ultimate objective is a lower bound (also known as a pessimistic bound) on the unclipped objective. Given this system, the change in probability ratio is ignored if the objective improves; conversely, it is only included when it makes the objective worse.

The PPO Actor-Critic algorithm outlined by (Schulman *et al.*, 2017) is defined below in pseudocode (see Algorithm 1). For each iteration, every agent adopts an initial policy (a set of action/state combinations) $\pi_{\theta_{old}}$ (as utilised in Formula 3.1) in the environment for T time steps. The advantage estimates $\hat{A}_t$ (as utilised in Formulas 3.1 and 3.2) are calculated for each time step. The algorithm then constructs the surrogate loss given the policy parameter θ on these NT time steps of data and optimises it with a Minibatch Stochastic Gradient Descent (this is a variation of gradient descent that splits training data into small batches that are used to calculate model error and update model coefficients) for K epochs.

The ability to deploy PPO in Unity as a decision-making framework for agents with relative ease and its performance against other learning algorithms outlined previously, made it the top contender of algorithms to adopt in this research. Similarly, being able to solve multiple problems with varying complexities without the need to tune training

Algorithm 1: PPO, Actor-Critic Style

initialization;
for *iteration=1,2...* **do**
  for *actor=1,2,...,N* **do**
    Run policy $\pi_{\theta_{old}}$ in environment for T timesteps
    Compute advantage estimates $\hat{A}_1$,...,$\hat{A}_T$
  end
  Optimise surrogate L wrt θ, with K epochs and minibatch size $M \leq NT$
  $\theta_{old} \leftarrow \theta$
end

parameters made it suitable for the experimental conditions we aim to conduct (Baker *et al.*, 2019; Juliani *et al.*, 2018; Schulman *et al.*, 2017).

In ml-agents (Juliani *et al.*, 2018) PPO uses an artificial neural network (ANN) to approximate the ideal function that maps an agent's observations to the best action an agent can take in a given state (Figure 3.14).

In the field of robotics, PPO has been applied, for example, to develop applications in which mobile robots learn how to navigate an unknown terrain, even without prior knowledge of the map (Zeng, 2018). This flexibility in PPO is further demonstrated in multi-agent environments because an actor-critic variant of PPO has been shown to learn to generate the action distributions through which the critic network predicted the discounted future returns (Sutton & Barto, 2018b). These yielded findings that were of fundamental importance to the success of our research, in that they provided further evidence for PPO's capability to support complex behavioural simulations (Baker *et al.*, 2019). By the term "valuable" here, the useful and relevant value that these outcomes bring to establish that PPO can indeed effectively model the type of complex behaviours that are the subject matter of this chapter is defined. These studies do show a key limitation, however, with regard to the fact that they did not further explore spatial patterns in greater depth due to environmental changes. As per literature, people's individual behaviours can differ and change from time to time (Juliani *et al.*, 2018; Sert *et al.*, 2020), and this further explores such dynamics. Its direct goal lies in illustrating and detailing the subtleties of individual behaviours in an effort to probe into environmental implications for these patterns, which tackle critical questions.

The drawbacks of PPO are: acquiring good results via Policy Gradient methods is demanding because they are sensitive to the choice of step size - too small, and progress is unbearably slow; too large, and the signal is overwhelmed by the noise (Schulman *et al.*, 2017). As model complexity increases, solving these problems using RL can become computationally intensive; therefore, in some cases, high-performance computing clusters may be required to adopt PPO in research. Lastly, trained agents may be unable to adapt to changes in the environment; this is commonly referred to as overfitting in machine learning. One mechanism to alleviate this and train more efficient agents is to expose agents to these changes during training (Juliani *et al.*, 2018). As a model becomes more complex and training environments become more dissimilar to test scenarios, overfitting becomes more likely. It is recommended to use separate training/test scenarios (with varying levels of stochasticity) while ensuring some similarities (Zhang *et al.*, 2018). This study proposes a simplistic model with only two agent types and a static environment, with relatively short training time than other large-scale models (Lopes *et al.*, 2018; Nawrocki & Sniezynski, 2018; Spatharis *et al.*, 2019). Therefore, overfitting is less likely to occur and could be identifiable when individual behaviours are quantitatively and qualitatively interpreted.

3.2.2 Unity and ml-agents

The integration of Unity, a game development platform, with advanced RL methodologies, illustrates a pivotal shift in the simulation and training of intelligent agents. This review aims to delve into the extensive body of work surrounding the Unity platform and its application with RL algorithms, highlighting significant contributions and developments that have propelled the field of ABM and intelligent systems.

Unity, through its ml-agents toolkit (Juliani *et al.*, 2018), has emerged as a transformative platform for developing complex simulations where agents can learn, adapt, and interact within a richly detailed virtual environment. This review encapsulates seminal and recent papers that have laid the groundwork and expanded the reach of ABM and RL applications, ranging from robotics and gaming to natural language processing and beyond.

Unity ml-agents: foundations and applications

- Juliani *et al.* (2018) introduced Unity as a versatile platform for intelligent agents,

emphasising its utility in creating complex, interactive simulations for AI research. Strengths: A major strength of this work is its pioneering approach to leveraging a popular game development engine for AI research, making sophisticated simulations more accessible to researchers. The Unity ml-agents toolkit opens up new avenues for experimentation in AI, offering a flexible and powerful tool for a wide range of applications from robotics to social behaviour simulations. Weaknesses: While the paper provides a comprehensive introduction to using Unity for AI research, it is more of an overview than a deep technical exploration. The specifics of implementing complex AI models within Unity environments are not detailed, which may leave readers requiring more granular technical guidance.

- Engelbrecht (2023) explores the interplay between neural networks and simulation space using the Unity ml-agents package. This work demonstrates how Unity can be used to bridge the theoretical aspects of AI, particularly neural networks, with practical application in simulated environments, enhancing the realism and applicability of AI research. Strengths: The book's strength lies in its focus on the practical application of neural networks within Unity, providing a valuable resource for researchers and developers looking to apply AI theories in simulated environments. It serves as a useful guide for understanding and implementing neural network-based models in Unity. Weaknesses: As a book focused on the introduction of concepts, it may lack in-depth technical details or advanced topics in neural network optimisation and architecture design within Unity. The broad scope might not satisfy readers looking for advanced or specific use cases.

- Youssef *et al.* (2019) showcases the use of imitation learning within Unity ml-agents for gameplay, highlighting the toolkit's potential in replicating human-like behaviour in agents. The research demonstrates a significant reduction in computational time and training data needed by using imitation learning over traditional RL methods. Strengths: A key strength of this work is its practical demonstration of imitation learning to create more realistic and varied agent behaviours with less computational overhead. This approach is particularly innovative in the context of game development and AI training. Weaknesses: The focus on a specific application of imitation learning might limit the generalisability of the findings. Additionally, the paper could benefit from a broader comparison with other learning methods beyond computational efficiency and training data re-

quirements.

Reinforcement learning: advances and applications

- Polydoros & Nalpantidis (2017) provides a comprehensive overview of model-based RL applications in robotics. It underscores the significant contributions of RL techniques to enhancing autonomous systems' ability to learn and adapt from interactions with their environment. Strengths: The article's strength lies in its breadth, covering a wide range of RL applications in robotics, and providing a solid foundation for researchers new to the field. It effectively highlights the adaptability and potential of RL in solving complex robotic tasks. Weaknesses: As a survey, it may not delve deeply into the specific technical challenges or limitations of the RL techniques discussed. The article could benefit from more detailed case studies or comparisons between different RL approaches within robotics.

- Sierla *et al.* (2022) reviewed RL applications in HVAC system control, highlighting the impact of RL on optimising energy consumption and maintaining indoor air quality. Strengths: This review is notable for its focus on a practical and critical application of RL in energy systems. It provides a thorough examination of how RL can be applied to optimise complex systems like HVAC, offering valuable insights for both researchers and practitioners. Weaknesses: The paper primarily focuses on the potential and theoretical applications of RL in HVAC systems without providing extensive empirical data or case studies demonstrating these applications' real-world effectiveness.

- Xiang & Foo (2021) discuss the recent advancements in deep RL for solving Partially Observable Markov Decision Processes (POMDP) problems. The paper highlights the broad applicability of Deep-RL (DRL) in various domains, including games, robotics, and natural language processing, and emphasises the method's effectiveness in environments where agents have limited information. Strengths: This work's strength lies in its detailed examination of DRL applications across different fields, showcasing the versatility of DRL in addressing complex, real-world problems. It provides a solid theoretical foundation and points to diverse applications, making it a valuable resource for researchers across disciplines. Weaknesses: The focus on a wide range of applications might come at the

expense of depth in any single domain. Additionally, the challenges and limitations of applying DRL to POMDP problems are not extensively discussed, which could be an area for further exploration.

Multi-agent systems and cooperative learning

- Orr & Dutta (2023) provide a survey of multi-agent deep RL applications in multi-robot systems. They emphasise the importance of cooperative learning among agents to achieve complex objectives, highlighting the potential for multi-agent DRL to advance collaborative robot tasks. Strengths: The paper is strong in its comprehensive coverage of multi-agent systems and the specific challenges and opportunities they present. It effectively synthesises current research in the field, providing a clear overview of the state of multi-agent DRL in robotics. Weaknesses: While the survey is broad and informative, it may lack specific technical details or guidelines for implementing multi-agent DRL in practice. Furthermore, the paper could benefit from more discussion on the limitations and challenges of scaling multi-agent systems.

- Vinyals *et al.* (2019) achieved a grand-master level in StarCraft II using multi-agent RL, showcasing the potential of RL in competitive gaming environments. Strengths: This paper is notable for its application of multi-agent RL to a highly complex and dynamic environment, showcasing the cutting-edge potential of RL techniques in achieving human-competitive performance. It provides valuable insights into the design and optimisation of RL algorithms for complex tasks. Weaknesses: The focus on StarCraft II, while impressive, may limit the direct applicability of the findings to other domains. Additionally, the complexity of the approach and the resources required might not be feasible for smaller scale or less resource-intensive projects.

Challenges and future directions

- (Gu *et al.*, 2022; Nunes, 2019) discuss the challenges and methodologies for ensuring safety in RL applications. They highlight the importance of developing algorithms that prioritise safety to prevent harmful outcomes during the learning process, especially in real-world, safety-critical tasks. Strengths: These works are significant for their focus on an often-overlooked aspect of RL applications:

safety. They contribute to a growing body of literature that seeks to address the ethical and practical challenges of deploying RL in sensitive or critical environments. Weaknesses: While they address crucial concerns, these papers might benefit from more concrete examples or case studies of safe RL being successfully applied in practice. The theoretical discussion of safe RL principles could be complemented with more practical guidance for implementing these principles.

- Kayhan & Yildiz (2021) provided a literature review on RL applications to machine scheduling problems, suggesting future research directions in industrial and manufacturing settings. Strengths: This review is particularly strong in its focus on a specific, practical application of RL, providing a thorough overview of current research and potential future developments in machine scheduling. It bridges the gap between theoretical RL models and their practical implementation in manufacturing. Weaknesses: The paper's focus on machine scheduling, while in-depth, might limit its interest to readers outside the manufacturing and industrial engineering fields. Additionally, the review could benefit from a more critical analysis of the challenges in applying RL to real-world scheduling problems.

This literature review underscores the versatility and robustness of Unity and RL in fostering the development of intelligent, adaptive agents capable of complex behaviours. The studies reviewed herein not only highlight the significant strides made in the field but also point towards future research trajectories that could further harness the power of Unity and RL in creating sophisticated ABM systems.

3.3 Model description

This model description adheres to an adapted version of the Overview, Design Concepts, and Details (ODD) protocol (Grimm *et al.*, 2006), a structured framework for describing agent-based and individual-based models. Whilst the full ODD protocol encompasses seven elements, this document focuses on the most appropriate components tailored to the context of the study: the purpose of the model, the agents, and the environment. These components align with the ODD sections on purpose, entities, state variables, and scales, as well as the design concepts. This adapted approach is taken to streamline the description and to hone in on the aspects most relevant to the emergent behaviours we aim to observe.

The model contains two agents: a predator agent to catch the prey agents; and prey agents that avoid the predator and forage points. The model simulates agents interacting in a simple three-dimensional environment that contains physical barriers that block the vision and movement of all agents. Through several experiments, this research explores how agents that adopt PPO devise complex adaptive behaviours, given the environmental surroundings they find themselves in.

Predator-prey models are not new; they have been previously used to simulate the interactions of wildlife in ecology (Colon *et al.*, 2015; Hawick *et al.*, 2008; Zhdankin & Sprott, 2010). As mentioned in the Introduction section, the predator-prey scenario is a simple scenario to model. Thus, this domain was chosen to promote explainability and reproducibility.

In this research, interest is centred around learned behaviours that emerge given the environmental factors the agents find themselves in. For example, barriers were added to the environment to test if prey agents utilise them in hiding from the predator. Prey agents have no inherent defence mechanism to deploy against the predator. The behaviours we aim to observe have been identified in past literature, especially those in fish biology research of predator-prey behaviours. Sullivan & Atchison (1978) describe the interactions among two types of fish; they describe the primary defence for prey fish as "detection of pursuit". This behavioural strategy includes "fish visually tracking the predator's movement, and when it swims within their line of sight, they then get away from it at the nearest opportunity". Given the behaviours observed in these predator-prey scenarios, it is expected that the following behaviours are observed in the model:

- prey evade the predator.
- prey forage rewards.
- prey use the environment to hide from the predator.

This research is interested in training agents to learn and enact simple to more complex spatial behaviours. Some of these complex behaviours are outlined above. Thus, the environment has been designed to motivate the above behaviours and allow prey to utilise these behaviours in new situations.

3.3.1 Purpose

The ABM was developed in Unity using the ml-agents software package (Juliani *et al.*, 2018), the model allows the modeller to simulate hypothetical activities of prey and predator agents in 3D space.

3.3.2 Agents

Prey agents are rewarded for foraging positive points placed on the environment and penalised for foraging negative points and being caught by the predator. The predator agent follows any prey that falls within its field of view.

The prey agents adopt the PPO RL framework for learning, while the predator agent's decision-making framework is a set of condition-action rules such as IF prey-within-view THEN chase-prey. The predator moves around the environment randomly and tries to catch prey agents while the prey agents learn how to adapt to this. Once the predator physically touches the prey agent, the prey agent is caught, leading to a penalty. The agents were created this way to reduce the impact on computational demand as having two types of agents, both applying RL, would be computationally expensive.

For the predator agent parameters, Table 3.6 and Table 3.7 for the prey agent parameters. The values for each agent type are relative to that agent's size and mass. For both agents, an appropriate view radius was chosen relative to the environment and the agent's physical features, preventing one from physically being superior.

In the model, prey agents share the same characteristics. This ensures no prey agent has abilities that can make it superior to others, e.g. speed or vision - as this would arguably introduce additional complexity in interpreting model outcomes during the two experiments (as described in the Introduction 3.1). The environment randomly distributes positive points (which prey collect) and negative points (which prey should avoid). These point objects are used as a training indicator; if the reward increases, agents are learning (Figure 3.6).

The physical representation of both agents can be seen in Figure 3.1. The predator field of view and viewing angle allows it to identify objects (Figure 3.2). Prey agents move on the surface of the environment and can perceive the world through ray-cast sensors collecting observations. A first-person observation can be found in Figure 3.15.

The predator moves randomly around the environment until a prey agent falls

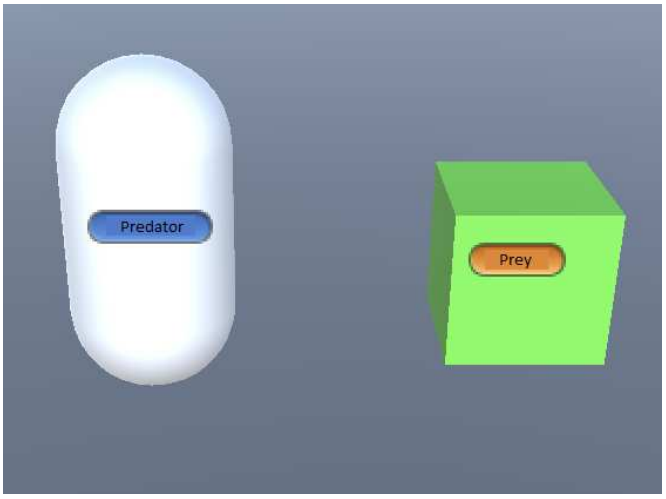


Figure 3.1: The predator agent (left) and a prey agent (right) in the Environment.

within its vision cone; this is the patrolling phase. The predator makes its movement unpredictable; thus, prey agents can be trained for all circumstances. As mentioned earlier, the condition-action rules for the predator are;

- Chase the prey if the prey is within view.
- While the simulation is running, move randomly on the surface of the environment.

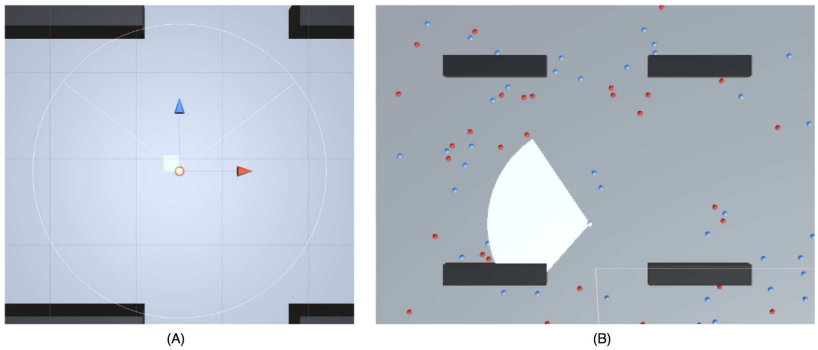


Figure 3.2: The area in which the predator can see (A). The model environment and the vision cone of the predator looking for prey agents (B).

3.3.3 Environment

The environment includes the following Unity components;

- Plane - a 3D flat surface area for agents to stand on.
- Wall - a 3D object that acts as a barrier stopping agents from falling off the plane.
- Camera - A camera pointing at the environment and agents.
- Directional light - A light ray pointing at the environment with soft shadows helps the observer see the environment.

The environment provides the prey agents with enough information to allow them to learn complex adaptive behaviours. If barriers were not present, the prey agents could not learn how to hide. Similarly, if the predator does not exist, there is no motivation for prey agents to learn how to evade capture. Each element of the environment has several customisable properties (Table 3.8) and can be changed depending on specific requirements.

The parameters described in Table 3.8, if implemented, would produce the 3D environment scene in Figure 3.3 (B).

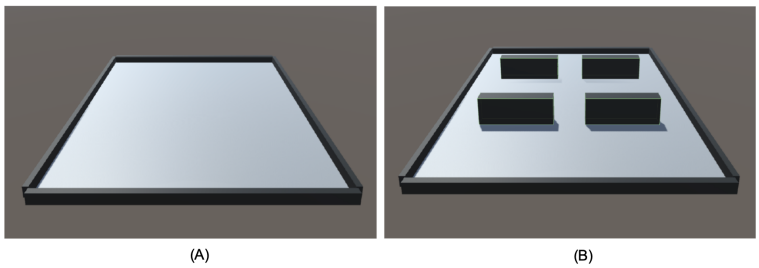


Figure 3.3: The initial state of the environment (A). The environment scene once parameters from Table 3.8 are applied (B).

The positive and negative points are randomly distributed on the environment surface so that the prey agents can forage them (Figure 3.4). If the points are collected, they re-appear at a random location within the environment. For every point collected, the prey agent is either rewarded or penalised.

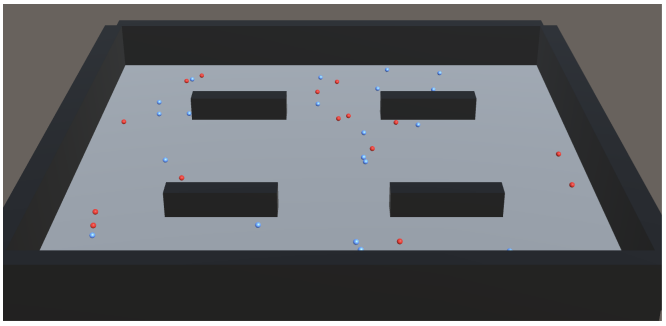


Figure 3.4: Environment with positive point objects (blue spheres) and negative point objects (red spheres).

3.4 Model verification

3.4.1 Introduction to model verification

Model verification is an essential step in the development of a reliable Reinforcement Learning (RL) model. It is the process of ensuring that the model accurately represents the intended system behaviours and assumptions. Verification aims to establish confidence in the model by demonstrating that it is logically correct and works as intended.

3.4.2 Verification strategies

Algorithmic correctness

Algorithmic correctness was achieved through code reviews and unit testing of the RL algorithms. The algorithms were tested in controlled environments to confirm that they function correctly and efficiently, adhering to the theoretical frameworks described in the literature.

Ensemble of simulation runs

An ensemble of simulation runs was conducted to test the model under various conditions and configurations. This approach helps in identifying any potential anomalies or outliers that could affect model performance. By running the model 50 times for each experiment, we could ensure the consistency and robustness of the results obtained.

Statistical hypothesis testing

Statistical hypothesis testing was employed to verify the model's performance. One-Way ANOVA tests were conducted to determine the significance of the differences observed in the model outcomes across various scenarios. Additionally, Cohen's d was utilised to measure the effect size, providing insight into the magnitude of the differences.

3.4.3 Results interpretation

The interpretation of the results from these simulation runs and statistical tests was then used to confirm the validity of the model. Increasing cumulative rewards and other performance metrics would indicate successful learning and provide quantitative evidence of the model's capabilities.

3.5 Training process

To successfully train an RL algorithm, training parameters are selected to ensure the performance of learning processes and quality of generated motions (Juliani *et al.*, 2018; Kim & Lee, 2019). An RL model is performing well if the cumulative reward is increasing during training (Poole & Mackworth, 2010). To conduct the experiments outlined in the Introduction 3.1, three neural network models will be trained using parameters that coincide with the experiment objectives, the differences between experiments are the training length and presence of a new stimulus Table 3.1.

3.5.1 Training parameters

The ml-agents package (Juliani *et al.*, 2018) simplifies the training process of artificial agents in Unity (Figure 3.5). The Learning Environment component contains the Unity scene, which includes the environment agents can act, observe and learn from. The "brain" component takes the observed data from agents (known as Vector Observations) and is trained using the Academy (Table 3.1). The Academy connects the brain to the python trainer, where the artificial neural network training commences. Once the training process ends (Figure 3.6), the output neural network is attached to the agents post-training, and thus, the agents can infer decisions from the trained model (Tables 3.2 and 3.4).

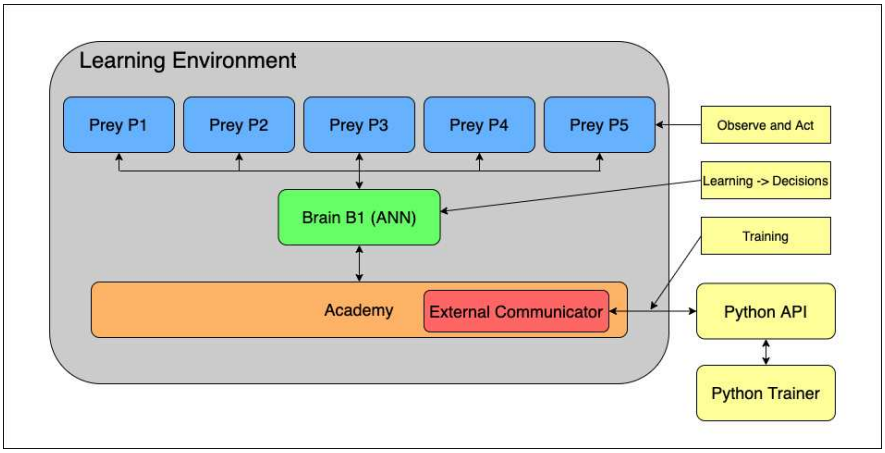


Figure 3.5: A block diagram of ml-agents

In this research, the default training parameters provided by ml-agents (Juliani *et al.*, 2018) were used for training; ml-agents researchers recommended these parameters as suitable for the majority of tested environments (Table 3.1). Furthermore, these parameters were within the recommended ranges (Table 3.9). These parameters led to successful learning sessions (Figure 3.6). The training process outputted three fully trained artificial neural network models (ANNs). For the model architecture, refer to Figure 3.14.

In Table 3.1, the only differences between the models are the max_steps variable for scenario one, which is 580,000-time steps, and in scenario three, the predator was not present during training. While training, the algorithm only started to successfully learn how to increase cumulative rewards at 100,000-time steps. This was identified as the lower bound. To then identify the upper bound and the point at which cumulative rewards stabilise, the model was trained for longer, and 1,000,000 time steps were identified as the point in which rewards were no longer being gained (scenarios two and three in Table 3.1). The reason behind this process is to produce models that coincide with the two experiments outlined earlier. For other models or research questions, the training max_steps would be different given the model complexity.

(Maqbool *et al.*, 2011) describe the adaptability of an RL agent by "making a major or minor change to the system after it has learnt the optimal policy and analyse the learning algorithm adapt to these changes". To test this theory, model training configuration three (scenario three in Table 3.1) was conceived. Configurations two

Parameter	Scenario 1 (w/predator)	Scenario 2 (w/predator)	Scenario 3 (wo/predator)
Trainer	PPO	PPO	PPO
Batch_size	1024	1024	1024
β	1.0e-2	1.0e-2	1.0e-2
Buffer_size	10240	10240	10240
ϵ	0.2	0.2	0.2
Hidden_units	128	128	128
GAE λ	0.95	0.95	0.95
Learning_rate	3.0e-4	3.0e-4	3.0e-4
Learning_rate_schedule	Linear	Linear	Linear
Max_steps	**580000**	**1.0e6**	**1.0e6**
Memory_size	256	256	256
Normalize	false	false	false
Num_epoch	3	3	3
Num_layers	2	2	2
Time_horizon	64	64	64
Sequence_length	64	64	64
Summary_freq	10000	10000	10000
Use_recurrent	false	false	false
Reward_signals	extrinsic: strength: 1.0, γ: 0.99	extrinsic: strength: 1.0, γ: 0.99	extrinsic: strength: 1.0, γ: 0.99

Table 3.1: PPO training parameters for all three scenarios.

and three were trained for the same time. The former was trained with the predator, while the latter was trained without the predator.

3.5.2 Training results

Figure 3.6 highlights the results from the training process outlined earlier in Table 3.1. Descriptions of the results presented in Figure 3.6 can be found below:

- Cumulative reward (Figure 3.6, top left) - the mean cumulative episode reward over all agents should increase during a successful training session.
- Policy loss (Figure 3.6, middle left) - the mean magnitude of the policy loss function. Correlates to how much the policy (process for deciding actions) is changing. The magnitude of this should decrease during a successful training session.
- Value loss (Figure 3.6, bottom left) - the mean loss of the value function update. Correlates to how well the model predicts the value of each state, should increase while the agent is learning, and then decrease once the reward stabilises.

Figure 3.6: Six graphs from the PPO training process each line corresponds to a scenario the model applied during training, outlined in Table 3.1. x-axis: number of time-steps in training, y-axis: value.

- Entropy (Figure 3.6, top right) - represents how random the decisions of the model are. Should slowly decrease during a successful training process. If it decreases too quickly, the β parameter should increase.
- Extrinsic reward (Figure 3.6, middle right) - this corresponds to the mean cumulative reward received from the environment per episode.
- Extrinsic value estimate (Figure 3.6, bottom right) - the mean value estimate for all states visited by the agent. Should increase during a successful training session.

In Figure 3.6, we observe that the training parameters for scenario three were the most successful compared to scenarios one and two. The only difference between scenarios two and three was the predator. Thus, these training outcomes show that the predator's presence impacts how the prey agents behave regarding rewards; when the predator is not present, the prey agents can forage more rewards. Furthermore, agents' policies frequently change over time, meaning prey agents identify good policies more frequently. The Policy Extrinsic Reward results suggest that for all three scenarios, prey agents tend to increase their rewards over time; this also means they are more likely to reduce their penalties, including predator avoidance. Ultimately, the results from these

data show that all three training scenarios and model setup was successful, i.e. prey agents were designed to successfully learn policies from their immediate environment, which they can apply post-training. Lastly, scenarios one and two led to different training outcomes (the only difference was the max_steps parameter); the reason for this is the stochasticity of the model; for each epoch, prey, points and predator are randomly distributed.

3.6 Results

Before examining the behaviours learnt by the prey agents operating under PPO, several experiments are developed. The training phase of PPO ensures agents learn to develop policies that will be subsequently used in the testing phase. Analysis of the trained models is conducted through the following experiments. These include the length of time agents train for (specified in time steps) and the stimuli presented to prey agents during the training phase (in this case, the presence or absence of the predator agent) and, in turn, the impact of this stimulus on their subsequent behaviour in the testing phase.

In the first experiment, identical initial populations of prey agents are compared across two conditions. In condition one, agents are trained for 580,000 cycles (Table 3.2, Model Condition one). In the second condition, they are trained for 1,000,000 cycles (Table 3.2, Model Condition two). The hypothesis is that agents that train for longer may develop more effective strategies which they utilise within the test scenario. A task-efficiency measure is conceived to evaluate agents under both configurations to assess if this is the case. This is achieved by measuring the amount of reward and penalties collected by agents under each configuration. A single dependent variable was produced by combining these measures, consisting of a formula containing the mean of the total positive points collected *PosTotal*. The mean of the total negative points *NegTotal*. Finally, the mean of the total number of times agents are caught *CaughtTotal* which produces the task efficiency formula:

$$(PosTotal \times 1) + (NegTotal \times -0.2) + (CaughtTotal \times -1) \tag{3.3}$$

The Task Efficiency formula is weighted by reward and penalty values. The rewards and penalties, set during the training parameter selection stage, allow prey agents to

know that getting caught by the predator and foraging negative points lead to negative outcomes while foraging positive point objects lead to positive outcomes. The rewards for foraging positive points are +1, and the penalty for foraging a negative point is -0.2, while the penalty for being caught by the predator is -1. The task efficiency formula was devised to distinguish between agents that forage positive points while avoiding negative points and the predator. Compared to agents with lower task efficiency, where agents were less capable in foraging positive points and avoiding penalties. In both experiments, rewards and penalties averaged over multiple model runs over two conditions are compared. These are short training and long training for experiment one and the predator's presence either pre or post-training or both for experiment two. This research is interested in the behaviours that emerge when prey agents interact with a predator; however, some extra elements such as points are added to the environment to introduce spatial complexity. This ensures prey agents train to achieve a goal such as foraging positive points (a metric used to identify how well they are doing) and not randomly roaming the environment waiting to encounter the predator. The penalty for foraging a negative point is smaller than being caught by the predator as we wish to introduce complexity toward agents' decision-making process. Similarly, if the penalty for foraging a negative point was -1, then the severity of foraging negative points and being caught by the predator would be equal. Consequently, prey agents may prefer being caught by the predator in certain situations.

In the second experiment, interest is centred around the notion of behavioural adaptation to stimuli. This idea is based on how RL agents behave when presented with stimuli in the testing phase, which were not present during the training phase (some sub-optimal measure of their adaptability). Furthermore, how this impacts agent decision-making relative to other agents exposed to the stimuli during training is quantified. As a result, these agents should have already developed behaviours to respond. Three model configurations are compared across the previously envisaged task-efficiency measure (Formula 3.3) to explore this. Using identical initial populations of prey agents, in the first model condition, the effectiveness of prey agents who train without the predator agent and complete their task in the testing phase without the predator is measured (Table 3.4, Model Condition one). This baseline experiment provides a comparative measure of the upper bounds of task efficiency in our model. In the second model condition, task efficiency for prey agents who train with the predator and subsequently

test with the predator present is measured (Table 3.4, Model Condition two). In the final condition, the task efficiency of prey agents who train without the predator but are tested with the predator is observed (Table 3.4, Model Condition three).

Due to the stochastic nature of each post-training test, to verify the experiments, each model condition is tested fifty times for the same duration across all conditions. A summary of results, including task efficiency, can be viewed in the following sub-sections.

3.6.1 Experiment one, exploring the impact of training length on task efficiency

Experiment one was developed to identify the impact training length has on how well agents complete tasks. Are agents that adopt RL as a decision-making process, if trained for longer, more effective at their task?

Model Condition	**1**	**2**
Training Cycles	580k	1 mil
Predator in Training	Present	Present
Predator in Testing	Present	Present
Positive Points	326.88 (34.231)	779.4 (57.434)
Negative Points	320.44 (35.595)	706.88 (53.059)
Caught by Predator	55.48 (15.931)	72.7 (11.784)
Task Efficiency	207.312 (33.835)	565.324 (53.164)

Table 3.2: Summary of the mean and (std) for each variable including task efficiency measure over all experiment one model conditions.

When inspecting model conditions one and two (Table 3.2), it becomes clear that training agents for an extended period lead to agents that can learn better policies such as foraging more positive points. The task efficiency for model condition two has done relatively better than model condition one. However, agents in model condition two still forage a relatively large amount of negative points. Furthermore, agents are caught more often compared to model condition one; this shows that positive point foraging in model condition two outweighs the penalties for negative points and being caught by the predator. Due to the weighting of Formula 3.3, we would expect to see

Caught_by_Predator occur less often than Negative_Points, while Positive_Points would be greater than the former two variables.

Statistics accompanying the quantitative results from Table 3.2, can be found in Table 3.3.

Variable	Condition 1	Condition 2	F score	p	Cohen's d
Task Efficiency (*Pos - Neg - Caught*)	207.312	565.324	1613.742	0.000	-8.034
Positive Point	326.88	779.4	2290.235	0.000	-9.571
Negative Point	320.44	706.88	1829.039	0.000	-8.553
Caught by Predator	55.48	72.7	37.758	0.000	-1.228

Table 3.3: Summary One-Way ANOVA and Cohen's d results over all experiment one model conditions.

The outcome of the statistical tests (Table 3.3) shows that the data collected from these two model conditions vary. As demonstrated by Cohen's d, the difference between the means of the task efficiency for the two groups is large; this confirms the earlier point that agents that train for longer are better at foraging positive points than agents who train for a shorter time. However, the same cannot be said for avoiding the predator. To conclude, agents that train for longer are better at finding optimal solutions for foraging but lack the same level of strategic behaviour to avoid the predator than agents trained for a shorter time.

3.6.2 Experiment two, exploring the impact of stimuli on task efficiency

Experiment two was conceived to determine how agents adapt to the presence of an unknown stimulus. For this example, the stimulus is the predator agent.

In this experiment, three model conditions are compared, the independent variables are; the presence of the predator either in training, testing or both. To examine how well agents perform relative to the predator, the task efficiency measure is used.

When comparing task efficiency across these conditions, several things become apparent. In model conditions where the predator is not present during training (Table 3.4, Model Conditions one and three), the difference between positive and negative points is large. Agents tend to collect more positive points while keeping the negative

points minimal. However, agents are caught more often in model condition three compared to model condition two, which suggests agents that have not learnt policies to deal with the predator are caught more often.

Model Condition	**1**	**2**	**3**
Training Cycles	1 mil	1 mil	1 mil
Predator in Training	Not Present	Present	Not Present
Predator in Testing	Not Present	Present	Present
Positive Points	1455.18 (111.235)	779.4 (57.434)	1476.62 (122.026)
Negative Points	491.24 (42.519)	706.88 (53.059)	504.98 (44.173)
Caught by Predator	0	72.7 (11.784)	102.08 (14.475)
Task Efficiency	1356.93 (108.803)	565.324 (53.164)	1273.544 (124.072)

Table 3.4: Summary of the mean and (std) for each variable including task efficiency measure over all experiment two model conditions.

Given all three model conditions, when agents train without the predator, they collect more positive points than agents that train with the predator present. Furthermore, agents trained with the predator weigh the risks of getting caught with the risk of foraging a negative point (Table 3.4).

Comparing the task efficiency of model conditions two and three (see Table 3.5) shows agents in model condition three produce statistically significantly higher task efficiency scores than those in model condition two. These results may seem counter-intuitive. However, one way to interpret this outcome is that agents in model condition two have likely devised policies that encourage predator avoidance to the detriment of foraging positive points. Conversely, in model condition three, agents are focused solely on foraging positive points. As emphasised by Cohen's d, the effect size of model condition two compared to model conditions one and three is large; this shows the predator significantly impacts how well prey agents forage.

Collectively the results of both experiments indicate that within the simulation:

- Agents that train for longer are more effective in devising goal-oriented strategies

Task Efficiency (*Pos - Neg - Caught*)	Condition 1	Condition 2	F score	p	Cohen's d
Model condition 1 vs 3	1356.93	1273.544	12.767	0.001	0.714
Model condition 2 vs 3	565.324	1273.544	1376.41	0.000	-7.420
Model condition 1 vs 2	1356.93	565.324	2136.585	0.000	9.244

Table 3.5: Summary of One-Way ANOVA and Cohen's d results for Task Efficiency over all experiment two model conditions.

(experiment one model condition two, experiment two model conditions one, two and three).

- Agents that weigh the risks between multiple penalties (negative points and being caught by the predator) perform sub-optimally in achieving a goal (experiment one, both model conditions and experiment two, model condition two) compared to agents that focus solely on a single reward and penalty (experiment two, model condition one).

Agents that train with the predator present weigh the risks between rewards and penalties. This cannot be said with certainty for agents that are trained without the predator present. The behaviours that emerge from agents in both experiments (experiment one, both conditions, experiment two conditions one and three) should differ. We expect to observe more sophisticated behaviours from experiment one as agents devise different ways to avoid the predator while acquiring positive points compared to experiment two model conditions one and three where agents ignore the predator.

The quantitative results above provide a means to make confident assertions regarding how well prey agents have done foraging and avoiding the predator. The spatio-temporal patterns that emerge from these behaviours must not be neglected. It would be helpful to compare the occupied spaces within the environment for both agent types against the different model conditions. These data should allow us to observe the effect the predator has on prey agents.

The Spatio-temporal movement patterns re-enforce the findings from Table 3.2 showing how well the prey agents do in avoiding the predator for model condition one (Figure 3.7, A and 3.7, B), compared to model condition two (Figure 3.7, C and 3.7, D) in experiment one. The prey agents in experiment one, model condition one learn to avoid the centre of the environment where the predator occupies; for model condition two, we do not see similar patterns; the prey agents move diagonally more of-

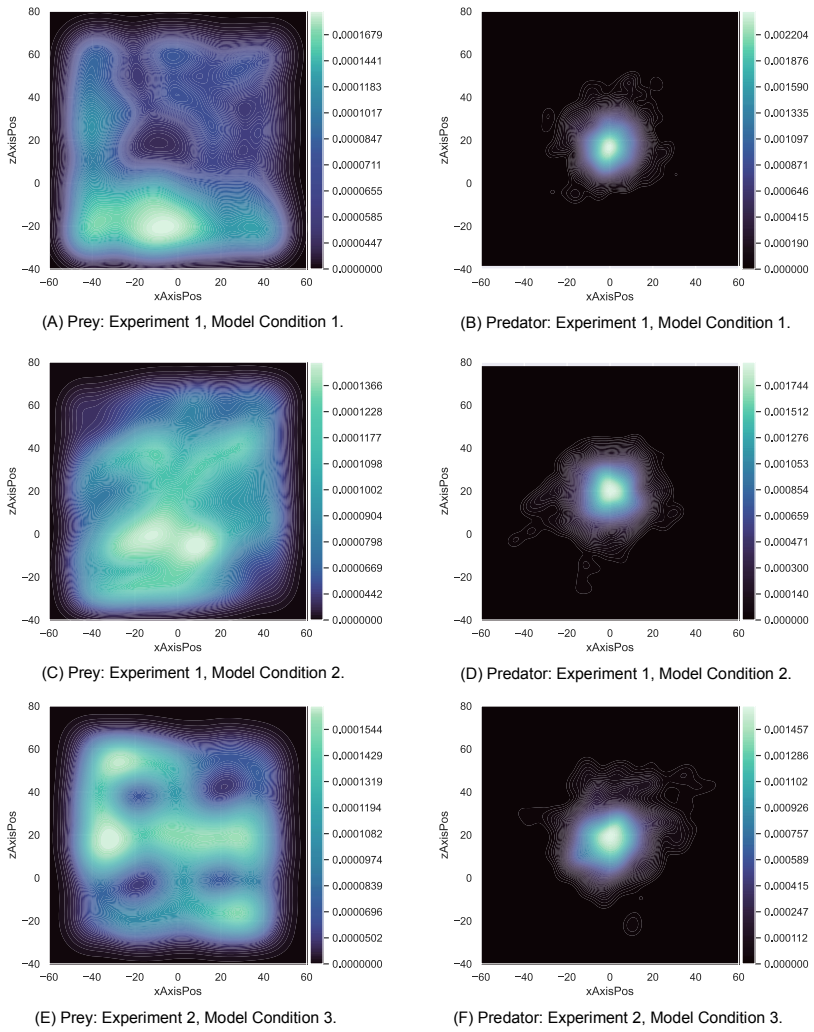


Figure 3.7: A sample of spatio-temporal KDE plots of prey and predator movement patterns.

ten and explore most of the environment. This is possibly the consequence of training for longer; thus, learning for longer makes the prey agents more goal-oriented (forage points) to the detriment of avoiding the predator (Figure 3.7, C and 3.7, D). In experiment two, model condition three (Figure 3.7, E and 3.7, F), the prey agents explore the entire environment while avoiding the barriers. Prey do navigate in areas occupied by the predator and get caught more often; however, compared to scenarios where prey are trained with the predator, agents are caught less often (Figures 3.7, C

and 3.7, D). To conclude, when agents are trained using RL under various conditions, the spatio-temporal patterns reflect these differences. These quantitative results show that emerging complex behaviours are likely to be observed at an individual level; we see that the two experiments, namely, training length and unknown stimulus, lead to different spatio-temporal patterns.

To answer the main research question, the agents' behaviours for these experiments must be analysed. Agents are more effective in avoiding the predator for some experiments. Conversely, in other experiments, agents perform more strongly in foraging positive points. To better understand these differences at an individual level, the behaviours of agents at an individual level need to be examined.

3.6.3 Individual behaviour analysis

This sub-section describes the behaviours traced given the previous experiments conducted to explore whether complex adaptive behaviours occur when agents operate under the PPO framework. To recap, agents trained with the predator weigh the risks of getting caught with foraging a negative point. Furthermore, agents that train for longer outperform agents that train for a shorter time. Lastly, agents trained without the predator focus solely on foraging points; thus, they outperform their peers trained with the predator. However, these agents are caught by the predator more often.

This sub-section contains behaviours traced from all model conditions, and they are described frame-by-frame following the systematic approach below;

1. Recording the experiment from start to finish.
2. Playing back the experiment recording and taking note of the model scenario, i.e. one, two or three.
3. Watching the movement of the predator agent to see if it interacts with a prey agent or if the predator is not present, only focusing on the prey.
4. Every time the predator interacts with the prey; the video is paused and played frame-by-frame.
5. The behaviour of the prey is captured frame-by-frame during the encounter until it has moved away and continues foraging points (known as the foraging behaviour).

6. The frame-by-frame interactions are inspected closely; the behaviours observed are noted and compared with the quantitative analysis from the sub-sections 3.6.1 and 3.6.2, to try to interpret how the prey agent has behaved.

At the current time, the only way learned, complex behaviours can be identified is by visualising the model runs and qualitatively interpreting these behaviours; therefore, this was the chosen methodology. An accompanying video of some behaviours can be found at the following link: https://youtu.be/-0bozJWC6l4.

The visually observed behaviours from the model are utilised, as these are the best indicators of complex adaptive behaviours in the model, as has been the case in past literature (Jalalimanesh *et al.*, 2017; Juliani *et al.*, 2018; Lopes *et al.*, 2018; Maqbool *et al.*, 2011; Olsen & Fraczkowski, 2015; Spatharis *et al.*, 2019).

Following the visual inspection approach, several important examples of behaviours traced during each experiment are described below. These results come in the form of figures; each frame in a figure represents a state of the model from 1 to N time steps. The **red box** in each figure focuses the viewer on where the specific behaviour in question is occurring. The **green circle** indicates a prey agent and the **orange circle** is the predator agent.

Hiding behaviour

In Figure 3.8, a prey agent in the top right of the environment; spots the predator in the second frame; then it moves towards the opposite side of the closest barrier and hides; this behaviour indicates that prey agents have learned that the predator cannot see through obstacles in the environment. Consequently, the prey agent positions itself with its back towards the wall. This hiding behaviour is also observed in Figure 3.10.

Co-operative behaviour

Prey agents are not designed to be adversarial nor cooperative; however, in experiment one, model condition one, what can be interpreted as cooperative behaviour is observed; however, this may be coincidental. Figure 3.9 depicts a scenario from model condition one, where prey agents recognise the predator and move in opposite directions to evade capture. It could be argued that a wide range of actions could have led to a rewarding outcome. However, these prey agents adopt a policy that involves them moving away from each other.

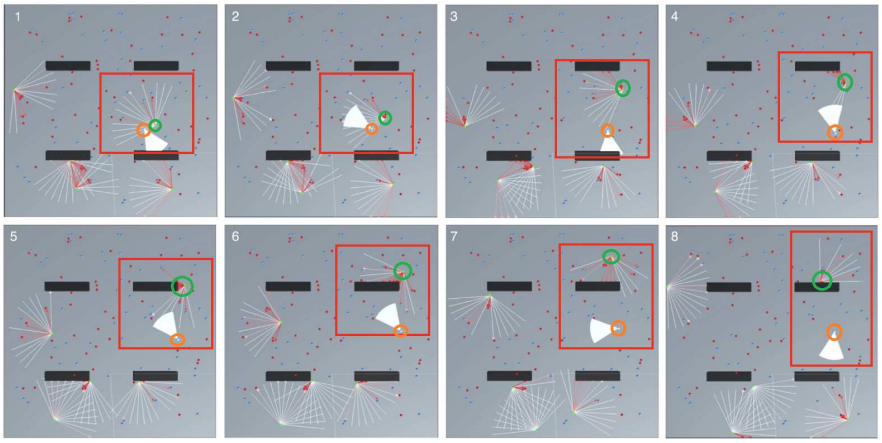


Figure 3.8: Experiment one, model condition one; a prey agent looking for a wall to hide behind.

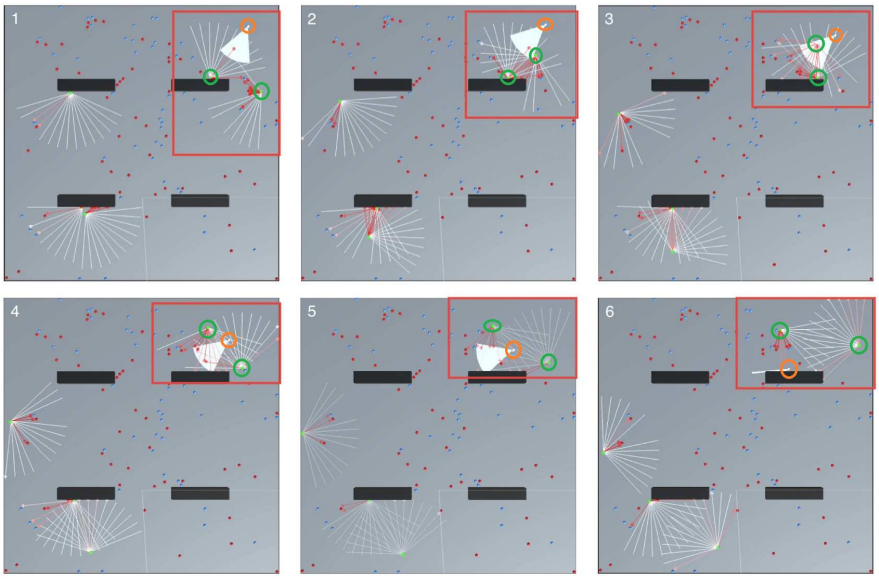


Figure 3.9: Experiment one, model condition one; two prey agents identify each other and move in opposite directions to avoid the incoming predator.

Evading behaviour

Figure 3.10 depicts a blocking behaviour where a prey agent recognises a barrier, then realises the predator moving towards it; it successfully evades the predator and passes it using the barrier to block the predator's field of view.

This learned behaviour considers the distance of the predator from the prey agent; as soon as the prey agent realises its presence, it takes immediate action to avoid it.

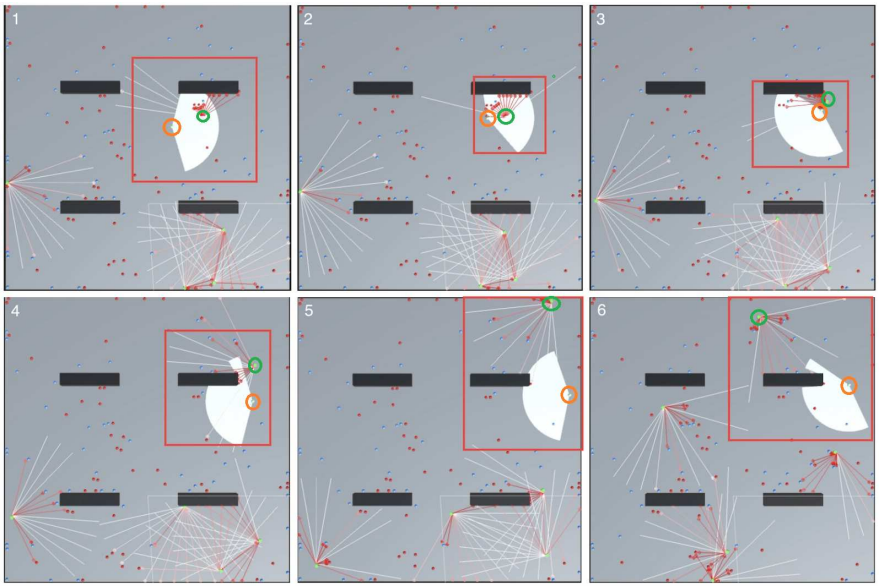


Figure 3.10: Experiment one, model condition two; a prey agent moves away from approaching predator and tries to use the barrier to evade the predator.

Figure 3.11 depicts a prey agent that intuitively dodges the incoming predator and allows it to collide with the barrier behind it. This behaviour might indicate prey agents have learned how quick predator agents move and thus can devise policies that take this information into account.

Foraging behaviour

Figure 3.12 depicts foraging behaviour captured during experiment two, model condition one. In this experiment, we recall that prey agents developed policies in the absence of the predator agent. As the predator is not present, prey agents learn behaviours that only entail foraging points.

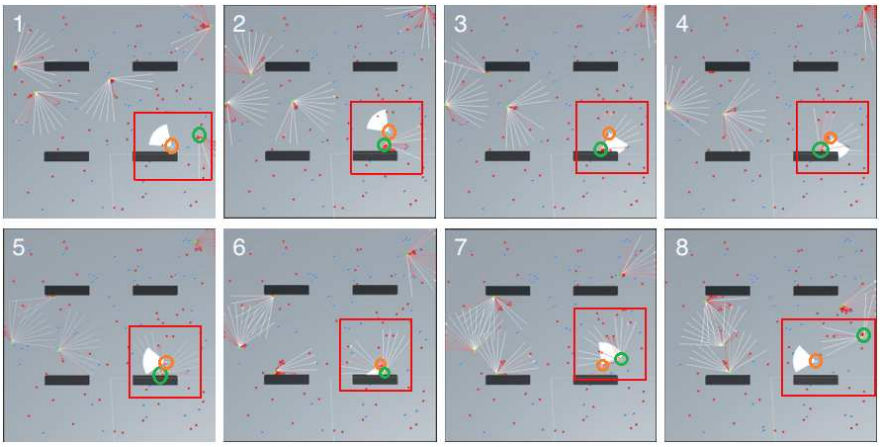


Figure 3.11: Experiment one, model condition two; a prey agent dodges incoming predator making it collide with the barrier.

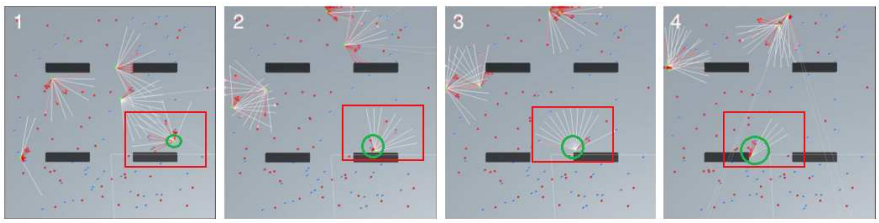


Figure 3.12: Experiment two, model condition one; the prey agents explore the environment foraging rewards.

Circling behaviour

In experiment two, model condition three (Table 3.4), prey agents are trained in a setting without a predator, then situated in an environment that contains a predator post-training. Agents appear to continue foraging positive points and avoiding negative ones. Agents move around in circles until they are close to a positive point; once they identify a positive point, they move to it (Figure 3.13).

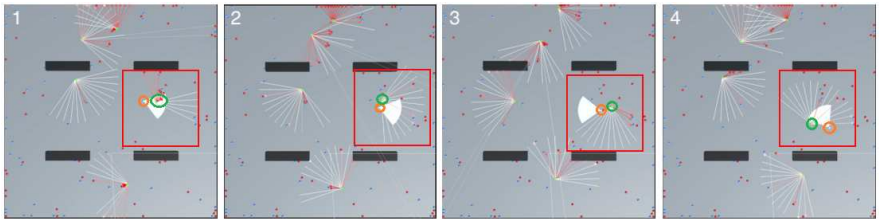


Figure 3.13: Experiment two, model condition three; prey agents move in a circular motion continuously while moving nearer to the closest positive point.

Prey agents are prone to being caught by the predator in experiment two, model condition three, as they have not devised strategies to deal with the predator (Table 3.4). However, prey agents have learned that moving in circles (Figure 3.13) increases their chances of foraging a positive point while avoiding incoming negative points. This behaviour was not observed in any of the other conditions. This behaviour may have developed due to the sphere-shaped points moving on the environment's surface more often during training than previous model conditions.

This sub-section analysed the behaviours and outcomes of model conditions for both experiments using a systematic approach. Various behaviours were identified. These behaviours can be interpreted as 'complex adaptive behaviours' such as hiding behind objects, evading the predator, using the environment to their advantage. The majority of these complex behaviours emerge from model conditions where the predator was present during training. These results show agents learn and apply policies that focus on foraging positive points while intuitively avoiding predators.

Agents trained without the predator focus on a specific task, i.e. forage points and avoid negative ones. There is a clear distinction between behaviours that emerge in conditions where the predator was present during training compared to conditions it was not. These results also show that agents trained using PPO can function in different situations/environments while achieving a goal.

3.7 Discussion and conclusion

This research set out to assess the effectiveness of developing an ABM in conjunction with the PPO technique as a behavioural framework. The research found that the RL technique supports agents in the behavioural evolution of complex decision making that resemble those observed in the real world. There have been several examples of ABMs that utilise RL. However, the majority of these have either deployed RL as an extension to extract the optimal set of steps in decision making, or they have lacked any investigation of the impact of RL algorithm features, i.e. the impact of new stimulus on the subsequent behaviours of agents (Olsen & Fraczkowski, 2015; Rahimiyan & Mashhadi, 2010; Tellidou & Bakirtzis, 2006; Thapa *et al.*, 2005). The model presented in this chapter suggests that training time directly impacts an agent's ability to improve its behavioural goals, i.e. foraging a large number of rewards compared to agents that have trained for a shorter period. However, agents that devise policies influenced by multiple penalties weigh the impact of these penalties and try to minimise the most impactful compared to the less impactful. Furthermore, the research highlights the ability for agents to operate under conditions in which they were never trained to encounter and continue to perform relatively well. While the points mentioned above seem apparent, this research attempted to quantify the degree to which such outcomes transpired. Finally, where this research diverges from past literature is the fact that individual-level behaviours that were procured from the experiments were subjectively interpreted by following a set of comprehensive steps from procurement to interpretation.

Limitations of this work include the difficulty of identifying training parameters that allow the algorithm to train the model efficiently. If the parameter values are not appropriate for the model configuration this can negatively impact the outcome of training; for example, agents may behave erratically and diverge from achieving any goal. In this research, the default parameters provided by the software library ml-agents (Juliani *et al.*, 2018) were adopted as these were tested extensively on multiple training environments; however, in future research, the impact of these choices should be assessed.

Another limitation of the research relates to the identification and interpretation of individual-level agent behaviours. Core to understanding the advantages and disadvantages of RL for developing realistic agent behaviours is assessing the behaviours

generated by the RL algorithms. When writing this chapter, the literature provides no agreed-upon methods for identifying and subsequently interpreting behaviours that agents enact during model testing. This limitation is an indicator of the lack of research done in this area. In response, this research attempted to identify a comprehensive set of steps in interpreting the observable behaviours that agents enact. However, this time-intensive technique may not be applicable for more complex models.

The accessibility of model development using new technologies such as Unity and the subsequent programming language C# is challenging. As the complexity of the ABM increases, so do the computational requirements. Due to rapid advancements in computing, this challenge is not insurmountable.

Despite the lack of literature and steep learning curve in developing an ABM in Unity with RL, these techniques together can lead to valuable outcomes. The research shows that if PPO is adopted as a decision-making mechanism, agents can organically grow behaviours through achieving rewards and punishments. Furthermore, agents can weigh multiple risks and rewards and act accordingly. These attributes that agents develop can be observed in real-world situations, i.e. when people weigh the risks of being captured by the police before attempting a crime, or predatory animals weighing the risks of the prey escaping before deciding to pursue or ignore. This research indicates that agents can portray behaviours that would be considered "complex" without any explicit prior knowledge of these behaviours, which is one of the core strengths of RL. Moreover, agents have been shown to adapt to non-deterministic dynamic changes within the environment; these agents can continue achieving their relative goals within these scenarios.

In light of the experimental findings, it is evident that RL can indeed lead to the development of sophisticated strategies, particularly in the context of extended training durations, as shown in the higher task efficiency scores for agents trained longer (Table 3.2). However, the results also underscore that longer training does not uniformly translate to better task performance across all metrics. For example, while longer-trained agents excel in positive point foraging, they do not necessarily outperform in predator avoidance (Table 3.3). This aligns with the insights from (Brearcliffe & Crooks, 2021), suggesting that learning, especially with regard to complex adaptive systems, does not always equate to more optimal outcomes. Such findings prompt a critical reflection on the value proposition of RL within social systems. They highlight

that while RL can optimise certain agent behaviours, the complexity of social dynamics marked by irrationality and bounded rationality—demands a careful consideration of what 'optimisation' entails. Thus, RL is worth the effort, provided its application is approached with an understanding of the agents' context and the multifaceted nature of optimisation in social systems, where the attributes valued by the modeller may not always align with the emergent priorities of the agents themselves.

An avenue for future research is using the techniques presented here to explore the relationships between people and enforced rules to prevent a contagious virus from moving through the population. This research is also applicable to population dynamics by simulating changing populations and the individual level interactions among populations.

To conclude, this research demonstrates that RL can provide a means for developing agents through training that exhibits 'complex' behaviours that evolve through space and time. This research shows that behaviours are encouraged by the environmental surroundings of agents. The experiments conducted highlight that training time impacts agent performance and that agents can adapt to environmental changes and behave sub-optimally when multiple penalties are considered. The spatial patterns of agent movement for each experiment condition vary, showing a strong spatial influence in decision making. The research demonstrates that agents with varying decision-making frameworks can co-exist within an environment, i.e. the predator agent applied simple if-then-else rules while prey agents utilised RL. This research shows that RL is a viable option as a decision-making framework for future ABMs and that the use of RL within ABM research could be revolutionary.

3.8 Summary

This chapter has demonstrated through experimentation and analysis of quantitative and qualitative results that RL does enable agents to learn and adapt which subsequently leads to emergent complex behaviours. As this chapter conducted experimentation on a proof-of-concept ABM, the natural subsequent task is to apply this approach to a real-world domain area, namely that of crime and simulating offender behaviour.

In the following Chapter 4, several objectives are met, the first of these is, to develop an ABM that embeds the three fundamental theories of environmental criminology, namely, the rational-choice perspective, routine activity theory and crime pattern

theory. The second objective is to train offender agents in scenarios where situational crime prevention interventions are applied which lead to heightened risks during criminal acts such as burglary. The last objective is to qualitatively and quantitatively demonstrate that offender agents can organically learn behaviours in agreement with findings from empirical and theoretical studies in criminology, such as, crime patterns are clustered, usually concentrated in a few places, few victims account for most of the victimisation and the journey to crime is typically short. These findings should subsequently lead to emergent complex behaviours that are empirically and theoretically sound, making the model more realistic, thus, the model can in the future be used to support policy-makers in making decisions as to where intervention measures should be placed to be most effective.

3.9 Notes

1. ML-agents Unity extension can be found here: here.
2. Unity can be downloaded here.
3. ML-agents architecture documentation can be found here.
4. PPO hyper parameter best practices can be found here.
5. ABM developed and trained on the following desktop: Intel Core i7-7700K, 32 GB RAM, 256 GB SSD Plus 2 TB HDD, 2 x NVIDIA GTX 1070 8 GB Graphics, Windows 10 Home.
6. ABM is run on the following laptop: MacBook Pro (15-inch, 2018), 2.6 GHz Intel Core i7, 32 GB 2400 MHz DDR4, Intel UHD Graphics 630 1536 MB

3.10 Appendix

3.10.1 Figures

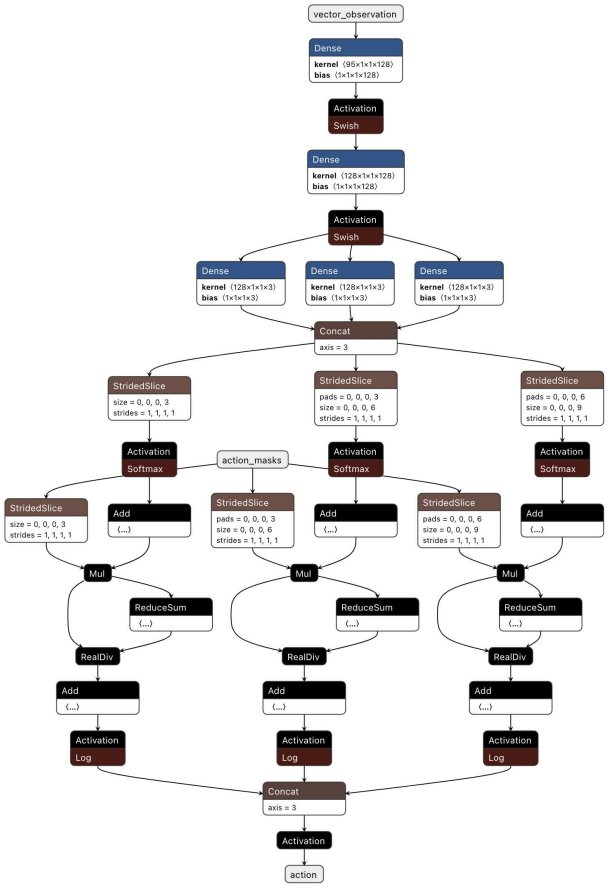


Figure 3.14: Artificial Neural Network Architecture.

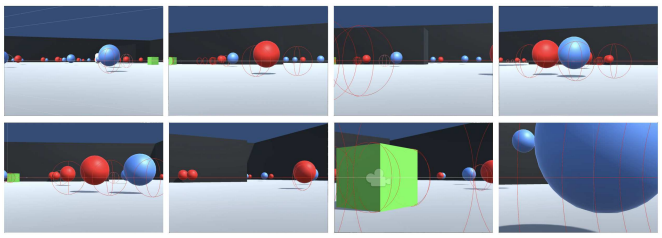


Figure 3.15: First person view from Prey agent's perspective.

3.10.2 Tables

Variable	Description	Value
Movement_speed	The speed of movement	20
Rigidbody	The 3D agent object has a rigid body property	
Capsule	The 3D agent's shape property	
Capsule_Collider	A component used to trigger an event when it comes into contact with another object	
AIPredator.cs	The C# script component allows the predator to perform actions in the environment	
View_camera	A camera component traces the movement of the agent in first person	
Velocity	The velocity of the agent (take both axes X, Y normalise them then multiply by Movement_speed)	
View_radius	The radius of the field of view for the agent	10.33
View_angle	The angle of view within the View_radius is between 0 and 360	80
Target_mask	A layer tag for objects the agent considers as targets	Target
Obstacle_mask	A layer tag for objects the agent considers as obstacles	Obstacle
Visible_targets	A list of all the targets the agent has seen	

Table 3.6: Predator agent's parameters.

Variable	Description	Value
Rigidbody	The 3D agent object has a rigid body property	
Cube	The 3D agent's shape property	
Box_Collider	A component used to trigger an event when it comes into contact with another object	
Prey.cs	The C# script component allows the agent to perform actions in the environment	
Camera	A camera component traces the movement of the agent in first person	
Velocity	The velocity of the agent	
Turn_speed	The speed of which the agent turns	300
Move_speed	The speed of which the agent moves on the X, Z axis	2
Normal_material	Normal material (agent is neither rewarded or penalised if it interacts with these)	GreenAgent
Good_material	Good material (agent is rewarded if it interacts with these)	PositivePoint
Bad_material	Bad material (agent is penalised if it interacts with these)	NegativePoint
Use_Vector_Obs	If checked, the agent will send information to the neural network during training	
Ray_Perception_Sensor	A sensor which identifies objects within a given perimeter	

Table 3.7: Prey agent's parameters.

Component	Property	Value
Plane	Position, Rotation and Scale along the X, Y, Z coordinates	P[0, 0 ,0], R[0, 0, 0], S[10, 1, 10]
Wall 1	Position, Rotation, Scale and Box Collider	P[-5.11, 0.8, 0], R[0, 0, 0], S[0.2, 1.7, 10], Box Collider [1, 1, 1]
Wall 2	Position, Rotation, Scale and Box Collider	P[5.11, 0.8, 0], R[0, 0, 0], S[0.2, 1.7, 10], Box Collider [1, 1, 1]
Wall 3	Position, Rotation, Scale and Box Collider	P[-0.04, 0.8, -5.07], R[0, 90, 0], S[0.2, 1.7, 10], Box Collider [1, 1, 1]
Wall 4	Position, Rotation, Scale and Box Collider	P[-0.04, 0.8, 5.07], R[0, 90, 0], S[0.2, 1.7, 10], Box Collider [1, 1, 1]
Camera	Position, Rotation, Scale, Clear Flags, Culling Mask and Projection	P[-1.13409, 32.80403, 125.6395], R[26.565, -180, 0], S[1, 1, 1], Clear Flags = Skybox, Culling Mask = Everything, Projection = Perspective
Directional Light	Position, Rotation, Scale, Type and Mode	P[0, 3, 0], R[50, -30, 0], S[1, 1, 1], Type = Directional, Mode = Realtime

Table 3.8: Environment parameters.

Parameter	Definition	Range
Batch_size	The amount of experiences that transpire in each iteration of gradient descent. This parameter must always be a fraction of buffer_size.	(Continuous): [512, 5120], (Discrete): [32, 512]
β	The strength of entropy regularisation. This parameter ensures that agents accurately explore the action space during training. Increasing this will ensure that more arbitrary actions are taken frequently.	[1e-4, 1e-2]
Buffer_size	The quantity of observed sensory information (experiences) to accumulate before updating the policy model. This parameter corresponds to how many experiences (observations, actions and rewards obtained) should be secured before learning begins or updating the model. This should be a multiple of batch_size.	[2048, 409600]
ϵ	This parameter determines how fast the policy can evolve during training. Epsilon corresponds to the adequate threshold of divergence between the old and new policies during gradient descent updating. Setting this value small will end in more stable updates but will also slow the training process.	[0.1, 0.3]
Hidden_units	The number of units in the hidden layers of the neural network.	[32, 512]
GAE λ	This parameter corresponds to the lambda parameter used when determining the Generalised Advantage Estimate (GAE). This can be considered how much the agent relies on its current value estimate when determining an updated value estimate. Low values correspond to relying more on the current value estimate (which can lead to bias), and large values correspond to relying more on the actual rewards received in the environment (which can be considerable variance).	[0.9, 0.95]
Learning_rate	The initial learning rate for gradient descent. This parameter relates to the strength of each gradient descent update step.	[1e-5, 1e-3]
Max_steps	The greatest number of simulation steps to run during a training session.	[5e5, 1e7]
Memory_size	The size of the memory an agent must keep; this is generally utilised if use_recurrent is true. The parameter relates to the size of the array of floating-point numbers used to store the hidden state of the recurrent neural network. This value must be a multiple of 4 and should scale with the amount of information the agent will need to remember to complete the task.	[64, 512]
Normalise	If true, will automatically normalise observations. This normalisation is based on the running average and variance of the vector observation. Normalisation can be effective in complex continuous control problems but may be harmful with more straightforward discrete control problems.	[true, false]
Num_epoch	The number of passes to make through the experience buffer when performing gradient descent optimisation. Decreasing this will ensure more stable updates at the cost of slower learning.	[3, 10]
Num_layers	The number of hidden layers in the neural network. This parameter corresponds to how many hidden layers are present after the observation input or after the Artificial Neural Network (ANN) encoding of the visual observation. For more minor problems, fewer layers are likely to train faster and more efficiently. More layers may be necessary for more significant control problems.	[1, 3]
Horizon (T)	Corresponds to the number of steps of experience to collect per-agent before appending it to the experience buffer. When this limit is reached before the end of an episode, a value estimate is used to predict the overall anticipated reward from the agent's current state. As such, this parameter trades off between a less biased but higher variance estimate (long time horizon) and more biased but less varied estimate (short time horizon). If there are many rewards within an episode or episodes are prohibitively large, a smaller number can be more ideal.	[32, 2048]
Sequence_length	Corresponds to how long the sequences of encounters must be while training (only used with Recurrent Neural Networks).	[4, 128]
Summary_freq	A Tensorboard specific parameter used to identify how often to log training statistics during a training session.	[1e3, 5e4]
Use_recurrent	Set to true if a Recurrent Neural Network is to be used, else a default Artificial Neural Network (ANN) is applied.	[true, false]
γ	This parameter corresponds to the discount factor for future rewards. This can be thought of as how remote into the future the agent should care about possible rewards. When the agent should be operating in the present to prepare for rewards in the distant future, this value should be large. In instances when rewards are more immediate, they can be smaller.	[0.8, 0.995]

Table 3.9: Formal definition of the training parameters and recommended range, where [x, y] inclusive, source: (Juliani *et al.*, 2018).

Chapter 4

Learning the Rational Choice Perspective: A Reinforcement Learning Approach to Simulating Offender Behaviours in Criminological Agent-Based Models

Over the past 15 years, environmental criminologists have explored the application of agent-based models (ABMs) of crime events and various theoretical frameworks applied to understand them. Models have supported criminological theorising and, in some cases, been applied to make predictions about the impact of interventions devised to reduce crime. However, decision-making frameworks utilised in criminological ABMs have typically been implemented through traditional techniques such as condition-action rules. While these models have provided significant insights, they neglect a crucial component of theoretical accounts of offending, the notion that offenders are learning agents whose behavioural dynamics change over time and space. In response, this chapter presents an ABM of residential burglary in which offender agents utilise reinforcement learning (RL) to learn behaviours. This solution enables

offender agents to learn from individual-level perceptions of the environment and, given these perceptions, develop behavioural responses that benefit themselves. The model includes conceptualisations of the Routine Activity Theory (RAT), Crime Pattern Theory (CPT) and a utility function, Target Attractiveness, which acts as a behavioural mould to nudge offender agents to learn behaviours in keeping with the Rational Choice Perspective (RCP). Trained behaviours are then tested by introducing crime prevention interventions into the model and examining the reactions of offender agents. In keeping with empirical studies of offending, experimental results demonstrate that offender agents utilising RL learn to offend at targets where rewards outweigh risks and effort, offend close to home, frequently victimise high-rewarding targets, and conversely learn to avoid offending in areas associated with high levels of risk and effort.

4.1 Introduction

Established among supervised and unsupervised learning, reinforcement learning (RL) allows artificial agents to learn how to behave within their environment. Agents learn behaviours by receiving feedback rewards when they perform an action. Under RL an agent's goal is to learn actions that maximise its cumulative reward (Sert *et al.*, 2020; Sutton & Barto, 2018b; Wiering & Van Otterlo, 2012).

RL algorithms can learn advanced problems to solve using neural networks (Islam *et al.*, 2019), such as playing complex games and defeating human players (Justesen *et al.*, 2020). In health research, RL was used to develop treatment plans for patients (Jalalimanesh *et al.*, 2017). In social sciences, researchers have integrated RL with Agent-Based Modelling (ABM) to demonstrate previously unidentified phenomena in agent's decision-making using a well-known ABM (Sert *et al.*, 2020). Lastly, RL was used to teach a system how to trade stocks on a stock market (Dang, 2020).

Given recent advances in RL research and open-source software, this chapter attempts to demonstrate the value of this approach in ABMs of environmental criminology - a facet of criminology focusing on environments and how they influence crime/victimisation. Our rationale in doing so is to support those engaged in modelling criminal behaviour and occurrence of crime events with more accurate models of offender behaviour (Johnson & Groff, 2014; Johnson *et al.*, 2014; Park & Buckley, 2016).

Rational choice perspective (RCP) (Cornish & Clarke, 1987) proposes a framework

for understanding offender decision-making processes. RCP states that offenders choose their behaviour and weigh whether rewards from committing an offence outweigh effort and risks in pursuing that offence. If this condition is satisfied, offenders will likely offend. Conversely, the likelihood of not offending is greater when risks and effort outweigh rewards. RCP provides practitioners and policymakers with a blueprint to understand offender behaviour and devise interventions aimed to reduce crime (Birks *et al.*, 2012; Clarke, 1997a; Cornish & Clarke, 2003; Hayward, 2007; Wortley, 2001). RCP underpins situational crime prevention intervention (SCPI) (Clarke, 1980), which seeks to reduce crime by manipulating the rewards, risks and effort calculus - i.e., making crime riskier requiring more effort or less rewarding. SCPIs have reduced crime (Eck & Clarke, 2019; Linden, 2007; Poyner, 1991), which is why many scholars believe the RCP is correct. RCP is an integral component when developing computational models of offender behaviour. Typically, models embed some measure (known as suitability in (Malleson *et al.*, 2010) or probability of offence (Birks *et al.*, 2012)) which define conditions where offences take place.

There are many studies where environmental criminologists adopt ABMs (Birks *et al.*, 2012; Bosse & Gerritsen, 2008; Gerritsen, 2015; Gialopsos & Carter, 2014; Groff, 2007; Joubert *et al.*, 2022; Malleson *et al.*, 2010, 2012). Most models propose traditional condition-action rules to respond to situations preventing agents from learning and adapting to change. This chapter presents an alternative approach to modelling behaviour. Here, offender agents are trained in various environmental configurations using RL to observe how they learn and adapt to intervention measures. We compare offender agent behaviours to RCP during testing to ensure expected behaviours are learned if the model has replicated these theoretical conceptualisations accurately (i.e., Routine Activity Theory and Crime Pattern Theory) (Gialopsos & Carter, 2014), then we expect offender agents to develop behaviours characterised by RCP (Cornish & Clarke, 1987).

Using this model, in this chapter, we explore three primary questions:

1. Do offender agents utilising RL portray behaviours in agreement with RCP, i.e., to what extent do they learn to offend when rewards outweigh risk and effort and vice versa?

2. Do offender agents utilising RL adapt to changes in their immediate environment given the introduction of simulated crime prevention interventions?

3. Do simulated crimes generated by offender agents utilising RL display patterns commonly observed in empirical studies of crime?

The chapter begins by providing a brief literature review of related work 4.2 critiquing the main theories and analysing ABMs in environmental criminology. Model description 4.3 section describes the model logic, including conceptualised theories. Results 4.4 section describes a series of experiments run using the model, and the outcomes of those experiments are presented. Lastly, discussion and conclusion 4.5 of findings, drawbacks and contributions to environmental criminology are presented.

4.2 Literature review

ABMs allow researchers to simulate a process at the individual level. Producing a disaggregation of complex systems split into components with individual characteristics (Epstein & Axtell, 1997; Heppenstall *et al.*, 2012). Most criminological applications of ABMs have been applied within the field of environmental criminology due to their spatial modelling capabilities (Groff *et al.*, 2019). As described by scholars, most ABMs in environmental criminology adopt static rules referred to as "condition-action rules", whereby a rule is triggered when an agent is situated in a state where conditions for that rule are satisfied (Johnson & Groff, 2014). These traditional methods have meant agents are not susceptible to changing and adapting their behaviour while learning from their surroundings, which some neurologically inspired computational studies argue are fundamental characteristics of the brain (Niv, 2009; Sutton & Barto, 2018b; Wiering & Van Otterlo, 2012). This chapter presents RL as a means to contribute to behavioural decision-making in these models. The following section critiques some of the foundational theories in environmental criminology which we aim to replicate through RL.

4.2.1 Critique of environmental criminology theories

While theories like Routine Activity Theory (RAT) (Cohen & Felson, 1979), Rational Choice Perspective (RCP) (Cornish & Clarke, 1987), and Crime Pattern Theory (CPT) (Brantingham & Brantingham, 2019) have significantly contributed to the field of environmental criminology, they are not without their critics. This section delves into some of the limitations and areas of contention surrounding these theories.

Routine activity theory (RAT)

RAT hypothesises that crime occurs when a motivated offender encounters a suitable target without a capable guardian in place. Critics argue that RAT oversimplifies complex social interactions and fails to account for the offender's background, motives, and other socio-psychological factors (Sampson, 1988). It also does not fully consider the broader social structures and inequalities that may influence criminal behaviour (Stevenson & Forsythe, 1998).

Rational choice perspective (RCP)

RCP suggests that offenders make calculated decisions based on a risk-reward analysis. However, this theory is often criticised for assuming a level of rationality in decision-making that may not align with the spontaneity and situational influences that affect real-world crime (Hayward, 2017). It has been pointed out that not all offenders engage in a conscious cost-benefit analysis, and many crimes are opportunistic rather than premeditated (Young, 2004).

Crime pattern theory (CPT)

CPT focuses on the spatial and temporal patterns of crime, assuming that offenders are primarily influenced by their awareness of space. Critics of CPT argue that it may not fully encapsulate the complexity of environmental and societal factors that shape crime patterns, such as urban design and socioeconomic conditions (Brantingham & Brantingham, 2016). Moreover, it may neglect the psychological and emotional states that can influence an individual's criminal behaviour (Godin, 2007).

General drawbacks

Across these theories, a common critique is their tendency to overlook the heterogeneity of offenders and the variability of human behaviour. They often assume a homogeneous offender population and do not account for individual differences in cognition, emotion, and social influence (Wikström, 2006). Additionally, these theories may not adequately explain why the same environmental factors can lead to different outcomes for different individuals or why crime rates vary significantly across similar physical settings (Johnson *et al.*, 2014). The following section describes some prominent applications of ABM

in environmental criminology.

4.2.2 ABMs in environmental criminology

According to Johnson & Groff (2014), most ABMs in criminology have either conceptualised one or more of the following theories; these are Routine Activity Theory (RAT) (Cohen & Felson, 1979), Rational Choice Perspective (RCP) (Cornish & Clarke, 1987) and Crime Pattern Theory (CPT) (Brantingham & Brantingham, 2019). RAT is concerned with the likelihood of crime occurring when a suitable target and motivated offender cross paths without a capable guardian. CPT focuses on when and where these convergences occur, how offenders perceive their environment and how these perceptions lead to offences. RCP describes the framework for thinking about offenders' decisions, where offenders are likely to offend in situations where rewards for offending outweigh risks and effort.

Most criminological ABMs reviewed here embed some form of condition-action rule inspired behavioural frameworks while embedding some of the above crime dynamics. Here, we review the frameworks utilised and the identified drawbacks. As a result of these frameworks, we believe the proposed RL framework can contribute.

Bosse & Gerritsen (2008) researched how offender behaviours, targets and guardians impact displacement of crime hot spots using the RAT. The model adopted a predicate logic framework. As behaviours are static, authors found their model led to unsatisfactory outcomes, such as "police always arrive too late" in every situation. Caskey *et al.* (2018) presented a similar ABM to (Bosse & Gerritsen, 2008). However, they opted for a more advanced decision-making framework called belief learning. In this approach, agents were able to learn and adapt to other agents' actions by modelling the RAT, while, in our research, offender agents will learn from spatial perceptions and adapt to the environment. Empirical research of offender behaviour suggests "specialised knowledge" transpires from offender-environment interaction (Taylor & Gottfredson, 2015; Topalli, 2005).

Birks *et al.* (2012) developed an ABM of residential burglary using condition-action rules. The article demonstrates how ABM can test hypothetical mechanisms explaining criminological phenomena. However, the model assumed equal weighting in decision-making processes, i.e., perceived utility and localised knowledge, which authors state "unlikely to reflect real-world offending" (Birks *et al.*, 2012, p. 244). In our model, we

develop a target attractiveness utility that is offender specific. Groff (2007) proposed an ABM with RAT to investigate street robbery dynamics. Researchers utilised condition-action rules. As a result of the study, authors argued, and we agree with, for individual-level perceptions of geographical localities to be incorporated into individuals' decision-making (Groff, 2007, p. 99).

Malleson *et al.* (2010) developed an ABM of residential burglary and opted for the Physical conditions, Emotional states, Cognitive capabilities and Social status (PECS) framework (Urban & Schmidt, 2001). Authors found that "the complexity of agents must be increased to allow them to perceive their environment correctly" (Malleson *et al.*, 2010, p. 248). Furthermore, they highlight the need to incorporate the decision to "not offend" as a viable option in enhancing the models' accuracy in replicating offender behaviours. Using RL, the proposed offender agents will use individual-level spatial perceptions to learn about the space and make decisions. Some of these decisions include the choice not to offend.

A recent article utilised RL as a decision-making framework (Joubert *et al.*, 2022). Researchers investigated street robbery dynamics in Cape Town (South Africa). The model successfully enhances the characteristics of behaviours, environments and agents using RL. During review, we found that "perpetrator reward signal" (Joubert *et al.*, 2022, p. 5) did not incorporate conceptualisations of effort and risk, which are crucial components impacting offender's behaviour with regards to target selection according to RCP(Cornish & Clarke, 1987). Furthermore, researchers do highlight the need to "investigate the effect within-episode variance of robbery opportunities, and whether endowing perpetrators with a risk appetite, along with other robbery dynamics such as guardianship effects", could allow models to replicate empirical robbery data (Joubert *et al.*, 2022, p. 17). In our model, we incorporate risk by introducing SCPIs mid-episode where interventions surrounding a target increase risk in victimising that target.

These contributions demonstrate how ABMs in environmental criminology span various applications. Some ABMs have been employed to predict crime events or model theoretical propositions of crime theory over various spatio-temporal resolutions, for example, (Groff, 2007) modelling the RAT in a street robbery context applied to Seattle (US) and (Joubert *et al.*, 2022) in Cape Town (South Africa) and burglary rates modelled in Leeds (UK) by (Malleson *et al.*, 2009). Some models have addressed challenges of testing criminological theory through ABM, these include (Birks *et al.*, 2012;

Bosse & Gerritsen, 2008; Bosse *et al.*, 2011; Caskey *et al.*, 2018).

We would argue that despite diverse applications, ABM in environmental criminology is still primarily in its infancy. Therefore, new contributions can focus on various challenges in modelling crime dynamics (Gerritsen, 2015). Most models lack behavioural heterogeneity; for example, offender agents have different characteristics - i.e., home location and propensity to offend. However, for the majority, all offender agents employ the same offending behaviour. While crime models typically use condition-action, which limits adaptive behavioural heterogeneity, other fields have started to explore more complex approaches to simulating agent behaviour (Littman, 2015; Lockwood & Klein-Flügge, 2021; Rahimiyan & Mashhadi, 2010; Rawal *et al.*, 2010). Consequently, a gap in behavioural representation in crime models has emerged. Cornelius *et al.* (2017) found that discrete rules such as "count nearest four agents if nearest four agents are criminals, do the following" are adopted by all offenders, which in most cases leads to scenarios where offender agents are more likely to behave similarly while navigating the environment. Some models consider relatively simple structures for learning agents such as geospatial awareness spaces (cognitive map) (Birks *et al.*, 2012; Groff, 2007; Malleson *et al.*, 2010, 2012) which allows agents to rudimentarily represent some level of heterogeneous awareness of the model environment. However, in these models, agents do not learn behaviours (in contrast, behaviours are defined explicitly).

In response, and following recent advances in other areas of social simulation (Baker *et al.*, 2019; Sert *et al.*, 2020), this chapter proposes that using RL in ABMs of crime event dynamics may increase the accuracy of modelled behaviours as agents learn through assessing their dynamic environment and adapting to changes. Thus overcoming challenges that may be introduced through the application of condition-action behaviours (Dahlke *et al.*, 2020, p. 13) typically used in previous studies.

To improve upon previous models, this research will present a model of burglary dynamics in which a RL algorithm is applied to model offender agent behaviour. We intend to explore if plausible adaptive behaviours can be organically learnt by agents, i.e., can agents learn the RCP and produce crime patterns indicative of empirical findings? Demonstrating RL as a viable decision-making option to model crime event dynamics more accurately.

4.3 Model description

This section details our proposed Agent-Based Model (ABM), which incorporates elements of the Overview, Design Concepts, and Details (ODD) protocol (Grimm *et al.*, 2006). The ODD protocol, a standardised framework for describing individual and agent-based models, enhances clarity and reproducibility. Here, we describe the model's purpose, the entities involved, the process overview, and the design concepts involved, particularly focusing on the aspects of the ODD protocol that most closely align with our research objectives. In Figure 4.1 we provide an activity diagram describing all the modelled entities and their relationships.

4.3.1 Purpose

Our ABM utilises Proximal Policy Optimisation (PPO) (Schulman *et al.*, 2017) to simulate the dynamics of residential burglary events. The choice of residential burglary was informed by the extensive literature and well-documented spatio-temporal features of this offence (Johnson *et al.*, 2007; Zhang & McCord, 2014; Zhang & Song, 2014), which provide a basis for model validation.

4.3.2 Model entities

Prior to introducing the model environment visualised in Figure 4.2, it is essential to define the key components comprising the model. These entities include:

- **Offender agents (B)**: Autonomous entities within the model that exhibit decision-making capabilities, which are influenced by their interactions with the environment and other model components.
- **Static targets (C)**: Representations of residential properties or other points of interest for the offender agents, which do not change location once the model begins.
- **Situational interventions (D)**: Measures implemented within the environment to influence the behaviour of offender agents, such as increased surveillance or physical barriers, intended to reduce the attractiveness of targets.
- **Nodes (E)**: Points in the model that form a network simulating a rudimentary transport system, guiding the navigation of agents through the environment.

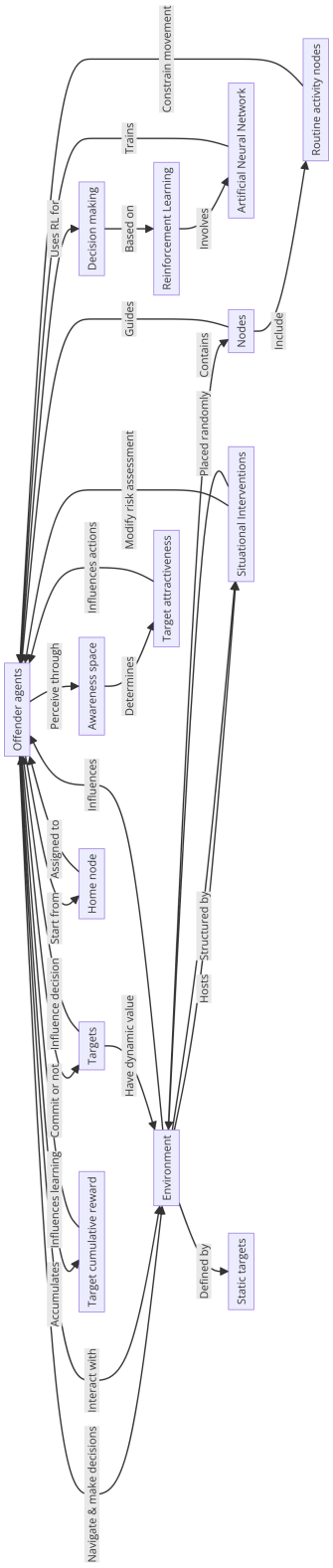


Figure 4.1: Activity diagram describing the modelled entities and their relationship.

- **Routine activity nodes (F)**: A subset of nodes that represent significant locations in the daily movements of offender agents, such as a 'home' or commonly visited areas.
- **Spatial localities**: Two distinct areas within the model environment where the aforementioned entities are situated and interact.

These entities are integral to the construction of the model environment and its capacity to simulate complex behaviours. They have been conceptualised based on seminal works in the field (Birks *et al.*, 2012; Malleson *et al.*, 2010; Park & Buckley, 2016).

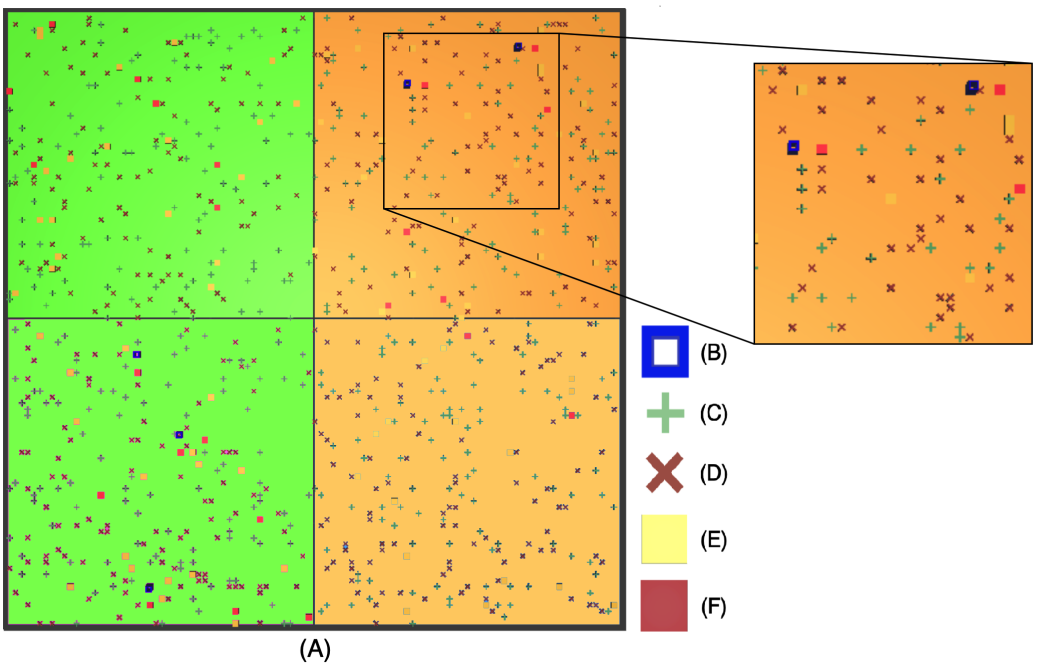


Figure 4.2: Example model environment, where (A) is the environment on a 100 x 100 grid which includes five offender agents (B), 100 targets (C) and interventions (D) at each of the two spatial localities, 100 nodes (E) of which, 23 are routine activity nodes (F), where each offender agent has five assigned nodes (the same node can be assigned to two or more offender agents).

4.3.3 Model environment

The model applies situational interventions (Clarke, 1980) to stimulate individual behavioural learning and adaptivity of offender agents using RL, where agents respond to spatial stimuli with differing responses, intending to output spatial patterns of crime in agreement with environmental criminology theory. Situational interventions are conceptualised as forms of property protection, such as increased guardianship through CCTV, which in turn modify the agents' risk assessment of targets.

The spatial configuration of the model environment is deliberately structured with defined boundaries rather than a toroidal space. This design choice is informed by the need to accurately observe and analyse the impact of environmental interventions on offender behaviour within a controlled setting. The bounded space allows for a clear delineation of edge effects and the central concentration of offences, providing valuable insights into the migration of offender activities in response to heightened risk perception. It is essential to consider that the opportunities to offend are equally distributed across the model environment. Offender agents are placed randomly, ensuring that each has an equal chance to encounter targets of varying reward levels, irrespective of their home location or routine activity nodes. This approach aligns with the principles of environmental criminology, allowing for the assessment of spatial dynamics and the effectiveness of situational interventions within the confines of a well-defined area.

Interventions

Interventions are materialised as static objects within the model environment, serving as protective measures for properties. Their placement is systematically randomised across the grid to mimic real-world distribution and to influence agent perception of risk. The number and allocation of these interventions can be precisely controlled, allowing for reproducible setups across simulation runs (Figure 4.2, D). The quantity can be manipulated for each of the two spatial localities (areas), for example:

$$Area = \{g, o\},$$

where

$$Area_g(I) = w, Area_o(I) = x,$$

g = green, o = orange for both localities, I = Interventions and w, x are the number of interventions where $0 \leq w \leq N, 0 \leq x \leq N$. N is the total number of unoccupied

grid cells available within the environment.

In order to distribute interventions spatially, unoccupied cells ($cell \in N$) are randomly selected. The purpose of the intervention is to increase $risk$ surrounding an adjacent target, which subsequently affects target attractiveness for that target. Interventions are a conceptualisation of SCPIs (Clarke, 1980).

Targets

Targets, which represent residential properties, have a dual nature in the model. They are static in placement but dynamic in their value attributes, with each target assigned a reward value within a pre-defined range. This reward value, combined with a distance-based effort calculation unique to each agent, forms the basis of a target's attractiveness, which is central to agent decision-making (Figure 4.2, C). More formally, the number of targets per locality can be specified. In keeping with RCP (Cornish & Clarke, 1987), each target has a $reward$ value:

$$T_i(reward) = [x, y],$$

where each target i has a randomly assigned floating-point value between x and y inclusive. Furthermore, the reward scale can vary for each locality, where:

$$Area_g(T_i(reward)) = [w, x], Area_o(T_i(reward)) = [y, z],$$

each target has an effort value. The $effort$ is offender agent specific, where effort for a target is the normalised Euclidean distance of the offender agent's home routine activity node to the target; thus, the furthest target to an offender agent's home has an effort value of 1.0 and the closest target, has an $effort$ where $0 \leq effort < 1$. Effort reflects the principle of least effort (Florence & Zipf, 1950) within offender agent decision-making regarding target selection; it also increases heterogeneity among offender agent behaviours, i.e., some offender agents living closer to rewarding targets may be less inclined to travel further.

The above three components are used to build a target attractiveness measure Formula 4.1 for each offender agent. A formal example of the logic behind $Reward$, $Risk$ and $Effort$ can be found in Appendix 4.8.1.

$$Target\_Attractiveness(T_i) = T_i(Reward) - (T_i(Effort) + T_i(Risk)) \tag{4.1}$$

While reward and risk are target-specific, effort is agent specific - thus, the target attractiveness measure is also agent-specific. Thus, behaviours learned will vary across offender agents depending on their routine activity spaces. This measure should enable offender agents to perceive targets in different ways and develop behaviours consistent with empirical patterns of offending, which research has shown to be situation specific (Brantingham *et al.*, 2006; Clarke, 1997b).

Nodes

Navigation nodes act as way-points in the agents' movement across the environment, distributed to create a rudimentary transport network. These nodes facilitate realistic agent navigation, aligning with the concept that offender agents are likely to encounter potential targets along their routine paths as found in (Birks *et al.*, 2012; Park & Buckley, 2016). Nodes are randomly distributed across the environment in unoccupied *cells* (Figure 4.2, E). If $Offender_X(RAN) > 0$ then $Nodes > 0$, where $Offender_X(RAN)$ are routine activity nodes assigned to Offender agent X and $Nodes$ are the number of nodes in the environment. In Figure 4.2 A, we see 100 nodes distributed, where 23 of these are routine activity nodes.

Routine activity nodes

Routine activity nodes are a crucial element that constrains and directs the spatial movement of offender agents. Each agent is assigned a set of these nodes, including a unique 'home' node, which serves as the origin point for its activities. This assignment follows the Routine Activity Theory (Cohen & Felson, 1979), ensuring that each agent's movement patterns and the potential for offending are closely related to its routine activities (Figure 4.2, F), where:

$$Offender_X(RAN) \subset Nodes,$$

Each offender agent has a number of routine activity nodes assigned to it:

$$Offender_X(RAN) = [2, Nodes],$$

therefore, $2 \leq Offender_X(RAN) < Nodes$. These routine activity nodes constrain spatial movement of offender agents, where they move between routine activity nodes and encounter potential targets during their travels. Each offender agent has a home

node from which they originate. No two offender agents can have the same home node; however, the home of one offender agent can be within the routine activity space of another offender agent. Offenders have daily routine activities (Brantingham & Brantingham, 2019); therefore, we expect offender agents to offend/not offend in locations they have previously encountered during their travels. Similar conceptualisations were adopted in (Birks *et al.*, 2012; Eck & Liu, 2004; Groff, 2007; Malleson *et al.*, 2010).

The network of routine activity nodes within the model is designed to reflect the stochastic nature of a real-world urban grid, functioning similarly to a random graph structure. This design choice models the unpredictability and complexity of an urban environment, where the connectivity between nodes represents potential paths an offender may traverse during their routine activities. This randomness is grounded in the theory that urban movements are often non-linear and subject to various social and environmental factors (Hillier, 2007). A practical illustration of this is the allocation of a 'home' node and additional nodes to each agent, simulating an individual's place of residence and frequented locations, such as workplaces or social hubs. For example, an agent might be assigned a home node at location (x_1, y_1) and additional nodes at (x_2, y_2) and (x_3, y_3), reflecting a simplified version of a person's daily commute and socialising spots. This arrangement allows the model to encapsulate the diversity of movement patterns and the potential for crime commission across different urban landscapes. The stochastic distribution of these nodes ensures that each simulation run generates unique movement patterns, thus providing robustness to the model by introducing variability in the agents' experiences and decision-making processes, much like in real life.

4.3.4 Offender agents

Agents within our model are known as offenders, Figure 4.2, B. These agents navigate the environment, perceive surroundings and make decisions influenced by spatial stimuli inferred by RL (sub-section 4.3.5). Agents can undertake three key actions: Move, Commit_Offence and Dont_Commit_Offence; RL is used to develop behaviours using these actions. The following paragraph will describe these actions, including their abstract and formal specification (found in Appendix 4.8.1).

Movement

Abstract definition: Offender agents start at their home routine activity node and follow the shortest straight line distance (Euclidean distance) to a selected routine activity node from their routine activity space. Once an offender agent reaches the next node, the above process repeats itself Figure 4.3. A formal definition of movement can be found in 4.8.1.

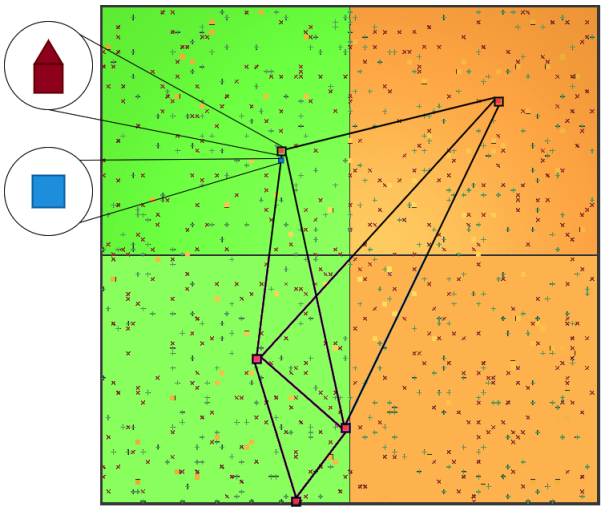


Figure 4.3: Example Model Environment: a single offender agent A navigating from its home node to a routine activity node i where $RAN_i \in Offender_A(RAN) - Offender_A(RAN_H)$.

Offend & dont_offend

Abstract definition: An offender agent can commit or not commit an offence when in the same cell as a target. The decision to choose the latter or former depends on 1) the immediate environment and data perceived during RL training and 2) the estimated reward outcome for deciding to offend or not offend. During training, offender agents are reinforced (negatively or positively) by receiving the target attractiveness utility associated with a given target when deciding to victimise it; this informs the RL algorithm of this decision during training. If the outcome is negative when an offence has been committed, risk and effort have outweighed the reward.

The decision to not offend was added to the model to 1) observe if-when offender

agents learn to not offend at targets where risks + effort outweigh rewards. 2) to demonstrate heterogeneity among offender agents when an offender agent lands in a cell containing a target, they make an active choice to offend or not offend, and both decisions lead to some reward or punishment, which influences the learnt behaviours through RL. Some offender agents may offend more than others, given their localised experiences, as presumed empirically. An offender agent that chooses not to offend accumulates zero rewards during testing; however, to ensure offender agents learn that not offending is a plausible outcome in situations where $Target\_Attractiveness < 0$, they are given a +1 training reward for choosing not to offend. Conversely, they are penalised -1 when they decide not to offend when $Target\_Attractiveness > 0$. By applying these rewards and penalties, we expect offender agents to make appropriate decisions given their circumstances.

Ultimately, offender agents should only offend when opportunity presents itself, and rewards outweigh risks and effort and not offend otherwise. Refer to 4.8.1 for a formal definition of the described process.

Offender perception

Abstract definition: Every offender agent has an "awareness space". These are sensory information captured from the offender agent's immediate environment. Offences occur when awareness spaces of offender agents converge with targets. This formalisation attempts to encapsulate propositions of CPT (Brantingham & Brantingham, 2019), which proposes that offences are likely to take place where rewarding opportunities intersect with offender awareness. In our model, offender agents perceive objects within their immediate space and capture data, including the distance to the object and its type. Offender agents utilise these data during training, enabling RL to train the Artificial Neural Network (ANN) (Islam *et al.*, 2019) to learn the most suitable conditions in which offences occur and vice-versa. Post-training, offender agents apply learned knowledge about the environment and make decisions. In Figure 4.4, offender agents have ten individual sensors. A formal definition of this process can be found in 4.8.1.

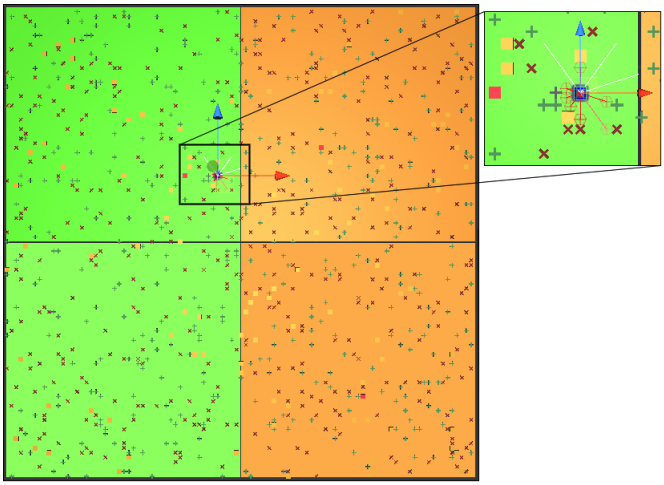


Figure 4.4: Example Model Environment: offender agent perception, a sensor can be either red or white. The former means object identified, and the latter means no object.

Target cumulative reward

Abstract definition: Every offender agent has a target wealth, target cumulative reward (TCR). If during training, an offender agent accumulates total cumulative reward $\geq$ TCR (where total cumulative reward is the sum of target attractiveness), they are rewarded a training reward to introduce eagerness.

To counterbalance eagerness, we introduce losses/costs. Every time step, each offender agent loses a small amount of accumulated total cumulative reward. Therefore, Some offender agents will reach and surpass their TCR, while others may not (these measures are analysed in section 4.4).

Our model setup allows agents to learn behaviours within their immediate space from individual perceptions. We compare outcomes at various spatio-temporal resolutions with patterns of crime characteristics in agreement with environmental criminology theory and empirically observed patterns, including:

- Offenders committing more offences at familiar locations compared to unfamiliar locations, CPT (Brantingham & Brantingham, 2019) and spatial concentration of crime (Weisburd *et al.*, 1993).
- Offending in areas closer to home, JTC (Rengert, 2002), least effort principle (Florence & Zipf, 1950).

- Victimising the same rewarding targets more frequently, assault reputation (Bosse & Gerritsen, 2008), repeat victimisation (Farrell & Pease, 2001).
- Not offending due to lack of rewards, known as offender discouragement (Clarke & Weisburd, 1994) and representing the variability in offender propensity to offend (Nadal *et al.*, 2010) - with some offenders offending a lot more than others.

4.3.5 RL for Decision-Making

Reinforcement Learning (RL) agents learn by interacting with their dynamic environment. At each timestep, agents perceive the current state of the environment, execute an action, and consequently, the environment transitions to a new state. A reward signal, often conceptualised as target attractiveness, evaluates the quality of each transition. The primary objective for these agents is to maximise their cumulative reward over time through effective interaction (Buşoniu *et al.*, 2010; Islam *et al.*, 2019; Kaelbling *et al.*, 1996; Sutton & Barto, 2018b; Wooldridge, 2020).

The term "behaviourally realistic" refers to the capability of RL agents to emulate complex behaviours observed in real-world entities or systems. This involves agents learning to make decisions and take actions that closely resemble those of humans, animals, or sophisticated systems in specific contexts, based on empirical data. Such capabilities are crucial for applications where realistic simulation of behaviour is essential for understanding, predicting, or enhancing real-world processes.

Applications of RL to create behaviourally realistic agents are varied and span multiple disciplines. For instance:

1. **In Finance**, Dang et al. (2020) demonstrated how RL can be used in algorithmic trading to simulate realistic market strategies that adapt to changing market conditions, mimicking the strategic decision-making of human traders (Dang, 2020).
2. **In Social Science**, Sert et al. (2020) applied RL to model social segregation dynamics, where agents learned to exhibit behaviours that closely resemble patterns observed in empirical studies of social interactions and movements (Sert *et al.*, 2020).
3. **In Urban Planning and Crime**, Joubert (2022) showed that RL could simulate the decision-making processes of offenders operating within South Africa

conducting petty crimes (Joubert *et al.*, 2022).

4. **In Environmental Science**, Liu (2020) reviewed applications of RL in managing natural resources and environmental challenges, where agents learn to make decisions that align with sustainable practices observed in real-world scenarios (Liu *et al.*, 2020).

These examples underscore the versatility and effectiveness of RL in creating agents that not only learn and adapt within a digital environment but also exhibit decision-making patterns that are indistinguishable from those observed in empirical data across various domains.

RL is split into two parts, training and testing. During training (Figure 4.5, A), agents are initialised, some state s of the environment is observed, and an initial action is taken. Objective function (Formula 4.8.3) measures how good the current policy π_θ (a set of state-action pairs) is compared to the previous policy $\pi_{\theta old}$. A reward or penalty is provided to the ANN, which updates future decision policies. For example, if an offender agent offends at a target near home, they will most likely receive a greater reward than offending at a target further from home as effort will be greater. Thus, the offender agent learns to offend closer to their home location as a better choice. Nevertheless, if a very rewarding target exists further away from home, it may also be considered. Thus, during training, offender agents learn to trade off the core measures of risk, reward and effort represented in the model to develop offending preferences. After this initial training phase, trained ANNs are assigned to each agent during testing, and the model is run. These agents use the ANN to infer decisions in the environment and adapt to potential changes. As the environment is stochastic, each run will be dissimilar to the previous; thus, agents should perform the most suitable action. To evaluate the behaviours, output data will be analysed. See Table 4.2, for a list of model outputs.

The PPO algorithm improves training stability by reducing search space when devising a new policy. It does this using a clipped surrogate objective (Queeney *et al.*, 2021; Schulman *et al.*, 2017). These underpinning formulae can be found in Appendix 4.8.2, Formulas 4.8.2 and 4.8.3. A simple illustration of these formulae is found below.

Empirical expectation $\hat{\mathbb{E}}_t$ is the ratio of the difference between the old policy and current policy distributions. $r_t(\theta)$ is greater than 1 when the action is more likely for the current policy than the old policy; it will remain between 0 and 1 when the action is

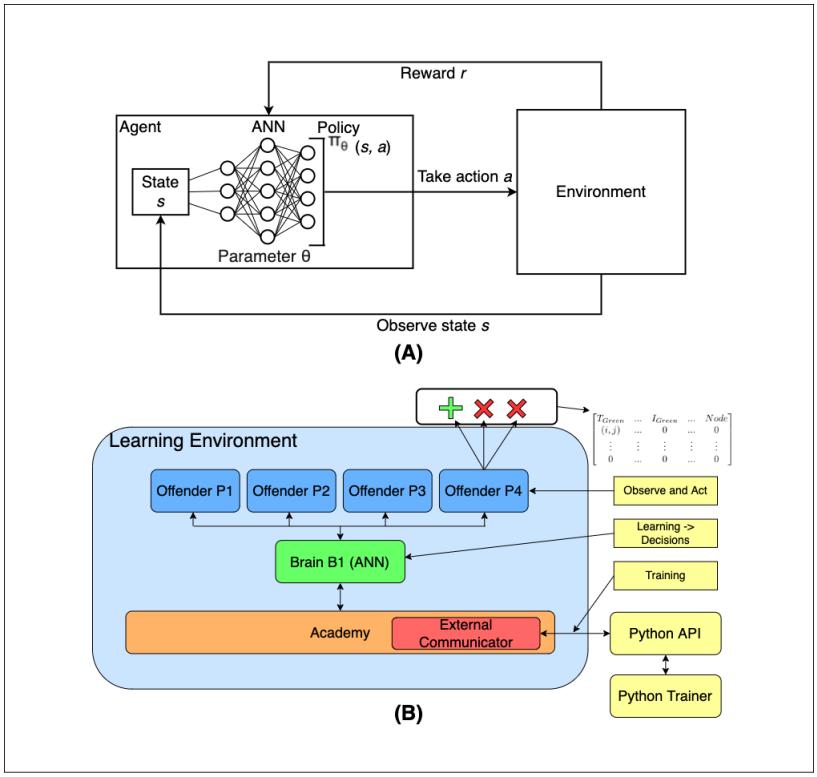


Figure 4.5: Block diagrams, where (A): basic reinforcement learning ANN training architecture. (B): RL life cycle in ml-agents package (Juliani *et al.*, 2018).

less likely for the current than the old policy. $\hat{A}_t$ is the advantage estimate; the higher the value, the better the agent's current actions are than the actions it started with. Thus, Formula 4.8.2, also known as the conservative policy iteration, tries to prevent substantial policy updates, which can cause unstable training outcomes. The L^{CLIP} Formula 4.8.3, known as the clipped surrogate objective, clips the distribution change of the policy ratio between 0.8 and 1.2, where $\epsilon = 0.2$ (Schulman *et al.*, 2017). It takes the minimum of the current policy ratio $r_t(\theta)\hat{A}_t$ and the clipped policy ratio. This removes the incentive for the policy change to move outside the bounds of 0.2, thus minimising instability in policy change, which was a drawback of PPO's predecessor Trust Region Policy Optimisation (TRPO) (Schulman *et al.*, 2015). Ultimately, the extent to which policies change is monotonic; thus, behaviours learnt at timestep t_1 will not significantly differ from those learnt at timestep t_5.

The PPO algorithm (Schulman *et al.*, 2017) uses an ANN to approximate a func-

tion mapping an agent's observations to the best action an agent can take in a given state (Figure 4.8.17). Each offender agent makes observations given localised information (sub-section 4.3.4). These observations train ANNs, which subsequently learn to match scenario/state to action by trying various configurations to maximise an objective function (Schulman *et al.*, 2017; Sutton & Barto, 2018b), Figure 4.5, B.

Proximal Policy Optimisation (PPO), which is the focus of this chapter, has been demonstrated to surpass its competitors in a multitude of evaluated environments. These competitors encompass algorithms such as Trust Region Policy Optimisation (Schulman *et al.*, 2015), Cross-Entropy Method (Szita & Lorincz, 2006), Advantage Actor-Critic (A2C) (Mnih *et al.*, 2016), and A2C with Trust Region (Wang *et al.*, 2016). PPO is commended for its approximation to an optimal policy structure (Martín-Guerrero *et al.*, 2008) and for its efficient data utilisation coupled with robust parallelism capabilities (Chu, 2018). The 'ml-agents' toolkit, introduced by Juliani et al. (Juliani *et al.*, 2018), facilitates the streamlined implementation of such algorithms, enhancing the reproducibility and explainability of resultant model behaviours. Moreover, the substantial citation count of over 8032 for PPO (Schulman *et al.*, 2017) on Google Scholar as of 03/08/2022 is reflective of its broad recognition and utilisation within the literature.

It is appropriate to note, however, that citation counts, while indicative of a paper's influence and the academic community's engagement with the work, are not absolute measures of its importance or correctness. They should be interpreted with caution. The case of Wakefield's study on MMR and autism serves as a stark reminder that high citation numbers can also result from controversy or refutation (Wakefield *et al.*, 1998). The rationale for referencing highly cited work in this domain is to highlight the discussion in well-regarded and extensively scrutinised studies, thereby ensuring academic rigour.

4.4 Results

Having described the model, we now set out a series of experiments conducted using it, where environmental parameters are manipulated (i.e., distribution of rewards and interventions) and offender agents are trained. Once offender agents are trained, they are tested post-training, where output data are used to evaluate whether crime patterns change as the environment changes.

To recap, in this chapter we aimed to explore three primary questions:

1. Do offender agents utilising RL portray behaviours in agreement with RCP, i.e., to what extent do they learn to offend when rewards outweigh risk and effort and vice versa?

2. Do offender agents utilising RL adapt to changes in their immediate environment given the introduction of simulated crime prevention interventions?

3. Do simulated crimes generated by offender agents utilising RL display patterns commonly observed in empirical studies of crime?

The RCP (Clarke & Cornish, 1985; Cornish & Clarke, 1987, 2017) suggests that offenders act in a particular way with regards to target selection. If the offender agents in our model act per RCP, we would expect to see the following:

- Offender agents will learn to commit offences at targets where target attractiveness is > 0 (i.e., where the rewards associated with victimising a particular target outweigh the risks and the effort involved in doing so).

- Conversely, offender agents will learn not to commit offences at targets where target attractiveness is < 0 (i.e., where risk and effort combined outweigh rewards).

The two spatial environment localities described in Section 4.3 are setup as treatment (left) and buffer (right) areas.

Experimental conditions are outlined in Table 4.1.

- **50 episodes** are run for each experiment.

- Each episode consists of **2000 discrete timesteps** (an episode is one execution of the model, and timesteps are the duration of that execution).

- Due to stochasticity, **10 simulation batches** were operated for each experiment condition with **50 simulations per batch** leading to 500 instances per experiment.

- In total, **500 episodes** per experiment condition were analysed.

- Model environment is made up of **100x100 grid** configured as a box.

- **1/4 of all cells** (2500) contain a potential target representing a residential property.
- 1% of these targets are offender agent homes; therefore **25 offender agents** are instantiated.
- Each offender agent has **5 routine activity nodes**.
- A total of **500 navigational nodes** are distributed.

The above list describes the instantiated state for each episode. In both the training and testing phase, the model simulates the setup of SCPIs (Clarke, 1980, 1997b) at a specific temporal point as a given simulation is running. In episode 25/50, **1250 interventions** are introduced into the treatment area. Thus, every target will have at least one intervention within one of the eight adjoining cells. These interventions test the adaptability of offender agents post-training under two significantly different environmental conditions to observe the impact increased risk has on learned behaviours.

Experiment Condition	Interventions[1]	Distribution of Target Rewards	Target Cumulative Reward
1	1250	1	5
2	1250	0	5
3	1250	U[0,1][2]	5

Table 4.1: Model experiment parameters for three conditions.

Output data (Table 4.2) from each experiment condition (Table 4.1) generates 25 rows of data (each offender agent) 2000 times (one for each timestep), capturing 2,500,000 rows of data per simulation across ten repeated simulation conditions (n = 25,000,000), these can be used to analyse individual behaviours of offender agents throughout a model experiment. With the utility of visualisations, these data evaluate how offender agents adapt to their environment, help to assess if they learn to behave per the RCP and reveal insights into heterogeneous behaviours. The following sub-sections illustrate each experiment condition and findings.

The findings from Figure 4.6 and summary statistics in Table 4.5 suggest a relatively narrow spread in the means of Target Attractiveness, standard deviation and coefficient of variation across different simulation runs for the same experiment conditions. This can be indicative of a few key points in the context of assessing the adequacy of the simulation runs:

Column Name	Type	Description
AgentID	Integer	A unique agent identifier.
Action	Categorical	The current action an agent has chosen, can be one of [OFFEND, DON'T OFFEND, MOVE].
Area	Categorical	The locality in which the above action has taken place.
Target_Attractiveness	Float	The target attractiveness value of the victimised property.
Target_Reward	Float	The victimised property reward.
Target_Risk	Float	The risk surrounding the victimised property.
Target_Effort	Float	The effort of the victimised property by the specific offender agent.
Total_Cumulative_Reward	Float	The total Target_Attractiveness acquired by the offender agent.
xAxisPos	Integer	The x-axis position of the cell the offender agent is currently in.
zAxisPos	Integer	The y-axis position of the cell the offender agent is currently in.
Zone_Travelled_To	Categorical	The locality the offender agent is currently travelling to.
Episode	Integer	The current episode.
Distance_To_Home	Float	The normalised Euclidean distance to the offender agent's home node from the victimised property.
Distance_To_Next_Node	Float	The normalised Euclidean distance to the next routine activity node from the victimised property.
Timestep	Integer	The current discrete time point.
Target_Cumulative_Reward	Float	The total amount of Target_Attractiveness the offender agent aims to achieve.

Table 4.2: Model output data, type and description.

Figure 4.6: Variation of mean across 10 batch runs, where (A)-(C) are Experiments 1 to 3 respectively.

- Stability and Consistency: The narrow range implies that the mean 'Target Attractiveness' is relatively stable and consistent across the different runs. This stability is a good sign that the number of runs (10) and episodes per run (50) might be sufficient to capture the inherent variability of the model.
- Low Variability Between Runs: The small variance between the runs suggests that increasing the number of runs might not significantly change the overall results, indicating that our choice of 10 batches of 50 simulation runs was potentially a good balance between accuracy and computational efficiency.

4.4.1 Experiment 1 - uniform distribution of rewards (1)

Before we delve into the experiment, it is worth defining a crucial mechanism in the experiments, namely 'interventions' within the context of this model refers to situational crime prevention measures designed to alter the environment in a way that deters criminal behaviour. These interventions increase the perceived risk of offending at a target location, thereby reducing its attractiveness to offenders. They are conceptualised based on situational crime prevention strategies which focus on making crime harder, riskier, and less rewarding to commit.

In this experiment, *Rewards* at each target is uniform set to 1, therefore, $0 \leq Target\_Attractiveness \leq 1$ pre-intervention and $-1 \leq Target\_Attractiveness \leq 1$

[1]These interventions are introduced at the 25^{th} episode.
[2]A uniform distribution of rewards [X, Y] inclusive.

post-intervention. If offender agents have learnt the RCP, they should offend substantially more pre-intervention and less post-intervention. Furthermore, post-intervention crime should spatially concentrate in the buffer area due to higher rewarding opportunities.

When analysing the results, we found that the spatial distribution of the number of offences is in agreement with the RCP. A high level of crime concentrates centrally pre-intervention (Figure 4.7, A). However, once interventions are introduced and *Risk* increases (treatment area), offences concentrate in the buffer area (Figure 4.7, B). Small pockets of offences in the treatment area still occur. In Figure 4.8.18a we observe some offender agents offending proportionally more post-intervention than they did pre-intervention. 11 (44%) offender agents committed proportionally more offences post-intervention in the treatment area compared to pre-intervention episodes (least deterred by SCPI (Clarke, 1980)). In contrast, 14 (56%) offender agents committed more offences pre-intervention than post-intervention in the treatment area (deterred by SCPIs). Overall, these early indicators show some level of behavioural heterogeneity among offender agent decision-making, showing signs of adaptation to environmental changes Figure 4.8.18a, B.

These results alone do not prove offender agents behave in ways that would be characterised by RCP. The target attractiveness at each target over time must be quantified and compared to validate behaviours. In Figure 4.7, D we observe a concentration of positive target attractiveness pre-intervention. If we compare these patterns to Figure 4.7, A, offences clustered at the centre and decreased as we branch out. When comparing Figures 4.7, A-B and D-E, offences were taking place more frequently in locations with greater target attractiveness compared to locations with lower target attractiveness. Showing signs of spatial concentration of crime (Brantingham & Brantingham, 1995; Weisburd *et al.*, 1993).

These results (Figures 4.7, A-B and D-E) indicate that offender agents have learnt to identify targets where rewards outweigh risks and effort - and when SCPIs are introduced to increase risk, making some targets less desirable than others, offender agents adapt to reduce their offending. However, how has this "best case" scenario for offender agents where every target has a relatively large amount of reward impacted total cumulative reward for each offender agent? Given the frequency of offences, we could expect most offender agents to achieve their TCR.

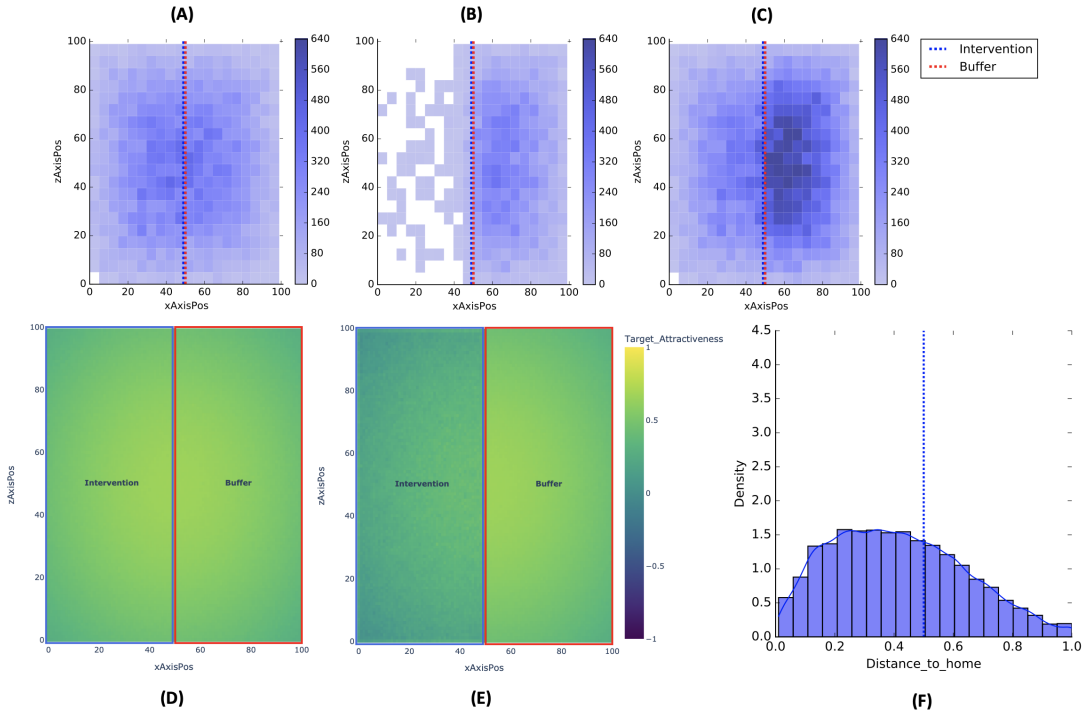


Figure 4.7: Distributions of offences, average target attractiveness and distance between home and offence locations, where (A-C): offences across targets pre, post and pre-post merged. (D-E): target attractiveness pre and post-intervention, respectively. (F): distance between home and offence locations (Intervention = Treatment area).

These results show offender agents successfully learned to offend at targets where rewards outweighed risk and effort, evidenced in Figure 4.8. This is true for both pre and post-intervention; however, post-intervention accumulated rewards drastically dropped as high rewarding opportunities decreased.

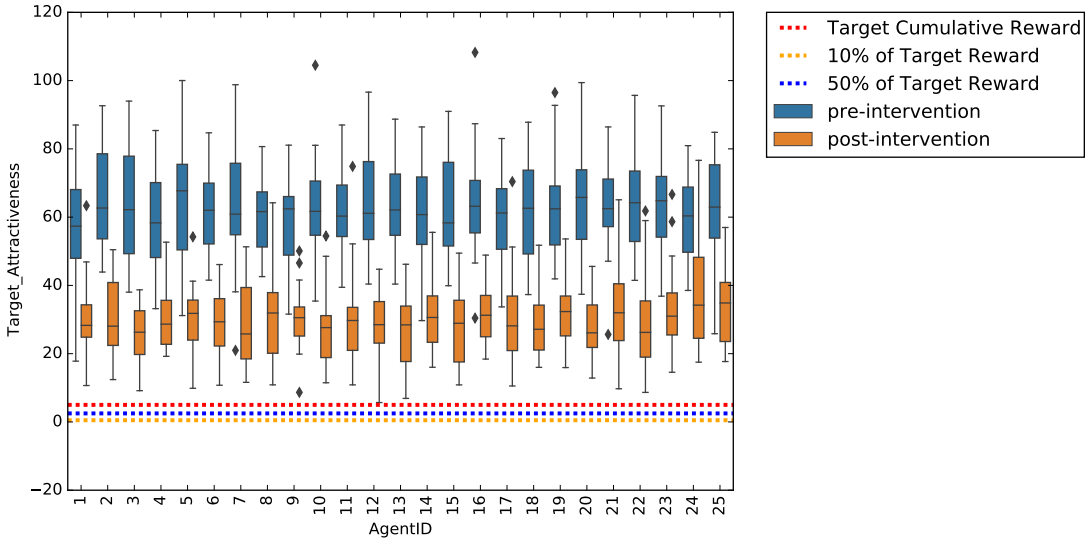


Figure 4.8: The cumulative reward distribution for each offender agent pre-post intervention across all episodes.

There is consensus among some scholars that the majority of offences take place in areas most common to an offender, such as places near home (Baudains *et al.*, 2013; Brantingham & Brantingham, 2019; Rengert, 2002). As our results are in agreement with patterns of crime described by RCP, presumably, offender agents make rational decisions and offend near their home (agreement with the least effort principle (Florence & Zipf, 1950)) rather than in less familiar places. Figure 4.7, F depicts the average JTC across each offender agent over all ten simulation runs. It shows offender agents have learnt to offend closer to home (positive skew), where the JTC for each offender agent is also positively skewed.

These early indicators present the "best case" scenario for offender agents, which given its unrealistic nature (a uniformly high level of rewards distributed across space), demonstrates high levels of crime, some risk-taking and spatial clustering of offences. Conversely, what happens if no target offers any reward? When risk and effort outweigh rewards at all targets, offender agents should rationally decide not to offend (according

to the RCP).

4.4.2 Experiment 2 - uniform distribution of rewards (0)

In this experiment, rewards are set to 0, therefore, $-1 \leq Target\_Attractiveness \leq 0$ pre-intervention and $-2 \leq Target\_Attractiveness \leq 0$ post-intervention. We explore if offender agents will change their behaviours and learn the rational decision not to offend as risk and effort outweigh rewards.

Results show (Figures 4.9, A-C) that the frequency of crime has drastically dropped. The majority of offender agents, fourteen (56%), chose not to offend; this can be observed in Figure 4.8.18b. Seven (28%) offender agents had committed at least one offence in the buffer area, while eighteen did not (72%). Four (16%) offender agents committed at least one offence post-intervention in the buffer area. In the treatment area, post-intervention, six offender agents committed at least one offence (24%), three offender agents committed an offence pre-intervention (12%) refer to Figure 4.8.18b, B. Sixteen offender agents did not offend in the treatment area (64%). The average number of offences pre-intervention was 2.54 per offender agent. Similarly, the average number of offences per agent post-intervention was 1.62. These small pockets of offences may have transpired from stochasticity. However, these results are significant as the overall pattern suggests offender agents have learnt to commit near-zero offences when risk + effort outweigh rewards.

Figures 4.9, D-E show that RCP was correctly followed in this instance as the number of no offence decisions is greater than offence decisions. Thus, offender agents learned that not offending was better.

The spatial distribution of target attractiveness shows no target contained $Target\_Attractiveness > 0$ as evidenced in Figures 4.9, G-H.

Due to the lack of rewarding targets, TCR could not be met. Thus, we expect those who offended to have a negative accumulated reward. The data in Figure 4.10 indicates that the accumulated reward is near zero for every offender agent with an average total cumulative reward of -0.38.

We have shown that offender agents adapt to two different environmental configurations and learn decisions in agreement with RCP. When every target contains a high reward, offender agents commit high number of offences. Conversely, offender agents learn not to offend when targets contain no rewards. The most impressive finding was

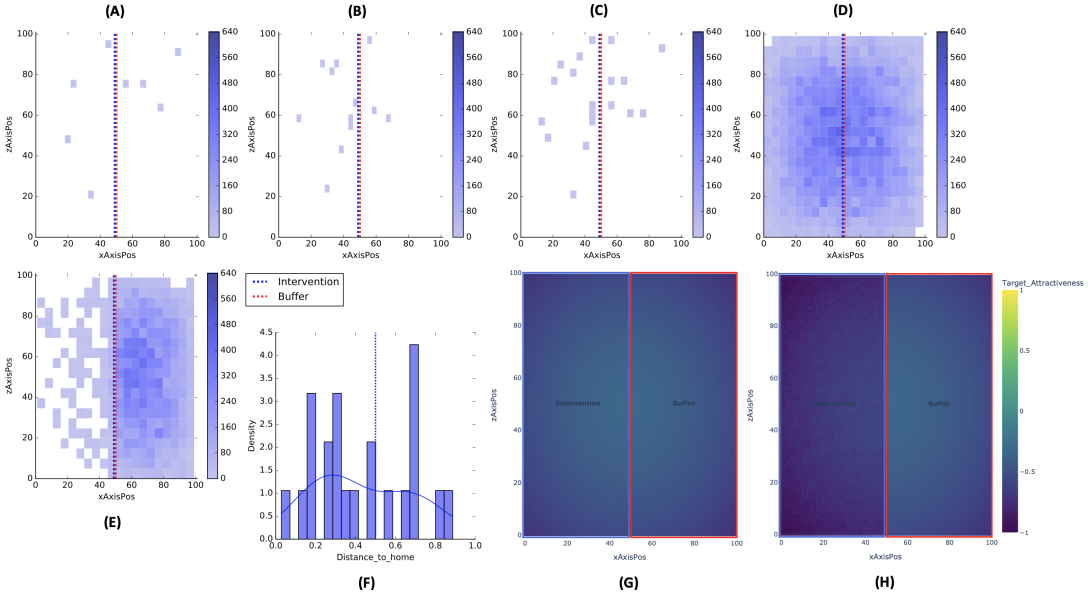


Figure 4.9: Distributions of offence and not to offend target locations, the distance between home and offence locations and average Target_Attractiveness, where (A-C): offences across targets pre, post and pre-post intervention merged. (D-E): no offence decisions across targets pre and post-intervention, respectively. (F): distance between home and offence locations. (G-H): Target_Attractiveness pre and post-intervention, respectively (Intervention = Treatment area).

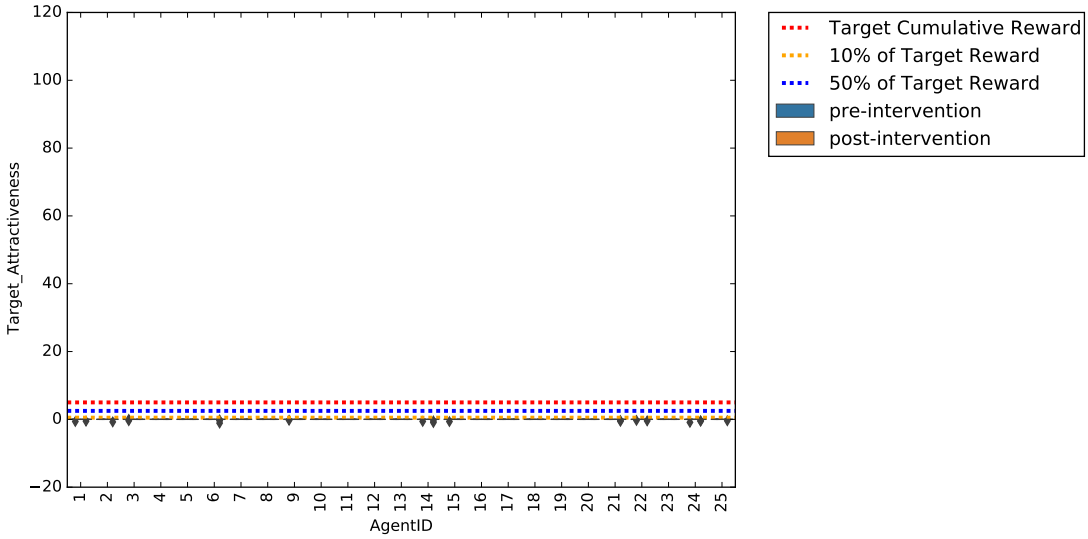


Figure 4.10: The cumulative reward distribution for each offender agent pre-post intervention across all episodes.

that some offender agents adapt to the environmental changes better than others during simulation run-time. These results demonstrate that RL offender agents can adapt their behaviours (by enabling agents to learn (Ramchandani *et al.*, 2017)) when SCPIs are introduced. The model produced mainly rational (majority of offender agents mainly offending in the buffer area) with some examples of irrational offender agent decisions (some offender agents continue to offend in the treatment area when risk increases).

In contrast, at the time of writing this chapter, there are no alternative condition-action frameworks that can achieve behavioural learning where agents can reflect on perceived past experiences and update their rules (behaviours) internally to adapt to new situations (previously unseen situations) which is an essential aspect of human cognition (Jipp, 2007; Sternberg & Gastel, 1989; Wong & Candolin, 2015). As the simulated environments present highly unrealistic scenarios in which all rewards are high or all are low, this does not reflect the real world; these are merely best-case and worst-case situations for offender agents. Presumably, target rewards would differ from place to place in the real world; thus, perceived wealth would vary from target to target.

4.4.3 Experiment 3 - random distribution of rewards ([0, 1])

In the final experiment, the reward at each target is randomly distributed between 0 to 1 inclusive. Therefore, $-1 \leq Target\_Attractiveness \leq 1$ pre-intervention and $-2 \leq Target\_Attractiveness \leq 1$ post-intervention. We expect some offender agents to achieve their TCR while others, on average, will not. Therefore, a more diverse range of TCRs per-agent should be observed.

Figures 4.11, A-C show that spatial patterns are similar to those in MC1 Figures 4.7, A-C. This would be expected as at least half of the targets will have $Target\_Attractiveness > 0$. Contrary to expectation, half of the treatment area has had fewer offences than the other half Figure 4.11, A. Upon detailed analysis, the average target attractiveness at this area (bottom left) was 0.05 where offences were committed. The second half (top left) was 0.09, and the buffer area was 0.1. Therefore, offender agents found less rewarding opportunities in the bottom left half of the treatment area pre-intervention, where 47% of targets had negative attractiveness on average. We expect offender agents to choose not to offend more frequently in these areas, as observed in Figure 4.11, D, focusing on more rewarding surrounding locations.

For reference, the average target attractiveness across episodes, Figures 4.11, G-H when compared with Figures 4.11, A-C show offender agents offending more frequently in locations maintaining positive target attractiveness compared to less rewarding locations.

Individual-level data show a drop in accumulated reward post-intervention. However, both pre and post-intervention accumulated reward is closer to zero Figure 4.12 compared to MC1, Figure 4.8.

These results show offender agents have acquired an average total cumulative reward greater than TCR pre-intervention Figure 4.12. Sixteen offender agents (64%) acquired an average total cumulative reward $\geq$ TCR post-intervention. In contrast, four offender agents acquired less than 50% of TCR post-intervention. In the buffer area pre-post intervention, the proportion of offences per offender agent share a similar pattern. However, some offender agents have committed more offences post-intervention than pre-intervention, demonstrating minor signs of heterogeneity Figure 4.8.18c, A. The proportion of offences per offender agent in the treatment area pre-intervention is dissimilar to post-intervention Figure 4.8.18c, B. Some offender agents offend substantially more post-intervention than they did pre-intervention and vice-versa. When

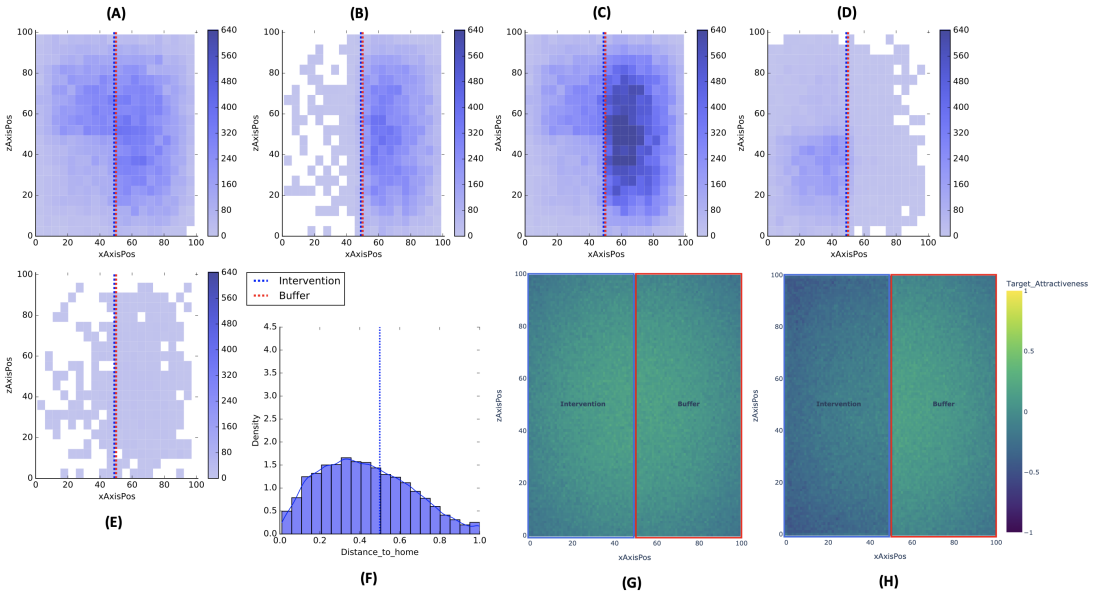


Figure 4.11: Distributions of offence and not to offend target locations, the distance between home and offence locations and average Target_Attractiveness, where (A-C): offences across targets pre, post and pre-post intervention merged. (D-E): no offence decisions across targets pre and post-intervention, respectively. (F): distance between home and offence locations. (G-H): Target_Attractiveness pre and post-intervention respectively (Intervention = Treatment area).

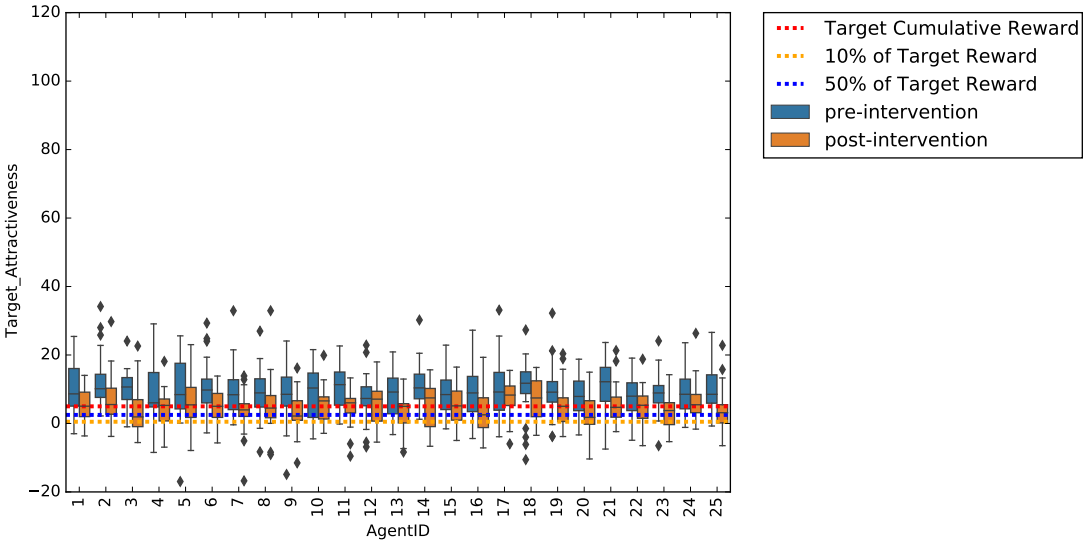


Figure 4.12: The cumulative reward distribution for each offender agent pre-post intervention across all episodes.

SCPIs are adopted, offender agent decision-making becomes more heterogeneous.

Model Condition	**Buffer Area**	**Treatment Area**	**Difference**$(+/-)$	**t-test**(p)[3]
1	(711, 214.11)	(574, 188.91)	-137	2.40(.02 $p < 0.05$)
2 [4]	(0, 0.0)	(0, 0.0)	-	-
3	(669, 196.86)	(615, 155.70)	-54	1.08(.28 $p > 0.05$)

Table 4.3: The (mean, std) of offences in the buffer and treatment areas by offender agents living in these areas across post-intervention episodes, including the mean difference (where - means drop and + means increase).

These results demonstrate that RL as a behavioural framework can incorporate complex adaptive decision-making in an environmental criminology ABM. These findings could allow those who model crime dynamics to utilise ABMs better reflective of real-world offender decision-making to support SCPI planning. Furthermore, these results show behavioural heterogeneity within offender agents' decision-making (exacerbated by SCPIs). Exhibiting impact of environment on learning, i.e., some offender

[3] H_0: that two independent samples have identical average (expected) values. H_a: the means of the distributions underlying the samples are unequal.

[4] No offences occurred in the buffer or treatment areas by local offenders for MC2.

agents living closer to or in the treatment area commit fewer offences than those living elsewhere, as observed in Table 4.3. Consequently, some offender agents can be "spatially better suited" to offending than others, evidenced by their net gains in Figure 4.12 and proportion of offences in Figure 4.8.18c. Some offender agents consistently maintained high crime levels in the treatment area across all post-intervention episodes Figures 4.8.18c, B.

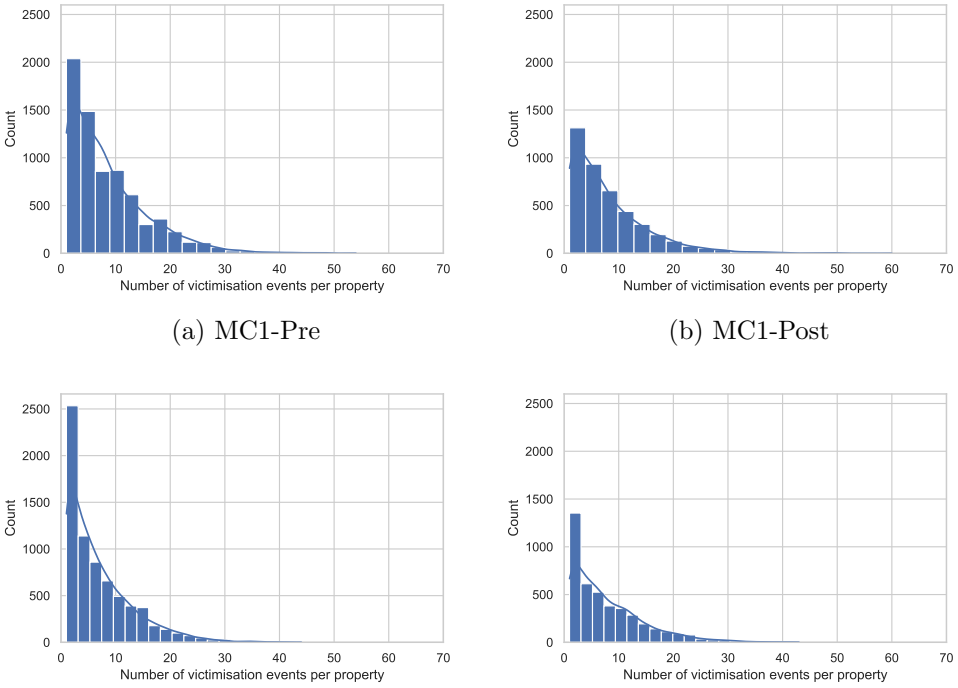
Overall, JTC (Rengert, 2002) is positively skewed, where offender agents learnt to minimise effort by offending closer to home Figures 4.7, 4.9, 4.11, F. Crime Concentration (Farrell, 2015; Weisburd *et al.*, 1993) is clustered to areas maintaining greater target attractiveness such as the centre and buffer areas Figures 4.7, A-C, 4.7, D-E, 4.11, A-C and 4.11, G-H. When no rewards exist, decision to not offend is vast and clustered in the centre Figures 4.9, D-E (similar empirical patterns of SCPI found in (Eck & Clarke, 2019)).

Crime does not spatially concentrate when rewarding opportunities are non-existent Figures 4.9, A and G.

Most importantly, our complex agents operating under RL produce simulated crime patterns that share some characteristics with empirical crime patterns as described in (Eck & Liu, 2004). For example, crime patterns should be clustered, crime concentrated in a few places, few victims accounting for most of the victimisation as shown in Figure 4.13 which was also observed in the following articles (Stokes & Clare, 2019; Tillyer *et al.*, 2018), journey to crime is typically short (positively skewed), and lastly, non-static patterns of crime over time.

4.5 Discussion and conclusion

This chapter presents an ABM investigating spatio-temporal dynamics of burglary. The purpose of the model is to overcome some of the behavioural deficiencies in ABMs of environmental criminology highlighted in literature (Groff *et al.*, 2019; Johnson & Groff, 2014). Our rationale in doing so is to support those engaged in modelling offender behaviour and occurrence of crime events with more accurate models of behaviour. Pre-existing literature in this domain neglected behavioural learning in modelling crime dynamics, as pointed out by (Johnson & Groff, 2014). The implications led to process models requiring pre-determined behaviours, which sometimes bound behaviour and lead to unrealistic outcomes (Arthur, 1994; Cornelius *et al.*, 2017; Manson, 2006)—re-



(a) MC1-Pre

(b) MC1-Post

(c) MC3-Pre

(d) MC3-Post

Figure 4.13: Number of victimisation events per residential property (n = 177,129, bins = 20) pre-post intervention episodes for experiment conditions one and three.

sulting in a deficiency of realism in offender behaviour, such as adapting to change. These decision-making frameworks fail to capture crucial components of real-world offender decision-making, such as the ability to learn behaviours, adapt to environmental stimuli and make goal-oriented decisions concerning target selection (Gialopsos & Carter, 2014; Sigurdsson *et al.*, 2008; Topalli, 2005). To provide a solution to the problem above, we proposed several research aims:

1. Do offender agents utilising RL portray behaviours in agreement with RCP, i.e., to what extent do they learn to offend when rewards outweigh risk and effort and vice versa?

2. Do offender agents utilising RL adapt to changes in their immediate environment given the introduction of simulated crime prevention interventions?

3. Do simulated crimes generated by offender agents utilising RL display patterns commonly observed in empirical studies of crime?

To achieve these aims, we developed a multi-agent RL approach to modelling offender behaviour. We showed how a neurologically inspired decision-making framework such as RL (Niv, 2009; Sutton & Barto, 2018b) (with the recent notable exception of (Joubert *et al.*, 2022) which was critiqued in sub-section 4.2.2) could integrate with ABMs in environmental criminology, by which, offenders can learn behaviours, adapt to SCPIs and try to fulfil personal goals such as a certain amount of wealth (TCR). We designed three experiments where offenders are tested in stochastic environment configurations, where SCPIs (Clarke, 1980; Eck & Clarke, 2019) are introduced mid-simulation and the reactions of offenders captured. We found that in all three experiments, offenders naturally learned to abide by the RCP criteria where they learned to offend at targets where rewards outweighed effort and risk in enacting the offence (Cornish & Clarke, 1987) and vice versa.

Experiment Condition	**Pre-Intervention**	**Post-Intervention**	**Difference**$(+/-)$	**t-test**(p)[5]
1	(2532.12, 96.96)	(1274.26, 266.28)	-1257.86	$21.98(p < 0.00)$
2	(1.5, 1.22)	(1.3, 0.48)	-0.19	-[6]
3	(2078.08, 104.08)	(1282.80, 180.58)	-795.28	$19.02(p < 0.00)$

Table 4.4: The (mean, std) of the number of offences committed for each experiment condition across pre-post intervention episodes.

This chapter shows that a model inspired by the three theoretical tenets of environmental criminology RAT, RCP and CPT (Brantingham & Brantingham, 2019; Cohen & Felson, 1979; Cornish & Clarke, 1987) and implemented using RL produced a number of outcomes compatible with empirical patterns of crime, such as, spatial concentration of crime (Brantingham & Brantingham, 1995; Weisburd *et al.*, 1993), journey to crime curve (Rengert, 2002), assault reputation (Bosse & Gerritsen, 2008), offender discouragement (Clarke & Weisburd, 1994) and repeat victimisation (Farrell & Pease, 2001). The model successfully showed signs (to varying degrees) of all crime patterns above, exposing some level of model validity.

Whilst possessing incomplete knowledge, such as rewards offender agents gain from committing an offence before acting out the offence, RL offender agents applied learned intelligence using ANNs and made heterogeneous decisions in agreement with RCP. This decision-making process led to generative crime patterns consistent with those found in empirical studies of burglary from various disciplines (Eck & Liu, 2004; Piquero & Rengert, 2006; Short *et al.*, 2011; Vandeviver *et al.*, 2015).

Most ABMs in environmental criminology adopt decision-making frameworks with logical foundations without cognitive faculties such as learning and adapting. A recent example, Cornelius *et al.* (2017) adopted a set of discrete rules, for example, "count nearest four agents" then "are three nearest four agents criminals DO the-following". These frameworks constrain models, as agent behaviour lacks crucial mechanisms such as the ability to learn from individual perceptions. Most studies have previously used similar condition-action rules (Birks *et al.*, 2012; Bosse & Gerritsen, 2008; Bosse *et al.*, 2011; Groff, 2007).

In this chapter, offender agents learned to adapt to spatial changes and, in doing so, developed adaptive behaviours. Introduction of interventions at a later time point (in the treatment area) led most offender agents to adapt their behaviours clustering at the buffer area (Figure 4.7, B) as this location maintained higher levels of rewarding opportunities (Figure 4.7, E). However, when opportunity is removed, crime drops (Figures 4.9, A-C) and does not concentrate; in fact, offender agents choose to stop committing offences (Figures 4.9, D-E). Similar outputs were found in other studies (Bernasco & Luykx, 2003; Eck & Clarke, 2019; Levy *et al.*, 2014; Short *et al.*, 2011) where lack of

[5] H_0: that two independent samples have identical average (expected) values. H_a: the means of the distributions underlying the samples are unequal.

[6] t-test could not be applied to condition two as the number of offences was small.

opportunity in an area successfully removed hot spots in those areas. When complexity increased, and rewards were no longer uniform, offender agents continually performed actions in agreement with RCP. When high rewarding opportunities decreased Figure 4.11, H, offender agents adapted to either not-offend Figures 4.11, D-E or continued offending in the buffer area Figure 4.11, B. Consequently, a heterogeneous pattern of learned behaviour emerged where some offender agents achieved their target reward while others did not Figure 4.12. A small number of victims experienced high levels of victimisation Figure 4.13, which agrees with findings from repeat victimisation literature (Farrell & Pease, 2001; Stokes & Clare, 2019; Tillyer *et al.*, 2018). These promising results highlight the wealth of opportunity for future environmental criminology ABMs.

The simulated experiments demonstrate that "lack of opportunities" where target attractiveness is ≤ 0 is more effective at reducing crime numbers compared to the adoption of SCPIs (this can be observed in Table 4.4 comparing experiment conditions 2 with 1 and 3). Although SCPIs have been very effective in deterring crime, for experiment conditions 1 and 3, crime dropped on average by 49.67% and 61.73% post-intervention, respectively.

A notable limitation experienced and worthy of discussing was a lack of accessible computational capabilities, given the complex nature of training a set of agents in millions of iterations (Ding & Dong, 2020; Faghri, 2021; Farkas *et al.*, 2020). This, unfortunately, led to a computationally tractable yet scaled-down version of model development.

Findings from this chapter in future will explore how SCPIs (Clarke, 1980) might affect offending behaviour in real-world cities. The questions to be answered are: does crime displace? If so, where does it displace? Is the alternative, diffusion of benefits, a more likely outcome as suggested by most researchers (Barr & Pease, 1990; Guerette & Bowers, 2009; Wortley, 2016). If a predictive ABM can utilise RL agents to learn displacement patterns, this could lead to accurate snapshots of where crime is likely to displace and which localities are likely to benefit from interventions supporting strategic decision-making.

This chapter shows that biologically inspired decision-making frameworks can help produce agents that learn theoretically sound and empirically valid behaviours through incomplete knowledge of the world without pre-defined behavioural rules. Moreover, these simulated behaviours can produce outcomes in keeping with multiple indepen-

dent empirical patterns that result from real-world behaviours. We believe that this approach to modelling crime may foster a range of new research opportunities and hopefully provide new and unique insights into the dynamics of crime - not only supporting environmental criminologists but ultimately those who seek to reduce crime.

4.6 Summary

This chapter demonstrated the applicability of RL to an ABM that reflects a real-world domain and problem. The investigation has proven that empirically observed and theoretically studied behaviours can organically be learnt by offender agents without a single hard-coded pre-determined decision rule. The investigation has shown the value of utilising these RL approaches in modelling emergent complex behaviour, moreover, given that in the real-world people sometimes make irrational decisions, the proposed model has also reflected this where some offenders have continued to offend in areas that have heightened security and risk associated with them.

The following chapter 5 will now investigate the practical usefulness of RL in making autonomous decisions to alleviate exogenous shocks to a housing market. The proposed approach assesses the capabilities of an RL "central bank" agent to learn real-time trends from a housing market ABM, such as house price fluctuations, number of entrants, the homelessness numbers, the number of properties in negative equity and so on, it should then be able to adjust measures like the interest rate to learn how it can tweak these measures to fulfil some goal criteria, such as a maintaining a certain level of homelessness and not allowing it to rise. The objectives for this investigation are to replicate a well-known, highly cited classic housing market model in the python programming stack. The second objective is to then validate the new model against the previous version, i.e., are the behaviours from the model similar or do they deviate. The third objective is to identify a "healthy housing market state" which the RL agent will try to maintain, prior to and during a shock to the market, such as an abrupt change. Lastly, the results are analysed which demonstrate the practical usefulness of RL in controlling and maintaining some state during a stochastic event that can if not addressed cause serious economic harm to people.

The three main chapters of this thesis tell a story, the first chapter of that story is the testing of RL and identifying a suitable software pipeline to conduct the subsequent research, some interesting results were observed from this activity (chapter 3).

The second chapter adopts RL in a real-world domain area, to observe the strengths and weaknesses in learning emergent complex behaviours that are empirically and theoretically sound. This activity shows that RL can be of use to environmental criminology research and in the future, these learning methods will be adopted by practitioners and policymakers to simulate behaviours that are more realistic (chapter 4). The third and final chapter demonstrates how RL can be practically useful in managing a market autonomously by learning from data and trying to adapt a market to achieve some policy such as the "healthy housing market state" (chapter 5).

4.7 Open-source model access

The agent-based model can be found at the following source with documentation CLICK HERE, accessed on 01/08/2022. The datasets and jupyter notebook used for analyses in this chapter can be found at the following link: CLICK HERE, accessed on 02/08/2022.

4.8 Appendix

4.8.1 Formal definitions

The logic behind risk, effort and reward

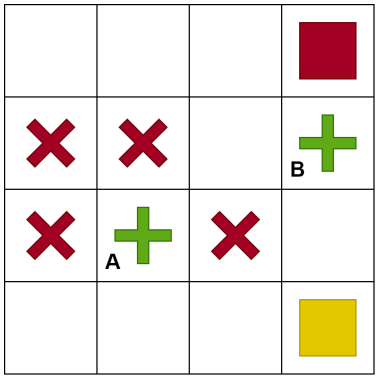


Figure 4.14: An Example Scenario, where 4 Interventions, 2 Targets, 1 Node and Routine Activity Node.

The $T_i(Risk)$ value depends on the number of interventions I surrounding the 8

cardinal directions of a target i. If Interventions $= 0$ then $T_i(Risk) = 0$. In relation to Figure 4.14, $T_A(Risk) = 4/8$ which would give $T_A(Risk) = 0.5$ and for $T_B(Risk) = 1/8$ which means $T_B(Risk) = 0.125$. Therefore, $Risk$ at any target can be at most 1.

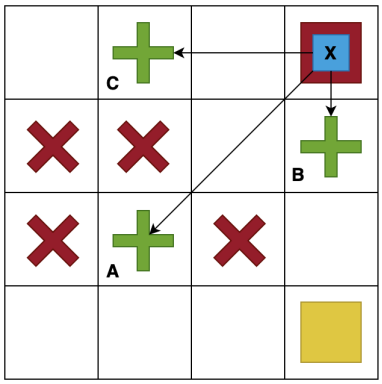


Figure 4.15: An Example Scenario, where 4 Interventions, 3 Targets, 1 Node, Routine Activity Node and Offender Agent.

Given Figure 4.15, lets assume $Distance(Offender_X, T_A) = 3$, $Distance(Offender_X, T_B) = 1$ and $Distance(Offender_X, T_C) = 2$ where:

$$Offender_X(T_A(effort)) = (3-1) \div (3-1),$$

and

$$Offender_X(T_B(effort)) = (1-1) \div (3-1),$$

and

$$Offender_X(T_C(effort)) = (2-1) \div (3-1),$$

This leads to:

$$Offender_X(T_A(effort)) = 1,$$

and

$$Offender_X(T_B(effort)) = 0,$$

and

$$Offender_X(T_C(effort)) = 0.5$$

Given Figure 4.16, lets assume:

$$Area_g(T(reward)) = [0, 0.5],$$

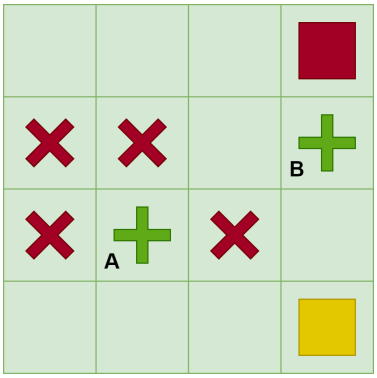


Figure 4.16: An Example Scenario, where Area is Green, 2 Targets, 4 Interventions, 1 Node and Routine Activity Node.

this would mean, $0 \leq T_A(Reward) \leq 0.5$ and $0 \leq T_B(Reward) \leq 0.5$. The three scenarios described above are purely for demonstrating the logic behind the conceptualisation of Reward, Effort and Risk values of target attractiveness.

Offender movement

At simulation start, offender agent A starts at home $Offender_A(RAN_H)$, a new node is selected from a set of routine activity nodes minus the home node, lets call this i, where, $RAN_i \in Offender_A(RAN) - Offender_A(RAN_H)$. The offender agent A then picks the next cell within the shortest path Euclidean distance to RAN_i, and moves into that cell; this process continues until the offender agent A arrives at i.

Offend & dont_offend

An offender agent A offends or does not offend at some target T_i, where:

$$Offend(Offender_A, T_i) \vee DontOffend(Offender_A, T_i),$$

only if A lands in the same grid cell as target i:

$$Position(T_i) = Position(Offender_A),$$

once an offence has been committed the offender agent is rewarded the $Target\_Attractiveness(T_i)$ as a reward. If this value is negative, then the $risk + effort$ outweighed the $reward$ (undesirable outcome). If it is positive then the $reward$

outweighed the $risk + effort$ (desirable outcome). If the offender agent does not commit an offence at target i then

$$DontOffend(Offender_A, T_i),$$

is only desirable if target attractiveness is less than 0. If, the target attractiveness is greater than 0 then not offending here is undesirable as the offender agent is no longer maximising utility. Therefore, we expect offender agents to learn to offend when:

$$Target\_Attractivness(T_i) > 0,$$

and don't offend when

$$Target\_Attractivness(T_i) < 0.$$

Offender perception

There are 5 spatial objects that offender agents can identify in the model, these are:

$$objects = T_{Green}, T_{Orange}, I_{Green}, I_{Orange}, Node$$

where T_{Green}, I_{Green} are Targets and Interventions located within the Green area. All Nodes including routine activity nodes are identifiable. An offender agent observation can only consist of specific object(s) information, only if objects fall within its line of sight (vision):

$$Distance(Offender_X, Object_j) \leq Offender_X(Vision(Length))$$

if this is not the case, then at some time t:

$$Offender_X(Observation_t) = \begin{bmatrix} T_{Green} & ... & I_{Green} & ... & Node \\ 0 & ... & 0 & ... & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & ... & 0 & ... & 0 \end{bmatrix},$$

if an observation is made and objects fall within the offender agents line of sight, then at time t a matrix containing information about the current state of the visual perception

is captured:

$$Offender_X(Observation_t) = \begin{bmatrix} T_{Green} & ... & I_{Green} & ... & Node \\ (i,j) & ... & 0 & ... & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & ... & 0 & ... & 0 \end{bmatrix},$$

where at time t the offender agent was able to identify a target within the green area T_{Green}, where $i = 1$ indicates object perceived or $i = 0$ no object and j is the normalised distance from $Offender_X$ to the object, where $0 \leq j \leq 1$. Each row in the matrix represents the individual sensor.

4.8.2 Formulae

$$L^{CPI}(\theta) = \hat{\mathbb{E}}_t \left[\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta old}(a_t|s_t)} \hat{A}_t \right] = \hat{\mathbb{E}}_t \left[r_t(\theta) \hat{A}_t \right] \tag{4.8.2}$$

$$L^{CLIP}(\theta) = \hat{\mathbb{E}}_t[min(r_t(\theta)\hat{A}_t, clip(r_t(\theta), 1-\varepsilon, 1+\varepsilon)\hat{A}_t)] \tag{4.8.3}$$

Where θ is the policy parameter, π is the policy, a and s are action and state respectively, $\hat{\mathbb{E}}_t$ is the empirical expectations over timesteps. r_t is the probability ratio under the new and old policies, respectively. $\hat{A}_t$ is the estimated advantage at time t. ε is a hyperparameter, where $0 \leq \varepsilon \leq 1$; the hyperparameter value is used to control the learning process. As described by Schulman *et al.* (2017), the first term inside the min is L^{CPI} (Formula 4.8.2). The second term, $clip(r_t(\theta), 1-\varepsilon, 1+\varepsilon)\hat{A}_t)$, adjusts the surrogate objective by clipping the probability ratio, which eliminates the incentive for moving r_t outside of the period $[1-\varepsilon, 1+\varepsilon]$. For a more detailed description, refer to (Schulman *et al.*, 2017).

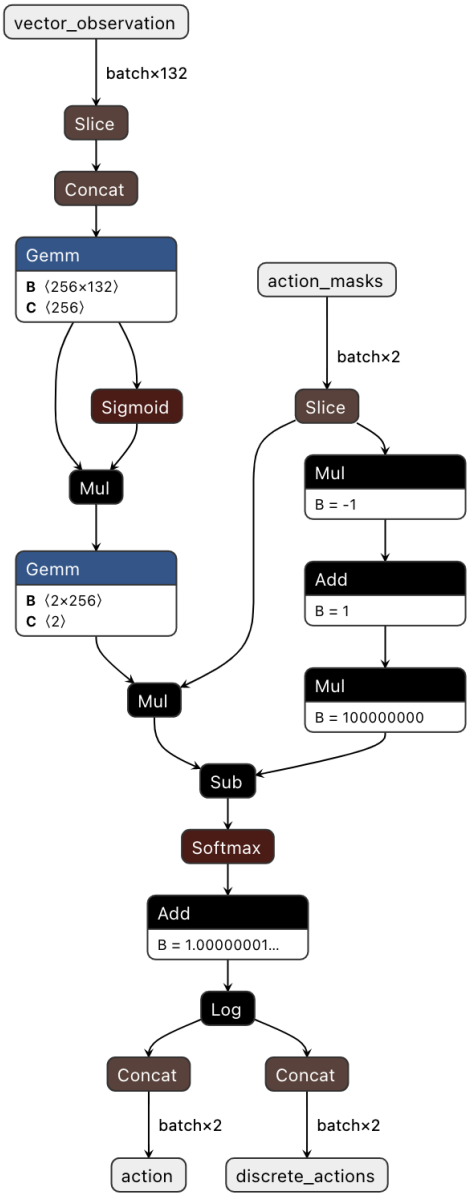
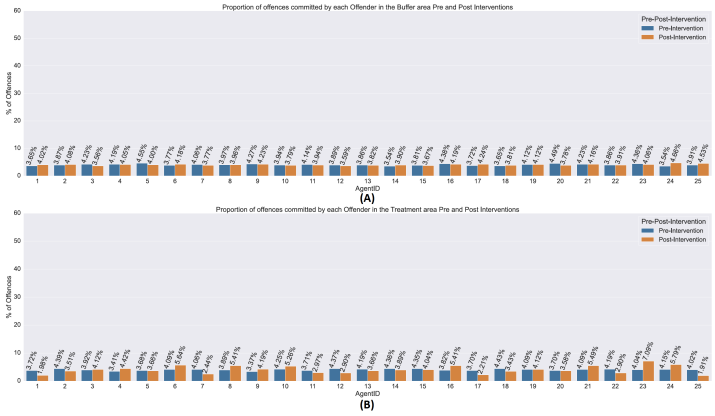
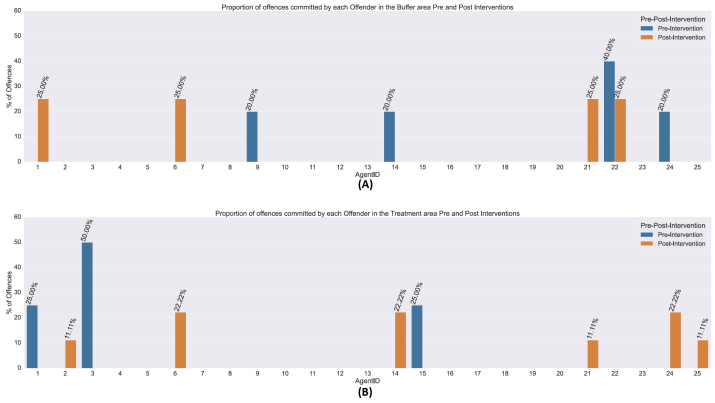


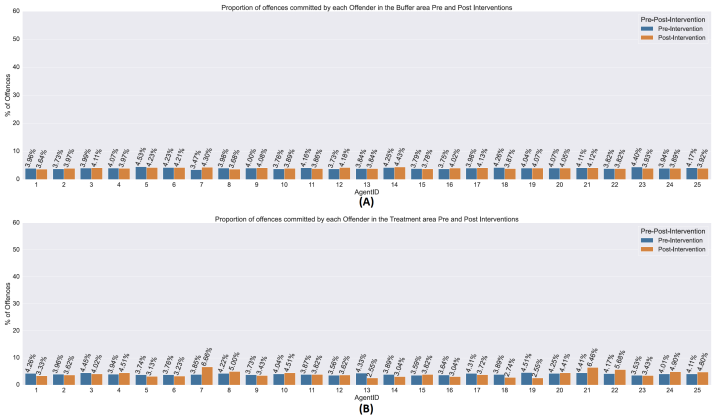
Figure 4.8.17: The Artificial Neural Network architecture for the Offender agents.



(a) MC1



(b) MC2



(c) MC3

Figure 4.8.18: The proportion of offences per model condition committed by each offender agent across all simulations pre-post interventions in the Buffer (A) and Treatment (B) areas.

4.8.3 Supporting figures

4.8.4 Supporting tables

Experiment Condition	**Mean**	**Standard Deviation**	**Coefficient of Variance**
1	0.51	0.22	The CV values are in the range of 0.42 to 0.44 approximately
2	-0.55	0.26	The CV values ranging approximately from -0.41 to -0.47
3	-0.11	0.38	The CV values are very large due to the means being close to zero, which greatly amplifies the CV. The CV is not a reliable measure of dispersion when the mean is near zero since it involves division by the mean, which can lead to inflated values.

Table 4.5: Statistics of variability across simulation runs per experiment condition

Chapter 5

Alleviating Housing Market Shocks in Real-Time: an Agent-Based Reinforcement Learning Approach

Research in modelling housing market dynamics utilising agent-based models (ABMs) as the primary methodological tool has grown over the years. One reason for this growth could be the rise of accessible individual-level data sources. Research has involved forecasting house prices, analysing urban regeneration and the impact of economic shocks on housing markets. There has been a push towards machine learning (ML) algorithms to enrich the decision-making frameworks in these ABMs to enable researchers to broaden the research trajectory. This chapter will investigate exogenous shocks to a UK housing market and demonstrate how an autonomous decision-making framework, namely, reinforcement learning (RL), is integrated and used to learn and adapt the dynamics of the housing market (simulated through an ABM). The results show that these agents can learn trends in real-time from the housing market and make decisions to overcome shock events and fulfil goal criteria such as increasing or decreasing the median house price without any pre-determined decision rules. This model is entirely transferable to housing markets in other countries or systems with similar complex characteristics. In our proposed model, the RL agent action space contains the adjustment of mortgage interest rates which the agent learns to change depending on the

market conditions. Most importantly, our experimental model shows how the central bank agent learned behaviours in line with the original article published in 2009. The central bank agent learned to behave conservatively in a more sensitive scenario and vice-versa, demonstrating emergent behavioural cues.

5.1 Introduction

Agent-based models (ABMs) have been adopted in various research areas since their inception in the late '90s to early 2000s (Filatova, 2015; Ge, 2017; Groff Elizabeth R. *et al.*, 2018; Heppenstall *et al.*, 2006; Kothari *et al.*, 2014; Tang & Bennett, 2010). ABMs enable researchers to simulate a complex system with autonomous agents interacting with each other within an environment. The main strength of ABMs over mathematical models is that they simulate, validate, and verify behavioural characteristics at granular spatio-temporal resolutions (Olmez *et al.*, 2022b; Secchi, 2015; Todd *et al.*, 2017). This allows researchers to analyze complexity and investigate how a studied phenomenon develops at the individual level (Epstein & Axtell, 1997). This chapter focuses on housing markets and investigates market shocks, which are unanticipated changes to economic variables that impact the market's health (Ramey, 2016).

ABM has been used in housing market research. Researchers investigated the emergence of housing bubbles (Axtell, 2014; Erlingsson *et al.*, 2014; Ge, 2014, 2017), the dynamics of urban regeneration (Jordan *et al.*, 2011, 2012; Picascia, 2014), and how real-world shocks such as the 2008 financial crash affected the housing market (Gilbert *et al.*, 2009; Hamill & Gilbert, 2015). The number of ABMs for studying housing and financial markets is growing (Bae *et al.*, 2019; Baptista *et al.*, 2016; Carstensen, 2015; Geanakoplos *et al.*, 2012). These models generally allow agents to make decisions in volatile scenarios, either to hedge against volatility or profit from it (Fischer & Riedler, 2014; Todd *et al.*, 2017; Westerhoff, 2010).

A research area less explored is applications of machine learning (ML) algorithms supporting decision-making in alleviating shocks once they have occurred, which central-bank policymakers can use to inform policy. Most models cited earlier examined how, when, and why shocks occur. However, developing techniques to counteract these shocks can reduce the impact on the economy and people's health (Oguibenine, 2011). This chapter proposes a hybrid model that integrates reinforcement learning (RL), with a housing market ABM. Conducting a series of experiments, we investigate if a com-

plex adaptive central bank agent (Almahamid & Grolinger, 2021; Deepanshu Mehta, 2020; Littman, 2015)) can learn trends from a housing market in real-time. During learning, this central bank agent makes decisions to fulfil a goal, for example, decreasing homelessness. In this chapter, "complex adaptive agent" is defined as: "systems or machines that utilise inferential or complex computational algorithms to modify or change control parameters, knowledge-bases, problem-solving methodologies, course of actions, or other objects in order to accomplish a set of tasks required by the user." (Imam & Kerschberg, 1997)

We identified several benefits of utilising RL in the housing market domain (1) researchers can test macroeconomic policies in a safe "sandbox" environment without real-world consequences. (2) researchers can adopt various RL goal criteria to test policy interventions in the housing market. (3) researchers can test various interventions in their housing markets and document the steps to counteract these interventions. (4) shocks (crashes) can artificially be induced to speed up learning, whereas market shocks are rare events in the real world.

This chapter replicates an ABM of the UK housing market (Gilbert *et al.*, 2009). Other notable housing market ABMs exist (Baptista *et al.*, 2016; Filatova, 2015; Ge, 2017; Rosenfield *et al.*, 2013; Yun & Moon, 2020). However, we found that either these articles were not open access and did not include download links to the models (Filatova, 2015; Ge, 2017) or the articles were open access. However, no documentation was provided to access the ABMs (Baptista *et al.*, 2016; Rosenfield *et al.*, 2013; Yun & Moon, 2020). We chose (Gilbert *et al.*, 2009) as it was well received by researchers (61 citations as of 20/05/2022 on Google Scholar), for its documentation. Furthermore, (Gilbert *et al.*, 2009) strikes a good balance between simplicity (where results are tractable) and realism (simulating important processes unique to the UK housing market, such as chain trade and can replicate empirical patterns).

To investigate whether RL can manipulate the housing market, this chapter reproduces two identical experiments conducted in (Gilbert *et al.*, 2009) as a comparator. Where exogenous shock events occur, and the decisions made by the central bank RL agent are observed. These results are compared to baseline scenarios where the RL intervention is removed. The model outputs reflect the consequences of RL decisions, and findings are compared with the original assertions made in (Gilbert *et al.*, 2009).

To summarise, this chapter will investigate whether (i) an RL agent can be inte-

grated with a housing market ABM and (ii) can an RL agent be trained using input data from the housing market ABM and make decisions to counteract shocks when they occur during run-time.

Section 5.2 reviews pre-existing studies with section 5.3 describing the ABM developed for this chapter, including the RL application of the central bank agent. The results section 5.4 defines the experiments conducted and the subsequent outcomes. Lastly, a discussion and conclusion section 5.5 discusses the findings from the experiments, limitations and strengths and concludes with future avenues to be explored.

5.2 Literature review

Economic crises sometimes take the form of debt crisis (where a government's debt increases while repayments decrease), banking crisis (when a large swathe of people withdraw their savings as confidence in the banks depletes), asset bubble burst (i.e., housing bubble bursts which leads to a sudden devaluation of houses, an example of this was the subprime mortgage credit crisis in 2007-2008 (Dou & Wang, 2014)) and balance of payment crisis (when a country cannot afford the price of imports or services). Regulatory policy is vital when a country tries to prevent or counteract an economic crisis (Malyshev, 2015), such as a central bank's monetary policy. (Martin *et al.*, 2022, p. 3) researched whether central banks can stabilise housing markets via interest rates. Researchers found that the ability of central banks to manage housing markets by increasing interest rates, which softens the demand pressure on house prices, is limited. However, they note that "central banks can significantly improve stability of housing markets by dynamically adjusting interest rates." Researchers agree that ML can be used to support decision-makers in alleviating economic crises (Chiriţă, 2011; Ho, 2020; Loukis *et al.*, 2020; Maghdid & Ghafoor, 2020; Nik *et al.*, 2016).

RL algorithms are a subset of ML approaches which enable artificial agents to learn. An agent tries to complete a task and, in doing so, maximises its internal rewards (Sutton & Barto, 2018b). Typically, these agents learn how to complete a task through trial-and-error by interacting with their environment (Kaelbling *et al.*, 1996). RL theory was derived from empirical observations of the psychological and neuroscientific studies in animal behaviours (Volodymyr Mnih *et al.*, 2016; Zhan-quan, 2006). RL has successfully demonstrated the ability of an agent to learn how to achieve long or short-term goals through interactions with the immediate environment, the

reflection of one's past knowledge and decisions influenced by rewards and penalties. Many applications of RL exist, including but not limited to (Liu *et al.*, 2020) where researchers optimise the choice of medications identifying the correct drug dosing and timing of interventions. Spatharis *et al.* (2019) developed a model where air traffic is managed through an RL agent that observes millions of data points and makes optimal decisions as to when and where planes should land.

Most RL applications in the housing market domain are related to "house price forecasting" and prediction techniques (Chen *et al.*, 2017a; Shankar *et al.*, 2020; Zhan *et al.*, 2020). Some studies have integrated deep neural networks to investigate housing markets, given the recent growth in data from websites like Craigslist, Rightmove, and Gumtree. Researchers trained a neural network using textual data to identify how the rental market dynamics were changing (Zhou *et al.*, 2019). Similarly, researchers implemented neural networks, to classify physical and socio-demographic characteristics, to assess how interrelated these factors are in the housing market of Budapest, Hungary (Norwegian, 2009). An article developed an early warning system that identified market volatility from house price training data (Park & Ryu, 2021). A drawback of this approach was that rich data sources are usually placed behind paywalls, and the neural network would have to be trained every time new data was accessible. In our research, the ABM of the housing market acts as a continuous data stream. Most importantly, in our approach, we can artificially introduce shocks (crashes) to the system to speed up learning, whereas market shocks are rare events in the real world.

Research articles such as Yamaguchi *et al.* (2018) show how RL identifies specific behaviours worms possess pre and post-feeding. Sali *et al.* (2021) used RL to deal with the feature selection problem, where researchers identified the most accurate and optimal features for reducing computation costs. As evidenced by the limited yet critical studies above, RL can learn to identify a particular phenomenon/pattern in data and develop effective interventions using neural networks to achieve a particular goal. Such as identifying the correct dosage for a patient's medication (Jalalimanesh *et al.*, 2017). Compared to the above studies, examples of applied ABM and RL in housing market research are rare (only four articles with the terms "housing market", "reinforcement learning", and "agent-based", source Web of Science). The articles (Cincotti *et al.*, 2005; Suzuki *et al.*, 2014; Zhou *et al.*, 2017) utilise RL as an optimisation method to identify the most efficient strategies in power-to-power (P2P) sharing of energy

between households and companies. Kang *et al.* (2019), on the other hand, uses data assimilation and RL to fit real-world Korean housing market data to an ABM. In light of these advances, this chapter contributes to the literature by integrating an RL decision-making algorithm in a housing market ABM focusing on shocks. It is worth noting that the work proposed here is purely experimental at this stage and acts as a proof of concept.

In this chapter, the artificial "central bank" agent observes data streams from the housing market ABM (Olmez, 2022) and makes dynamic decisions that impact the market (such as raising, holding or reducing interest rates), demonstrating how RL can be used to stabilise a market effectively in real-time in simulation. The opportunities for using RL and ABM are considerable. For example, this chapter demonstrates how RL can support decision-making in stabilising the housing market during volatile times. However, in future studies, it may be used to identify early signals of a recession or a financial crisis and alleviate the negative impact of exogenous shocks such as pandemics.

5.3 Model description

This model adheres to the Overview, Design concepts, and Details (ODD) protocol (Grimm *et al.*, 2006) for describing agent-based models, ensuring clarity and reproducibility of the simulation framework. The following overview provides a recap of the model's purpose, highlighting its aim to simulate the nuanced interactions within the UK housing market. In the following sub-sections, we delve into the behavioural rules and interactions that govern agent dynamics, encapsulating the agent's adaptation strategies, goals, and learning mechanisms. Furthermore, the details of the model outline the initialisation process, input parameters, and the step-by-step progression of the simulation, offering comprehensive insights into the inner workings of the model. By integrating the ODD protocol, this model description facilitates a transparent and structured presentation that aligns with best practices for agent-based modelling.

The housing-market model simulates the characteristics of the UK housing market. The model contains agents that are either buyers, sellers, estate agents or houses. An aggregate distribution of these agents interact in the environment where agent-environment and agent-agent interactions grow micro and macro emergent properties. The model simulates the interactions between buyers and sellers, who utilise information from local estate agents Figure 5.3.1. Buyers make offers depending on budget

and successful acquisition of mortgages, while sellers depend on valuations from estate agents, who evaluate a property's price depending on past sales and a markup.

The proposed ABM in Figure 5.3.2 is a reproduced version of (Gilbert *et al.*, 2009). The purpose of this reproduction in the Python programming language (Olmez, 2022) was multifaceted. Python offers unparalleled support for machine learning (ML) and reinforcement learning (RL), with a vast array of libraries and frameworks that are specifically tailored for these fields. The choice was further influenced by the requirement to integrate complex ML algorithms and tools, which are far more accessible within the Python environment compared to NetLogo. This ease of access is crucial when applying sophisticated algorithms such as Proximal Policy Optimisation (PPO), facilitated by the TensorForce library (Kuhnle *et al.*, 2017), which is particularly capable for the stable and efficient learning policies necessary for the housing market model in question.

Furthermore, the syntactical transition from NetLogo to Python is considerably less complex than to C, which is the primary language for Unity. This ease of translation significantly reduces the potential for errors and misinterpretations in the model transference process. It also saves considerable time and effort, which can instead be directed towards model refinement and analysis. Python's syntax is renowned for its readability and conciseness, which not only aids in code translation but also enhances the understandability and reproducibility of the research. Given that reproducibility is a cornerstone of scientific inquiry, this aspect was deemed highly important. MESA, being a dedicated ABM framework within Python, naturally complements this approach, providing a structured yet flexible platform for model development and experimentation.

5.3.1 Model environment

The environment generates a 60 x 60 grid, which can be changed depending on computing power — producing 3600 cells that can either be a house, occupied house, unoccupied space or an estate agent. Houses are randomly distributed and, depending on density, in the case of Figure 5.3.2 70% of the space is occupied. The initialVacancyRate sets the proportion of unoccupied houses at the start, making these houses available to buy. The price of these unoccupied houses follows the same rules outlined earlier. Estate agents find the highest valuation from previous sale records. The house

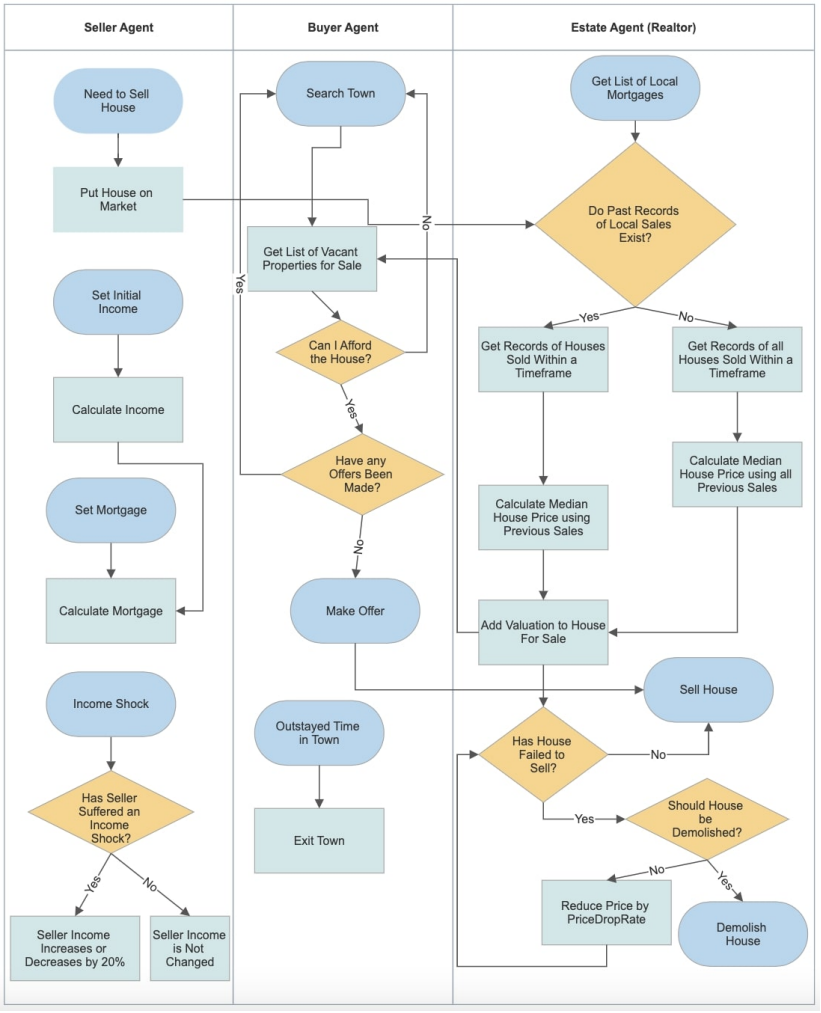


Figure 5.3.1: Flowchart presenting decisions the Seller, Buyer and Realtor (Estate) agents undertake.

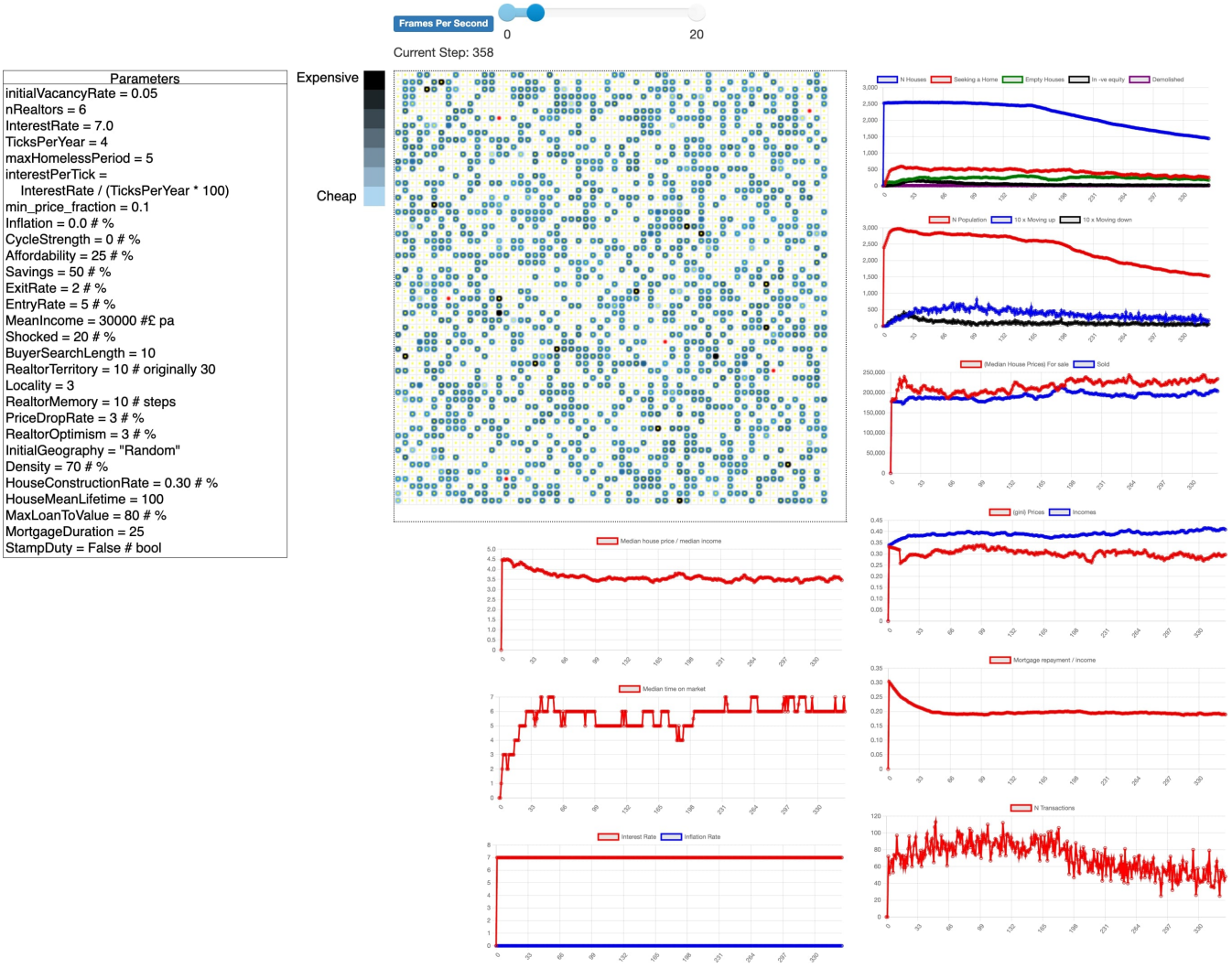


Figure 5.3.2: The user interface of the model, the parameters that can be changed on the left, the visual representation of the ABM in the centre where the small squares represent houses. Yellow dots are occupants, red dots are estate agents, white grid cells represent free space. Output plots are on the right.

prices are randomly distributed using a uniform random distribution, as is the case in Figure 5.3.2. Each house has a quality index calculated upon initialisation. This measure is a ratio of the average price of other houses within the locality of the constructed house's sale price. The process mentioned above adheres to Tobler's first law of geography (Tobler, 1970) which states that nearer things are more likely to be similar than those farther apart. Every output parameter is described in Table 5.1.

5.3.2 Seller agents

Every step, the model moves forward in time; a step is 3 months defined by the TicksPerYear parameter Figure 5.3.2. A percentage of homeowners (ExitRate parameter) vacate and try sell the house at a price set by the estate agent valuation. If the house does not sell at the current timestep, it remains on the market for the next period. Every homeowner agent has an initial income determined randomly using a gamma distribution from parameters 1.3 and 5×10^{-5} multiplied by the MeanIncome parameter. Furthermore, a mortgage is calculated by the ratio of the Affordability parameter, divided by interestPerTick multiplied by the owner's income. Initially, the mortgage duration is 25 years. However, this can be changed. People borrowing money must have some deposit from their capital determined by MaxLoanToValue parameter and their mortgage. At every step, a percentage of homeowners suffer income shocks determined by the Shocked parameter, which is +20, and the same percentage suffers a shock of -20%. This leads to some homeowners having their income increase or decrease by this percentage permanently. When the ratio of the mortgage repayment is higher than twice the Affordability, the homeowner trades down. Conversely, they trade up when the ratio is less than half the Affordability.

5.3.3 Estate agents

The term estate agents is used interchangeably with realtors. Every realtor agent has a coverage radius called the RealtorTerritory. Any house outside a realtor's territory is assigned the closest realtor calculated by the Euclidean distance. Each realtor keeps records of the previous sale. These records contain the following information: record ID, the house sold, selling price and date of sale. At the start, the mortgage value of each house is sent to one local realtor, providing realtors with a starting point for their valuations (Gilbert *et al.*, 2009). When a seller asks for a valuation, the realtor

looks through their records within the last RealtorMemory timesteps and gathers all the house prices of houses sold locally multiplied by the quality index of these houses. It then calculates the median house price of these previous sales as a valuation. If no sales have been made within the locality and period, any past sales made within the locality are considered regardless of time. Every valuation made is increased by the RealtorOptimism percentage, allowing realtors to try to sell a house more than the going rate. Lastly, if a house fails to sell at timestep N, the selling price of this house is reduced by PriceDropRate %, and it remains on the market for N+1 until it is sold or demolished.

5.3.4 Buyer agents

At each timestep, people arrive in town. The amount depends on the EntryRate parameter, which is a percentage of the current population Figure 5.3.2. New entrants and sellers who remain search the whole town for several timesteps defined by BuyerSearchLength parameter. Looking for vacant properties for sale that they can afford and which no offers have yet been made. Any accumulated capital from buying and selling can be put towards the costs of the new property. Subsequently, buyers choose the nearest property in price to their maximum budget and make an offer at a price set by the seller. The first buyer to make an offer has their offer accepted.

5.3.5 Sales

A sale is only successful if the chain of buyers and sellers remains intact. A successful chain can only occur if the house being bought is either empty or the seller succeeds in purchasing a new house and moving to that house. The people leaving town move out, and potential buyers move into these vacant properties if their offers are successful. Once all sales down the chain are complete, the model moves forward one step in time. When a house is sold, the seller receives the sale price and uses as much of it necessary to pay off any remaining mortgage. If money is in excess, this is added to capital and can be used as partial or complete payment of the house being bought. Conversely, if the sale price is less than the amount remaining to pay off the mortgage, the seller is in negative equity and withdraws the house from the market. The estate agent records successful sales and uses these records (as discussed above) to value houses within the same area. Finally, if an offer falls through, it lapses (Gilbert *et al.*, 2009).

5.3.6 Building new houses

New houses are constructed at random empty grid cells at every timestep. The number of houses depends on the HouseConstructionRate Figure 5.3.2 % of the total number of constructed houses unless there are no empty cells.

5.3.7 Demolition of houses

Every house has a lifetime set when it is created. This is drawn from a random exponential distribution with a mean of HouseMeanLifeTime. When a house reaches its lifetime, it is demolished, and the cell becomes vacant and available for new construction. If a house's sale price falls below one-tenth of the median price of all houses, it is demolished. If someone occupies a house that is being demolished, they attempt to purchase a new home, and if they fail after MaxHomelessPeriod, they leave town.

5.3.8 Model outputs

The data produced from the model are presented below. Visually, houses are assigned a colour that reflects their current value. The lighter the shade, the cheaper the house and vice versa Figure 5.3.2. The quantitative model outputs are the following:

- Number of houses, empty houses, and demolished houses. Number of people searching for a home, the number of people occupying a home in negative equity, and number of transactions.
- The number of people in the model.
- The median house price of houses for sale and sold.
- The Gini index of the median house prices and median incomes.
- The ratios between median house price to median income, and mortgage repayments to median income.
- The mortgage interest rate, inflation rate and median time houses have spent on the market.

5.3.9 Quantifying model similarities (validation)

The proposed model (Olmez, 2022) was reproduced by interpreting (Gilbert *et al.*, 2009) source code, refer to Figure 5.6.1 for class diagram. The behaviour of our model must be compared and deemed similar to the original, whereby model outputs produce similar trends in data. If the model outputs differ, we have deviated from the original at some point in the development. Suppose the model produces similar trends in the data outputs. We can be confident in our model's behaviour in producing realistic trends of the UK housing market. Model replication is an important topic in ABM literature, as discussed by (Donkin *et al.*, 2017, p. 1) "model replication remains rare, yet is vital to assessing the repeatability of existing ABMs".

Two housing market scenarios were simulated for both models using input parameters in Table 5.2. In scenario one, no shock is introduced, and in scenario two, a shock is introduced mid-simulation run. This allows us to compare behaviours in two unrelated scenarios to quantify two completely different outcomes.

To compare both models, we adopt a visual statistical approach known as quantile-quantile (Q-Q) plot, the benefits of which have been thoroughly discussed in the following literature (Dhar *et al.*, 2014; Oldford, 2016). Where two probability distributions are compared using their quantiles. In our case, one variable in (Olmez, 2022) is compared to the same variable in (Gilbert *et al.*, 2009). Furthermore, over 100 model runs over 100 simulation years are drawn for each scenario providing a large sample size to quantify the stochasticity produced and output-variability (Bogdoll *et al.*, 2012; Lelei & McCalla, 2019). A one-degree gradient (45°) reference line is plotted to compare variables. If $x = y$, each point sits on the reference line, then both variables compared are identical and vice versa.

Due to the large volume of output variables (18) Figure 5.3.2. We select a sub-set of Q-Q plots to use in the results. These include parameters that capture the housing market's health, for example, the median house price to income ratio and the median price of houses for sale.

Figure 5.3.3 presents (Gilbert *et al.*, 2009) on the x-axis and (Olmez, 2022) on the y-axis. Each column and row represents the scenario and variable respectively. These results show a normal distribution and results are correlated which quantitatively replicate similar trends, that sometimes deviate due to stochasticity, such as (Olmez, 2022) overestimating (Figures 5.3.3c, 5.3.3f) or underestimating (Figures 5.3.3g, 5.3.3h).

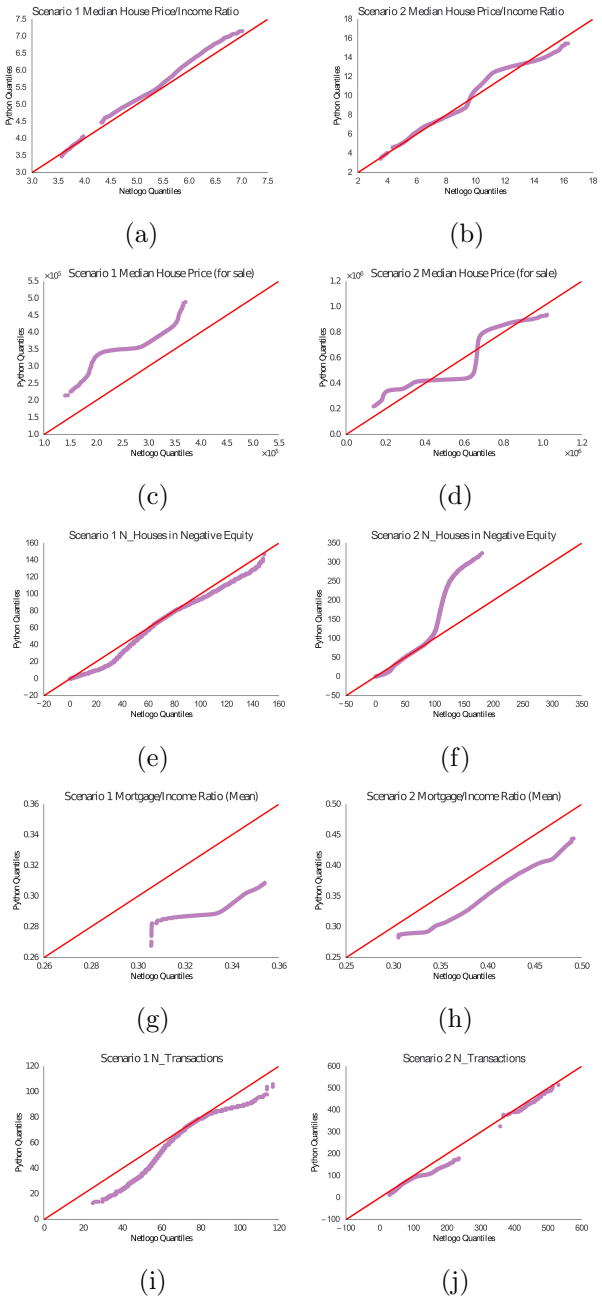


Figure 5.3.3: Q-Q plots comparing the distributions of model output variables. The solid line indicates $x = y$ for reference. Where (a-b): Median house price to income ratio (Scenarios 1-2), (c-d): Median house price for sale (Scenarios 1-2), (e-f): Number of households in negative equity (Scenarios 1-2), (g-h): Mean mortgage to income ratio (Scenarios 1-2), (i-j): Number of transactions (Scenarios 1-2).

Note, due to model architectures, frameworks and other factors, models cannot be replicated perfectly as highlighted by (Donkin *et al.*, 2017; Xiong Yingfei, 2009).

The descriptive statistics, Figure 5.6.2 show that 79% of the output variables returned a positively correlated Pearson's correlation coefficient, where 11 of these from scenario two had $r > 0.60$ which were all statistically significant $p < 0.01$. In scenario one, eight variables had $r > 0.60$ and $p < 0.01$.

In scenario two, all variables returned an $0.659 \leq r \leq 0.992$, all statistically significant, Figure 5.3.3. In scenario one, $0.235 \leq r \leq 0.839$ four variables were statistically significant. The only variable that was not statistically significant was the mean mortgage-to-income ratio for scenario one. The most likely reason could be stochasticity, where the algorithm has more steps to process (Donkin *et al.*, 2017).

In replicating the ABM presented in Figure 5.3.2, meticulous care was taken to interpret the original source code and align the model's behaviour with established benchmarks. The differences observed in Figure 5.3.3 are not indicative of errors but are attributable to the intrinsic disparities in how different programming languages compile code and represent data structures. NetLogo, the language used for the original model (Gilbert *et al.*, 2009), handles code compilation and data structures differently from Python. Moreover, discrepancies in random number generation between platforms can lead to divergent results, especially over a multitude of iterations.

The complexity of translating code from NetLogo to Python is notably less than the translation from NetLogo to C, the language used in Unity (used in prior chapters). This reduced complexity minimises the risk of translational inaccuracies. Python's syntax and structural paradigms bear a closer resemblance to NetLogo, facilitating a more steady replication of the model's logic and processes. The use of descriptive statistics and Q-Q plots in Figures 5.3.3 and 5.6.2 fortifies the argument that the differences in model outputs are statistically significant and not the result of implementation errors. The QQ plots, in particular, illustrate the correlation of outputs between the models, highlighting normal distribution and corroborating the overall replication of trends, albeit with expected stochastic variations.

Overall, the proposed model demonstrates the UK housing market characteristics observed in the original (Gilbert *et al.*, 2009). Trends develop when external tweaks to the market are made, showing that indicators are sensitive to these changes in both models. The next stage is to integrate RL, to test whether an complex adaptive

observer agent can learn to identify shocks to the market and deploy countermeasures to minimise their effects in real-time in simulation.

5.3.10 Reinforcement learning agent

RL allows agents to learn without explicitly telling the agent what the task is or how it is completed. A feedback reward allows the agent to learn through trial-and-error by performing actions for each state in the environment. If the reward is positive, the agent has enacted a desirable action. If the reward is negative, the action is undesirable (Sutton & Barto, 2018b).

Given how well policy-gradient methods have performed (Agarwal *et al.*, 2020; Schulze *et al.*, 2017), this was an applicable approach. Put simply, we denote a policy as π, where $\pi\theta(a|s)$ is the probability of taking action a in state s and θ are the parameters of our policy. Our goal is to update θ_t to θ_{t+1} such that we reach the optimal policy. In our model, the optimal policy would be the state where the "healthy housing market" criteria (described below) are met. If we assume $a*$ is the optimal action, i.e., raise interest rates by 0.01 at time t, then we want to perform gradient ascent on $\pi\theta(a*|s)$ (ascent as we want to increase our cumulative reward). Therefore, at each iteration, we update θ in the following way $\theta_{t+1} = \theta_t + \alpha\nabla\pi_{\theta_t}(a*|s)$ this can be described as we keep "pushing" towards more of action $a*$ in our policy, which is indeed what we want as raising the interest rate by 0.01 will mean we are closer to achieving our "healthy housing market" criteria.

This chapter proposes an application of RL to identify and counteract market shocks in the housing market in real-time in simulation. Several steps were taken to integrate RL with the housing market ABM:

1. Re-producing the well-known housing market (Gilbert *et al.*, 2009) in a new framework (Olmez, 2022) to use as our experimental sandbox.

2. Replicating two experiments originally described and subsequently investigated in (Gilbert *et al.*, 2009, p. 5) known as the loan-to-value experiments where, the MaxLoanToValue parameter is set to 80% and 100% respectively. During these experiments, an exogenous shock known as "ratefall" where a sharp increase 7% to 10% in interest rates is triggered. These two experiments were selected, as in the original article, Gilbert et al. demonstrated how the impact varying interest

rates on the market were prone to being less sensitive when loan-to-value was reduced compared to loan-to-value being 100%. These experiments are poised to test RL's ability to adapt its behaviour in two similar initial conditions but with very different outputs.

3. Training the RL agent on the housing market scenarios over 100 episodes, this process can be observed in 5.3.4.

Due to computational complexity of training RL agents with neural networks (Baker *et al.*, 2019; Juliani *et al.*, 2018; Olmez *et al.*, 2022a; Sutton & Barto, 2018b), the experiments are kept concise. It is worth noting that the proposed model is a proof-of-concept used to demonstrate how the adaptive qualities of cognitive models such as RL are suited to modelling housing market dynamics and supporting decision-making to counteract shock events.

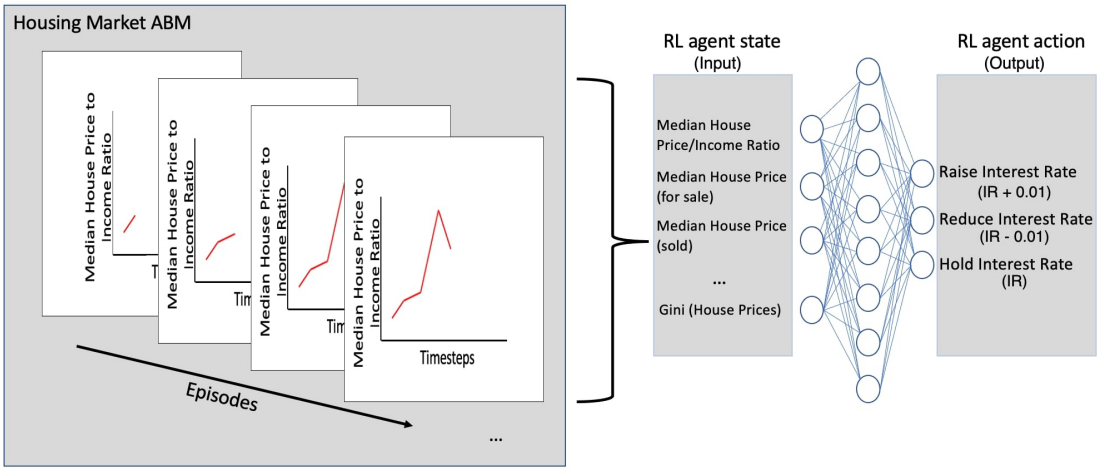


Figure 5.3.4: RL central bank agent neural network, that determines the central bank agent's decision regarding interest rates, in the current instance, an action with high probability may be to raise interest rates as a sharp increase in house price to income ratio is observed.

This chapter models simplistic behaviours of a central bank agent. As discussed in Section 5.2. In reality, a central bank has more policy tools and goals to achieve beyond housing market stability. This simplicity is necessary to demonstrate a proof-of-concept. In future research, these behaviours will become more advanced.

To train the RL agent, we first identify the healthy housing market indicators. This way, the agent can learn to differentiate between an undesirable state and a desirable one. For simplicity, we identified three conditions which should be satisfied for our goal. These are:

- Stable median house prices for sale with small fluctuations up to $\leq 400,000$.
- Median house price to income ratio ≤ 7.
- Number of people in negative equity is $\leq 5\%$ (123) where $N\_people = 2466$.

If the above conditions are met, we have a desirable state (reward returned to the central bank agent, 0). If all but one of these conditions are not met, we are in an undesirable state (reward returned, -1).

The results from the RL process are presented and compared to base case scenarios in the following section. The RL outputs illustrate the central bank agent's learning process during training and how the RL agent adapts to the LTV scenarios. We compare findings to those discussed in the original article to demonstrate how RL has or has not benefitted the housing market in alleviating shocks and fulfilling goal criteria.

5.4 Results

This section describes the market shock, provides an overview of central bank decisions to deal with shocks, and outlines the healthy housing market conditions. This is followed by a detailed analysis of findings from original housing market research conducted in 2008 (Gilbert *et al.*, 2009) emphasising the experiments we aim to conduct in this chapter to gauge the strengths and weaknesses of RL. Lastly, experiments and subsequent results show how complex adaptive RL agents can learn and make decisions in real-time in simulation to counteract induced shocks. Note that actions the RL agent can undertake simplify how decisions are made in the real world. We aim to explore the strengths and weaknesses of this methodology in a simplified housing market, hoping to identify the potential for future applications.

Shocks come in various forms, as described in Section 5.2. To manage shocks, central banks often enforce a monetary policy which stabilises and counteracts the aftermath of the shock (Martin *et al.*, 2022). From 2007 to 2009, the world endured a financial shock resulting in a crash in the UK housing market (Whitehead & Williams, 2011).

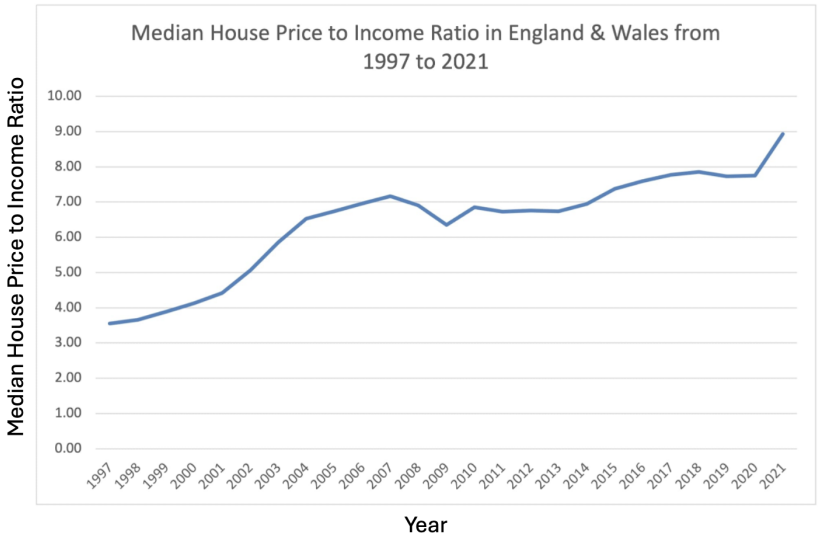


Figure 5.4.1: The median house price to income ratio in England and Wales from 1997 to 2021 (source: (Office for National Statistics (ONS), 2024)).

In Figure 5.4.1, the house price to income ratio in England and Wales dropped from 7.17 in 2007 to 6.35 in 2009. Furthermore, a drop of 18.7% in house prices (Munro, 2018), from Q3 2007 to Q1 2009. To counteract the pressures, the central bank reduced interest rates from 5.25 in Jan 2007 to 1.50 in Jan 2009 (Tse *et al.*, 2014). In contrast, a healthy housing market trend may look like Figure 5.3.2, where mortgage repayment to income ratio is 20%, and the median house price to income ratio oscillates between 3.5 and 4.0. House prices increase gradually as people move in and out of the market. The number of people in negative equity is as small as possible (Been *et al.*, 2021; Melzer, 2010; Morescalchi *et al.*, 2018).

Gilbert *et al.* (2009) conducted several experiments using their UK housing market model in 2008, where the model showed how properties of the UK housing market are emergent. Some crucial findings that aligned with empirically observed behaviours of the UK housing market were:

- House price to income ratio showed a stable relationship given mortgage interest rates and the loan-to-value ratio. For example, if interest rates are reduced or loan to value increased, house price to income rises in response.

- When the loan-to-value ratio is 100%, and the market experiences an exogenous

> interest rate hike from 7% to 10%, a sharp drop in the house price-to-income ratio is observed. However, if loan-to-value is set at 80% and the same increase in the interest rate is observed, the effect is much weaker (Gilbert *et al.*, 2009, p. 5).

These findings were also explored in other housing market research, such as (Narayan & Narayan, 2011; Tse *et al.*, 2014; White, 2015). The experiments (Gilbert *et al.*, 2009) make for a well-documented comparator for this chapter, where the strengths and weaknesses of RL can be tested in relation to the earlier assertions. Given these original experiments and results, in this chapter, we expect RL to behave in a certain way when adjusting interest rates in the 100% LTV scenario compared to the 80% scenario. The goal is to test if the RL central bank agent can adapt to these scenarios and fulfil its goals. In the next paragraph, we describe the experiments in detail.

We replicate two experiments; these are "loan-to-value A and B". In the A experiment, the MaxLoanToValue parameter (refer to Table 5.1) is 100, and in the B experiment, it is set to 80. In both experiments, an exogenous shock occurs at timestep 200, where mortgage interest rates suddenly increase from 7% to 10%. In the base case experiments (no RL), we observe findings from the (Gilbert *et al.*, 2009). In the RL experiments, we observe behavioural differences and consequences of actions taken by the RL agent in achieving the "healthy housing market" criteria. As our results show, we believe there is value in utilising these contemporary methods to support future research in modelling housing markets.

Given the computational complexity in these models and the nature of RL training, we run the experiments for 100 iterations to capture a distribution of results quantifying model stochasticity. Moreover, during training, we found that the "healthy housing market" criteria were met.

To recap, the experiments were run for 400 simulation timesteps. At 200^{th} timestep, a shock impacts key indicators to varying extents as observed in (Gilbert *et al.*, 2009, p. 4) and presented in Figure 5.4.2. The ratefall shock severity is greater in 100% LTV compared to 80% LTV; this can be observed in Figures 5.4.2g and 5.4.2h. Similar behaviour is observed for median house prices for sale and the number of households in negative equity (Figures 5.4.2k, 5.4.2l and 5.4.2o, 5.4.2p).

Interestingly, the RL agent learns to approach the 100% LTV experiment conservatively with slight adjustments but mainly holding interest rates evidenced in Figure 5.4.2b compared to the 80% experiment Figure 5.4.2a. This is a response to the

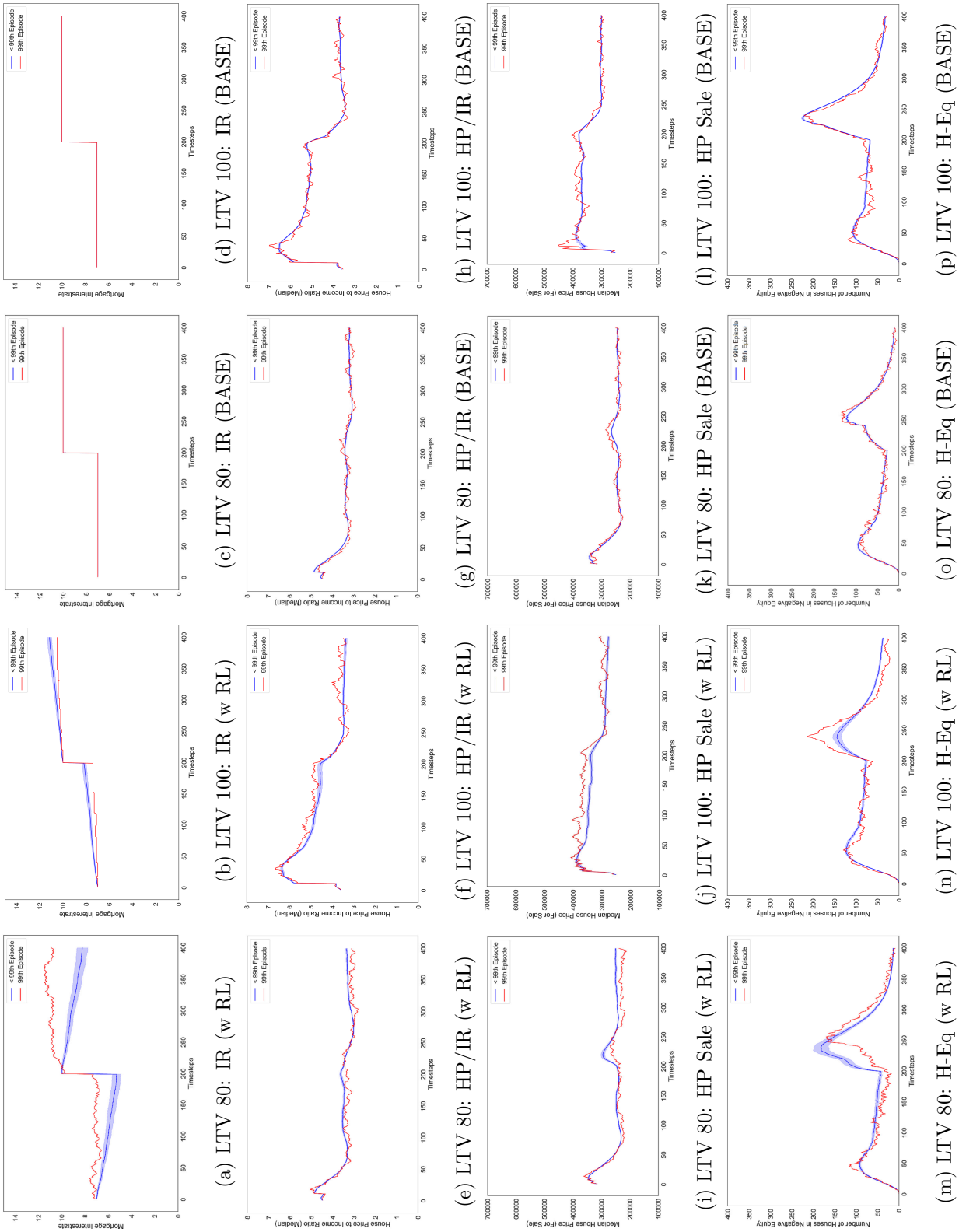


Figure 5.4.2: Line graphs showing the last model run (99^{th} due to index starting at 0) and the average with a confidence interval for all previous runs (< 99) aggregated for each experiment condition, including base case conditions. Each row is a tracked variable, and the column is the experiment, where IR = Interest Rates, HP/IR = House Price to Income ratio and H-Eq = Houses in Negative Equity.

sensitivity of indicators of the housing market to prevailing interest rates. This particular finding demonstrates the adaptive capabilities of RL, where slight environmental changes in a model can be responded to by learning and experiences. Furthermore, on average the RL agent reduces interest rates for the LTV 80% scenario (see Figure 5.4.2a), conversely, in the 100% LTV scenario, it increases interest rates on average (see Figure 5.4.2b).

Another finding from Figure 5.4.2 when analysing RL behaviours, in particular, is that leading up to and right after the shock at timestep 200, the confidence intervals are much wider (further from the mean). This means the RL agent has explored more state space at these crucial points during training. This behaviour can be observed most clearly in Figures 5.4.2a and 5.4.2b. In the 100% LTV experiment, the variance statistic for interest rates is 2.691 (mean 9.095, std 1.640) compared to the 80% experiment 4.521 (mean 7.607, std 2.126), where RL agent is exploring more and subsequently adjusting interest rates more often in the 80% compared to the 100% LTV experiment.

RL trains iteratively across several simulation runs, also known as episodes. The latest episode in the training phase is the most recent behavioural output, usually representing the stage at which RL has learned the most optimal set of behaviours to achieve some goal. Conversely, the most recent episodes are those in which RL is yet to learn effective behaviours (Sutton & Barto, 2018b). The data observed at the 99^{th} model-run (episode) are the outputs for when the RL was most trained. Therefore, we compare these to our goal criteria for a healthy housing market state. For some indicators, the RL agent was better at achieving the healthy housing market goals than others, which would be expected as some indicators are more sensitive to interest rates than others. For both loan-to-value experiments, RL successfully ensured the house price-to-income ratio was below seven even after the ratefall shock, refer to Figures 5.4.2e and 5.4.2f. The median-house-price-for-sale indicator shows that RL was more effective in achieving the $\leq 400,000$ goal in the 80% experiment (Figure 5.4.2i) compared to the 100% experiment (Figure 5.4.2j) which was only above 400,000 for a short time at the earlier timesteps from 0 - 100. A similar outcome was also observed in the base case Figure 5.4.2l. The data show that the houses-in-negative-equity was a more complex indicator for RL to achieve the goal of ≤ 123 where the shock exacerbated the complexity as presented by the wide confidence intervals post-shock Figures 5.4.2m and 5.4.2n. However, pre and post-shock, we observe a downward trend where the

number of households in negative equity is less than 123, even achieving less than 50 near the end of the simulation. It is worth noting that while differences exist between the RL and base case scenarios, these can be considered small. However, this can result from the chosen healthy housing market goal conditions, and differences may be more significant if other goal conditions or a combination of conditions were chosen. These results also demonstrate the housing market's ability to settle after a shock.

Overall, these experiments show that RL can achieve healthy housing market goal criteria and alleviate the shock effect on the market, which varies in effectiveness across the different indicators. Given these results, we presume that as the number of goal indicators increases from 3 to $3 + i$ for some i, complexity in achieving these goals also increases. Given this complexity, we believe that RL's housing market goals may be unachievable at some point. This may be due to equilibrium whereby increasing one indicator, the goal is met, but another indicator is reduced; thus, the goal is not met and vice-versa. This study has demonstrated that RL is a valuable technique that should be welcomed by housing-market and macroeconomic researchers interested in utilising autonomous decision-making methods to aid policy making in dealing with uncertainty like economic shocks. In the next section, we break down the learning process of RL and describe the strengths and weaknesses in utilising RL within this domain.

5.5 Discussion and conclusion

This chapter reproduces a well-known ABM of the UK housing market to integrate a RL algorithm that learns to counteract housing market shocks in real-time. We answer the following research questions: can RL be integrated with housing market ABMs? Moreover, can agents learn trends from the housing market and adapt to economic shocks by counteracting the impact of these shocks in real time? Findings show RL can be integrated with housing market ABMs, as evidenced in sub-section 5.3.10 and does well, in learning to counteract shocks through monetary policies such as interest rate adjustments.

This chapter shows how RL could adapt its behaviours and, over time, through training, learn behaviours that enable it to achieve the goal state. Furthermore, the RL agent portrayed characteristics of the original model (Gilbert *et al.*, 2009). One example is the effects of interest rates in the 80% loan-to-value (LTV) compared to the 100% loan-to-value environment. Responding to the impact of interest rates being more

sensitive in the 100% LTV case compared to 80%. The RL agent learned to explore a greater range of interest rates in the less-sensitive scenario (80%) compared to the more sensitive scenario (100%) intended to counteract the market's sensitivity in these two conditions, which was purely learnt and not hard-coded.

A drawback of our approach is that the RL agent's tools are limited. This is not indicative of a real-world central bank, which has more policies to counteract crises, such as regulatory, monetary, and fiscal policies. However, macroeconomic policies such as adjusting interest rates is a critical intervention central banks make (Martin *et al.*, 2022; Popescu, 2014; Valadkhani *et al.*, 2019) with the most recent example tackling inflation in the UK (Inman, 2022; J. Lynch & Adam, 2022; Luhnow & Colchester, 2022). Another caveat is that our model is, a simplified version of the real world. This would ensure computational tractability. Thus, this chapter only focuses on a single policy tool to demonstrate the application of RL in the housing market research domain and is exploratory in nature.

There are several weaknesses in RL methods, including overfitting, the exploration-exploitation trade-off (Sledge & Principe, 2017), and computational demand. To address the exploration-exploitation issue, we used an objective function that was not greedy but balanced both aspects (Silver *et al.*, 2014; Sutton & Barto, 2018b). While computational demand was not a problem for our model, it could become an issue for more complex environments with more agents and action spaces. In these cases, advanced computational resources may be required.

There are several exciting directions for future research based on this work. For example, the findings can support housing market modelling, where researchers forecast the potential for exogenous shocks and identify policy decisions to alleviate economic downturns. The technique can also be adapted to simulate a realistic case where the goal is to optimize the current state of the housing market through RL. Additionally, given the recent release of the 2021 UK census, researchers can enhance the model with this data to study the dynamics of the housing market. This is the first example in the literature that uses RL algorithms within housing market agent-based models to develop a methodology for autonomously counteracting exogenous shocks to the market. There is also value in exploring this application in macroeconomics, where artificial intelligence-assisted policy-making and signal detection can have a significant impact. For example, it may help a central bank to detect a recession or financial crisis

and take action early.

Notes

The model code can be acquired at the following source: (Olmez, 2022). The datasets can be found at https://doi.org/10.6084/m9.figshare.21719879.v1

5.6 Appendix

Parameter	Description
InitialVacancyRate	proportion of empty houses.
nRealtors	number of realtors.
InterestRate	mortgage interest rates.
TicksPerYear	number of timesteps per year.
MaxHomelessPeriod	maximum number of timesteps a person can be homeless.
InterestPerTick	interest rate, after a cyclical variation has been applied.
MinPriceFraction	if a house price falls below this fraction of the median price, it is demolished.
Inflation	inflation percentage per year.
CycleStrength	percentage of interest rate change periodically.
Affordability	owners trade down or up depending on the affordability ratio.
Savings	the amount of capital a person has.
ExitRate	percentage of people exiting the market.
EntryRate	percentage of people entering the market.
MeanIncome	average annual income.
Shocked	percentage of person's income increased or decreased.
BuyerSearchLength	length of time a buyer searches for a home.
RealtorTerritory	the locality a realtor covers.
Locality	distance between houses used for valuations.
RealtorMemory	historical collection of records used for valuations.
PriceDropRate	percentage a house drops in price.
RealtorOptimism	a fee added to the valuation.
InitialGeography	type of geography.
Density	how clustered the agents are spatially.
HouseConstructionRate	the rate at which houses are constructed.
HouseMeanLifetime	the average lifetime of a house.
MaxLoanToValue	the amount of deposit.
MortgageDuration	the timeframe for mortgage repayment.
Stampduty	a tax levied on house sales (UK only).

Table 5.1: Model input parameters and description (source (Gilbert *et al.*, 2009)).

Parameter	Input Value
NumberOfRuns	100
Timesteps	400
CycleStrength	0%
MeanIncome	£30000 pa
ExitRate	2%
Density	70%
StampDuty	False
TicksPerYear	4
HouseMeanLifetime	101
MortgageDuration	25
HouseConstructionRate	0.30%
MaxLoanToValue	100%
Shocked	20%
BuyerSearchLength	10
EntryRate	5%
MaxHomelessPeriod	5
InitialGeography	"Random"
PriceDropRate	3%
InterestRate	7%[1]
RealtorOptimism	3%
Inflation	0%
Affordability	25%
Savings	50%
RealtorTerritory	30
RealtorMemory	10
Locality	3

Table 5.2: Model input parameters for similarity testing.

[1]For scenario two, the interest rate will drop from 7% to 3% initiating a shock at timestep 200.

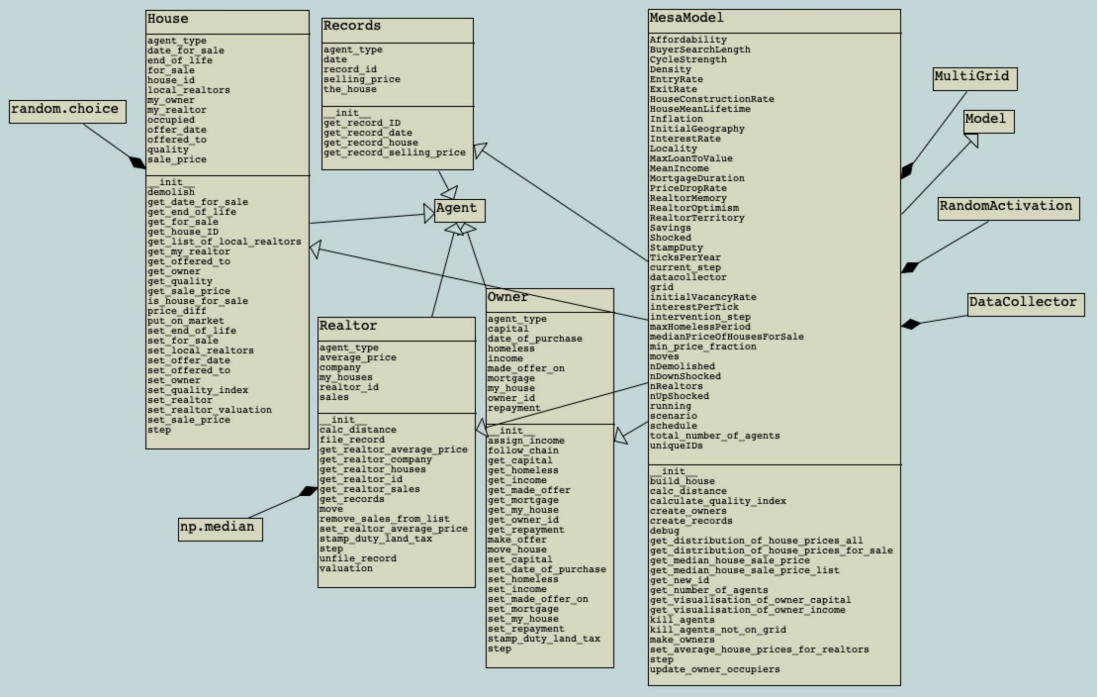


Figure 5.6.1: A UML class diagram of the core model framework (Olmez, 2022)

	Simulation	Python_2	Netlogo_2	Python_1	Netlogo_1
Houseprice/Income_Ratio(Median)	mean	8.575	8.226	5.244	5.094
	std	3.385	3.204	0.494	0.466
	min	3.444	3.552	3.474	3.569
	25%	5.305	5.265	5.004	4.841
	50%	6.682	6.401	5.139	5.001
	75%	11.655	10.700	5.354	5.264
	max	15.483	16.314	7.156	7.017
	pearson (corr)		0.947 ($p < 0.01$)		0.839 ($p < 0.01$)
Median_Houseprice_for_sale	mean	540784.707	492637.040	376547.125	303944.850
	std	205852.457	209494.168	21553.608	28655.851
	min	220618.472	140373.576	214807.564	140800.961
	25%	370983.716	306518.909	366592.465	297064.457
	50%	406323.037	344237.304	377485.583	308279.053
	75%	793104.838	694714.404	388691.637	319404.545
	max	939038.900	1024133.025	489329.873	371404.589
	pearson (corr)		0.809 ($p < 0.01$)		0.323 ($p < 0.01$)
Median_Houseprice_for_sold	mean	468766.092	372691.395	289877.291	227158.909
	std	200050.365	159258.916	26848.436	20349.625
	min	138245.853	118400.328	138660.344	118881.129
	25%	286752.743	231172.403	285557.479	225151.212
	50%	314505.776	252806.798	294130.012	231070.367
	75%	674358.036	515474.594	302448.009	236600.114
	max	811625.267	688298.713	340736.414	263754.427
	pearson (corr)		0.961 ($p < 0.01$)		0.837 ($p < 0.01$)
Gini_index_house_prices	mean	0.307	0.283	0.290	0.279
	std	0.037	0.037	0.034	0.035
	min	0.193	0.192	0.205	0.199
	25%	0.278	0.255	0.265	0.252
	50%	0.304	0.280	0.282	0.271
	75%	0.337	0.310	0.314	0.304
	max	0.432	0.387	0.394	0.382
	pearson (corr)		0.641 ($p < 0.01$)		0.832 ($p < 0.01$)
Gini_index_incomes	mean	0.394	0.374	0.388	0.373
	std	0.014	0.016	0.013	0.015
	min	0.331	0.301	0.329	0.302
	25%	0.386	0.364	0.379	0.364
	50%	0.395	0.376	0.388	0.373
	75%	0.403	0.385	0.397	0.383
	max	0.445	0.412	0.456	0.416
	pearson (corr)		0.534 ($p < 0.01$)		0.667 ($p < 0.01$)
Mean_Mortgage_repayment_to_income_ratio	mean	0.353	0.401	0.295	0.340
	std	0.059	0.065	0.004	0.005
	min	0.283	0.306	0.268	0.306
	25%	0.296	0.340	0.293	0.338
	50%	0.306	0.353	0.295	0.340
	75%	0.417	0.473	0.298	0.342
	max	0.444	0.491	0.309	0.354
	pearson (corr)		0.992 ($p < 0.01$)		-0.150 ($p = 2.026$)
Number_of_transactions	mean	62.960	70.694	53.087	60.676
	std	24.733	26.553	14.230	9.463
	min	13.000	29.000	13.000	25.000
	25%	52.000	59.000	42.000	54.000
	50%	61.000	67.000	54.000	60.000
	75%	70.000	76.000	64.000	67.000
	max	515.000	531.000	106.000	117.000
	pearson (corr)		0.659 ($p < 0.01$)		0.235 ($p < 0.01$)
Median_time_on_market	mean	8.733	9.287	9.007	9.566
	std	1.871	2.316	1.588	1.796
	min	1.000	1.000	1.000	1.000
	25%	9.000	9.000	9.000	9.000
	50%	9.000	10.000	9.000	10.000
	75%	10.000	11.000	10.000	10.000
	max	13.000	15.000	12.000	14.000
	pearson (corr)		0.865 ($p < 0.01$)		0.824 ($p < 0.01$)
N_houses	mean	2034.573	2421.156	2083.285	2475.416
	std	460.394	80.058	397.827	44.420
	min	1074.000	2196.000	1267.000	2354.000
	25%	1600.000	2373.000	1712.000	2442.000
	50%	2219.000	2423.000	2223.000	2473.000
	75%	2438.000	2481.000	2436.000	2511.000
	max	2568.000	2587.000	2567.000	2615.000
	pearson (corr)		0.746 ($p < 0.01$)		-0.038 ($p = 1.868$)
N_people_seeking_home	mean	380.377	371.058	369.355	377.195
	std	122.045	109.796	108.300	79.574
	min	23.000	0.000	21.000	2.000
	25%	275.000	286.000	284.000	330.000
	50%	393.000	367.000	365.000	357.000
	75%	458.000	438.000	426.000	407.000
	max	701.000	628.000	694.000	634.000
	pearson (corr)		0.907 ($p < 0.01$)		0.882 ($p < 0.01$)
N_empty_houses	mean	307.697	536.020	378.740	558.276
	std	152.558	226.519	104.894	180.824
	min	2.000	3.000	12.000	5.000
	25%	222.000	418.000	328.000	482.000
	50%	299.000	565.000	383.000	605.000
	75%	439.000	720.000	448.000	683.000
	max	632.000	1054.000	626.000	915.000
	pearson (corr)		0.329 ($p < 0.01$)		0.423 ($p < 0.01$)
N_houses_in_negative_equity	mean	111.932	86.326	63.244	66.777
	std	66.820	27.779	22.645	22.442
	min	0.000	0.000	0.000	0.000
	25%	72.000	71.000	49.000	53.000
	50%	88.000	86.000	63.000	64.000
	75%	133.000	104.000	77.000	77.000
	max	324.000	181.000	147.000	149.000
	pearson (corr)		0.674 ($p < 0.01$)		0.776 ($p < 0.01$)
N_demolished_houses	mean	6.518	7.242	6.228	7.004
	std	3.073	2.938	3.016	2.774
	min	0.000	0.000	0.000	0.000
	25%	4.000	5.000	4.000	5.000
	50%	6.000	7.000	6.000	7.000
	75%	8.000	9.000	8.000	9.000
	max	23.000	31.000	21.000	23.000
	pearson (corr)		0.117 ($p = 1.406$)		0.109 ($p = 1.364$)
N_people	mean	2107.253	2256.194	2073.900	2294.336
	std	544.515	362.434	485.047	261.922
	min	934.000	1522.000	1142.000	1859.000
	25%	1596.000	1930.000	1640.750	2119.000
	50%	2253.000	2240.000	2146.000	2200.000
	75%	2487.000	2494.000	2408.000	2398.000
	max	3149.000	3092.000	3136.000	3092.000
	pearson (corr)		0.897 ($p < 0.01$)		0.830 ($p < 0.01$)

Figure 5.6.2: Descriptive statistics of output data from all four model runs for both scenarios. Where rows with red borders are those described in sub-section 5.3.9

Chapter 6

Conclusion

The work within this thesis aimed to explore the emergence of complex behaviours across several distinct agent-based models using reinforcement learning algorithms. In doing so, it makes several contributions to the literature. Firstly, the work makes a contribution by critically setting out the theory and methodology of emergent complex behaviours within the context of agent-based models, additionally, a broad review of traditional decision-making frameworks are critically assessed in relation to strengths and weaknesses when compared with RL. Secondly, the thesis presents three completely different agent-based simulations from predator-prey interactions to burglary behaviour and housing markets, where the first two were developed using less-explored game engine software, contributing to future agent-based modelling programming stack. Thirdly, the thesis shows that behavioural theories derived from social science such as rational-choice can be organically learnt by RL agents without explicit decision-rules, currently, this is the only model to do this explicitly. Fourthly, this thesis shows that RL agents are able to adapt to changing environments and continue to perform sub-optimally when experiencing new situations (situations they never experienced during training), which has been a concept that ABM practitioners have described as a weakness using traditional decision-making algorithms, refer to Chapter 2. Finally, the thesis contributes to decision-support system literature by demonstrating how an RL agent can learn trends and behaviours of a complex system, i.e., the housing market and make real-time decisions to counteract and alleviate exogenous shocks to the market.

This chapter concludes the thesis and provides a summary of the research undertaken. Section 6.1 summarises the thesis and demonstrates the extent to which the

aim and objectives, detailed in Chapter 1, have been met. The limitations of the thesis are discussed in Section 6.2, in Section 6.3 the scenarios by which RL is most suitable compared to less suitable is discussed. In Section 6.4 notes recommendations for future work. An outlook on both producing and utilising RL in future agent-based models and concluding remarks are presented in Section 6.5.

6.1 Thesis summary and contribution to the literature

The aim of this thesis, as stated in Chapter 1, is to explore how reinforcement learning algorithms can be incorporated into agent-based models to generate complex agent behaviours across several domains. To fulfil this aim, five research objectives were established. This section revisits these objectives and assesses the extent to which they have been met by the work in this thesis.

Objective One: Review the existing literature describing methodological applications of behavioural frameworks in ABMs to simulate emergent complex behaviours.

Objective One was fulfilled through a review of the literature in Chapter 2. Chapter 2 reviews the underlying theory and practice of emergent complex behaviours in the context of agent-based modelling. The review provided a chronological description of emergent complex behaviours from the initial theoretical underpinning work in the 80s to applied simulation models in the early 2000s. The review also addressed the broad strengths and weaknesses in modelling emergent complex behaviours, highlighting research articles that focused specifically on the issues of predictability, verification, and validation. The second part of the review provides a critical discussion of computational simulations of emergent complex behaviours. This part discusses the use of ABMs with RL to simulate emergent complex behaviours and study different systems in various research domains. Several examples are provided, such as traffic control, group behaviour, and environmental management. The section also discusses the limitations of using RL in ABMs. The modelling of social and cultural factors in decision-making is highlighted as a limitation, and some studies' strengths are presented, such as cooperative and competitive problem-solving. The third part describes traditional decision-making frameworks adopted in ABM applications. Here, an in-depth analysis of the framework itself, its strengths and weaknesses recorded in literature including examples are provided. Furthermore, advanced behavioural frameworks such as RL is critiqued and compared to other decision-making frameworks. The last part of the review delves

into domain specific applications of ABMs in environmental criminology and housing market research. The review describes applications of ABMs in the aforementioned domains highlighting the areas in which RL can contribute compared to traditional decision-making frameworks. Additionally, Chapter 2 conveys that while RL has been shown to be effective in solving complex decision-making tasks, the limitations need to be considered, and more research needs to be done to compare findings across different disciplines.

Chapter 2 highlights the fact that as ABMs are adopted across disciplines, it is difficult to pinpoint a specific approach for a given problem, therefore, literature in various research domains is exhausted including traffic control system optimisation, group behaviour modelling, and environmental management systems. The chapter explores the strengths and limitations of using RL as the decision-making framework in ABMs, highlighting studies where RL has produced promising results, and those that show sub-optimal behaviours, including when agents had conflicting goals or when the environment was highly stochastic. One significant limitation observed was that agents were not able to communicate with each other, limiting the complexity of the behaviours that could be simulated. As mentioned before, the chapter also highlights areas where more work is needed, i.e., the modelling of social and cultural factors and the absence of a universal design and development framework for RL, limits its generalisability.

Chapter 2 has successfully met its objectives, not only has the chapter described the methodological applications of behavioural frameworks in ABMs to simulate emergent complex behaviours, but it's delved into the theoretical underpinnings of the term and developed a holistic review of the chronology of the term from animal behaviour to computer models. The RL decision-making approach was critiqued and compared to other traditional decision-making approaches including in the environmental criminology and housing market domains where its contributions to the literature were discussed.

Objective Two: Develop a proof-of-concept ABM using reinforcement learning where complex behaviours organically learnt by agents are quantified and analysed, assessing the extent by which emergent complex behaviours are observed.

Objective Two was achieved in Chapter 3 where the predator-prey ABM was developed using RL as a behavioural framework. The chapter discusses the importance of understanding complex adaptive behaviour of agents in complex systems and the role of ABM in simulating such behaviours. The chapter highlights the fact that, identify-

ing individual-level behaviour rules to embed within ABM remains a challenge. The chapter explores how agents can learn to exhibit complex adaptive behaviours that vary across space and time through RL. RL algorithms aim to learn how to apply an action in a given situation by mapping situations to actions through trial-and-error and measuring rewards. The chapter aims to evaluate the use of a new RL algorithm, proximal-policy optimisation, for simulating adaptive behaviours in an ABM of predator-prey interactions. Two experiments are designed to investigate the impact of training duration and the presence of an unknown stimulus on task efficiency and agent adaptation. The chapter utilises new game-engine software and packages to train and examine the models quantitatively and qualitatively. A framework is proposed to be used to record, analyse, and interpret real-time behaviours displayed by agents during simulation runs.

Chapter 3 proposes an agent-based model with reinforcement learning, where the agent population is made up of prey and predator agents. The prey agents use RL decision-making with simple actions such as turn and move, they are rewarded if they gather food and penalised if they get caught. The predator agent uses a production-rule system with simple rules such as if prey in line of sight, catch prey, else move randomly around the environment. We observe emergent complex behaviours post-experiments where the prey agents learn to spatially avoid areas in which the predator spends most of its time navigating, i.e., the centre of the environment space. Overall, the results show that prey agents that train for longer, are more effective in devising goal-oriented strategies. Moreover, prey agents that weight the risks between multiple penalties perform sub-optimally in achieving a goal compared to agents that focus solely on a single reward and penalty. Some qualitative behaviours that emerged were hiding, evading, foraging, and circling. The aforementioned behaviours were learnt organically and not hard coded in any form.

The discussion and conclusion sections of Chapter 3 evaluates the effectiveness of developing an ABM in conjunction with the PPO technique as a behavioural framework. The research concludes that the RL technique supports agents in the behavioural evolution of complex decision making that resembles those observed in the real-world. However, the research also identifies several limitations, including the difficulty of identifying appropriate training parameters and interpreting individual-level agent behaviours. Despite these limitations, the research demonstrates that the combination of game-

engines (Unity) and RL techniques can lead to valuable outcomes, and future research could explore the application of these techniques to a range of fields including population dynamics and the relationships between people and enforced rules to prevent the spread of a contagious virus.

Objective Three: Assess whether reinforcement learning can be used to investigate theoretical perspectives in social science using criminology as an example.

Objective Three was achieved in Chapter 4 where an ABM of burglary dynamics is proposed where potential offender agents are modelled with RL decision-making.

The chapter discusses the application of RL to ABMs of environmental criminology. The chapter explores the extent to which RL-trained offender agents behave in accordance with environmental criminology theories such as the rational choice perspective (RCP), subsequently generating crime patterns observed in empirical studies of crime. The chapter includes a literature review of ABMs in environmental criminology, a description of the model logic and methodology, results from a series of experiments using the model, and a discussion of findings and contributions to the field. The chapter argues that ABM in environmental criminology is still in its infancy, and the proposed use of RL can contribute to further advances. The proposed model incorporates spatial perceptions where each agent can observe the environment, and spatially there are risks distributed which symbolises Situational Crime Prevention Interventions (SCPIs). Through experimentation, these interventions are increased or decreased mid model-run to increase perceived risk and observe adaptive behaviours.

The results in Chapter 4 show that using RL as a behavioural framework can improve the accuracy of ABMs that simulate crime dynamics. The results also demonstrate behavioural heterogeneity among offender agents and the impact of the environment on their decision-making. Some offender agents are spatially better suited to offending than others. The study finds that crime patterns are clustered in areas with higher target attractiveness and that crime concentration is positively skewed. Crime patterns in the simulations also share some characteristics with empirical crime patterns. A crucial finding was that offenders organically learned to abide by the Rational Choice Perspective criteria, where they learned to offend at targets where rewards outweighed effort and risk in enacting the offence, and vice versa.

The results from Chapter 4 showed that offenders learned the following characteristics of crime, such as spatial concentration of crime, the journey to crime curve, assault

reputation, offender discouragement, and repeat victimization. The model demonstrated that "lack of opportunities" was more effective at reducing crime numbers compared to the adoption of SCPIs. The chapter also discussed the limitations of the model, such as a lack of computational capabilities and stochasticity. The chapter concluded that future studies could explore how SCPIs might affect offending behaviour in real-world cities and where crime is likely to displace.

Objective Four: Assess the capabilities of reinforcement learning agents to emulate the behaviours of policymakers who are tasked to make informed decisions about the dynamics of a complex system.

Objective Four was achieved in Chapter 5 which discusses the use of ABMs in housing market research and proposes a hybrid model that integrates RL with a housing market ABM. The chapter presents a series of experiments to investigate if an complex adaptive central bank agent can learn trends from a housing market in real-time and make decisions to fulfil a goal, such as decreasing homelessness, or counteracting abrupt interest rate hikes. To investigate whether RL can manipulate the housing market, the chapter reproduces two identical experiments conducted in a previous study as a comparator, and the model outputs reflect the consequences of RL decisions.

Chapter 5 presents the experiment results to achieve healthy housing market goals in the context of a simulated shock to the market. The experiments were run for 400 simulation timesteps, and at the 200^{th} timestep, a shock impacted key indicators to varying extents. The results showed that the RL agent could adapt to environmental changes in the model and adjust its behaviour accordingly. The RL agent learned to approach the 100% LTV experiment conservatively with slight adjustments but mainly holding interest rates, which was a response to the sensitivity of indicators of the housing market to prevailing interest rates. The study also found that the confidence intervals were much wider leading up to and right after the shock at timestep 200, which meant that the RL agent had explored more state space at these crucial points during training.

The chapter compared the RL agent's outputs to the goal criteria for a healthy housing market state. The results showed that the RL agent successfully ensured the house price-to-income ratio was below seven even after the shock. However, the median-house-price-for-sale indicator showed that RL was more effective in achieving the goal in the 80% LTV experiment than the 100% LTV experiment. The houses-in-negative-

equity indicator was a more complex indicator for RL to achieve the goal of ≤ 123, but pre- and post-shock, there was a downward trend where the number of households in negative equity was less than 123, even achieving less than 50 near the end of the simulation.

Overall, the chapter demonstrated that RL could achieve healthy housing market goal criteria and alleviate the shock effect on the market in real-time. The chapter highlights the importance of choosing appropriate goal conditions when applying RL to complex systems such as the housing market. Finally, the chapter contributes to the literature on ABMs and RL in housing market research and provides insights for policymakers looking to mitigate the impact of market shocks autonomously.

6.2 Limitations of the research

This thesis has explored the emergence of complex behaviours in agent-based models using reinforcement learning algorithms. The research has successfully developed three conceptual agent-based models to demonstrate how reinforcement learning can enable agents to learn different behaviours and adapt to new situations. These models were built in chronological order where the predator-prey model was a proof-of-concept demonstrating RL's ability to learn behaviours. The burglary model showed how theoretical behaviours were learnt and empirical patterns reproduced by RL agents organically and lastly the housing market model showed how RL can be utilised in a real-world scenario where central banks make decisions to alleviate economic turmoil. However, there are several limitations of the research which are discussed in this section.

6.2.1 Calibration and validation

Calibration and validation of ABMs are crucial steps undertaken by researchers to establish a degree of confidence in the accuracy of model outputs in representing the real-world systems they were designed to simulate. However, in the present thesis, calibration and validation procedures were limited in scope as the primary objective was to demonstrate the efficacy of RL in generating emergent complex behaviours and adapting to new situations without pre-determined agent rules or behaviours. To ensure that output results were measurable using contemporary data analysis techniques, the models were kept simple in architecture, and all were proof-of-concept models with

some underlying theoretical foundation, as illustrated in Chapters 4 and 5. The housing market model was the only model subjected to a validation exercise, which aimed to verify that the replicated model performed similarly to the original Netlogo version. Given the limitations of RL algorithms described below, future calibration and validation exercises will require more advanced computational resources.

Moving forward, the logical progression of this research is to develop more sophisticated models based on empirical data and perform extensive calibration and validation exercises to enhance confidence in the model outputs. Such exercises are essential in establishing the effectiveness of these approaches in generating insights that represent real-world systems accurately, making them more powerful tools for decision-making in various fields.

6.2.2 Quantifying reinforcement learning behaviours

Traditional decision-making approaches in ABM have embedded some underlying decision constraints such as what beliefs look like in BDI frameworks or physical conditions in the PECS framework. These behavioural constraints allow researchers to quantify the qualitative behaviours that are produced in the model as the agent behaviours are bounded. Conversely, RL agents can be trained to optimise complex and highly nonlinear reward functions, which can make it difficult to understand why the agent is behaving in a certain way. Moreover, the internal decision-making processes of the agent can be opaque, making it challenging to identify the factors that are driving its behaviour. In Chapters 3, 4 and 5 we discuss how these limitations can be overcome and identify formal verification algorithms as an area of future research.

6.2.3 Overfitting

Overfitting is a common limitation of machine learning algorithms; this also includes RL. Specifically in RL, overfitting can occur when an agent is trained for many iterations, resulting in the agent learning to perform well on the training environment but failing to generalise its learnt behaviours to new, unseen environments. This can happen when the agent learns to exploit specific features of the environment that are unique to the training environment, but not present in other environments.

This limitation was handled by utilising the proximal-policy optimisation (PPO) algorithm. PPO uses a surrogate objective function to update the policy (the policy in

PPO is represented as a neural network, where the input is the state of the environment, and the output is a probability distribution over the possible actions), which is designed to ensure that the policy update is not too large. This helps to prevent the policy from overfitting to the current batch of training data.

6.2.4 The exploration vs exploitation problem

Exploration and exploitation are two important concepts in RL that refer to the trade-off between gathering information about the environment (exploration) and using the information that has already been gathered to maximise the reward (exploitation). Specifically, exploitation refers to choosing actions that the agent believes will lead to the highest reward based on its current knowledge of the environment. While exploration refers to choosing actions that the agent is uncertain about in order to gain more information about the environment. The challenge is to balance exploration and exploitation in order to find an optimal policy that maximises the reward. If the agent only exploits, it may miss out on better strategies that it has not yet discovered. Conversely, if the agent only explores, it may not maximise its reward in the short term.

The presented models deal with this limitation through the PPO algorithm, which balances exploration and exploitation by using a clipped surrogate objective function (refer to Appendix 4.8.2) that limits the size of policy updates, and by adjusting the clipping parameter adaptively. This helps to prevent drastic changes in the agent's behaviour and promotes exploration in areas where the agent is uncertain.

6.2.5 Computational complexity

Computational complexity is a fundamental concept in computer science that measures the amount of computational resources required to solve a given problem. The complexity of a problem can depends on several factors, these include the size of the input, the nature of the problem, and the algorithms used to solve it.

Put simply, the limitation for RL is that it can be computationally expensive to train an agent to find an optimal policy in a complex environment. This is because the agent must learn from experience by interacting with the environment, which can require a large number of trials and a significant amount of time and resources. Moreover, the computational complexity of RL can increase exponentially as the number of states and actions in the environment increases. This can make it difficult to scale up RL

algorithms to handle more complex tasks and environments. The models presented in this thesis were computationally tractable, however, if these models contained empirical fidelity such as real cities, then training millions of agents would make it less tractable. This is an area of research that RL practitioners are focusing on.

6.3 Applicability of Reinforcement Learning

Reinforcement Learning (RL) is a powerful tool within machine learning, offering the ability as proven in this thesis to learn optimal behaviours through trial and error. However, the decision to utilise RL in modelling must be made judiciously, taking into account the strengths and limitations of the approach. Here we discusses various scenarios where the application of RL is advantageous and contexts where its use might not be recommended based on (Sutton & Barto, 2018b).

When to use RL:

- Problem Space with Clear Reward Signals: RL is suitable for environments where the objective can be quantified via clear reward signals. These are scenarios where the consequences of actions can be directly linked to a numerical reward.
- Dynamic and Complex Environments: RL excels in situations that require adaptive behaviour in complex and dynamic settings where traditional decision-making algorithms might fall short.
- Situations Requiring Exploration: RL is beneficial in tasks that involve a significant amount of exploration to learn about the environment, especially when explicit programming of all potential scenarios is unfeasible.
- Sequential Decision-Making: Scenarios that involve a sequence of decisions leading to a long-term goal align well with RL's sequential decision-making framework.

When not to use RL:

- Simple or Static Problems: For tasks that are static or can be solved with simple rule-based methods, RL may introduce unnecessary complexity without additional benefit.

- Limited Computational Resources: RL can be computationally intensive, particularly in high-dimensional spaces. If computational resources are a concern, RL may not be the most efficient approach.
- Lack of Sufficient Data: RL requires extensive interaction with the environment or a simulation for learning. In cases where such interactions are limited or costly, RL may not be practical.
- Environments with Poorly Defined Rewards: In situations where it is challenging to define what 'good' behaviour entails quantitatively, RL might struggle to learn effectively.

The decision to utilise RL should be informed by a careful assessment of the specific requirements and constraints of the problem at hand. The use of RL must be underpinned by a solid understanding of its theoretical foundation and practical implications within the context of agent-based modelling.

6.4 Recommendations for future work

There are several avenues for potential research identified by the work in this thesis. The most beneficial avenue for future work may be the further development of the models produced in Chapters 4 and 5. More specifically, empirical data may be used to calibrate the offenders and central bank agents to mimic real-world scenarios, this could be an activity that validates the learnt behaviours of agents in a real-world context which could make RL more powerful and realistic. Furthermore, with a calibrated and validated RL model, new insights may be identified such as the way in which offender's learnt to exploit loopholes in security measures or central bank agents utilising new monetary policy strategies previously unbeknown to economists. Another area that can be explored for both models is to incorporate more realistic reward functions. Reward functions are crucial in RL, as they define the objective that the agent is trying to achieve. In social processes such as offending and banking, however, it can be challenging to define a reward function that accurately captures the complexity of human behaviour. Researchers could explore new methods for defining reward functions that better capture the nuances of social processes, such as incorporating social norms and values.

Another area discussed on several occasions as a limitation of RL which needs to be researched further is to improve the scalability of RL algorithms. Many reinforcement learning algorithms are computationally expensive and require large amounts of data to train. As such, it can be challenging to apply these algorithms to complex social processes that involve many agents. Researchers could explore new algorithmic techniques or adapt existing algorithms to improve their scalability and make them more feasible for use in large-scale agent-based models.

Future research could focus on interpretability and transparency. RL agents can be challenging to interpret, which can make it difficult to understand why they make the decisions they do. Researchers could focus on developing methods for interpreting and visualising RL models, which would improve their transparency and allow for better insights into the underlying social processes. Furthermore, these findings can then help evaluate the impact of RL agents on social processes. It's important to evaluate the impact that RL agents have on social processes, i.e., to what extent does an RL agent impact the system positively or negatively. While these agents have the potential to improve society, they could also have unintended consequences or exacerbate existing social issues. Researchers could conduct simulations to evaluate the impact of RL agents on different social processes and identify any potential issues. An example scenario that may benefit from this, is environmental criminology. Researchers could create an ABM where the agents represent both criminals and police officers. The model could also include environmental factors such as lighting, landscaping, and the layout of buildings. The researcher could then introduce RL agents to represent the police officers. These agents would learn over time which areas of the neighbourhood are more likely to experience crime and adjust their patrol patterns accordingly. Once the model has been run for a sufficient number of iterations, the researcher could evaluate the impact of the RL agents on crime rates. They could compare the crime rates in the model with RL agents to those in a model without RL agents (similar to the work conducted in Chapter 3). They could also evaluate the impact of different parameters such as the number of RL agents, the amount of data they are given, and the types of crimes they are trying to prevent. By conducting these simulations, the researcher could identify any unintended consequences or social issues that might arise from using RL agents in this context. For example, the RL agents might focus their patrols on certain areas of the neighbourhood while excluding others, leading to a perception of

bias or unfairness. By identifying these issues, the researcher could modify the model to mitigate their impact and create a more effective tool for reducing crime.

6.5 Outlook and concluding remarks

The findings of this study provide opportunities for future research to explore the integration of RL algorithms into hybrid-ABMs further. Firstly, future studies could investigate how the proposed models could be refined to enhance their predictive capabilities. Secondly, this research could be extended to other complex phenomena such as social dynamics, healthcare, and environmental systems. Finally, future research could also investigate approaches to evaluate RL algorithm's effectiveness in modelling accurate real-world decision-making.

In conclusion, this thesis has contributed to agent-based modelling by demonstrating how RL algorithms can enhance the decision-making accuracy of ABMs by introducing learning and adaptability. The proposed hybrid-ABMs were tested on three distinct complex phenomena, and the findings indicate that the RL agents can learn complex behaviours from their environment, adapt to previously unbeknown situations, and perform relatively well. The contributions of this research will inform future research that should have practical implications. Furthermore, this research highlights the potential of RL algorithms for enhancing the predictive capabilities of ABMs and opens up new avenues for future studies in this area. Overall, this thesis has advanced the understanding of ABMs and reinforced the potential of RL algorithms for improving complex real-world decision-making in future models.

Appendix A

Peer-reviewed articles published during collaborative research

The following publications resulted from collaborative projects conducted by like-minded post-graduate researchers and academics from various institutions globally. None of the materials presented henceforth are used in the main thesis; however, it showcases collaboration instigated by the thesis author during the PhD programme.

Article one A.1, *Olmez, S.; Douglas-Mann, L.; Manley, E.; Suchak, K.; Heppenstall, A.; Birks, D.; Whipp, A. Exploring the Impact of Driver Adherence to Speed Limits and the Interdependence of Roadside Collisions in an Urban Environment: An Agent-Based Modelling Approach. Appl. Sci. 2021, 11, 5336. https://doi.org/10.3390/app11125336*

Article two A.2, *Olmez, S.; Thompson, J.; Marfleet, E.; Suchak, K.; Heppenstall, A.; Manley, E.; Whipp, A.; Vidanaarachchi, R. An Agent-Based Model of Heterogeneous Driver Behaviour and Its Impact on Energy Consumption and Costs in Urban Space. Energies 2022, 15, 4031. https://doi.org/10.3390/en15114031*

A.1 Exploring the Impact of Driver Adherence to Speed Limits and the Interdependence of Roadside Collisions in an Urban Environment: An Agent-Based Modelling Approach

Authors: Sedar Olmez, Liam Douglas-Mann, Ed Manley, Keiran Suchak, Alison Heppenstall, Dan Birks and Annabel Whipp

Abstract: Roadside collisions are a significant problem faced by all countries. Urbanisation has led to an increase in traffic congestion and roadside vehicle collisions. According to the UK Government's Department for Transport, most vehicle collisions occur on urban roads, with empirical evidence showing drivers are more likely to break local and fixed speed limits in urban environments. Analysis conducted by the Department for Transport found that the UK's accident prevention measure's cost is estimated to be £33bn per year. Therefore, there is a strong motivation to investigate the causes of roadside collisions in urban environments to better prepare traffic management, support local council policies, and ultimately reduce collision rates. This study utilises agent-based modelling as a tool to plan, experiment and investigate the relationship between speeding and vehicle density with collisions. The study found that higher traffic density results in more vehicles travelling at a slower speed, regardless of the degree to which drivers comply with speed restrictions. Secondly, collisions increase linearly as speed compliance is reduced for all densities. Collisions are lowest when all vehicles comply with speed limits for all densities. Lastly, higher global traffic densities result in higher local traffic densities near-collision sites across all adherence levels, increasing the likelihood of congestion around these sites. This work, when extended to real-world applications using empirical data, can support effective road safety policies.

Keywords: agent-based model; traffic simulation; urban environment; autonomous agents; data analysis; collisions; speed adherence

A.1.1 Introduction

A lack of adherence to speed limits can have serious consequences and pose a significant risk to life for drivers, passengers and members of the public. According to UK Government reports, car users account for the largest proportion of casualties across all categories of injury. A total of 736 car passengers/drivers suffered fatal collisions

in 2019 (Balendra, 2020; Murphy, 2020). Furthermore, on 30 mph (miles per hour) roads, 54% of cars exceeded the speed limit in the first quarter of 2019. In addition, 6% of these cars exceeded the speed limit by over 10 mph. This increased in the second quarter of 2019 to 56%. Similarly, 37% of fatalities among car passengers/drivers in 2019 occurred on urban roads—an increase of 1% since 2018. An additional 57% of fatalities occurred in rural roads, down 3% since 2018. These trends are evident outside of the UK. According to Pammer *et al.* (2021), almost half of the reported driving offences in the Northern Territory of Australia are regulatory; these include speeding and non-adherence to road rules. In Norway, a longitudinal study conducted on 145 young drivers (up to 25 years old) found that speeding behaviour was the main factor (80%) in causing motor vehicle collisions (Breen *et al.*, 2020).

Driving speeds are a vital component in exploring the factors that lead to collisions. Several empirical studies in speed and collision rates found evidence that crash rates increase faster given the increase in speed in minor roads compared to major roads. Two important factors related to collision rates are traffic density and traffic flow (Aarts & Van Schagen, 2006). Furthermore, the authors in Szumska *et al.* (2020) found that population density is a contributory factor in accident frequency. The authors suggest that population densities in cities are higher than in rural areas; thus, people are more exposed to vehicle collisions. Similarly, the authors in Maycock *et al.* (1999); Quimby *et al.* (1999) collected empirical data from drivers in the form of surveys to conduct studies in driver behaviour; this method was also adopted by (Fildes *et al.*, 1991). These studies found that an increase in speed led to an increase in collision rates and that fast moving vehicles have a higher crash rate than slow moving vehicles. Maycock *et al.* (1999); Quimby *et al.* (1999) both reported a power function to describe this relationship, while the authors in (Fildes *et al.*, 1991; Kloeden *et al.*, 1997, 2001) reported an exponential function. These latter three studies also found that the crash rate increases faster with increasing speed on urban than on rural roads. Methodological differences in the operationalisation of variables, and the influence of coincidental factors, all may account for differences in results at a detailed level (Aarts & Van Schagen, 2006).

Self-driving cars play an important role in vehicle collision research. Emuna *et al.* (2020) found that human-like driving policies are necessary to ensure the safety of passengers in these vehicles. The authors apply deep-reinforcement learning algorithms to simulate collision avoidance in dynamic settings. By adopting human expert knowl-

edge data and feeding these data into the model, the authors found that human-like driving policies can be achieved. Similarly, reference Pusse & Klusch (2019) developed a hybrid online POMDP planning and deep-reinforcement learning algorithm to enable self-driving cars to avoid collisions, including pedestrians. The research aims to deploy a collision-free navigation system such that vehicles are better equipped at handling high-risk scenarios. The authors found that their hybrid solution outperforms each applied technique, POMDP and deep-reinforcement learning on average. Moreover, the author in Spectrum & 2016 (2016) attempts to explore how and if ethics can be adopted in self-driving cars by comparing real-world scenarios where self-driving cars fail to adopt human intuition to avoid a collision with a pedestrian as doing so would result in the vehicle breaking its own intrinsic rules. The author in Spectrum & 2016 (2016) points out that self-driving cars cannot be sure that a road ahead is clear, such that it should cross and avoid hitting a person that it may encounter. These vehicles will estimate the confidence interval at 98 to 99 per cent, which ultimately means engineers would have to decide how high the confidence interval must be. Thus, engineers would need to consider what object is ahead, i.e., plastic bag or a person making this an essential line of enquiry in this field of research. While most of the research in self-driving car technologies is in its infancy, developing new technologies to handle collision is welcomed, which could, in turn, be adopted by regular non-self-driving vehicles.

Safety intervention policies in reducing variations in speed play an essential role in reducing collision rates. Interventions include speed humps, roundabouts, road markings, signposts and traffic lights. However, the measures that have been found to increase speed limit adherence are those that physically prevent a vehicle from driving faster than necessary, such as speed humps (Martens, 1997). Safety measures deployed in vehicles have also lead to a decrease in collisions. The European Union has made it mandatory through legislative requirements for vehicles to be fitted with advanced emergency braking systems and other measures, which lead to a decrease of 5000 fatal collisions on European roads per year (Moravčík & Jaśkiewicz, 2018). Alternatively, signposts that show the expected speed limit do not automatically imply that drivers will match the indicated speed limit. The authors in (Richards & Dudek, 1986) refer to static speed limit signposts as passive speed control. They argue that passive control alone is generally only sufficient at sites where the hazards are obvious, and drivers understand and accept the speed limitation (Martens, 1997). Ultimately, the authors

in Oei & Polak (1992) found that active signalling, using dynamic signs informing a driver that they are exceeding the speed limit, has more effect than passive control since drivers may also interpret the signal as an indication for impending danger. In addition, the authors in Galizio *et al.* (1979) found that, when comparing a signpost with a marked police car in increasing speed adherence, the police car had a significant effect on driving speed the drivers were in active fear of being reprimanded. According to (Kennedy *et al.*, 2005), features such as edge markings that visually narrow the road, the vicinity of buildings, reduced carriageway widths, barriers in the carriageway and pedestrian activity all tend to reduce speed.

The concepts discussed above provide insight into the literature on empirical evidence of collisions in Great Britain, the impact of speeding, safety intervention policies and road design. However, the decision to speed or comply with speed limits comes down to the individual driver. Speeding is a major contributory factor to roadside accidents (Peden *et al.*, 2004). To date, the majority of research in this area has investigated an extensive range of important factors from the viewpoint of those who exceed speed limits. This focus is understandable, given that faster vehicle speeds increase both risks of crash involvement and severity of crash outcomes (Fildes *et al.*, 2005).

The earlier statistics show that work needs to be done to curb the number of accidents on urban roads. This study will utilise a novel 3D Urban Traffic Agent-Based Model (Olmez *et al.*, 2021b) to conduct several experiments by testing multiple traffic density and speed limit adherence parameters to illustrate how these measures impact vehicle collisions among a heterogeneous agent population of vehicles in a simplification of an urban environment. This study adopts the widely accepted definition of collisions, defined by Saunier *et al.* (2010) as: "an observational situation in which two or more road users approach each other in space and time to such an extent that a collision is imminent if their movements remain unchanged". The study ultimately aims to assess the impact traffic density and speed limit adherence have on collision rates by utilising a novel agent-based model to provide recommendations on reducing these rates and inform policy.

Sub-section A.1.2 introduces the agent-based modelling methodology by describing what it is and how it has been adopted in similar research. Sub-section A.1.3 will describe the agent-based model using established protocols such as the Overview, Design concepts and Details (ODD) (Grimm *et al.*, 2006). This sub-section outlines

the purpose of the model, agents and environment characteristics. Sub-section A.1.4 describes the number of experiments conducted, the rationale behind them and an analysis of the subsequent outcomes. Lastly, Sub-section A.1.6 consists of the studies' initial aims, which was found after experimentation, and the recommendations made for future urban road infrastructure planning.

A.1.2 An Individual-Based Modelling Approach to Traffic Simulation

The traffic system is characterised by multiple individual actors (drivers) and a street network made up of individual rules such as right of way and speed limits. Given the nature of this system's individual-level components, it is evident that these systems are perfectly poised to be studied using individual-based modelling methods. According to Huston *et al.* (1988), individual-based modelling refers to simulation models that treat individual entities as unique and discrete components with at least one property, for example, age, height, position and these properties change during the life cycle of these entities. Therefore, in this study, vehicles can be thought of as individual heterogeneous entities with their own rules, while the urban street network is the environment in which these vehicle entities are observed from within. The aim is to test various interventions in this simplified world and collect observational data from these entities to assess the impact of these interventions.

Agent-Based Modelling (ABM) is a tool that allows the study of emergent behaviour of a system by simulating the actions and interactions of a collection of autonomous agents. It is used in a wide variety of disciplines such as ecology (McLane *et al.*, 2011), crime (Birks *et al.*, 2012) and sociology (Bianchi & Squazzoni, 2015). Implementing simple rules for the agents can lead to the reproduction of complex phenomena observed in the real world. Like all models, an agent-based model simplifies the isolated study of the effect of particular agent behaviour. In light of these advancements, several scholars advocate for contemporary simulation models as better suited in studying the underlying mechanisms of crash occurrence. Furthermore, these methods represent a richer and more detailed set of alternatives than statistical models (Davis & Morris, 2009).

Traffic in an urban space is a complex system that includes the environment (a road network with a plethora of features like intersections, traffic lights, roundabouts, hills and weather conditions) and drivers' behaviour as individuals. Urban traffic man-

agement has reached utmost importance worldwide as cities battle congestion and its impacts on public health and fossil fuel emissions. Many computational models exist—SUMO (Behrisch *et al.*, 2011), AIMSUN (Casas *et al.*, 2010), ARCHISIM (Bonte *et al.*, 2006) and PARAMICS (Cameron & Duncan, 1996), to name several—which aim to simulate traffic flow and aid the design and layout of urban roads and thereby help to minimise the impact of congestion. However, these models are typically explicitly collision-free; driver behaviour is formulated to prevent collisions. However, some contemporary academic research has focused on various aspects of roadside collisions. The authors in (Salim *et al.*, 2007) applied data mining techniques on data captured from intersection accidents to support real-time collision detection systems at intersections. A review of near-collision driver behaviour models by (Markkula *et al.*, 2012) found that most research has mainly been interested in the details of control in near-crash and crash-phases and have thus not needed to provide an account of why these states were reached in the first place. Furthermore, the authors in (Markkula *et al.*, 2012) argue that some authors have modelled reactions to collision warnings in various ways (Exposition *et al.*, 2002; Fitch *et al.*, 2008; Lee *et al.*, 2002), while none of the models has addressed the phenomena of behavioural adaptation to long-term system exposure. The model adopted in this study (Olmez *et al.*, 2021b) allows for the vehicle's life-cycle to be observed at an individual level while also observing the global patterns that emerge overtime at the street network level. The authors in (Markkula *et al.*, 2012) also found that almost all papers focused on a narrow set of collisions, namely rear-end collisions. Thus, they recommend looking at a more diverse range of pre-crash scenarios to achieve full credibility. The model adopted in this research deploys an environment that can observe multiple vehicle behaviours while applying the laws of physics. A variety of collisions among vehicles can then be observed. An example of this is where vehicles tip over if a collision occurs with a heavier, faster-moving vehicle, which ultimately captures a more realistic array of possible collisions and repercussions. Ultimately, collisions occur and contribute to various undesirable circumstances on roads, such as traffic jams and congestion. Furthermore, the rate, type, and severity of these collisions are emergent properties of the system, impacted by driver interaction, driver behaviour, and the environment.

Driver behaviour can be observed in many ways; these include surveys, camera footage, police reports, to name but a few. A prominent method within the liter-

ature is the Driver Behaviour Questionnaire (DBQ) introduced in 1990 by (Reason *et al.*, 1990). This questionnaire consists of 50 items describing various problems and violations while driving, which members of the public can fill out. After surveying 520 drivers, the authors in (Reason *et al.*, 1990) identified that errors are statistically distinct from violations, indicating that different psychological mechanisms trigger errors and violations. The authors in (De Winter & Dodou, 2010) found that violations were more prevalent among young drivers compared to senior drivers. On the other hand, errors decreased for younger drivers but remained constant with age among older drivers. The differences between attitudes among drivers in rural areas and urban areas reflect the significant difference among collisions in these areas. The authors in (Jones, 2007) identified that urban road network design consisted of higher lengths of road and traffic volume, which in turn increased the collision rates. The authors in (Nordfjærn *et al.*, 2010) add to this by highlighting the strongest predictors of fatality rates due to vehicle collisions as being age and number of residents in the geographical areas. The authors in (Pantangi *et al.*, 2019) adopted Naturalistic Driving Study (NDS) data which contains driver, trip and vehicle specific information. These data represent driver behaviour before, during and after the adoption of high-visibility enforcement programs. Furthermore, the study focused specifically on aggressive driving behaviour; these include speeding and tailgating to explore the intensity and duration of these behavioural patterns. The study found that high-visibility enforcement programs are likely to reduce speeding only in some instances. A survey result showed that drivers in rural areas are more likely to drive without a seat belt on or while intoxicated with alcohol compared to drivers in urban areas (Author *et al.*, 2007). The author in (Eiksund, 2009) identified two components that impact traffic risk. These are system risk and risk culture. The latter consists of factors independent of the driver, such as vehicle condition, weather and road plans. The former are human factors such as norms, feelings, attitudes and perceptions of risk. Adding to this, the authors in (Zhou *et al.*, 2021) attempted to analyse taxi driver speeding behaviours captured by GPS trajectory data. These data captured the hourly speeding frequency and average speeding severity of each driver. Their study concludes that aggressive driver behaviour among taxi drivers are linked to longer trips, short delivery time, high monetary value, driving at night, and, lastly, forced low-speed limits. Given all of the above, the authors in (Nordfjærn *et al.*, 2010) highlight the impact physical changes to road networks can have by enforc-

ing slower speeds such as road humps, while also indicating that driver behaviour may also be altered indirectly by influencing the public's attitudes and norms which links to the literature on "self-explaining" roads (Fildes & Jarvis, 1994) mentioned earlier in the Introduction sub-section.

Accident reduction is a crucial aim of transport management. It has been hypothesised that higher congestion leads to fewer road fatalities (Shefer, 1994) as congestion leads to lower overall speeds, and therefore collisions are less likely to occur. While some evidence of this relationship has been found in some scenarios such as on single-carriage rural roads in the UK (Baruya, 1998), results are much less conclusive in other scenarios such as in cities such as London (Noland & Quddus, 2005) or on highways such as the M25 motorway around London (Wang *et al.*, 2009). Furthermore, it has been argued that, in many empirical studies, congestion is evaluated using proxy variables such as volume over capacity ratio or employment density (Wang *et al.*, 2013) and that, to fully understand the impact of congestion, data with high levels of spatial and temporal resolution are needed (Retallack & Ostendorf, 2019). Microscopic traffic simulations track all vehicles' positions and velocity in the simulated road network in small time steps, allowing traffic dynamics to be observed in high spatial and temporal resolution. These simulations could complement empirical investigations and yield further insight into the interactions between road environment, speed, traffic density, congestion and accidents, adding to the debate regarding the impact of vehicular congestion on the frequency of road accidents (Cabrera-Arnau *et al.*, 2020).

The ABM described in this study has been designed to investigate the relationship between driver adherence to speed limits and the subsequent impact on the number of collisions. Unlike the previously described models, it utilises a physics engine provided by the Unity development platform. This feature allows physical collisions to occur between vehicles. The interaction between traffic density, adherence level and collisions can be studied by increasing the number of agents. The study will ultimately aim to argue for various policy interventions such as reducing or increasing density to reduce collision rates and regulate speeding in dense road networks and, by utilising the agent-based model, show the extent to which these interventions impact the system as a whole.

A.1.3 Model Description

This section describes the agent-based model adopted for this study. A general description of the model can be found at (Olmez *et al.*, 2021b). The model description includes the purpose of the model, the parameters that can be selected, the output variables from the model post simulation run, overview of the model workflow, and, lastly, a detailed description of the vehicle agents and environment. The Overview Design and Details (ODD) protocol will be utilised to explain the model (Grimm *et al.*, 2006).

Purpose

The agent-based model used in this research is the 3D Urban Traffic Simulator in Unity (Olmez *et al.*, 2021b), this includes the data produced during model experiments, found in the Supplementary Materials. The model was designed to provide researchers with the ability to simulate hypothetical vehicle drive–cycle activity scenarios in a 3D urban environment. The model utilises heterogeneous autonomous vehicle agents with granular control parameters such as vehicle mass, velocity, traction and downforce to name but a few. Similarly, the street network is developed around a built-up urban environment that contains the foundations of a dense urban street network with varying road speed limits and intersection rules.

Variables

The model requires input variables to run an experiment and output results that can later be analysed. The parameters that can be modified are listed in Table A.1.

The model consists of two entities: the vehicle agents and model environment. The vehicle parameters are:

- The vehicle mass parameter, each vehicle can weigh up to 7500 kg; the model distributes vehicles arbitrarily across the environment with varying weights, from small cars to large goods vehicles (LGVs) to capture heterogeneity, every vehicle must have a mass of at least one such that the laws of gravity apply during the simulation experiment. Mass only becomes significant when collision severity is of importance; however, all collisions are considered in this study.

- The top speed measure is between 30 and 45 mph, and is only applied to vehicles that do not adhere to speed limits (break speed limit rules), for example, vehicles

Entity	Parameter	Values
Vehicle	Mass	[1, 7500] (kg)
	Top Speed	[30, 45] (mph)
	Ray-cast Length	[1, 20] (m)
Environment	N. Of Vehicles	[1, 500]
	Speed Adherence	[0, N]
	Roads	1295
	Intersections	354

Table A.1: Model entities and parameter values, where [X, Y] are a random uniform distribution of values (inclusive) (Olmez *et al.*, 2021b).

that are driving on a 20 mph road can bypass the speed limit and drive at 45 mph which is more than double the speed limit. This measure is applied only if Speed Adherence is $\geq$1 (source (Balendra, 2020)).

- The ray-cast length parameter can be between 1 to 20. The variable assigns a distance between two vehicles in meters (source (Driving, 2022)).

The environment specific parameters are:

- The number of vehicles in the model, N; this can be between 1 and 500.
- The speed adherence variable can be between $0 \leq x \leq N$. This assigns the proportion of vehicles that will not adhere to the speed limits (vehicles that break the local and fixed speed limits) applied to the road which they are driving on during simulation.
- The urban road network consists of 1295 roads which vehicles drive on and 354 intersections which consist of right of way rules. The street network has been developed to depict a small urban town.

The parameters mentioned above, once selected, are used to initialise the experiment (model-run) which lead to output variables. These variables observe data points every step of the simulation experiment. Table A.2 describes the output variables that the model produces.

The model outputs thirteen variables (refer to Table A.2). The agent ID variable allows for a micro-level analysis of the agent behaviours during model execution at the street level, and this helps identify specific agents in the environment. The collisions variable is a cumulative number that increases each time the vehicle collides with another; this includes contact made between two or more vehicles on all road types and intersections. Top speed is the speed limit associated with the road that the vehicle is currently on, and the vehicle is trying to match the speed; however, in scenarios where some vehicles do not adhere to speed limits, the top speed for those vehicles would be a value between 30 and 45 mph ultimately breaking the speed limit. The current speed variable is the vehicle's speed at the current time of the simulation run. The distance of travel is in meters which tracks the vehicle's distance from the starting position on the road network up until the current simulation step. The ray-cast length variable is the distance the vehicle can identify objects ahead, for example, other vehicles. Traction control is either 1 (on) or 0 (off). If the traction control is on, the vehicle has full traction capability such that each wheel can adapt to the surface; however, it is not utilised for this study as not all vehicles have access to traction control. The velocity magnitude is a scalar value demonstrating the rate of motion at a specific time. The vehicleMass variable assigns a weight to the vehicle between 1 to 7500 in kilograms to capture heterogeneity. The physics engine in Unity requires that every object has a mass assigned to it to ensure gravity is applied. Downforce coefficient is between 0.1 and 10; for this research, it is left at 0.1 to have no impact. Lastly, date-time stamps are included in each row of data recorded such that time-series analysis can be applied (Olmez *et al.*, 2021b).

Model Overview

The agent-based model was developed using Unity. Unity is a 3D software development platform consisting of a rendering and physics engine and graphical user interface. Unity has received wide-spread acceptance in several industries, including gaming, automotive and film (Juliani *et al.*, 2018).

The following workflow diagram describes the processes that the model (Olmez *et al.*, 2021b) undergoes during run-time.

The Urban Traffic Simulator (Olmez *et al.*, 2021b) workflow (refer to Figure A.1.1) takes input values for the five variables described earlier (refer to Table A.1). The

Variable	Output Type	Example Value
AgentID	Integer	−38,572
xAxisPos	Float	75.94560
zAxisPos	Float	20.1927
collisions	Integer	12
topSpeed(mph)	Float	20.0
currentSpeed(mph)	Float	18.0
distanceOfTravel(meters)	Float	13.0
raycastLength	Integer	6
tractionControl	Integer	0
velocityMagnitude(BETA)	Float	0.195808
vehicleMass	Integer	1500
downforce	Float	0.1
date-time	DateTime	18 January 2021 13:05:40

Table A.2: Model output variables, source (Olmez *et al.*, 2021b).

software then resets all parameters to start the simulation scene, producing the agents and environment. Once the model has reset, the model produces all agents, starting locations, and environment parameters before the simulation starts. Now, the model runs each frame, and every change that occurs is stored with a time-stamp. Fixed Update is used to compute any physics elements such as vehicle wheels, mass and velocity. The Update method computes variables in each frame. The model utilises Fixed Update due to the number of physics components used; therefore, multiple changes occur during simulation run-time for each frame, and these changes are captured to output the thirteen variables' (Table A.2) post-simulation run; once this is done, the model is stopped (destroyed).

Agent

The vehicles are classed as autonomous agents; the vehicle population is heterogeneous (every vehicle has distinguishable characteristics). These agents apply similar characteristics to real-world motor vehicles; they have four wheels, steering angle, traction, mass, and drag. Each vehicle agent applies the following set of rules during its drive cycle, refer to Algorithm 2.

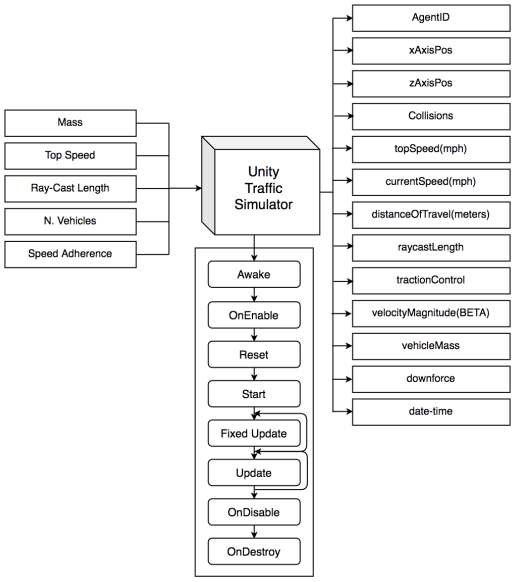


Figure A.1.1: Workflow diagram depicting processes that the Urban Traffic Simulator undergoes during run-time.

The rules described in Algorithm 2 allow vehicle agents to navigate the environment and collect data. Each vehicle follows the same rules. However, the features vary and depend on the input values from Table A.1. These vehicle agents are a simplification of real vehicles. Therefore, it is not expected to perfectly simulate real-world vehicles but includes the fundamental features that all vehicles retain.

If a vehicle is not adhering to speed limits, it can increase its speed between 30 to 45 mph. If vehicle X is ahead of Y, Y given the rules in Algorithm 2 should decrease speed to match vehicle X's speed. When a vehicle arrives at an intersection, if it has the right of way, i.e., on a horizontal lane and no vehicles are at the intersection, it reduces its speed to 10 mph and drives through the intersection. If the vehicle is at the intersection and does not have the right of way, it should wait until the intersection is cleared. If the vehicle is at an intersection, it does not have the right of way, and there are no vehicles at the intersection, the vehicle is free to reduce speed to 10 mph and drive through the intersection. Lastly, all vehicles that adhere to the speed limit increase or decrease speed to match the road's speed limit (Olmez *et al.*, 2021b).

Algorithm 2: Vehicle agent rules in pseudocode.

```
while Model running do
    Drive;
    if not_adherence == true then
        accelerate to top speed [30, 45];
    else
        accelerate matching road speed limit;
    end
    if vehicle_ahead == true then
        match speed of that vehicle;
    else
        continue at current speed;
    end
    if at_intersection == true AND vehicle_present == false AND right_of_way
      == true then
        reduce speed and drive out of intersection;
    else if at_intersection == true AND vehicle_present == true AND
      right_of_way == false then
        halt till intersection_clear == true;
    else if at_intersection == true AND vehicle_present == false AND
      right_of_way == false then
        reduce speed and drive out of intersection;
    else
        halt till intersection_clear == true;
    end
end
```

Environment

The vehicle agents described earlier require an environment to function within. The model (Olmez *et al.*, 2021b) deploys an urban street network that is described as a T-type network (Han *et al.*, 2020). This street network contains similar characteristics to downtown Philidelphia (Boeing, 2020) and San Francisco (Porta *et al.*, 2006). T-network patterns are like grid-shaped networks but include t-junctions. Several added features such as eight-lane intersections described in (Filocamo *et al.*, 2020) are also utilised to add complexity. The street network contains 1295 roads and 354 intersec-

tions, which were arbitrarily generated to cover a small town. The individual roads, speed limits and intersection rules are described in the following Figure A.1.2.

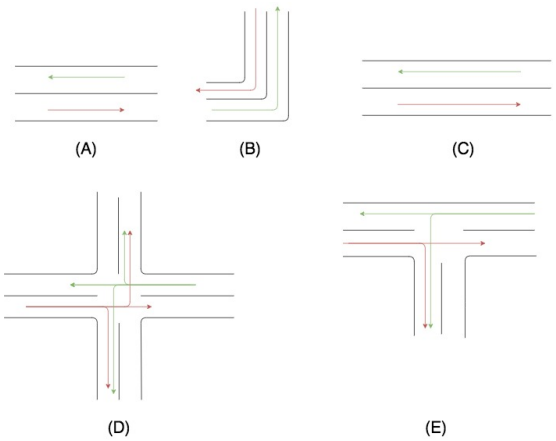


Figure A.1.2: Urban Street Network roads and intersections, A: two-way local road, B: two-way corner road, C: two-way fixed road, D: eight-way intersection and E: two-way T-junction.

The environment contains three road types with varying fixed and local speed limits and intersections with right of way rules. The environment is a simplification of the real world. Therefore, it does not utilise all intersection types. Moreover, overtaking is not utilised in the model as passing-lanes (overtaking lanes) do not exist in the street network and are commonly found in motorways or multi-lane highways (Clarke *et al.*, 1998). However, it does contain the basic features of an urban street network which have also been observed in several cities across the United States (Boeing, 2020; Porta *et al.*, 2006). The following list describes each road, intersection and the speed limits assigned to these roads from Figure A.1.2:

- (A) Two-way local road with a speed limit rule of 20 mph.
- (B) Two-way corner road with a speed limit rule of 10 mph.
- (C) Two-way fixed road with a speed limit rule of 30 mph.
- (D) Eight-way intersection, where right of way is for traffic on horizontal lanes, speed limit rule of 10 mph.

- (E) Two-way T-junction, right of way is for horizontal lanes, speed limit rule of 10 mph.

The speed limits for the three road types (Figure A.1.2A–C) were derived from UK government sources such as (Balendra, 2020), where urban streets consist of local 20 mph and fixed 30 mph zones; however, corner roads sometimes require lower speeds such as 10 mph as vehicles require more room to turn. A UK Government report identifies roads in built-up areas as having a fixed speed limit of 30 mph. However, for dense areas—usually city centres—this may be designated 20 mph by local councils to keep pedestrians safe from collisions (Department for Transport, 2006).

For comparison, the urban street network in the Urban Traffic Simulator (Olmez *et al.*, 2021b) is roughly the same size as the town of Morley, UK (refer to Figure A.1.3). Morley has 1526 roads compared to 1295, which the urban street network possesses.

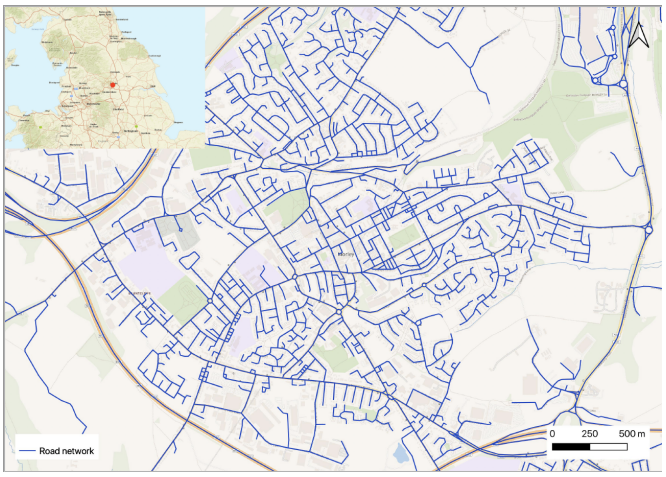


Figure A.1.3: Urban Street Network of Morley, UK (data source: (Survey, 2021)).

Summary

The model description section describes the rules vehicle agents follow for every road type and intersection it encounters. Five rules govern the vehicle's behaviour; these broadly involve increased or reduce speed depending on road or speed adherence, interacting with intersections in a safe way to reduce the risks of collisions. The environment comprises three road types and two intersections, with varying local and fixed speed limits taken from empirical data via UK government sources. Lastly, the town of Mor-

ley, UK happens to be very similar in size to the urban street network applied in the model; this provides a realistic snapshot of the global scale of the street network involved. In the next section, the model is used to run nine hypothetical scenarios. The output data from these scenarios will be quantitatively analysed in several ways.

A.1.4 Experimental Results

As mentioned previously, this study aims to quantify the relationship between speed limit adherence within different population sizes and the subsequent impact on collisions. The experiments will conduct multiple model execution scenarios under nine conditions, refer to Table A.3. The goal is to quantitatively identify the best and worst-case scenarios concerning the number of collisions in an urban street network. More specifically, low, mid and full adherence to speed limits will be compared across low, mid and high traffic density (number of vehicles); these are identified as the independent variables, while the dependent variable is the number of collisions. All other parameters will remain constant to ensure a heterogeneous population of vehicles across all experiments. The model is still in its infancy and can be thought of as a proof of concept. Thus, factors such as weather and time of day have not yet been implemented but will be considered for future extensions. The main variables of interest at this current time for this study are adherence to speed limit, vehicle density and collisions.

As described in the background section, the relationship between collision rate and traffic density has been theorised but not empirically verified. Since collisions in the real world can be caused by many factors, we will focus on collisions caused by speeding. Our question is: do higher traffic densities suppress the higher collision rates caused by speeding in an urban environment?

Independent Variable Measure	Low Adherence	Mid Adherence	High Adherence
Low traffic density	50 vehicles (25%) and 15 adherence (30%)	50 vehicles (25%) and 30 adherence (60%)	50 vehicles (25%) and 50 adherence (100%)
Mid traffic density	100 vehicles (50%) and 30 adherence (30%)	100 vehicles (50%) and 60 adherence (60%)	100 vehicles (50%) and 100 adherence (100%)
High traffic density	200 vehicles (100%) and 60 adherence (30%)	200 vehicles (100%) and 120 adherence (60%)	200 vehicles (100%) and 200 adherence (100%)

Table A.3: Experiment conditions.

Each experiment ran five times with different random seeds for five minutes due to the computational demand required to render 3D agents. Final collision values were averaged across runs and normalised by the number of vehicles, with the standard deviation displayed in the error bars. The results are shown in Figure A.1.4.

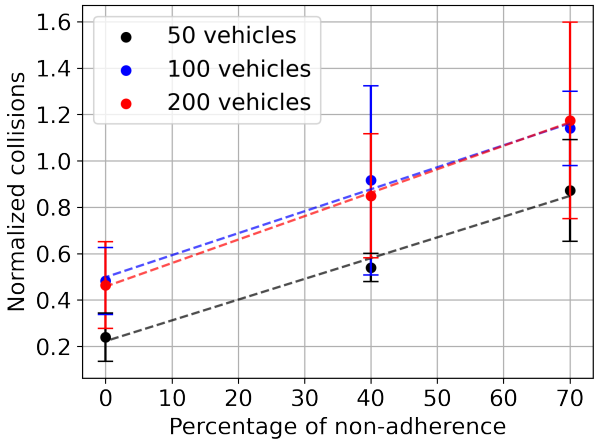


Figure A.1.4: Number of collisions (normalised by number of vehicles) against the percentage of non-adherence to speed limits, refer to Supplementary Materials for data used.

Firstly, it is important to note the size of the error. While some variance in model runs is expected, the extent of the overlap between scenarios makes drawing firm conclusions from these experimental results difficult. However, in future studies, this will be taken into account.

While keeping account of this variance, it is still clear that there is a greater difference in collision rates between 50 and 100 vehicles than between 100 and 200. This suggests that there exists a critical density at which the number of collisions begins to scale linearly with traffic density; prior to this critical point, an increase in vehicles results in a disproportionately large increase in collisions. Similar patterns were found in empirical data collected in the subsequent studies (Fildes *et al.*, 1991; Maycock *et al.*, 1999; Quimby *et al.*, 1999). In Figure A.1.4, we see little evidence of reduction (either proportional or absolute) in the number of collisions as traffic density increases. Higher traffic density also does not appear to suppress the effects of low-speed limit adherence on collisions. As can be seen from the trend lines in Figure A.1.4, collisions increase at a near-identical rate as a function of the percentage of non-adherence. Higher traffic densities also appear to loosely correlate with greater variance in collisions between runs.

While collision prevention is a primary goal of traffic management, the prevention

of congestion—and its impact on public health and CO_2 emissions—is equally crucial (Frey *et al.*, 2001; Mohandas *et al.*, 2009). As described in the background, it has been suggested that these goals could conflict (Retallack & Ostendorf, 2019; Shefer, 1994).

There is no single definition of congestion—several different definitions have been developed for different congestion scenarios (Vickrey, 1969) or for identifying congestion from the available data (Wan *et al.*, 2017). In this study, congestion will be understood both as a decrease in the overall traffic speed in the system and as an increase in the number of vehicles with speeds under 5 mph at a given time. Local traffic density (the number of vehicles in a particular area of the network) will also be considered under the assumption that this correlates with congestion and being of interest in its own right.

To compare traffic flow for the different scenarios, the average speed of all vehicles was calculated. The results are shown in Figure A.1.5. Lower adherence to speed limits leads to higher average speed for systems of varying density. However, the average speed of systems with low adherence is impacted more by increasing the density. For example, when increasing the number of vehicles from 50 to 200 for 100% speed limit adherence, the average speed decreases by 3.6%. When increasing the number of vehicles from 50 to 200 for 30% speed limit adherence, the average speed decreases by 13.3%; therefore, as density increases, the average speed of vehicles decreases.

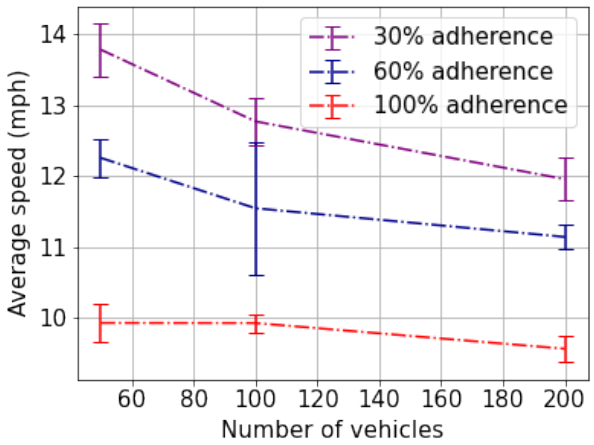


Figure A.1.5: Average speed of vehicles against number of vehicles for each adherence scenario, refer to Supplementary Materials for data used.

A.1 Exploring the Impact of Driver Adherence to Speed Limits and the Interdependence of Roadside Collisions in an Urban Environment: An Agent-Based Modelling Approach

While the average speed of traffic is an essential factor for considering the overall efficiency of the system (at least concerning average journey times), it does not communicate the distribution of speeds (for example, some vehicles may enjoy short journey times while others are stuck in congestion) or how these are spatially located.

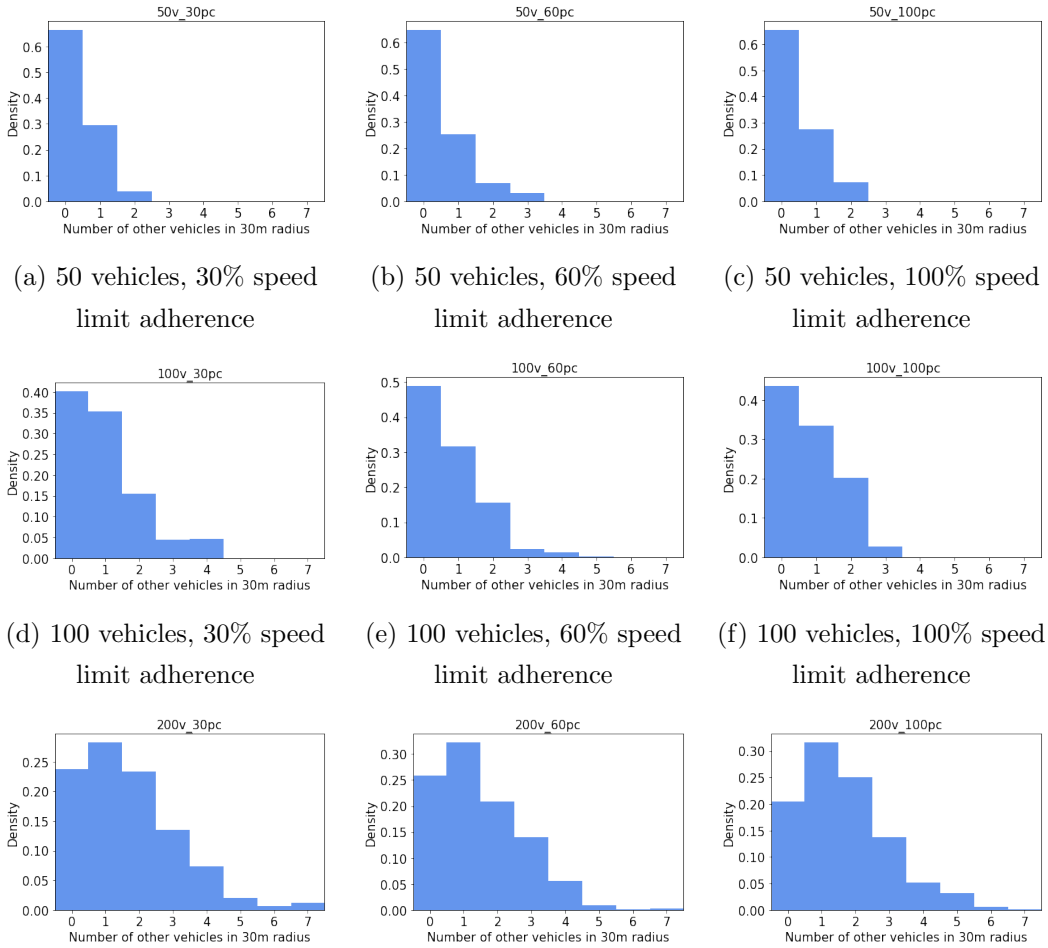
The average agents' speed, the spread of agents' speed, and the percentage of vehicles below 5 mph at the final time step of each scenario are presented in Table A.4. The spread of speed, which is the standard deviation of all agents' speeds, increases with lower adherence by a factor of more than 1.6 from 100% adherence to 30% adherence for all traffic densities. This increase is expected since non-adhering drivers can access a broader range of speeds up to 45 mph while adhering drivers cannot exceed 30 mph. The fraction of vehicles below 5 mph includes vehicles that have collided and cannot move, including vehicles stuck behind these collisions. There is an increase in this fraction for 100 and 200 vehicles as adherence decreases. This increase is not evident for 50 vehicles.

Scenario	**Speed (mph)**	**Spread (mph)**	**Vehicles under 5 mph (%)**
50 v, ad 30%	13.49	5.69	7.6
50 v, ad 60%	11.74	5.8	12.0
50 v, ad 100%	9.96	3.34	6.0
100 v, ad 30%	12.24	6.93	17.0
100 v, ad 60%	11.21	5.84	13.4
100 v, ad 100%	9.9	3.56	6.2
200 v, ad 30%	11.59	6.85	20.6
200 v, ad 60%	11.17	5.52	12.7
200 v, ad 100%	9.42	4.08	11.1

Table A.4: Average speed, spread of speeds, and fraction of vehicles moving below 5 mph for each scenario (where v = vehicles and ad = adherence percentage).

The above Table A.4 shows that higher traffic densities and lower speed adherence result in a greater fraction of vehicles travelling at very low speeds at any given point in time, even though the average speed is higher. Similarly, low-speed adherence with low traffic densities increases the average speed without increasing the fraction of vehicles at very low speeds.

This study is concerned with the spatio-temporal analysis of the whole urban street network. However, Figure A.1.6 shows that local micro-level phenomena can also be observed. We hope to conduct a comprehensive analysis of micro-level interactions between density, congestion and collisions for future studies.



(a) 50 vehicles, 30% speed limit adherence

(b) 50 vehicles, 60% speed limit adherence

(c) 50 vehicles, 100% speed limit adherence

(d) 100 vehicles, 30% speed limit adherence

(e) 100 vehicles, 60% speed limit adherence

(f) 100 vehicles, 100% speed limit adherence

(g) 200 vehicles, 30% speed limit adherence

(h) 200 vehicles, 60% speed limit adherence

(i) 200 vehicles, 100% speed limit adherence

Figure A.1.6: Distribution of number of additional vehicles involved in simulated collisions based on number of vehicles in system and proportion of vehicles adhering to speed limits.

The local traffic densities within 30 metres of a collision site, one second before the collision takes place, is shown in Figure A.1.6. Higher global traffic densities result

in higher local traffic densities near-collision sites across all adherence levels. This is highlighted when comparing Figure A.1.6c,f,i, where the modal value of the number of additional vehicles present near a collision increases from 0 to 1; that is to say that, for lower global vehicle densities, we typically observe that no additional vehicles are present at a collision site. Whilst for a large population of 200 vehicles, we observed that it is common for at least one other vehicle to be present. Furthermore, lower adherence results in higher local traffic densities near collision sites for 100 vehicles (Figure A.1.6d–f) and 200 vehicles (Figure A.1.6g–i) but not 50 vehicles (Figure A.1.6a–c).

Collisions, when they occur, appear to be more likely to take place in the presence of other vehicles both when global traffic density is increased and when adherence level is lowered. However, an increase in local traffic density alone does not appear to cause an increase in collisions; a similar pattern was observed in (Clark, 2003). This can be seen by comparing local traffic density results for 100 vehicles and 200 vehicles, which Figure A.1.4 shows to have a near-equal collision rate despite Figure A.1.6 showing that 200 vehicles have a higher local traffic density near collision sites.

However, according to Figure A.1.6, there is also an increase in local traffic density as adherence decreases, which always results in more collisions. This indicates that local traffic density may have a contributory effect towards collisions when combined with low adherence to speed limits, a higher average speed, or greater speed variance.

A.1.5 Summary

To conclude, the experiments found that a higher traffic density results in more vehicles travelling at lower speeds through space and time. This is the case even when 70% of vehicles do not adhere to speed limit rules, i.e., driving between 30 to 45 mph. Furthermore, collisions increase linearly as the non-adherence measure is increased. This is the case for all traffic densities; however, lower densities lead to fewer collisions. Lastly, collisions are at their lowest amount when all vehicles comply with speed limits for all densities.

In the next section, an overview of the results is provided. Furthermore, the findings from the paper will be validated by comparing the results with empirical findings. Additionally, recommendations for reducing collisions will be made, and, lastly, future avenues for research will be discussed.

A.1.6 Discussion

The goal of this study was to understand the relationship between traffic density and the number of collisions. Moreover, the paper aimed to look at the concept of higher traffic density serving to suppress collisions by regulating driver speed, especially if drivers were not adhering to prescribed speed limits. Previous studies indicate that higher levels of congestion can result in fewer road accidents. This theory was the case for a single highway segment in Detroit (Zhou & Sisiopiku, 1997). Similarly, the authors in (Dickerson *et al.*, 2000) found this to be the case on two to three-lane motorways in France. However, this is not true at intersections (Retallack & Ostendorf, 2020), or on urban roads in London (Dickerson *et al.*, 2000; Noland & Quddus, 2005), where the number of accidents was found to increase linearly at low to mid-levels of traffic and nonlinearly at high levels of traffic.

This study found that higher levels of traffic density do not reduce the frequency of collisions. Furthermore, higher traffic levels do not suppress the increased collision rates caused by non-adherence to speed limits. Empirical findings found that traffic congestion has little or no impact on the frequency of road accidents; however, it should be noted that the results are constrained to the M25 London motorway (Wang *et al.*, 2009). The author in (Jones, 2007) found that an increase in collision rates resulted from the road network design of urban roads, which consisted of higher lengths of road and high traffic density. This study aims to contribute to the ongoing debate as to whether traffic congestion impacts the frequency of road accidents (Cabrera-Arnau *et al.*, 2020).

This study found that high-density systems are affected to the same degree as low-density systems and provide no protective effect. This would suggest that the traffic management goals of congestion-reduction and accident-reduction are not in conflict for urban road networks.

This study also suggests that lower traffic density on average leads to fewer collisions regardless of adherence levels, as was observed in (Retallack & Ostendorf, 2020). However, as adherence decreases, this leads to increased collisions relative to the number of vehicles in the urban environment. These findings were also observed in (Vickrey, 1968).

Empirical evidence from UK government sources in 2019 shows that, on average, 55% of vehicles in 2019 exceeded the speed limit on urban roads. During this time,

63% of all collisions occurred on these urban roads (Balendra, 2020; Murphy, 2020). This study also shows that increases in collisions are more likely as more vehicles break speed limit rules.

This study's results do not reflect the same linear-to-nonlinear relationship between accidents and traffic levels as (Dickerson *et al.*, 2000; Retallack & Ostendorf, 2020). At low to mid traffic densities, collisions increase disproportionately as traffic density increases. Collisions begin to increase proportionately at a critical point in density, so an individual vehicle's risk of colliding does not increase as traffic increases. However, the results reflect the global number of collisions against an urban road network's global traffic density rather than studying micro-level intersections or specific urban roads. The study also attempted to provide a micro-level analysis to supplement the findings of the research. The analysis observed the distribution of the number of additional vehicles involved in collisions based on the global number of vehicles and the proportion of vehicles adhering to speed limits. We found that higher global traffic density resulted in higher local traffic density near collision sites. Furthermore, we found that lower adherence results in higher local traffic densities near collision sites for 100 to 200 vehicles; this is not the case for 50 vehicles. This indicates that local traffic density may contribute to collisions when combined with low adherence to speed limits, a higher average speed, or greater speed variance. Lastly, this micro-level analysis shows that additional vehicles are present within 30 m of a collision, ultimately leading to congestion at the local level.

The results in this study do show that higher traffic density results in higher levels of congestion. Even when maximum adherence is achieved, increased density resulted in reduced average vehicle speed, and this effect was greater for systems with lower adherence. Therefore, with high density, non-adhering vehicles are more likely to reduce their speeds more often as they find slower-moving vehicles ahead of them.

It should also be noted that conclusions drawn from this study are from the tested traffic densities. Studying a greater range of densities may reveal a more complex relationship. Conducting this study with a greater number of model runs per scenario may yield more precise insights into the relationship between collisions, traffic density, speed adherence, and speed distribution.

Since non-adherence to speed limits was found to increase collision rates for all traffic densities, this study recommends implementing measures to increase adherence to speed

limits on all roads regardless of their traffic level. This may include the introduction of more speed cameras, which have been found to reduce speeding significantly (Bener & Alwash, 2002). Feedback signs which broadcast the percentage of drivers who have stayed within the speed limit of an area have also been found to be effective at reducing speeding and resulting accidents (Van Houten *et al.*, 1985).

Another important finding from this study suggests that, if fewer vehicles occupy a street network, the total number of collisions is reduced. The most dramatic reduction in collisions may be areas that shift from medium traffic density to low traffic density. These findings support pedestrianisation policies, as these policies should reduce collision rates among vehicles in these urban environments and reduce CO_2 exposure. A report titled "The effect of pedestrianisation and bicycles on local business" published in 2017 found that: According to the 2012 Economic Impact Study, pedestrian activity has risen by 11%, with 35% fewer accidents with pedestrians and 63% fewer traffic accidents in New York Times Square (Pere, 2017).

A.1.7 Conclusion

This study aimed to explore the relationship between vehicle density and adherence to speed limits with collision rates through agent-based modelling. This area of research is still in its infancy but has shown that agent-based modelling is a powerful method that can provide the means to simulate hypothetical yet realistic properties of the real world and produce insight into these properties that can be empirically validated. Thus, this study will allow traffic practitioners and safety scientists to test their hypotheses through agent-based modelling in a safe, low-cost way prior to advising real-world policies.

In this study, the severity of collisions is not quantified. Since each vehicle's momentum is recorded in the model, this is a potential avenue for further study. Quantifying collision severity would allow future studies to categorise severe collisions (life-threatening) to mild collisions (dent in a vehicle), thus providing a more realistic snapshot of collision types. Some past studies have tried to quantify collision severity using alternate means such as the ordered logit model and the ordered probit model (O'Donnell & Connor, 1996). Similarly, empirical research found that mild collisions such as those that are not fatal were more likely to occur in cities across the UK (Cabrera-Arnau *et al.*, 2020); this can be a future avenue to explore using ABMs.

Another avenue to explore would be to incorporate changing weather into the agent-based model. Weather plays a significant role in having an impact on driver behaviour, which can, in turn, lead to higher collision rates, i.e., vehicles are more likely to collide during snowy conditions (Andrey *et al.*, 2003; Malin *et al.*, 2019; Qiu & Nixon, 2008). Given the ABMs drag, traction control and downforce parameters, the phenomena mentioned above can be modelled in future studies.

A.2 An Agent-Based Model of Heterogeneous Driver Behaviour and Its Impact on Energy Consumption and Costs in Urban Space

Authors: Sedar Olmez, Jason Thompson, Ellie Marfleet, Keiran Suchak, Alison Heppenstall, Ed Manley, Annabel Whipp, Rajith Vidanaarachchi

Abstract: By 2020, over 100 countries had expanded electric and plug-in hybrid electric vehicle (EV/PHEV) technologies, with global sales surpassing 7 million units. Governments are adopting cleaner vehicle technologies due to the proven environmental and health implications of internal combustion engine vehicles (ICEVs), as evidenced by the recent COP26 meeting. This article proposes an agent-based model of vehicle activity as a tool for quantifying energy consumption by simulating a fleet of EV/PHEVs within an urban street network at various spatio-temporal resolutions. Driver behaviour plays a significant role in energy consumption; thus, simulating various levels of individual behaviour and enhancing heterogeneity should provide more accurate results of potential energy demand in cities. The study found that (1) energy consumption is lowest when speed limit adherence increases (low variance in behaviour) and is highest when acceleration/deceleration patterns vary (high variance in behaviour); (2) vehicles that travel for shorter distances while abiding by speed limit rules are more energy efficient compared to those that speed and travel for longer; and (3) on average, for tested vehicles, EV/PHEVs were £233.13 cheaper to run than ICEVs across all experiment conditions. The difference in the average fuel costs (electricity and petrol) shrinks at the vehicle level as driver behaviour is less varied (more homogeneous). This research should allow policymakers to quantify the demand for energy and subsequent fuel costs in cities.

Keywords: agent-based model; electric vehicles; traffic simulation; energy intake; urban environment; fuel costs; public policy

A.2.1 Introduction

According to Bretzke (2013), by 2050, 70% of the world's population will live in urban areas, accounting for roughly 6.3 billion people. Battery-powered electric vehicle sales increased from 5.3 million sales in 2019 and are projected to reach over 39.9 million units by 2030 (Dhakal & Min, 2021). Given that the majority of people live in urban areas

and infrastructure development is targeted at these areas (LondonAssembly, 2021), it could be assumed that the majority of electric vehicles (EVs) will be driven in these areas. An environmental benefit that EVs present is the ability to consume energy from renewable energy sources (e.g., wind turbines and solar). Furthermore, the total energy use among EVs is 3.4 times lower than ICEVs that rely on petroleum, diesel or gas, which emit CO_2 that is harmful for the environment. During a well-to-wheel (WTW) analysis of ICEV and EV efficiency, Albatayneh *et al.* (2020) found that EVs, when using renewable energy, can reach an efficiency level of 40 to 70% depending on the location and environmental factors. In contrast, gasoline- and diesel-powered ICEVs had an WTW energy efficiency of 11–27% and 25–37%, respectively. Almost all vehicle manufacturing companies have started building and testing EV/PHEVs for the commercial market (Sarlioglu *et al.*, 2017; Sierzchula *et al.*, 2012). Governments are facilitating benefits to persuade people to replace ICEVs with EVs through economic incentives or legislation. However, not all countries have renewable technology to power these vehicles; some countries, such as China, still depend on coal to power the majority of their electric grid infrastructure (Tan *et al.*, 2018; Wang & Ke, 2018). In Australia, only 24% of electricity is generated from renewable sources (Angus, 2021). In their review of EVs and their impact on the climate, Hawkins *et al.* (2012) found that vehicles using electricity from sources with lower global warming potentials (GWP) (Yang *et al.*, 2021) are better than ICEVs. In contrast, Hawkins *et al.* (2013) found it was counterproductive to promote EV uptake in countries where electricity is produced from fossil fuels. The statistics mentioned above reaffirm the need to explore the impact these technologies have on future cities.

This study demonstrates how agent-based modelling (ABM) can be harnessed to quantify energy demand in cities from electric-powered vehicles at various spatio-temporal resolutions. To test the model, two variables are configured across multiple test scenarios to demonstrate the subtle differences in outcomes. These variables are the speeding behaviour (known as adherence to speed limits) and the number of vehicles on the street network (density of vehicles). Through experimentation, we show that individual vehicle behaviours and the number of vehicles on the street network impact the total energy usage (the amount of energy required by the vehicles to complete their drive cycle in kWh). Drive cycle is defined as a series of data points representing the speed of a vehicle versus time.

This article contributes an energy calculation extension (Figure A.2.1) which can be used in conjunction with the agent-based model (Olmez *et al.*, 2021b) to quantify EV energy usage. While the focus is centred around electric-powered vehicles, to demonstrate the effectiveness of the model, we illustrate how ICEV vehicles can also be incorporated by converting energy to petrol (L/km), allowing a direct comparison between individual-level behaviours/patterns using two types of vehicles and their relative impact on costs and efficiency. The novelty of this article is three-fold. Firstly, an agent-based method for quantifying energy demand from vehicle behaviour at the individual level is presented. Secondly, heterogeneity among driver behaviour and road characteristics is included, directly impacting the energy required, which is the case in the real world. Finally, the proposed model enables practitioners to quantify the potential energy costs these vehicles incur and compare scenarios such as high traffic to low traffic densities. For clarity, driver behaviour is defined as the interactions of the human driver and the impact those interactions have on the vehicle being driven. This includes, for example, the driver's foot dynamics and its impact on acceleration (Xing *et al.*, 2020b). This is represented as the speed limit adherence and non-speed limit adherence behaviours, enhancing heterogeneity.

A.2.2 Background

A traffic system is characterised by multiple individual actors (e.g., drivers) and a street network made up of individual rules characterised by (for example) traffic lights and posted speed limits. Given this system's individual-level components, it is amenable to being studied using individual-based modelling methods. According to Huston *et al.* (1988), individual-based modelling refers to simulation models that treat individual entities as unique and discrete elements with at least one property (e.g., age, height, speed), and these properties change during the life cycle of the entities. Therefore, in this study, vehicles can be thought of as individual heterogeneous entities with their properties and rules, while the urban street network is the environment within which these vehicle entities are observed.

Agent-based modelling (ABM) is an individual-based modelling method. It provides the means to plan, design and experiment with micro-heterogeneous agents in an artificial, computational environment. ABMs have been utilised in various domains to explain complex phenomena such as those that occur in crime (Birks *et al.*, 2012; Malle-

son *et al.*, 2010), ecology (Filatova *et al.*, 2013; Heckbert *et al.*, 2010; McLane *et al.*, 2011), economics (Dawid & Neugart, 2011; Olner *et al.*, 2015), sociology (Bianchi & Squazzoni, 2015; Squazzoni, 2012), geography (Crooks, 2015; Heppenstall *et al.*, 2012) and transportation (Olmez *et al.*, 2021a; Thompson *et al.*, 2020a). One advantage of using ABMs is that they are able to represent a richer and more detailed set of individual actors leading to potential policy alternatives and outcomes compared to the alternative, statistical models (Davis & Morris, 2009).

Several agent-based models have focused on electric vehicle research. Kangur *et al.* (2017) developed an agent-based model that measured consumer needs and decision strategies by policymakers to shift from ICEVs to EVs. They found that effective policy requires a long-lasting implementation of a combination of monetary, structural and informational measures. Similarly, Eppstein *et al.* (2011) developed a spatially explicit agent-based vehicle consumer choice model to identify the various influences that can affect the uptake of PHEVs. The study found that providing consumers with ready estimates of expected lifetime fuel costs associated with other vehicle types, including the rise of petrol costs, can generate preferences for purchasing EV/PHEVs over ICEVs.

Several studies have also explored the total cost of ownership between EVs and ICEVs from a consumer perspective to quantify the economic differences in ownership between vehicle types. Findings differ geographically due to international differences in the price of petrol, diesel, and electricity. In a study focused on New Zealand, Hasan *et al.* (2021) estimated that the per-kilometre cost of ownership (PCO) for a used EV was twelve percent lower than that of a used petrol-powered car over twelve years (25.5 NZ cents and 31.5 NZ cents for petrol vehicles). Although this study primarily focused on the differences in fuel costs, others have included additional factors such as insurance, vehicle depreciation and maintenance. Palmer *et al.* (2018) analysed these factors between 1995 and 2015 and found that in the UK, USA and Japan, owners of both mid-size battery EVs (BEVs) and hybrid EVs (HEVs) incurred lower costs than owners of ICEVs during the same period.

Fuel and electricity prices need to be estimated beyond the current year to provide insight into the future costs in ownership between EVs and ICEVs. This is difficult given the inherent fluctuation in oil and electricity markets. However, when investigating the relationship between oil and electricity prices, Bencivenga *et al.* (2010) found that

the Engle–Granger co-integration method identified a short-term relationship between these fuel types. Hasan *et al.* (2021), on the other hand, assumed that changes in fuel prices would follow the past decade trends, which exhibit a 1.4% per year increase for petrol and a 1.1% increase for electricity. Their findings for New Zealand, therefore, cannot be easily transferred to an international context because user-end electricity costs differ drastically between countries, with higher household electricity costs in Germany, Denmark and Italy and lower costs in Mexico, Korea, and Turkey (Iea, 2020). Such discrepancies in findings are reflected in international studies (Letmathe & Suares, 2017), which found that without subsidies, limited models of BEVs and HEVs incurred lower running costs than ICEVs at the time. Given the complexities mentioned above of integrating fluctuating costs of petrol and electricity into our analyses, we will use the most recent cost of electricity kWh per km and petrol per L/km in the UK.

As the discussion above indicates, EV modelling is a relatively new area of research. Prior studies also focused on a narrow set of issues such as market penetration and charging infrastructure, which may ultimately be driven by price considerations made by individual prospective owners. We, therefore, contend that planning and developing forecasts of electric energy consumption alongside pricing in urban street networks is of critical importance because electricity demand and pricing will influence uptake.

A.2.3 Model Description

This section describes the agent-based model adopted for this study. The overview design and details (ODD) protocol will be utilised to explain all aspects of the model (Grimm *et al.*, 2006).

Purpose

The agent-based model used in this research is the 3D Urban Traffic Simulator (UTS) in Unity (Olmez *et al.*, 2021b). The model was developed to allow researchers to simulate hypothetical vehicle drive cycle scenarios in a 3D urban environment. The model delivers heterogeneous autonomous vehicle agents with granular features such as mass, velocity and traction control. Similarly, the road network is designed around a built-up environment that contains all the characteristics of a dense urban street network with varying speed limits and intersection rules adopted from the UK Speed Limits (Highway Code, 2022). Lastly, the model was used by researchers looking at how driver

behaviour impacts collision rates (Olmez *et al.*, 2021a).

Variables

The model requires input parameters to run an experiment and produces output results for later analysis. The parameters that can be tuned are listed in Table A.5.

Entity	Parameter	Values
Vehicle	Mass	[1000, 3000] (kg)
	Top speed	[30, 45] (mph), [48, 72] (km/h)
	Gap acceptance	[1, 10] (m)
Environment	N. of vehicles	[1, 500]
	Speed adherence	[0, N]
	Roads	1295
	Intersections	354

Table A.5: Model entities and parameter values (source: (Olmez *et al.*, 2021a)).

The model has two entities: the vehicle agents and the model environment in which these agents are based. The vehicle parameters are:

- The vehicle mass parameter, drawn from a random uniform distribution between 1000 and 3000 kg (inclusive), allows the model to simulate a wider variety of vehicle types, from sedans to SUVs and hatchback. The rationale behind this distribution was to try intersect the EV and ICEV vehicle types, which larger vehicles such as vans or trucks are not part of; the model distributes vehicles arbitrarily across the environment with varying weights (source (Sellén, 2022)).
- The top speed measure is between 30 and 45 mph (48, 72 km/h) and is only applied to vehicles that do not adhere to speed limits. This measure is applied only if Speed Adherence is $\geq$1 (source (Balendra, 2020)).
- The gap acceptance parameter can be between 1 to 10 for each vehicle. The variable assigns a distance between two vehicles in meters. This ensures a wider variety of visual impairment is captured as some people with healthier eyes keep a fair distance from vehicles in front usually adhering to the 2-s rule compared to people with worse vision. Furthermore, the distance had to be relative to the average road distance in the model.

The environment-specific parameters are:

- The number of vehicles generated in the model, N. This can be between 1 and 500. However, this can be adjusted depending on the compute power accessible. The hardware accessible at the time of writing this article could only efficiently simulate up to 500 vehicles in 3D space while yielding valid results (refer to the Conclusions section for more on this limitation).
- The speed adherence variable can be between $0 \leq x \leq N$. This quantifies the proportion of vehicles that will not adhere to the speed limits applied to the road they are driving on.
- The urban road network consists of 1295 roads which vehicles drive on and 354 intersections which consist of traffic rules (Algorithm 1, (Olmez *et al.*, 2021a)). The road network has been designed to depict a small urban town.

The parameters above are used to produce output variables that observe various data points at every step of the simulation run, collecting individual-level data from each vehicle. Table A.6 describes the output variables that the model produces.

The ABM outputs thirteen variables that can be used for analysis (refer to Table A.6). As the agent ID variable is present, a micro-level analysis of the agent behaviours during model execution (e.g., observing individual drive cycles) can be explored. The collisions variable tracks the number of times a vehicle has collided with another. Top speed is the speed limit associated with the road that the vehicle is driving on, which the vehicle tries to match. However, in scenarios where some vehicles do not adhere to speed limits, this would be a value between 30 and 45 mph (48, 72 km/h). The current speed value is the vehicle's speed at the current time step of the model. The distance of travel tracks the vehicle's distance from the starting position on the road network at each time step in metres. The gap acceptance length is the distance the vehicle keeps from vehicles ahead. The velocity magnitude is a scalar value indicating the rate of motion at that specific time step. The vehicle mass variable assigns a weight to the vehicle between 1000 to 3000 in kilograms. The physics engine requires that every object have a mass assigned to it to ensure gravity is applied. The downforce coefficient is set to 0.1; for this research, it is left at 0.1 to have no impact on the vehicles. Lastly, date-time stamps are included in each row of data recorded such that time-series analysis can be applied. These output data are then

used as input to the energy calculation extension, which calculates energy intake and outputs energy-specific data, as seen in Table A.10.

Variable	Output Type
AgentID	Integer
xAxisPos	Float
zAxisPos	Float
collisions	Integer
topSpeed(mph)	Float
currentSpeed(mph)	Float
distanceOfTravel(meters)	Float
gapAcceptance (raycastLength)	Integer
tractionControl	Bool
velocityMagnitude	Float
vehicleMass	Integer
downforce	Float
date-time	DateTime

Table A.6: Model output variables.

Model Overview

The agent-based model was developed using the Unity development stack. Unity is a 3D game engine consisting of a rendering and physics system and a graphical user interface. The primary programming language is C#. Unity has received widespread adoption in several industries, including gaming, automotive, and film (Juliani *et al.*, 2018).

The following workflow diagram (Figure A.2.1) describes the processes that the model (Olmez *et al.*, 2021b) undergoes during run-time. In addition, the energy consumption calculation extension is also depicted.

The UTS (Olmez *et al.*, 2021b) workflow (Figure A.2.1) starts by taking input values for the five variables described in Table A.5. The software then resets all settings to launch the simulation scene to render the agents and environment. Once the reset process is complete, the model processes all agents, their starting locations and environment parameters. Next, the model can run each frame, and every change that occurs is captured and stored with a time-stamp in a CSV file. Fixed Update is used to compute physics elements such as vehicle wheels, mass, velocity. Update, on the

other hand, computes variables for each frame. The model uses Fixed Update due to the sheer number of physics components involved; these variables are tracked multiple times each frame. Once the user stops the model, the sixteen output variables are saved in a directory, and the model is destroyed (stopped). The output dataset is then used as input to an energy calculation notebook (Figure A.2.1), which uses the outputs to calculate F from Equation (A.2.3), with vehicle parameters from Tables A.7 and A.9. The output from this calculation is then used to calculate Equations (A.2.6) and (A.2.7) (Section A.2.4).

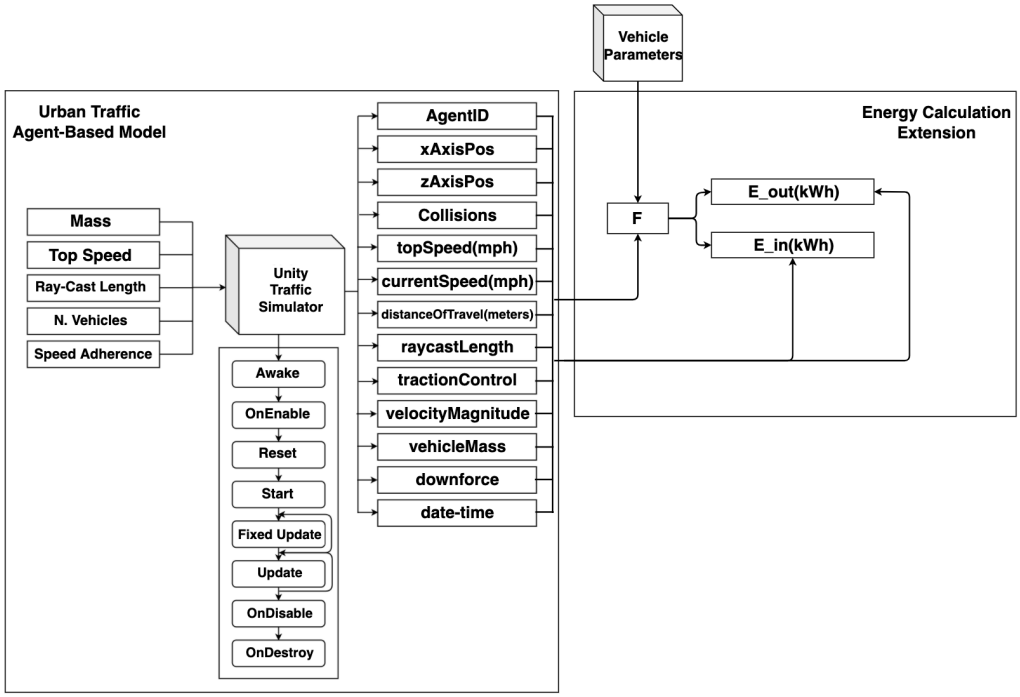


Figure A.2.1: Workflow diagram depicting processes the UTS undergoes during run-time including the Energy Calculation Extension.

Agent

The vehicles in the model are classed as autonomous agents; the vehicle population is heterogeneous, meaning every vehicle will have varying features. These agents inherit similar characteristics as real-world vehicles; they have four wheels, a steering angle, traction, mass and drag. Each agent applies a set of rules outlined in the article (Olmez *et al.*, 2021a).

The rules described in (Olmez *et al.*, 2021a) allow autonomous vehicle agents to navigate the environment and act as data collectors. Each vehicle follows the same condition-action rules. However, the parameters vary and depend on the input values from Table A.5. These vehicle agents are a simplification of real-world vehicles. Therefore, they are not expected to mimic the actions and behaviours of real-world vehicles perfectly, but they do include the essential behaviours that all vehicles demonstrate, such as stop/start and give-way behaviour.

If a vehicle is not adhering to the speed limits, it can increase its speed between 30 and 45 mph (48, 72 km/h). If vehicle A is ahead of B, B should decrease speed to match vehicle A's speed. When a vehicle arrives at an intersection, if it has the right of way (i.e., on a horizontal lane and no vehicles are on the intersection), it drives through the intersection at 10 mph (16 km/h). If the vehicle is at the intersection and does not have the right of way, it should wait until the intersection is cleared. If the vehicle is at an intersection and does not have the right of way, and there are no other vehicles at the intersection, the vehicle is free to reduce speed to 10 mph (16 km/h) and drive through the intersection. Lastly, all vehicles that adhere to the speed limit increase or decrease speed to match the road's speed limit.

Environment

The agents described in the last sub-section require an environment to function within. The UTS (Olmez *et al.*, 2021b) deploys an urban street network that is described as a T-type network pattern in (Han *et al.*, 2020) which contains similar characteristics as downtown Philadelphia, PA (Boeing, 2020) and San Francisco (Porta *et al.*, 2006). T-network patterns are like grid-shaped networks but include t-junctions. Several added features such as the eight-lane intersections described in (Filocamo *et al.*, 2020) also exist. The street network contains 1295 roads and 354 intersections, arbitrarily generated to cover a small town. The individual roads, speed limits and intersection rules are described in the following Figure A.2.2.

The environment consists of three road types with varying speed limits and intersections with right-of-way rules. The model environment is a simplification of the real world. Therefore, it does not capture all intersection types. However, it does contain the basic characteristics of an urban street network which have also been observed in several cities across the United States (Boeing, 2020; Porta *et al.*, 2006; Thompson

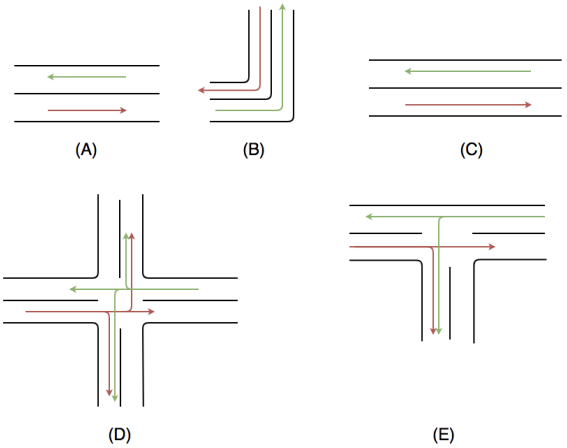


Figure A.2.2: Urban Street Network roads and intersections (source: (Olmez *et al.*, 2021a)).

et al., 2020b). The vehicles also adhere to stop-go rules (conceptualisation of traffic lights) enforced at junctions. These rules are present in the vehicle's decision-making algorithm (Olmez *et al.*, 2021a). The following list describes each road and intersection in Figure A.2.2:

- (A) A two-way local road with a speed limit of 20 mph (32 km/h);
- (B) A two-way corner road with a speed limit of 10 mph (16 km/h);
- (C) A two-way fixed road with a speed limit of 30 mph (48 km/h);
- (D) An eight-way intersection. Right-of-way is for traffic on horizontal lanes, and the speed limit is 10 mph (16 km/h);
- (E) A two-way t-junction. Right-of-way is for horizontal lanes, and the speed limit is 10 mph (16 km/h).

The speed limits for the three types of roads (Figure A.2.2A–C) were derived from UK government sources such as (Balendra, 2020), where urban street networks consist of local 20 mph (32 km/h) and fixed 30 mph (48 km/h) zones; however, corner roads sometimes require lower speeds such as 10 mph (16 km/h) as vehicles require more room

to turn. The 'setting local speed limits' report by the UK Government's Department for Transport outlines that most urban streets (roads in built-up areas) have a fixed speed limit of 30 mph (48 km/h). However, for dense areas—usually city centres—this may be designated 20 mph (32 km/h) by local councils to keep pedestrians safe from collisions (Department for Transport, 2006; Olmez *et al.*, 2021a).

A.2.4 Results

This section will analyse the experiments designed to quantify electric energy consumption across multiple vehicle densities and adherence levels. Once this is achieved, the model will quantify fuel consumption by simulating an ICEV drive cycle as a direct comparator between PHEV/EV and ICEV fuel consumption. The aforementioned comparator experiment will present novel insight by comparing drive cycle, fuel consumption and costs of ICEV and compare these patterns to the alternative PHEV/EV outputs. The output data from the energy calculation extension notebook can be found in Table A.10.

Before running the experiments and analysing outputs, the model must be tested against either (1) empirical data, which entails vehicle drive cycle and energy consumption in kWh over km travelled, or (2) model outputs from a different model utilised in research by the research community. Without a baseline comparator, there is no way in knowing if the model utilised in this research, namely (Olmez *et al.*, 2021b), outputs energy consumption accurately. Almost all agent-based models are validated using the former or latter processes (Benenson *et al.*, 2008; Heppenstall *et al.*, 2006; Kothari *et al.*, 2014; Malleson *et al.*, 2010; Olmez *et al.*, 2021a; Sert *et al.*, 2020; Thompson *et al.*, 2019).

The data used to compare model outputs were adopted in the following study (Gaete-Morales *et al.*, 2020). This study utilised German automotive statistics from empirical sources to generate drive cycles of EV journeys using a mathematical model. The variables of interest are kilowatt-hour over distance travelled in kilometres. The specific dataset used contains the drive cycle of 200 vehicles, where input parameters are derived from the statistics mentioned above and the physical properties of vehicles used in Germany (Gaete-Morales, 2021). The main drawbacks of this model are:

- That it produces outputs at a time resolution of 15 min; our model, on the other hand, has a time resolution of 1 s. This way, we can capture finer detail such

as the impact of traffic lights on acceleration/deceleration and momentary traffic congestion;

- That it only captures trips that are split into commuters and non-commuters. Thus, the modelled scenarios revolve around two profiles of drivers. Our model manipulates the entire system from the street network to traffic rules and vehicles; thus, no single driver profile is modelled, but a heterogeneous set of behaviours are captured.

These factors play an essential role in the electric energy consumption post-simulation run. The main strength of (Gaete-Morales *et al.*, 2020) is that variables such as heat transfer, weather, road condition and slope are all introduced as parameters to produce more robust energy consumption results. In our study, we make some basic assumptions, such as: our road surfaces are flat, and no weather parameters are introduced; these variables add complexity to the agent-based model and can hamper computation which, in turn, can affect outputs. We have, however, introduced rolling resistance (Equation (A.2.1)) and braking energy recovery (Figure A.2.11) which both impact electric intake. Adding additional levels of complexity is an area for future development.

Electric Energy Consumption Calculation

As the vehicle agents are not configured to mimic a specific vehicle, the goal is to adopt parameters from empirical statistics to ensure our findings are consistent with those within the UK given the environmental parameters adopted, such as local and fixed speed limits (Figure A.2.2). Currently, the most popular EV/PHEV in the UK is the Mitsubishi Outlander (source (Lilly, 2021)), with over 46,400 units sold as of June 2020. Therefore, this is the chosen vehicle in our analyses. However, the model can be readily adapted to other vehicles and can replicate a heterogeneous fleet.

Equations (A.2.3), (A.2.6) and (A.2.7) were applied to the model outputs to calculate electricity intake:

- For Equation (A.2.3), F is calculated by using the following parameter variables: $\theta = 0$ as the surface area is flat, $C_D = 0.33$, $A = 3.078\ m$ (where height = 1.71

[1] The official engine efficiency statistic is not provided by the vehicle manufacturer; therefore, an average engine efficiency for PHEVs was acquired from the cited academic source.

Parameter	Value
Height (m)	1.71
Width (m)	1.80
k	65% (source (Brslica, 2011)) [1]
m (kg)	1925
C_D	0.33

Table A.7: Vehicle parameters (PHEV).

m, width = 1.80 m), $m = 1925$ kg (Table A.7), $a = \frac{\Delta v}{\Delta t}$, where Δv is the velocity change over time period Δt, and lastly, v = velocityMagnitude (Table A.6).

- For Equation (A.2.6), E_{out} is calculated by multiplying the output from Equation (A.2.3) with total_distance (d) travelled in meters per second for each agent; see Table A.6.

- Lastly, Equation (A.2.7) is calculated by multiplying the output from Equation (A.2.3) (F) with the distance travelled d divided by the engine efficiency $k = 0.65$. E_{in} is then divided by 3.6×10^{-6} to convert from joules to kilowatt-hours (kWh).

To compare both emobpy (Gaete-Morales, 2021) with UTS (Olmez *et al.*, 2021b), we ran UTS for one hour, where fifteen vehicles were present. To add complexity, we set five of these vehicles to break speed limits; this allows us to capture the subtle differences between rule followers and rule breakers and their relative energy efficiency and energy consumption over a drive cycle. In comparison, fifteen vehicles were taken from emobpy and plotted against our model outputs.

We aggregated the fifteen vehicles from (Olmez *et al.*, 2021b) to five gradient colours to simplify the legend in Figure A.2.3.

Figure A.2.3 distinguishes between the rule followers and rule breakers. We can see the clustered lines at the bottom of the graph; these are vehicles that fully adhere to speed limits. On average, these vehicles consumed roughly 0.15–0.25 kWh/km. The five vehicles that broke speed limit rules travelled further as they were speeding and, on average, consumed more energy and were less energy-efficient, as expected. Furthermore, the urban environment has impacted the distances these vehicles could cover (differences in distance travelled). The furthest a vehicle has travelled is 28 km. We also

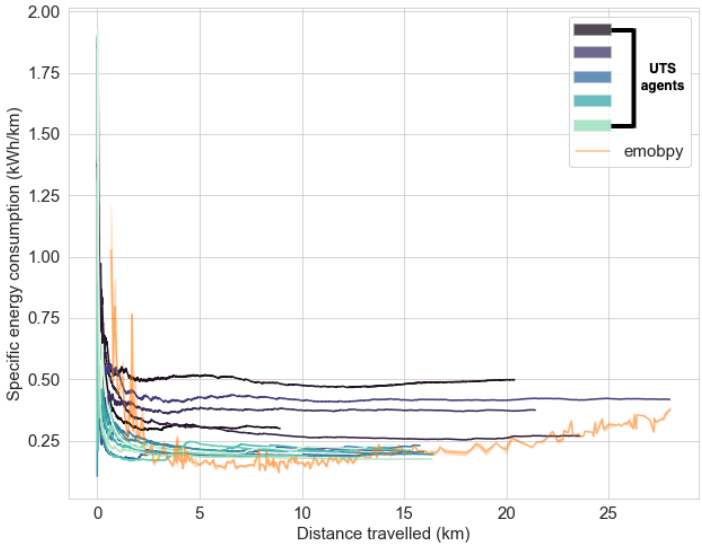


Figure A.2.3: Model output comparison: electric energy consumption (kWh) against distance travelled (km), with 15 vehicles over a 1 h drive cycle (5 vehicles break speed limits).

found that the least energy efficient vehicle was consuming roughly 0.50 kWh/km. It is evident that the model outputs from emobpy and UTS are in agreement, where the longer and faster a vehicle drives the less efficient it becomes. These characteristics of energy efficiency and consumption are evident in empirical literature (Moriarty & Wang, 2017; Wager *et al.*, 2016).

This small sample shows that the confidence intervals for emobpy are small. This means that vehicles are likely to follow similar drive cycle patterns and configurations, leading to similar energy consumption outputs. However, due to heterogeneity, our model captures a more diverse range of outputs from the same environment, which is a strength of the ABM approach over standard mathematical models.

The energy consumption (kWh/km) patterns are similar for both emobpy and UTS; see Figure A.2.4. These preliminary results are promising as they show that UTS is capable of producing behaviours of realistic drive cycles of electric vehicle energy consumption that have also been observed in a completely different model (Gaete-Morales *et al.*, 2020). Now that we have shown that UTS produces valid estimates of electric energy consumption, we can devise experiments to quantify the effects of

speed limit adherence and vehicle density on electric energy (kWh/km) and petrol consumption (L/km).

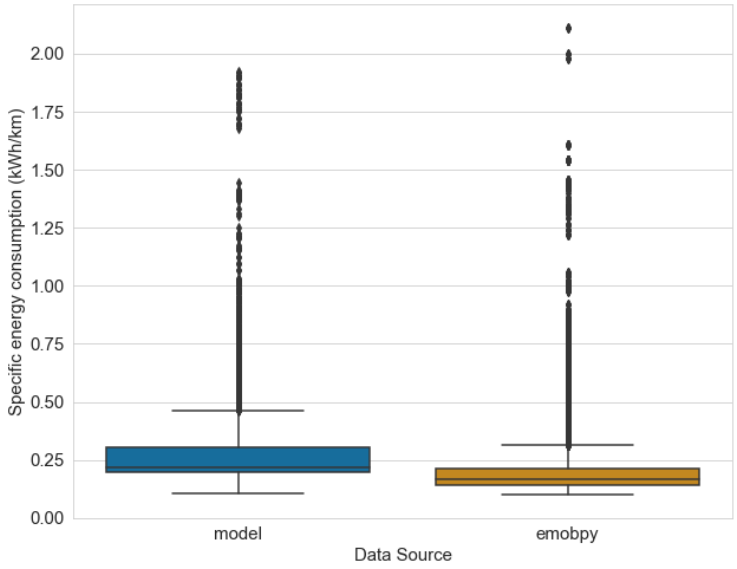


Figure A.2.4: Model Output comparison: electric energy consumption (kWh) for both UTS (model) (Olmez *et al.*, 2021b) and emobpy (Gaete-Morales, 2021).

Experiments

Due to the computational processes required to render 3D vehicles through space and time (Olmez *et al.*, 2021b) and the hardware capacity at hand, nine computationally cheaper experiments were designed. The independent variables were density and adherence to speed limits. These experiments are formally described in Table A.8.

Variable	Low Adherence	Medium Adherence	High Adherence
Low Density	Condition 1, 10 vehicles, 10 non-adherence	Condition 2, 10 vehicles, 5 non-adherence	Condition 3, 10 vehicles, 0 non-adherence
Mid Density	Condition 4, 50 vehicles, 50 non-adherence	Condition 5, 50 vehicles, 25 non-adherence	Condition 6, 50 vehicles, 0 non-adherence
High Density	Condition 7, 100 vehicles, 100 non-adherence	Condition 8, 100 vehicles, 50 non-adherence	Condition 9, 100 vehicles, 0 non-adherence

Table A.8: Experiment conditions.

These experiments should, in theory, allow us to explore energy consumption in different environmental and behavioural scenarios. The experimental conditions should yield an array of patterns that quantify energy consumption under these conditions. To explore these data, we produce several visualisations and later interpret outcomes.

As more vehicles adhere to speed limits, we see that the cumulative energy consumption is at its lowest 0.1–2 kWh (Figure A.2.5c,f,i). On the contrary, as more vehicles break speed limits, we see cumulative energy consumption at its highest (Figure A.2.5a,d,g). As density increases, the cumulative energy consumed also increases (Figure A.2.5g,h,i) regardless of speed limit adherence.

As adherence to the speed limit increases, it was observed that the overall distance travelled by vehicles was smaller; thus, energy consumption decreased Figure A.2.6.

According to official Mitsubishi statistics Outlander (2021), the range of the Outlander (kWh/km) is 0.169. To compare, the outputs in Figure A.2.6 show that as adherence increases (Experiment Conditions 3, 6, 9), the cumulative energy consumed is closer to the manufacturer's statistics. The similarity in high adherence cases and energy consumption compared to official statistics can be explained because these statistics were based on fixed/local speed limits in urban environments and do not account for speeding behaviour or consumption levels resulting from motorway speeds. Furthermore, energy efficiency increases as driver behaviour becomes more homogeneous (high adherence), which is what we would expect to see. These data show that the Energy Calculation Extension notebook seen in Figure A.2.1 enables the UTS (Olmez *et al.*, 2021b) to quantify the electric energy consumption, so long as the parameters outlined in Table A.7 are provided.

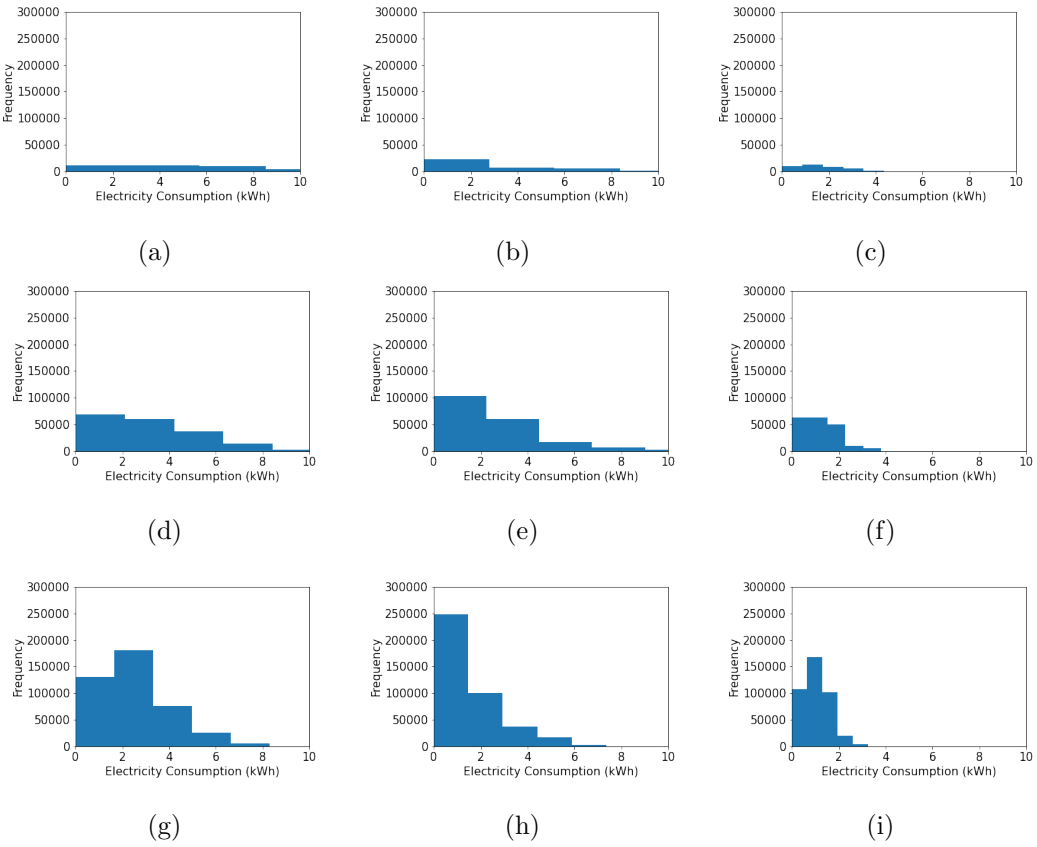


Figure A.2.5: Distribution of the cumulative energy consumption (kWh) for each experiment condition: (**a**) 10 vehicles, 10 non-adherence; (**b**) 10 vehicles, 5 non-adherence; (**c**) 10 vehicles, 0 non-adherence; (**d**) 50 vehicles, 50 non-adherence; (**e**) 50 vehicles, 25 non-adherence; (**f**) 50 vehicles, 0 non-adherence; (**g**) 100 vehicles, 100 non-adherence; (**h**) 100 vehicles, 50 non-adherence; (**i**) 100 vehicles, 0 non-adherence.

To recap, the previous Section A.2.4 developed experiments comparing outputs from the UTS (Olmez *et al.*, 2021b) to a mathematical model of energy fuel intake, emobpy (Gaete-Morales *et al.*, 2020), for validation. We found the UTS and Energy Calculation Extension notebook Figure A.2.1 produced results consistent with the mathematical model of driver behaviour (Gaete-Morales *et al.*, 2020); see Figures A.2.3 and A.2.4. The subsequent section, across nine experimental conditions, compared the effect that adherence to speed limits and vehicle density had on electric energy consumption by modelling a specific vehicle type (Table A.7) and subsequent results (Figures A.2.5

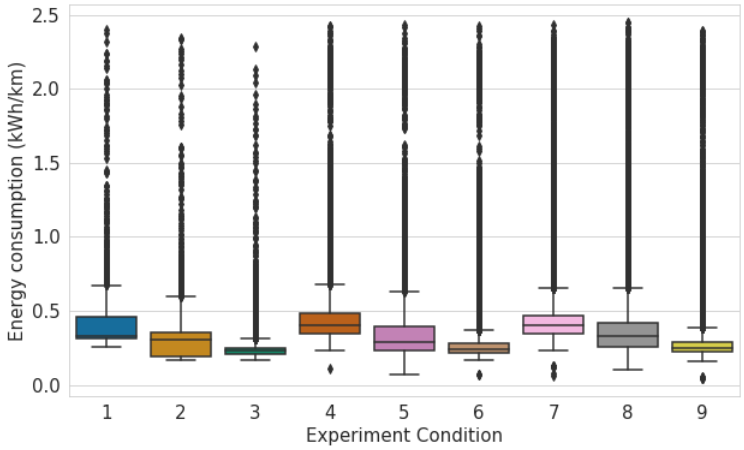


Figure A.2.6: Box plots of energy consumption per kilometer (kWh/km) across all experiment conditions

and A.2.6).

Fuel Consumption of Internal Combustion Engine Vehicle Fleet

Following the aims set out in Section A.2.1, to produce estimates of petrol consumption (L/km), an ICEV must be modelled. This process is straightforward, as the UTS is not vehicle-specific; other fuel types such as petrol and diesel can be modelled using the formulae from Section A.2.8.

As previously discussed in Section A.2.4, a specific type of vehicle must be identified to model energy consumption. From January 2020 to December 2020, the Ford Fiesta ST was purchased (registered) 49,174 times in the UK, making it the most-purchased ICEV according to Crooks (2020); therefore, it was the chosen ICEV. The vehicle parameters can be observed in Table A.9.

Parameter	Value
Height (m)	1.469
Width (m)	1.941
k	0.33% ((Automotive, 2022)) [1]
m (kg)	1635
C_D	0.341

Table A.9: Vehicle parameters (source (Thomasen, 2018)) (ICEV).

Equations (A.2.3), (A.2.6) and (A.2.7) were applied to the model outputs to calculate petrol intake:

- For Equation (A.2.3), F is calculated by using the following parameter variables: $\theta = 0$ as the surface area is flat, $C_D = 0.341$, $A = 2.851\ m$ (where height = 1.469 m, width = 1.941 m), $m = 1635$ kg (Table A.9), $a = \frac{\Delta v}{\Delta t}$ where Δv is the velocity change over time period Δt, and lastly, v = velocityMagnitude (Table A.6).

- For Equation (A.2.6), E_{out} is calculated by multiplying the output from Equation (A.2.3) with total_distance (d) travelled in meters per second for each agent; see Table A.6.

- Lastly, Equation (A.2.7) is calculated by multiplying the output from Equation (A.2.3) (F) with the distance travelled d divided by the engine efficiency $k = 0.47$. E_{in} is then multiplied by 2.923 to convert joule to gasoline/petrol (L).

To ensure comparability, the experiment conditions in Table A.8 will be re-run using the vehicle-specific parameters from Table A.9.

As engine efficiency for ICEVs is relatively low compared to EV/PHEVs, the amount of petrol converted to power that moves the vehicle is also low. Roughly 70% of energy is lost during this process (Automotive, 2022). Given the aforementioned, it is likely that the fuel consumption outputs seen in Figure A.2.7 deviate from the true value by some margin.

As expected, the energy calculation extension notebook used in conjunction with UTS (Olmez *et al.*, 2021b) produced outputs for the nine experimental conditions in Table A.8 with a different fuel type (Table A.9). The drive cycle patterns for the ICEV scenario (Figure A.2.7) are similar to those of the PHEV scenario (Figure A.2.6). At an individual level, the outliers for both conditions (Figures A.2.6 and A.2.7) could be due to traffic congestion, vehicle weight and routes travelled. According to the vehicle's technical specification, (Thomasen, 2018), in extra-urban environments, the vehicle is claimed to do 5.1 L/100 km. Therefore, on average, the expected fuel consumption would be roughly 0.05 litres per kilometer (19.608 km/L), as described in the PHEV case, as adherence increases and driver behaviour is more compliant with rules. The energy consumption patterns are closer to the manufacturers' specification in Figure A.2.7

[1] The official engine efficiency statistic is not provided by the vehicle manufacturer; therefore, an average engine efficiency for ICEVs was acquired from the cited source.

for Experiment Conditions 3, 6 and 9. However, as adherence decreases, distance travelled increases, and the fuel consumption levels deviate from the vehicle manufacturer's technical specifications. This would be an expected behaviour as every vehicle is driving at various speeds (heterogenous driver behaviour), which is an unrealistic observation of driver behaviour. The vehicle manufacturer may not have considered this when estimating fuel consumption. It could be that the tests carried out to measure the fuel economy are conducted using specific drive cycles such as high adherence.

Two examples of simulated fuel types have been quantified with a reasonable degree of accuracy. This has resulted in enough data to quantify the costs of running these vehicles in the hypothetical urban street network.

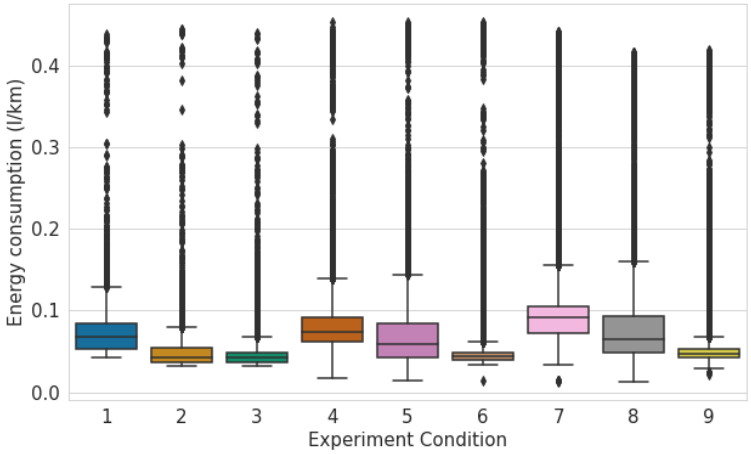


Figure A.2.7: Box plots of energy consumption per kilometer (L/km) across all experiment conditions.

Monetary Costs of Fuel and Electric Consumption

The domestic cost of fuel per litre (L) and electricity per kilowatt-hour (kWh) fluctuates over time. The price for each varies, depending on vehicle fuel efficiency, distance travelled, weight and the fuel price (Shafiei *et al.*, 2017). Therefore, to quantify the cost in Great British Pounds (GBP), the current cost of petrol and electricity is adopted. These are GBP 1.43 (per L/km) (Petrol Prices, 2021) and GBP 0.17 (per kWh/km) (Point, 2021), respectively.

To calculate the cost of petrol, the amount of petrol intake calculated for each vehicle was multiplied by 1.428, while the amount of electric energy intake was multiplied by

0.17. This produced a rough estimate of the fuel costs per car for each experiment condition for the UK.

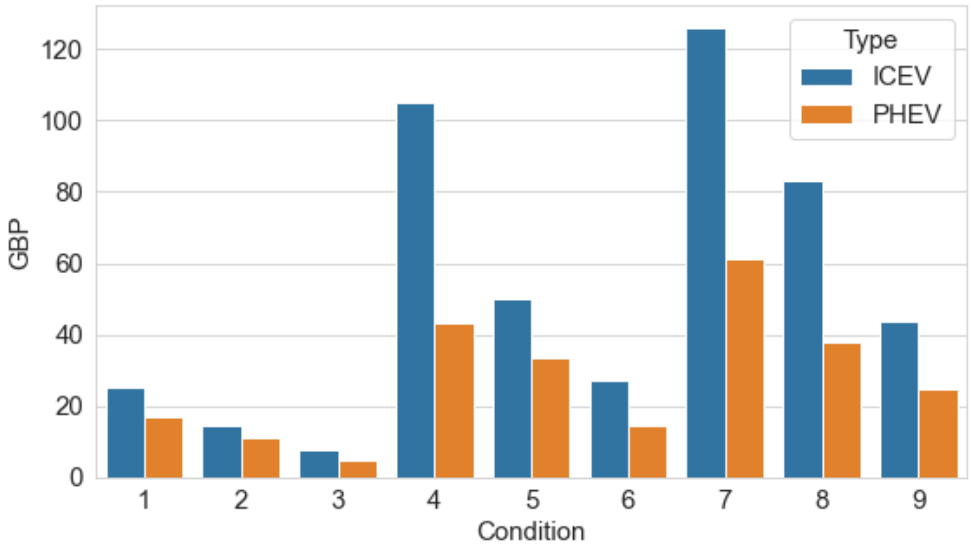


Figure A.2.8: The total sum of petrol/electric costs (GBP) for each experiment condition across all vehicles.

The total cost of electric and petrol across all experiment conditions were £248.76 and £481.89, respectively (These costs are merely estimates produced from the vehicle-specific parameters seen in Tables A.7 and A.9 and drive cycle scenarios from Table A.8 in an urban street network. These costs will not be the same in different types of street networks such as highways, motorways and rural roads). Overall, it is £233.13 cheaper to run PHEVs over ICEVs. As engine efficiency is greater for EV technologies (Prud'homme & Koning, 2012), it is likely that these vehicles would be more fuel-efficient and thus cost less than both PHEVs and ICEVs. In Figure A.2.8, a pattern emerges, where as driver adherence decreases (speeding increases) the cost of electricity and petrol increases (experiment conditions 1, 4 and 7) relative to vehicle density. Furthermore, as adherence increases, the difference in price between the fuel types shrinks (experiment conditions 3, 6 and 9) as vehicles have travelled roughly the same distances Figures A.2.6 and A.2.7. Similarly, these trends are broken down in the average cost per km in Table A.11. Overall, we observe that PHEVs on average are cheaper per kilometer travelled than ICEVs. Furthermore, when the total cost of each model condition results are compared, we see a clear distinction between costs,

and overall, PHEVs are cheaper (Figure A.2.8).

A core strength of ABMs over mathematical models, as specified earlier, is the spatio-temporal resolution variability of data. For instance, individual-level (vehicle-level) data are attained and should provide a more enhanced snapshot of the impact behaviour had on fuel costs.

An individual-level break-down of the fuel expenditure costs across all vehicles and types are discerned in Figures A.2.9 and A.2.10 and in Appendix A.2.10, Figures A.2.12–A.2.15. These results re-enforce our earlier made suppositions. For instance, as more vehicles regulate speed, the difference in average costs decreases, as seen in Figures A.2.9C and A.2.10C. Despite these trends, it could be argued that as variability amongst acceleration/deceleration increases (heterogeneity), PHEV owners will save more money compared to ICEV owners. The net benefits may not be substantial sums of money, but the environmental benefits (which have not been modelled) could be a bonus for consumers. Furthermore, as empirical evidence indicates, autonomous electric vehicles (AEVs) are more efficient than PHEVs/ICEVs (Lee & Kockelman, 2019). Consequently, we could expect a greater extent of financial savings and environmental benefits for owners of AEVs.

To conclude, when vehicle fleet is heterogeneous (more variability among speeds), this leads to greater savings in adopting PHEVs over ICEVs as the engine efficiency is greater. Consequently, more energy is converted to power than ICEVs (Experiment Conditions 4, 7). In contrast, as speeds become more regulated and similar, the average monetary costs between the two vehicle types reduce, as seen in Figures A.2.9C and A.2.10C. Furthermore, these findings are consistent for all fleet sizes, as seen in Figures A.2.12F, A.2.14F, A.2.13I and A.2.15I.

A.2.5 Discussion

This article set out to quantify the energy consumption by a fleet of vehicles in an urban street network using agent-based modelling. Given the current global agenda on climate change through various institutions and policies (i.e., COP 26, Paris Climate Accord 2015, Organisation for Economic Co-operation and Development, Green Economic Recovery (UK)), political discourse worldwide has shifted focus to green agendas, particularly renewable technologies, to facilitate a reduction of carbon emissions. The work conducted in this article plays a significant role in aiding national governments

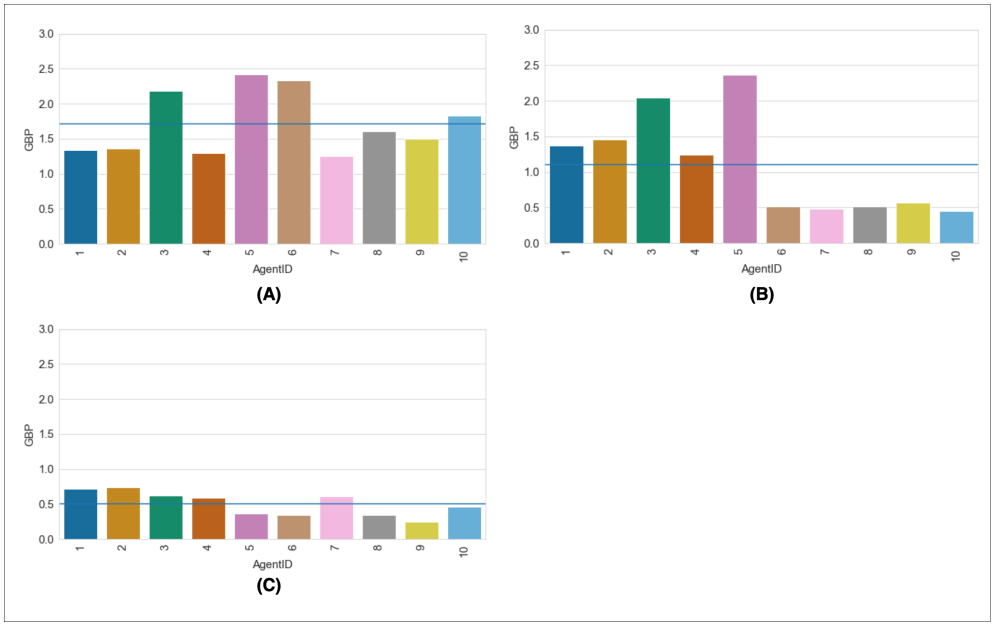


Figure A.2.9: The total sum of electric costs (GBP) for each PHEV, model conditions 1 to 3. Where (A): 10 vehicles, 10 non-adherence, (B): 10 vehicles, 5 non-adherence and (C): 10 vehicles, 0 non-adherence.

in modelling the potential landscape of energy consumption by EV/PHEVs. Previous literature has focused on mathematical models; however, this method is limited. Agent-based models are better suited to modelling individual-level drive cycle behaviours than mathematical models. The former typically provides an aggregated average of energy expenditure, while the latter provides a finer spatio-temporal resolution of individual-level energy consumption, highlighting the immediate environment's impact on a vehicle's drive cycle, which affects energy consumption and efficiency. The model presented has demonstrated vehicle energy efficiency patterns that have been identified in empirical literature (Moriarty & Wang, 2017; Wager *et al.*, 2016), namely that the longer distances travelled at speed reduces battery/engine efficiency, and conversely, vehicles that abide by rules in the long-term are more efficient. Another crucial finding was that as speed adherence increased, the energy consumption per kilometer better reflected vehicle manufacturer statistics for consumption; this was the case for both PHEV and ICEV tests, suggesting that vehicle manufacturers might be testing their vehicles at constant, more regulated speeds. Finally, we found that PHEVs overall are cheaper vehicles to run compared to ICEVs. This is due to several reasons, but primarily that

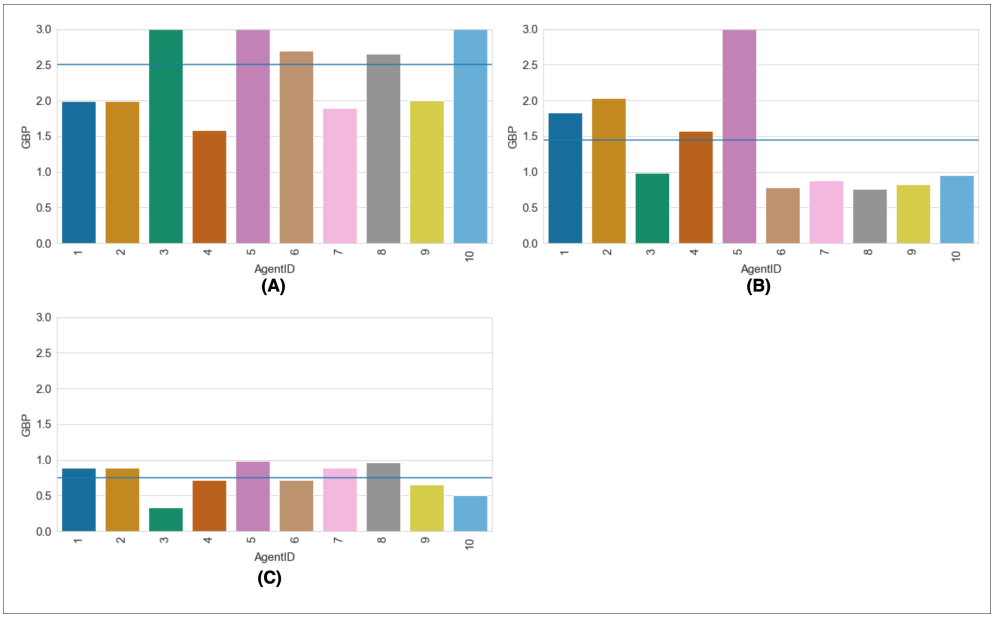


Figure A.2.10: The total sum of petrol costs (GBP) for each ICEV, model conditions 1 to 3. Where (A): 10 vehicles, 10 non-adherence, (B): 10 vehicles, 5 non-adherence and (C): 10 vehicles, 0 non-adherence.

PHEVs are more engine-efficient compared to ICEVs.

This work makes a novel methodological contribution to the modelling of vehicle energy consumption using agent-based modelling. While several attempts to model vehicle activity to quantify energy have been made, the majority of these models have hard-coded driver behaviour, which is bounded by constant speeds (Butler *et al.*, 1999; Gaete-Morales, 2021). This, we believe, is not informative of energy consumption in urban spaces, where the stochastic environment plays a more significant role in affecting energy intake, such as urban speed limits, stop-go rules, and other vehicles in the street network.

Another vehicle technology that is beginning to gain traction is autonomous vehicles (AVs). UK Government projections show a net gain of 823,000 jobs and over £82 billion from the manufacturing and shipping of AVs (Places Catapult, 2019). AVs are likely to be electric, so their ascendance may also be considered in relation to electric energy usage of vehicles of the future. Lee & Kockelman (2019) outlines how AVs and electric vehicles may reduce energy consumption due to their connected environment, as route choice can be optimised to avoid congestion, undertake routes with fewer

stops and ensure multiple passengers are catered to at once due to dynamic ride-sharing capabilities. Additionally, AVs may be able to drive in an optimally efficient manner due to the incorporation of traffic conditions received through communication and sensors. Lee & Kockelman (2019) estimates the energy consumption from these smoother driving cycles would decrease current energy use by between -10% and -20%.

In more revolutionary scenarios in which proportions of fully autonomous vehicles outweigh human-operated vehicles, vehicle-to-vehicle (V2V) communication could enable velocity synchronisation and shorter spaces between vehicles (i.e., platooning). Li *et al.* (2017); Talebpour & Mahmassani (2016) outline how this can improve string stability and increase network capacity as vehicles will operate with decreased acceleration noise and maintain closer distances to nearby vehicles, thus reducing aerodynamic drag. Lee & Kockelman (2019) outlines the energy and emissions reduction from autonomous vehicle platooning to be between 7% and 35%. Although sophisticated autonomous fleets (levels 4 and 5) are yet to be technologically perfected, as currently, the highest level of autonomy achieved in vehicles on sale is level 3 (Innovation, 2020), these findings provide insight into the combined benefits of electric autonomous vehicle (EAV) fleets in the future.

A.2.6 Conclusions

A future avenue to explore would be to extend the model environment to cater for AVs. This should be achievable as the model can currently model specific vehicles and adherence levels. By modelling AVs, the variation of energy consumption scenarios of typical EVs and AVs can be compared. Furthermore, this could allow policymakers to model charging infrastructure in cities to test how these different vehicles can adapt optimally to these environmental changes.

A significant limitation of this work is computational tractability. The compute demand exponentially grows as we increase the number of vehicles or induce complex environmental settings. This can prevent users from simulating a greater capacity of vehicles or more complex cities. Secondly, we assume the world to be a flat plane in the model. However, this diverges from the real world. This assumption was due to computational demand, and we tried to configure the most simple environmental setting to ensure computation was not hampered. However, as cloud computing technologies become mainstream, this problem can be overcome.

Another limitation of this article is that it utilises an urban road typology only. While necessary, as most EV infrastructure is focused on cities, this does not necessarily imply that energy consumption from urban street networks would remain identical to other road typologies (e.g., motorways, dual-carriageways, rural roads). Therefore, in the future, the model can be extended by analysing vehicle behaviours and their subsequent impact on energy consumption in various road typologies.

This study highlights the importance of individual-based modelling methods such as ABMs in investigating future transport systems in cities. As some of the most important global policy agendas focus on the diffusion of low carbon-emitting technologies, this research is well-timed and crucial in planning for the future city.

A.2.7 Open-Source Code and Data

(1) the datasets for both PHEV and ICEV can be accessed at the following source (Olmez & Heppenstall, 2021); (2) the Urban Traffic Simulator Agent-Based Model can be accessed at the following link click here (accessed on 20 May 2022); (3) the Energy Calculation Extension can be found at the following link click here (accessed on 30 May 2022).

A.2.8 Energy Calculation

To calculate the electricity energy consumption required to move the vehicles, the application of classical mechanics (Kibble & Berkshire, 2004) including the drag equation from fluid dynamics (Batchelor & Batchelor, 2000) were adopted for this article.

When a vehicle is moving at a constant velocity, its forces are balanced (i.e., the forces driving it forward are equal to those resisting). However, vehicle velocity is not constant when driver behaviour changes over the drive cycle period (e.g., halting at traffic lights, matching the speed of vehicles ahead). Therefore, we assume velocity v is not constant in this model. A vehicle travelling at a non-constant speed results from an imbalance in the forces acting on it, i.e., the net force acting on the vehicle is non-zero. Considering the drive force from the engine, the force of gravity, the drag force and rolling resistance opposing motion, the net (or total) force acting on a vehicle, F_{total}, can be calculated as:

$$F_{total} = \overbrace{F}^{drive} - \overbrace{mg \times \sin(\theta)}^{gravity} - \overbrace{\frac{1}{2}\rho C_D A v^2}^{drag} - \overbrace{C_{rr} mg \times \cos(\theta)}^{rollingresistance}, \tag{A.2.1}$$

where:

- F is the force provided by the engine driving the vehicle forward (N);
- m is the mass of the vehicle (kg);
- g is the gravitational acceleration (m/s^2);
- θ is the angle of the surface on which the vehicle is driving on;
- ρ is the density of air (1.225 kg/m^3);
- C_D is the drag coefficient;
- A is the reference area of the vehicle (m^2) (width $\times$ height);
- v is the velocity (m/s), and;
- C_{rr} is the coefficient of rolling resistance.

For this investigation, the value of the coefficient of rolling resistance is taken to be 0.012 based on the assumption that all vehicles in the system are passenger vehicles and the road surfaces are made of smooth asphalt (Palasz *et al.*, 2019). The total force acting on the vehicle can be expressed as the product of the vehicle's mass and its acceleration, i.e., $F_{total} = ma$, and consequently, we can write Equation (A.2.1) as:

$$ma = F - mg \times \sin(\theta) - \frac{1}{2}\rho C_D A v^2 - C_{rr} mg \times \cos(\theta) \tag{A.2.2}$$

In this investigation, we are concerned with the force produced by the engine, F, and the associated energy expended to produce this force. As a consequence, we may wish to rearrange Equation (A.2.2) as:

$$F = ma + mg \times \sin(\theta) + \frac{1}{2}\rho C_D A v^2 - C_{rr} mg \times \cos(\theta). \tag{A.2.3}$$

In the scenario where the road surface is flat, i.e., the vehicle is not travelling uphill or downhill, $\theta = 0$. This results in the gravitational aspect of the forces resisting motion being zero, i.e., $mg\sin(\theta) = 0$; it also results in the rolling resistance being $C_{rr} mg \times \cos(\theta) = C_{rr} mg$. Equation (A.2.3) therefore becomes:

$$F = ma + \frac{1}{2}\rho C_D A v^2 + C_{rr} mg, \tag{A.2.4}$$

which returns the force output by the car's engine to accelerate at rate a. In cases when the vehicle is travelling at a constant speed, Equation (A.2.4) simplifies to:

$$F = \frac{1}{2}\rho C_D A v^2 + C_{rr} m g. \tag{A.2.5}$$

Once the force exerted by the engine, F, and the distance of travel over which it is being exerted d are both known, the energy expended by the engine, E_{out}, can be calculated:

$$E_{out} = F \times d. \tag{A.2.6}$$

In this case, E_{out} is the energy output by the engine. To find the energy provided to the engine in the form of fuel, the engine efficiency, k, is needed. Assuming that the efficiency of the engine is constant, i.e., that it has the same efficiency for all scenarios, the energy that needs to be provided to the engine can be found using the following equation:

$$E_{in} = F \times \frac{d}{k}, \tag{A.2.7}$$

A.2.9 Tables

Variable	Output Type
VelocityChange	Float
Acceleration	Float
Deceleration	Float
Braking Energy (kWh) [1]	Float
Drag_Force	Float
Acceleration_Force	Float
Total_Force	Float
Drag_Work	Float
Acceleration_Work	Float
Total_Work	Float
Energy_Input (kWh)	Float
Energy_Input_Sum (kWh)	Float

Table A.10: Energy Calculation Extension notebook output data (EV/PHEV example).

[1] An amount of energy is generated every time a vehicle brakes (decelerates), also known as regenerative braking. This is accounted for in the notebook using the braking energy formula from the following source: (Ram, 2020).

MC	PHEV	ICEV
1	0.60	0.92
2	0.43	0.56
3	0.22	0.34
4	1.93	4.42
5	1.50	2.25
6	0.67	1.22
7	2.79	5.66
8	1.72	3.71
9	1.12	1.98

Table A.11: Average cost (£) per km for both vehicle types.

A.2.10 Figures

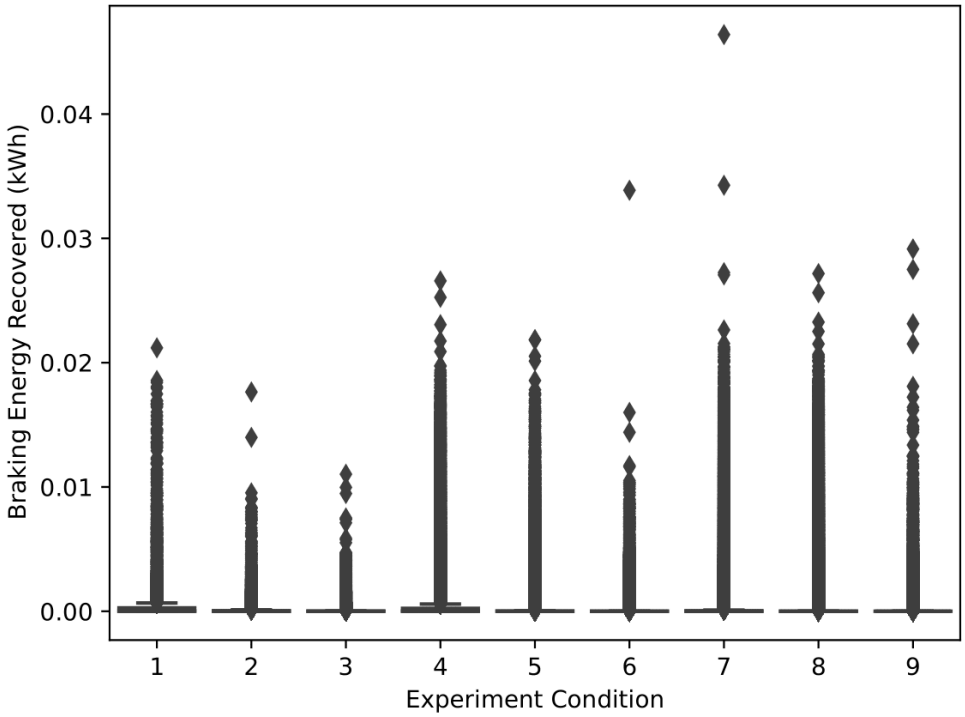


Figure A.2.11: Box plots of braking energy recovered in kWh for each experiment condition.

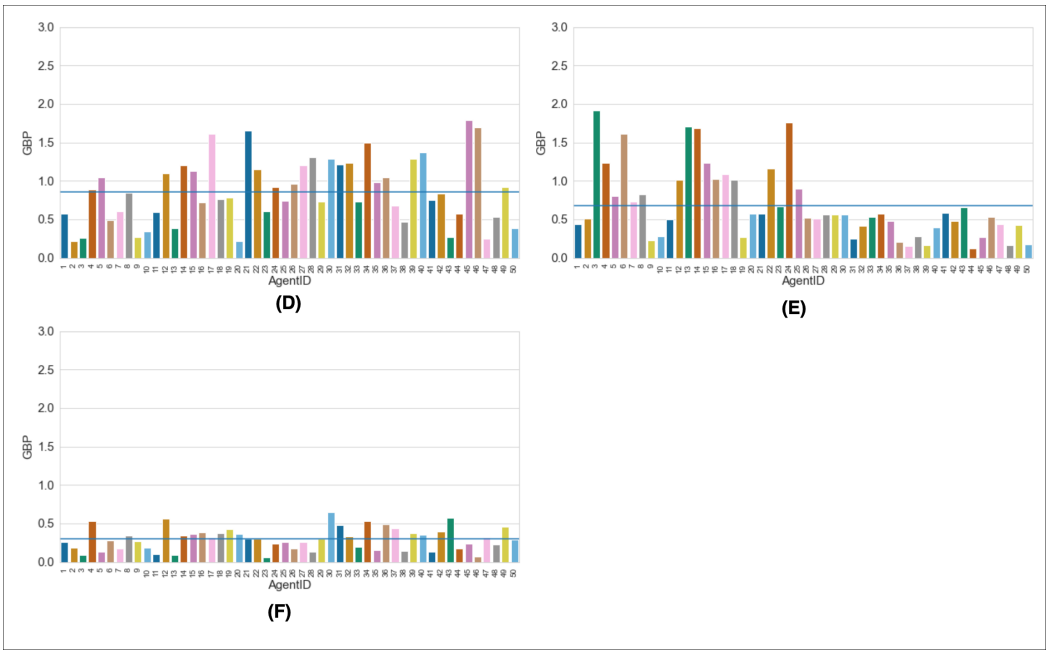


Figure A.2.12: The total sum of electric costs (GBP) for each PHEV, model conditions 4 to 6. Where (D): 50 vehicles, 50 non-adherence, (E): 50 vehicles, 25 non-adherence and (F): 50 vehicles, 0 non-adherence.

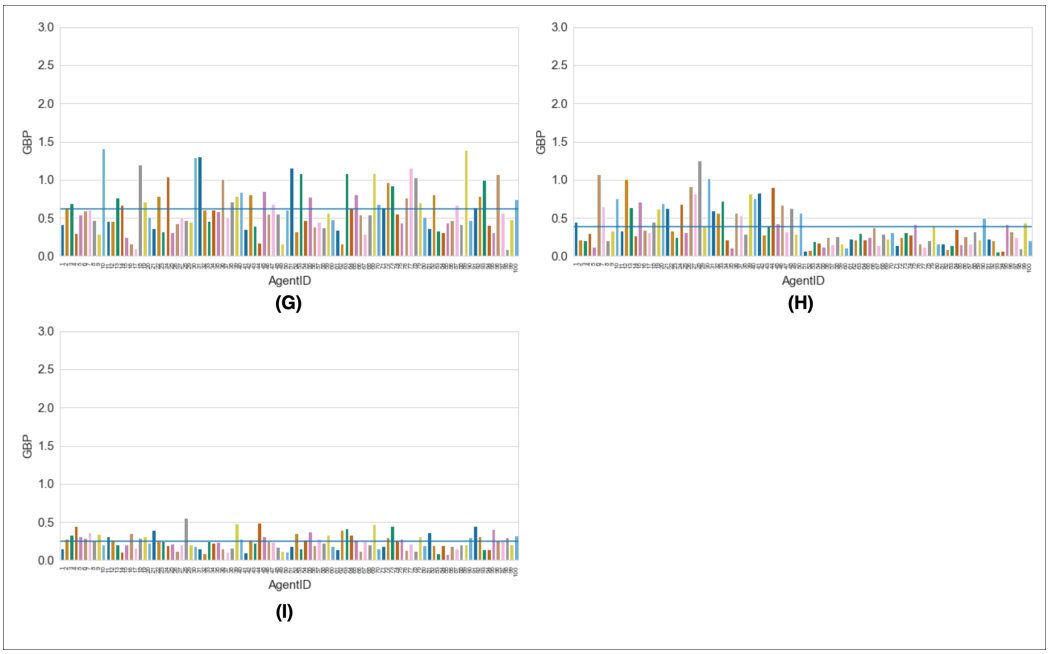


Figure A.2.13: The total sum of electric costs (GBP) for each PHEV, model conditions 7 to 9. Where (G): 100 vehicles, 100 non-adherence, (H): 100 vehicles, 50 non-adherence and (I): 100 vehicles, 0 non-adherence

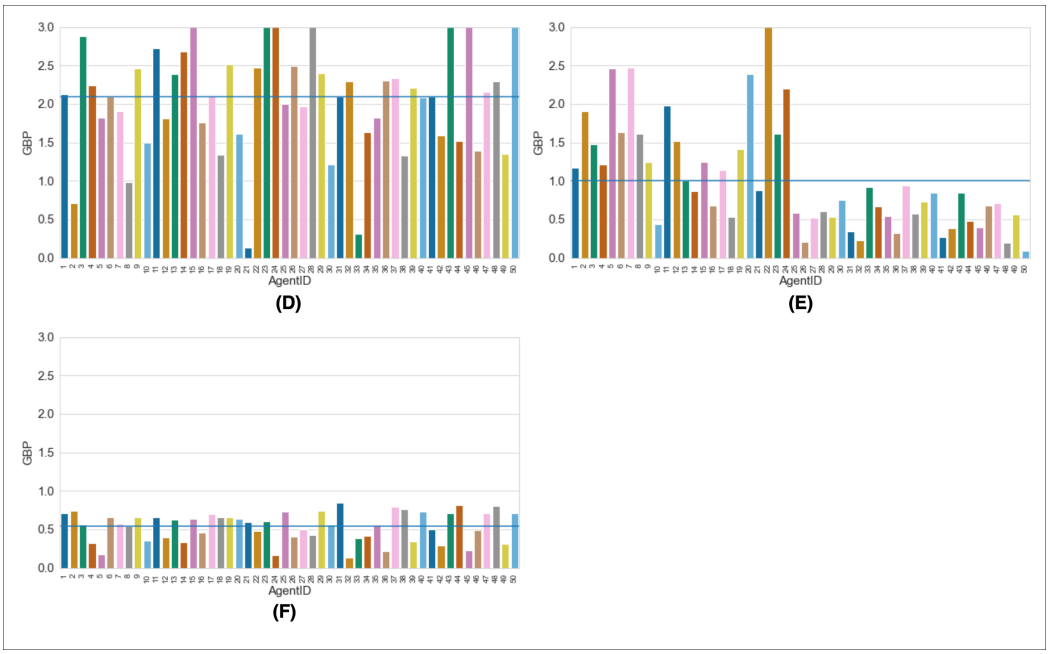


Figure A.2.14: The total sum of petrol costs (GBP) for each ICEV, model conditions 4 to 6. Where (D): 50 vehicles, 50 non-adherence, (E): 50 vehicles, 25 non-adherence and (F): 50 vehicles, 0 non-adherence.

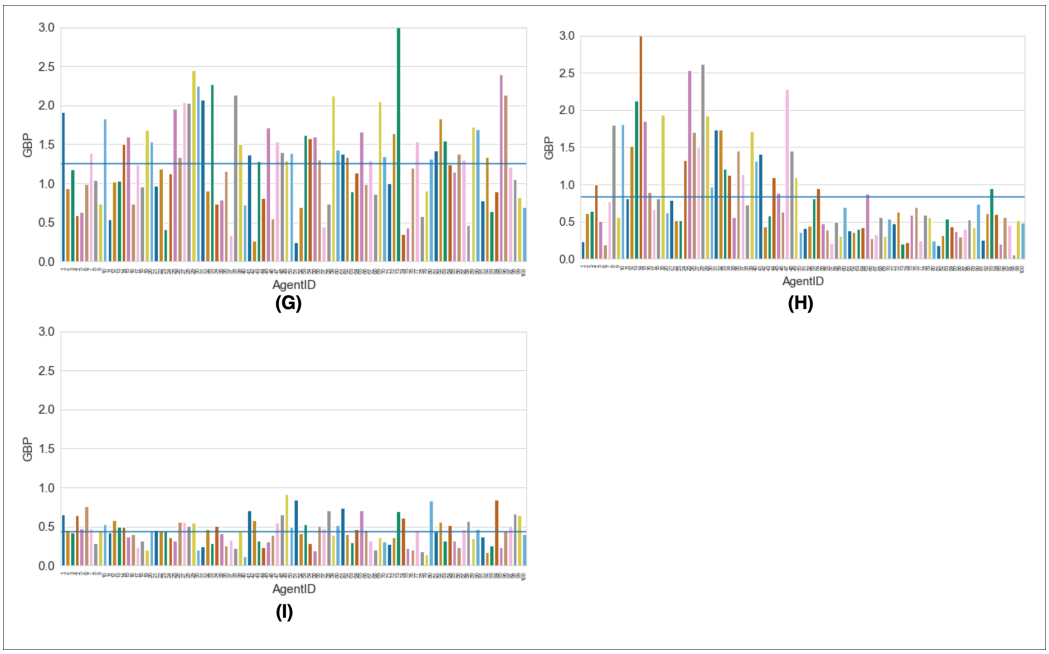


Figure A.2.15: The total sum of petrol costs (GBP) for each ICEV, model conditions 7 to 9. Where (G): 100 vehicles, 100 non-adherence, (H): 100 vehicles, 50 non-adherence and (I): 100 vehicles, 0 non-adherence

References

A Gulli, S.P. (2012). *Deep Learning with Keras*. Packt Publishing Ltd. 4

Aarts, L. & Van Schagen, I. (2006). Driving speed and the risk of road crashes: A review. *Accident Analysis and Prevention*, **38**, 215–224. 183

Agarwal, A., Kakade, S.M., Lee, J. & Mahajan, G. (2020). Optimality and Approximation with Policy Gradient Methods in Markov Decision Processes. In *Conference on Learning Theory*. 154

Albatayneh, A., Assaf, M.N., Alterman, D. & Jaradat, M. (2020). Comparison of the Overall Energy Efficiency for Internal Combustion Engine Vehicles and Electric Vehicles. *Environmental and Climate Technologies*, **24**, 669–680. 209

Almahamid, F. & Grolinger, K. (2021). Reinforcement Learning Algorithms: An Overview and Classification. *Canadian Conference on Electrical and Computer Engineering*, **2021-September**. 141

Anderson, C.A. & Titler, M.G. (2014). Development and verification of an agent-based model of opinion leadership. *Implementation Science*, **9**. 2

Anderson, J.R. (2013). *The adaptive character of thought*. Psychology Press. 38

Andrey, J., Mills, B., Leahy, M. & Suggett, J. (2003). Weather as a chronic hazard for road transportation in Canadian cities. *Natural Hazards*, **28**, 319–343. 207

Andrighetto, G., Conte, R., Turrini, P. & Paolucci, M. (2007). Emergence in the loop: Simulating the two way dynamics of norm innovation. In *Dagstuhl Seminar Proceedings*. 34

Angus, T. (2021). *2021 Australian Energy Statistics (Electricity) — Ministers for the Department of Industry, Science, Energy and Resources*. 2021-10-05, https://www.minister.industry.gov.au/ministers/taylor/media-releases/2021-australian-energy-statistics-electricity. 209

Arthur, W.B. (1994). Inductive Reasoning and Bounded Rationality: The El Farol Problem. *American Economic Review*. 37, 125

Asadi, K. (2015). Strengths, weaknesses, and combinations of model-based and model-free reinforcement learning. *Department of Computing Science University of Alberta*. 21

Asai, T. & Ikeda, M. (2008). Comparison of multi-attribute decision making and reinforcement learning for resource allocation in a manufacturing system. *IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans*, **38**, 545–552. 39

Author, ., Rakauskas, M.E., Ward, N.J., Gerberich, S.G., Alexander, B.H. & Program, H. (2007). Rural and Urban Safety Cultures: Human-Centered Interventions Toward Zero Deaths in Rural Minnesota. Tech. rep., University of Minnesota. 188

Automotive, A. (2022). How Efficient is Your Cars Engine — AAA Automotive. 226, 227

Axelrod, R. (1980). Effective Choice in the Prisoner's Dilemma. *Journal of Conflict Resolution*. 31

Axtell, R.L. (2014). An agent-based model of the housing market bubble in metropolitan Washington, D.C. 6, 44, 45, 140

Back, T. (1996). *Evolutionary algorithms in theory and practice: evolution strategies, evolutionary programming, genetic algorithms*. Oxford university press. 18, 24

Badham, J., Chattoe-Brown, E., Gilbert, N., Chalabi, Z., Kee, F. & Hunter, R.F. (2018). Developing agent-based models of complex health behaviour. *Health & Place*, **54**, 170–177. 2, 7

Bae, J.W., Paik, E., Dongoh, K., Jung, J. & Lee, C.H. (2019). Simulation framework for self-evolving agent-based models: A case study of housing market model. *Proceedings - Winter Simulation Conference*, **2018-December**, 1120–1131. 140

Bagheri-Jebelli, N., Crooks, A. & Kennedy, W.G. (2020). Capturing The Effects of Gentrification on Property Values: An Agent-based Modeling Approach. In *Conference of the Computational Social Science Society of the Americas*, 245–264. 44

Baker, B., Kanitscheider, I., Markov, T., Wu, Y., Powell, G., McGrew, B. & Mordatch, I. (2019). Emergent tool use from multi-agent autocurricula. *arXiv preprint arXiv:1909.07528*. 53, 97, 155

Balendra, P. (2020). Vehicle Speed Compliance Statistics, Great Britain: January - June 2020. Tech. rep., Department for Transport. 183, 191, 197, 205, 213, 218

Balke, T. & Gilbert, N. (2014). How do agents make decisions? A survey. *JASSS*. 2, 31, 37, 38, 39, 49

Baptista, R., Farmer, J.D., Hinterschweiger, M., Low, K., Tang, D. & Uluc, A. (2016). Macroprudential Policy in an Agent-Based Model of the UK Housing Market. *SSRN Electronic Journal*. 140, 141

Barr, R. & Pease, K. (1990). Crime Placement, Displacement, and Deflection. *Crime and Justice*. 129

Baruya, A. (1998). Speed-accident relationships on European roads. In *9th International Conference on Road Safety in Europe*, 1–19. 189

Batchelor, C. & Batchelor, G. (2000). *An introduction to fluid dynamics*. Cambridge University Press. 234

Batty, M. (2013). *The new science of cities*. MIT press. 48

Baudains, P., Braithwaite, A. & Johnson, S.D. (2013). Target choice during extreme events: A discrete spatial choice model of the 2011 london riots. *Criminology*. 118

Baumann, O. (2015). Models of Complex Adaptive Systems in Strategy and Organization Research. *SSRN Electronic Journal*. 15

Been, V., Ellen, I., Figlio, D.N., Nelson, A., Ross, S., Schwartz, A.E., Stiefel, L. & Ellen, A. (2021). THE EFFECTS OF NEGATIVE EQUITY ON CHILDREN'S EDUCATIONAL OUTCOMES. *NBER*. 157

Beer, R.D. & Gallagher, J.C. (1992). Evolving Dynamical Neural Networks for Adaptive Behavior. *Adaptive Behavior*, **1**, 91–122. 18

Behrisch, M., Bieker, L., Erdmann, J. & Krajzewicz, D. (2011). SUMO {–} Simulation of Urban MObility: An Overview. In S..U. of Oslo Aida Omerovic, R.T.I.I.R.T.P.D.A. Simoni & R.T.I.I.R.T.P.G. Bobashev, eds., *Proceedings of SIMUL 2011, The Third International Conference on Advances in System Simulation*, ThinkMind. 187

Bellemare, M.G., Naddaf, Y., Veness, J. & Bowling, M. (2013). Autonomous driving using deep reinforcement learning. In *International conference on machine learning*, 1942–1950. 25

Bencivenga, C., Sargenti, G. & D'Ecclesia, R.L. (2010). Energy markets: crucial relationship between prices. *Mathematical and Statistical Methods for Actuarial Sciences and Finance*, 23–32. 211

Benenson, I., Martens, K. & Birfir, S. (2008). PARKAGENT: An agent-based model of parking in the city. *Computers, Environment and Urban Systems*. 48, 219

Bener, A. & Alwash, R. (2002). A perspective on motor vehicle crash injuries and speeding in the United Arab Emirates. *Traffic Injury Prevention*. 206

Berglund, E.Z. (2015). Using agent-based modeling for water resources planning and management. *Journal of Water Resources Planning and Management*, **141**, 4015025. 1

Bernasco, W. & Luykx, F. (2003). EFFECTS OF ATTRACTIVENESS, OPPORTUNITY AND ACCESSIBILITY TO BURGLARS ON RESIDENTIAL BURGLARY RATES OF URBAN NEIGHBORHOODS. *Criminology*, **41**, 981–1002. 128

Bianchi, F. & Squazzoni, F. (2015). Agent-based models in sociology. *WIREs Computational Statistics*, **7**, 284–306. 186, 211

Birkin, M. (2021). Microsimulation. *Urban Book Series*, 845–864. 4

Birks, D. & Davies, T. (2017). STREET NETWORK STRUCTURE AND CRIME RISK: AN AGENT-BASED INVESTIGATION OF THE ENCOUNTER AND ENCLOSURE HYPOTHESES. *Criminology*. 41

Birks, D., Townsley, M. & Stewart, A. (2012). Generative explanations of crime: Using simulation to test criminological theory. *Criminology*. 1, 42, 43, 92, 95, 96, 97, 100, 103, 104, 128, 186, 210

Boeing, G. (2020). A multi-scale analysis of 27,000 urban street networks: Every US city, town, urbanized area, and Zillow neighborhood. *Environment and Planning B: Urban Analytics and City Science*, **47**, 590–608. 195, 196, 217

Bogdoll, J., Hartmanns, A. & Hermanns, H. (2012). Simulation and Statistical Model Checking for Modestly Nondeterministic Models. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, **7201 LNCS**, 249–252. 151

Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L., Monfort, M., Muller, U., Zhang, J. & others (2016). Limitations of deep reinforcement learning for autonomous driving. *arXiv preprint arXiv:1608.01230*. 27

Bonte, L., Espié, S. & Mathieu, P. (2006). Modélisation et simulation des usagers deux-roues motorisés dans ARCHISIM. *JFSMA*, **6**, 17. 187

Bosse, T. & Gerritsen, C. (2008). Agent-Based Simulation of the Spatial Dynamics of Crime: On the Interplay between Criminal Hot Spots and Reputation. In *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems - Volume 2*, AAMAS '08, 1129–1136, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC. 6, 41, 42, 92, 95, 97, 108, 128

Bosse, T., Jonker, C.M., Van Der Meij, L. & Treur, J. (2007). A language and environment for analysis of dynamics by simulation. *International Journal on Artificial Intelligence Tools*. 41

Bosse, T., Gerritsen, C., Hoogendoorn, M., Jaffry, S.W. & Treur, J. (2011). Agent-based vs. population-based simulation of displacement of crime: A comparative study. *Web Intelligence and Agent Systems*. 6, 97, 128

Brantingham, P. & Brantingham, P. (1995). Criminality of Place: Crime Generators and CrimeAttractors. *European Journal on Criminal Policy and Research*, **13**, 5–26. 41, 116, 128

Brantingham, P.J. & Brantingham, P.L. (2016). The geometry of crime and crime pattern theory. In *Environmental criminology and crime analysis*, 117–135, Routledge. 94

Brantingham, P.L. & Brantingham, P.J. (2019). Environment, Routine, and Situation: Toward a Pattern Theory of Crime. In *Routine Activity and Rational Choice*, 259–294, Routledge. 42, 93, 95, 104, 106, 107, 118, 128

Brantingham, P.L., Brantingham, P.J. & Taylor, W. (2006). Situational Crime Prevention as a Key Component in Embedded Crime Prevention. *http://dx.doi.org/10.3138/cjccj.47.2.271*, **47**, 271–292. 103

Brearcliffe, D.K. & Crooks, A. (2021). Creating Intelligent Agents: Combining Agent-Based Modeling with Machine Learning. *Springer Proceedings in Complexity*, 31–58. 4, 82

Breen, J.M., Næss, P.A., Hansen, T.B., Gaarder, C. & Stray-Pedersen, A. (2020). Serious motor vehicle collisions involving young drivers on Norwegian roads 2013–2016: Speeding and driver-related errors are the main challenge. *Traffic Injury Prevention*, **21**, 382–388. 183

Bretzke, W.R. (2013). Global urbanization: A major challenge for logistics. *Logistics Research*, **6**, 57–62. 208

Brintrup, A. (2010). Behaviour adaptation in the multi-agent, multi-objective and multi-role supply chain. *Computers in Industry*, **61**, 636–645. 16

Broersen, J., Dastani, M., Hulstijn, J., Huang, Z. & van der Torre, L. (2001). The BOID architecture: conflicts between beliefs, obligations, intentions and desires. In *Proceedings of the fifth international conference on Autonomous agents*, 9–16. 33

Broersen, J., Dastani, M., Hulstijn, J. & van der Torre, L. (2002). Goal generation in the BOID architecture. *Cognitive Science Quarterly*, **2**, 428–447. 33

Brslica, V. (2011). Plug-in Hybrid Vehicles. *Electric Vehicles The Benefits and Barriers*. 221

Bryant, J.W. (2004). Drama theory as the behavioural rationale in agent-based models. 3

Buşoniu, L., Babuška, R. & De Schutter, B. (2010). Multi-agent reinforcement learning: An overview. *Studies in Computational Intelligence*. 23, 27, 108

Butler, K.L., Ehsani, M. & Kamath, P. (1999). A matlab-based modeling and simulation package for electric and hybrid electric vehicle design. *IEEE Transactions on Vehicular Technology*, **48**, 1770–1778. 232

Byrne, M.D. (2000). The ACT-R/PM project. In *Simulating Human Agents: Papers from the 2000 Fall Symposium*, 1–3. 36

Cabrera-Arnau, C., Curiel, R.P. & Bishop, S.R. (2020). Uncovering the behaviour of road accidents in urban areas. *Royal Society Open Science*. 189, 204, 206

Cai, X., Huang, Y. & Liu, X. (2019). Autonomous driving with natural language and reinforcement learning. *IEEE Access*, **7**, 113–944. 25

Cameron, G.D.B. & Duncan, G.I.D. (1996). PARAMICS—Parallel microscopic simulation of road traffic. *The Journal of Supercomputing*, **10**, 25–53. 187

Cao, S. & Liu, Y. (2011). An Integrated Cognitive Architecture for Cognitive Engineering Applications. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 55, 1075–1079, SAGE Publications. 36

Capasso, A.P., Bacchiani, G. & Broggi, A. (2020). From Simulation to Real World Maneuver Execution using Deep Reinforcement Learning. *IEEE Intelligent Vehicles Symposium, Proceedings*, 1570–1575. 17

Card, S.K. (1981). The model human processor: A model for making engineering calculations of human performance. In *Proceedings of the Human Factors Society Annual Meeting*, vol. 25, 301–305. 35

Card, S.K., Moran, T.P. & Newell, A. (1983). *The psychology of human-computer interaction*. Lawrence Erlbaum Associates. 38, 39

Carstensen, C.L. (2015). An agent-based model of the housing market: steps toward a computational tool for policy analysis. 140

Casas, J., Ferrer, J.L., Garcia, D., Perarnau, J. & Torday, A. (2010). Traffic simulation with aimsun. In *Fundamentals of traffic simulation*, 173–232, Springer. 187

Caskey, T.R., Wasek, J.S. & Franz, A.Y. (2018). Deter and protect: crime modeling with multi-agent learning. *Complex & Intelligent Systems*. 41, 42, 95, 97

Castelfranchi, C., Dignum, F., Jonker, C.M. & Treur, J. (2000). Deliberative normative agents: Principles and architecture. In *Intelligent Agents VI. Agent Theories, Architectures, and Languages: 6th International Workshop, ATAL'99, Orlando, Florida, USA, July 15-17, 1999. Proceedings 6*, 364–378. 33

Charpentier, A., Élie, R. & Remlinger, C. (2021). Reinforcement Learning in Economics and Finance. *Computational Economics*, 1–38. 6

Chen, J.H., Ong, C.F., Zheng, L. & Hsu, S.C. (2017a). Forecasting spatial dynamics of the housing market using Support Vector Machine. *International Journal of Strategic Property Management*, **21**, 273–283. 143

Chen, Y., Lu, J. & Kockelman, K.M. (2017b). A deep reinforcement learning approach for traffic signal control. *Transportation Research Part C: Emerging Technologies*, **78**, 611–623. 25

Chiriţă, M. (2011). Usefulness of Artificial Neural Networks for Predicting Financial and Economic Crisis. *Economics and Applied Informatics*, 61–66. 142

Choshen, L., Fox, L., Aizenbud, Z. & Abend, O. (2019). On the weaknesses of reinforcement learning for neural machine translation. *arXiv preprint arXiv:1907.01752*. 21

Chu, X. (2018). Policy Optimization With Penalized Point Probability Distance: An Alternative To Proximal Policy Optimization. *ArXiv*, **abs/1807.00442**. 111

Cincotti, S., Guerci, E. & Raberto, M. (2005). Price dynamics and market power in an agent-based power exchange. *Noise and Fluctuations in Econophysics and Finance*, **5848**, 233. 143

Clark, D.E. (2003). Effect of population density on mortality after motor vehicle collisions. *Accident Analysis and Prevention*. 203

Clarke, D.D., Ward, P.J. & Jones, J. (1998). Overtaking road-accidents: Differences in manoeuvre as a function of driver age. *Accident Analysis and Prevention*, **30**, 455–467. 196

Clarke, R. (1980). "SITUATIONAL" CRIME PREVENTION: THEORY AND PRACTICE. *The British Journal of Criminology*, **20**, 136–147. 42, 92, 101, 102, 113, 116, 127, 129

Clarke, R.V. (1997a). *Situational Crime Prevention: Successful Case Studies*. Harrow and Heston. 92

Clarke, R.V. (1997b). *Situational Crime Prevention: Successful Case Studies*. 103, 113

Clarke, R.V. & Cornish, D.B. (1985). Modeling Offenders' Decisions: A Framework for Research and Policy. *Crime and Justice*, **6**, 147–185. 112

Clarke, R.V. & Weisburd, D. (1994). Diffusion of crime control benefits: Observations on the reverse of displacement. *Crime Prev. Stud.*. 108, 128

Cocea, M., Gutierrez-Santos, S. & Magoulas, G.D. (2012). Case-based reasoning approach to adaptive modelling in exploratory learning. *Studies in Computational Intelligence*, **376**, 167–184. 2, 7, 15

Cochet, H. & Byrne, R.W. (2015). Complexity in animal behaviour: towards common ground. *Acta Ethologica*, **18**, 237–241. 14

Cohen, L.E. & Felson, M. (1979). Social Change and Crime Rate Trends: A Routine Activity Approach. *American Sociological Review*. 42, 93, 95, 103, 128

Colon, C., Claessen, D. & Ghil, M. (2015). Bifurcation analysis of an agent-based model for predator-prey interactions. *Ecological Modelling*. 59

Conte, R., Gilbert, N., Bonelli, G., Cioffi-Revilla, C., Deffuant, G., Kertesz, J., Loreto, V., Moat, S., Nadal, J.P., Sanchez, A. & others (2012). Manifesto of computational social science. *The European Physical Journal Special Topics*, **214**, 325–346. 3

Cornelius, C.V.M., Lynch, C.J., Modeling, V. & Gore, R. (2017). AGING OUT OF CRIME: EXPLORING THE RELATIONSHIP BETWEEN AGE AND CRIME WITH AGENT BASED MODELING. *ADS '17: Proceedings of the Agent-Directed Simulation Symposium*. 6, 97, 125, 128

Cornish, D. & Clarke, R. (2003). A Reply to Wortley's Critique of Situational Crime Prevention. *Crime Prevention Studies*. 92

Cornish, D.B. & Clarke, R.V. (1987). UNDERSTANDING CRIME DISPLACEMENT: AN APPLICATION OF RATIONAL CHOICE THEORY. *Criminology*. 91, 92, 93, 95, 96, 102, 112, 127, 128

Cornish, D.B. & Clarke, R.V. (2017). *The reasoning criminal: Rational choice perspectives on offending*. 112

Crooks, A. (2015). Agent-based Models and Geographical Information Systems. In *Agent-based modeling and geographical information systems.*, 63–77, SAGE Publications Ltd. 1, 211

Crooks, A. (2020). Best-selling cars in the UK 2020 — Auto Express. 226

Crooks, A.T. & Hailegiorgis, A.B. (2014). An agent-based modeling approach applied to the spread of cholera. *Environmental Modelling and Software*. 49

Curreli, C., Pappalardo, F., Russo, G., Pennisi, M., Kiagias, D., Juarez, M. & Viceconti, M. (2021). Verification of an agent-based disease model of human Mycobacterium tuberculosis infection. *International Journal for Numerical Methods in Biomedical Engineering*, **37**. 2

da Silva Assis, L., da Silva Soares, A., Coelho, C.J. & Van Baalen, J. (2016). An Evolutionary Algorithm for Autonomous Robot Navigation. *Procedia Computer Science*, **80**, 2261–2265. 18

Dahlke, J., Bogner, K., Mueller, M., Berger, T., Pyka, A. & Ebersberger, B. (2020). Is the Juice Worth the Squeeze? Machine Learning (ML) In and For Agent-Based Modelling (ABM). *arXiv preprint arXiv:2003.11985*. 97

Dang, Q.V. (2020). Reinforcement Learning in Stock Trading. *Advances in Intelligent Systems and Computing*, **1121 AISC**, 311–322. 6, 91, 108

Darbandsari, P., Kerachian, R., Malakpour-Estalaki, S. & Khorasani, H. (2020). An agent-based conflict resolution model for urban water resources management. *Sustainable Cities and Society*, **57**, 102112. 1

Davis, G.A. & Morris, P. (2009). Statistical versus Simulation Models in Safety: Steps Toward a Synthesis Using Median-Crossing Crashes. *Transportation Research Record*, **2102**, 93–100. 186, 211

Daw, N.D., O'Doherty, J.P., Dayan, P., Seymour, B., Dolan, R.J. & Schultz, W. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, **441**, 876–879. 25

Dawid, H. & Neugart, M. (2011). Agent-based models for economic policy design. *Eastern Economic Journal*. 211

Dawson, R.J., Peppe, R. & Wang, M. (2011). An agent-based model for risk-based flood incident management. *Natural Hazards*. 48

De Jong, K. (2012). Evolutionary computation: A unified approach. In *GECCO'12 - Proceedings of the 14th International Conference on Genetic and Evolutionary Computation Companion*. 23, 24

De Winter, J.C. & Dodou, D. (2010). The driver behaviour questionnaire as a predictor of accidents: A meta-analysis. 188

DeAngelis, D.L. & Diaz, S.G. (2019). Decision-making in agent-based modeling: A current review and future prospectus. 2, 31, 49

Deepanshu Mehta (2020). State-of-the-Art Reinforcement Learning Algorithms. *International Journal of Engineering Research and*, **V8**. 3, 141

Department for Transport, U. (2006). Setting Local Speed Limits. Tech. Rep. July, UK Government Department for Transport. 197, 219

der Hoek, W. & Wooldridge, M. (2003). Towards a logic of rational agency. *Logic Journal of IGPL*, **11**, 135–159. 32

Developers, T. (2023). TensorFlow. 28, 29

Dhakal, T. & Min, K.S. (2021). ¡p class="text"¿Macro Study of Global Electric Vehicle Expansion¡/p¿. *Foresight and STI Governance*, **15**, 67–73. 208

Dhar, S.S., Chakraborty, B. & Chaudhuri, P. (2014). Comparison of multivariate distributions using quantile–quantile plots and related tests. *https://doi.org/10.3150/13-BEJ530*, **20**, 1484–1506. 151

Dickerson, A., Peirson, J., Vickerman, R., Retallack, A.E., Ostendorf, B. & Martin, J.L. (2000). Road accidents and traffic flows: an econometric investigation. *Economica*, **67**, 1393. 204, 205

Dignum, F., Boissier, O. & Rodriguez-Aguilar, J.A. (2006). Using BRIDGE for modeling and simulation of organizational behavior. In *Proceedings of the 5th international joint conference on Autonomous agents and multiagent systems*, 307–314. 33

Dignum, F., Boissier, O. & Rodriguez-Aguilar, J.A. (2009). Using BRIDGE for modeling and simulation of organizational behavior: An update. *Journal of Autonomous Agents and Multi-Agent Systems*, **19**, 99–131. 33

Dignum, V. & Dignum, F. (2015). Contextualized planning using social practices. In *Coordination, Organizations, Institutions, and Norms in Agent Systems X: COIN 2014 International Workshops, COIN AAMAS, Paris, France, May 6, 2014, COIN@ PRICAI, Gold Coast, QLD, Australia, December 4, 2014, Revised Selected Papers 10*, 36–52. 33

Ding, Z. & Dong, H. (2020). Challenges of reinforcement learning. *Deep Reinforcement Learning: Fundamentals, Research and Applications*, 249–272. 129

Donkin, E., Dennis, P., Ustalakov, A., Warren, J. & Clare, A. (2017). Replicating complex agent based models, a formidable task. *Environmental Modelling & Software*, **92**, 142–151. 151, 153

Dou, X. & Wang, J. (2014). Asset Securitization and Bubbles: An Illustration of Subprime Mortgage Default Crisis. *Advances in Economics and Business*, **2**, 112–119. 142

Doya, K. (2010). Reinforcement learning: Computational theory and biological mechanisms. *https://doi.org/10.2976/1.2732246/10.2976/1*, **1**, 30–40. 5

Driving, T.S. (2022). Safe separation distances and what you should know. 191

Durfee, E.H. (1993). Cooperative distributed problem solving between (and within) intelligent agents. In *Neuroscience: From Neural Networks to Artificial Intelligence: Proceedings of a US-Mexico Seminar held in the city of Xalapa in the state of Veracruz on December 9–11, 1991*, 84–98. 32

Eck, J.E. & Clarke, R.V. (2019). Situational Crime Prevention: Theory, Practice and Evidence. *Handbooks of Sociology and Social Research*, 355–376. 92, 125, 127, 128

Eck, J.E. & Liu, L. (2004). Routine activity theory in a RA/CA crime simulation. *American Society of Criminology, Nashville, TN*. 104, 125, 128

Eiksund, S. (2009). A geographical perspective on driving attitudes and behaviour among young adults in urban and rural Norway. *Safety Science*, **47**, 529–536. 188

Emuna, R., Borowsky, A. & Biess, A. (2020). Deep Reinforcement Learning for Human-Like Driving Policies in Collision Avoidance Tasks of Self-Driving Cars. *Arxiv*. 183

Engelbrecht, D. (2023). *Introduction to Unity ML-Agents: Understand the Interplay of Neural Networks and Simulation Space Using the Unity ML-Agents Package*. 55

Eppstein, M.J., Grover, D.K., Marshall, J.S. & Rizzo, D.M. (2011). An agent-based model to study market penetration of plug-in hybrid electric vehicles. *Energy Policy*, **39**, 3789–3802. 211

Epstein, J.M. (1999). Agent-based computational models and generative social science. *Complexity*. 1, 49

Epstein, J.M. & Axtell, R. (1997). Artificial societies and generative social science. *Artificial Life and Robotics*, **1**, 33–34. 1, 93, 140

Erlingsson, E.J., Teglio, A., Cincotti, S., Stefansson, H., Sturluson, J.T. & Raberto, M. (2014). Housing market bubbles and business cycles in an agent-based credit economy. *Economics*, **8**, 2014–2022. 6, 140

Exposition, D.S.I.A.t.A.M., & 2002, u. (2002). RECMODELER: EVALUATING COOPERATIVE COLLISION AVOIDANCE. *trid.trb.org*. 187

Faghri, F. (2021). Training Efficiency and Robustness in Deep Learning. *ArXiv*, **abs/2112.01423**. 129

Fan, J., Wang, Z., Xie, Y. & Yang, Z. (2020). A theoretical analysis of deep Q-learning. In *Learning for dynamics and control*, 486–489. 19

Farkas, A., Kertesz, G. & Lovas, R. (2020). Parallel and Distributed Training of Deep Neural Networks: A brief overview. *INES 2020 - IEEE 24th International Conference on Intelligent Engineering Systems, Proceedings*, 165–170. 129

Farrell, G. (2015). Crime concentration theory. *Crime Prevention and Community Safety 2015 17:4*, **17**, 233–248. 125

Farrell, G. & Pease, K. (2001). *Repeat victimization*, vol. 12. Criminal Justice Press. 108, 128, 129

Filatova, T. (2015). Empirical agent-based land market: Integrating adaptive economic behavior in urban land-use models. *Computers, Environment and Urban Systems*, **54**, 397–413. 140, 141

Filatova, T., Parker, D. & der Veen, A. (2009). Agent-based urban land markets: agent's pricing behavior, land prices and urban land use change. *Journal of Artificial Societies and Social Simulation*, **12**, 3. 44

Filatova, T., Verburg, P.H., Parker, D.C. & Stannard, C.A. (2013). Spatial agent-based models for socio-ecological systems: Challenges and prospects. *Environmental Modelling and Software*. 211

Fildes, B. & Jarvis, J. (1994). *Perceptual Countermeasures: Literature Review*. Transportation Research Board. 189

Fildes, B., G., R. & A., L. (1991). Speed Behaviour and Drivers' Attitudes to Speeding. Tech. rep., Monash University Accident Research Centre. 183, 199

Fildes, B.N., Langford, J.W., Andrea, D.J. & Scully, J.E. (2005). *Balance between harm reduction and mobility in setting speed limits: A feasibility study*. Austroads, Australia, ap-r272/05 edn. 185

Filocamo, B., Ruiz, J.A. & Sotelo, M.A. (2020). Efficient management of road intersections for automated vehicles-the FRFP system applied to the various types of intersections and roundabouts. *Applied Sciences (Switzerland)*. 195, 217

Fischer, T. & Riedler, J. (2014). Prices, debt and market structure in an agent-based model of the financial market. *Journal of Economic Dynamics and Control*, **48**, 95–120. 140

Fitch, G.M., Rakha, H.A., Arafeh, M., Blanco, M., Gupta, S.K., Zimmermann, R.P. & Hanowski, R.J. (2008). Safety benefit evaluation of a forward collision warning system: final report. *NHTSA DOT HS*, **810**, 910. 187

Fleetwood, M.D. & Byrne, M.D. (2002). Modeling icon search in ACT-R/PM. *Cognitive Systems Research*, **3**, 25–33. 36

Florence, P.S. & Zipf, G.K. (1950). Human Behaviour and the Principle of Least Effort. *The Economic Journal*. 102, 107, 118

Foerster, J.N., Nardelli, F., Farquhar, G., Afouras, T., Torrance, H., Pritzel, A., Houthooft, R., Schaul, T., Sifre, L., Kalchbrenner, N. & others (2016). Multi-Agent Communication with Deep Reinforcement Learning. In *Advances in neural information processing systems*, 2137–2145. 26

Foerster, J.N., Klimov, O., Zhang, Y., Cho, K. & Silver, D. (2017). On the limitations of deep reinforcement learning for multi-agent systems. *arXiv preprint arXiv:1709.04326*. 26

Forero, D.S., Ceballos, Y.F. & Torres, G.S. (2019). Simulation of consumers decision-making process using agent-based model approach. *International Journal of Modeling, Simulation, and Scientific Computing*, **10**, 1950037. 35

Francès, G., Rubio-Campillo, X., Lancelotti, C. & Madella, M. (2015). Decision making in agent-based models. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, **8953**, 370–378. 2, 31

Frey, H.C., Rouphail, N.M., Unal, A. & Colyar, J.D. (2001). Emissions reduction through better traffic management: An empirical evaluation based upon on-road measurements. Tech. rep., Transportation Research Board. 200

Funes, P., Orme, B. & Bonabeau, E. (2003). Evolving emergent group behaviors for simple humans agents. In *Proceedings of the seven european conference on artificial life*, 76–89. 17

Gaete-Morales, C. (2021). emobpy: application for the German case. xv, 219, 221, 223, 232

Gaete-Morales, C., Kramer, H., Schill, W.P. & Zerrahn, A. (2020). An open tool for creating battery-electric vehicle time series from empirical data – emobpy. *Scientific Data*, **8**. 219, 220, 222, 225

Galizio, M., Jackson, L.A. & Steele, F.O. (1979). Enforcement symbols and driving speed: The overreaction effect. *Journal of Applied Psychology*. 185

Ge, J. (2014). Who creates housing bubbles? An agent-based study. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, **8235 LNAI**, 143–150. 6, 44, 140

Ge, J. (2017). Endogenous rise and collapse of housing price: An agent-based model of the housing market. *Computers, Environment and Urban Systems*, **62**, 182–198. 6, 44, 140, 141

Geanakoplos, J., Axtell, R., Farmer, J.D., Howitt, P., Conlee, B., Goldstein, J., Hendrey, M., Palmer, N.M. & Yang, C.Y. (2012). Getting at Systemic Risk via an Agent-Based Model of the Housing Market. *American Economic Review*, **102**, 53–58. 140

Gerritsen, C. (2015). Agent-based modelling as a research tool for criminological research. *Crime Science*. 92, 97

Gialopsos, B.M. & Carter, J.W. (2014). Offender Searches and Crime Events:. *http://dx.doi.org/10.1177/1043986214552608*, **31**, 53–70. 6, 92, 127

Gilbert, N. & Troitzsch, K. (2005). *Simulation for the social scientist*. McGraw-Hill Education (UK). 38, 44

Gilbert, N., Hawksworth, J.C. & Swinney, P.A. (2009). An Agent-Based Model of the English Housing Market. In *Technosocial Predictive Analytics, Papers from the 2009 AAAI SpringSymposium, Technical Report SS\-09\-09, Stanford, California, USA, March23\-25, 2009*, 30–35, AAAI. xviii, 6, 140, 141, 145, 148, 149, 151, 153, 154, 156, 157, 158, 161, 164

Godin, P. (2007). Beyond the risk society: Critical reflections on risk and human security. *Health, Risk & Society*, **9**, 343–344. 94

Goldberg, D.E. (1989). *David E. Goldberg-Genetic Algorithms in Search, Optimization, and Machine Learning-Addison-Wesley Professional (1989).pdf*. 23

Gómez-Marín, C.G., Arango-Serna, M.D. & Serna-Urán, C.A. (2018). Agent-based microsimulation conceptual model for urban freight distribution. *Transportation Research Procedia*, **33**, 155–162. 4

Grimm, V. & Railsback, S.F. (2012). Pattern-oriented modelling: a 'multi-scope'for predictive systems ecology. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **367**, 298–310. 6

Grimm, V., Berger, U., Bastiansen, F., Eliassen, S., Ginot, V., Giske, J., Goss-Custard, J., Grand, T., Heinz, S.K., Huse, G., Huth, A., Jepsen, J.U., Jørgensen, C., Mooij, W.M., Müller, B., Pe'er, G., Piou, C., Railsback, S.F., Robbins, A.M., Robbins, M.M., Rossmanith, E., Rüger, N., Strand, E., Souissi, S., Stillman, R.A., Vabø, R., Visser, U. & DeAngelis, D.L. (2006). A standard protocol for describing individual-based and agent-based models. *Ecological Modelling*. 58, 98, 144, 185, 190, 212

Groeneveld, J., Müller, B., Buchmann, C.M., Dressler, G., Guo, C., Hase, N., Hoffmann, F., John, F., Klassert, C., Lauf, T., Liebelt, V., Nolzen,

H., Pannicke, N., Schulze, J., Weise, H. & Schwarz, N. (2017). Theoretical foundations of human decision-making in agent-based land use models – A review. 2, 49

Groff, E.R. (2006). *Exploring the geography of routine activity theory: a spatio-temporal test using street robbery*. University of Maryland, College Park. 42

Groff, E.R. (2007). Simulation for theory testing and experimentation: An example using routine activity theory and street robbery. *Journal of Quantitative Criminology*. 6, 41, 42, 43, 92, 96, 97, 104, 128

Groff, E.R., Johnson, S.D. & Thornton, A. (2019). State of the Art in Agent-Based Modeling of Urban Crime: An Overview. *Journal of Quantitative Criminology*. 6, 93, 125

Groff Elizabeth R., Johnson, S.D. & Thornton Amy (2018). State of the Art in Agent-Based Modeling of Urban Crime: An Overview. *Journal of Quantitative Criminology*. 140

Gu, S., Yang, L., Du, Y., Chen, G., Walter, F., Wang, J., Yang, Y. & Knoll, A. (2022). A Review of Safe Reinforcement Learning: Methods, Theory and Applications. *ArXiv*, **abs/2205.10330**. 57

Guerette, R.T. & Bowers, K.J. (2009). ASSESSING THE EXTENT OF CRIME DISPLACEMENT AND DIFFUSION OF BENEFITS: A REVIEW OF SITUATIONAL CRIME PREVENTION EVALUATIONS*. *Criminology*, **47**, 1331–1368. 129

Guerini, M. & Moneta, A. (2017). A method for agent-based models validation. *Journal of Economic Dynamics and Control*, **82**. 2

Gulden, T., Harrison, J.F. & Crooks, A.T. (2011). Modeling cities and displacement through an agent-based spatial interaction model. In *The Computational Social Science Society of America Conference*. 1

Haase, D., Lautenbach, S. & Seppelt, R. (2010). Modeling and simulating residential mobility in a shrinking city using an agent-based approach. *Environmental Modelling & Software*, **25**, 1225–1240. 45

Hachicha, M., Halima, R.B. & Kacem, A.H. (2017). Modeling and verifying self-adaptive systems: A refinement approach. *2016 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2016 - Conference Proceedings*, 3967–3972. 16

Hamill, L. & Gilbert, N. (2015). Agent-Based Modelling in Economics. 1, 6, 140

Han, B., Sun, D., Yu, X., Song, W. & Ding, L. (2020). Classification of urban street networks based on tree-like network features. *Sustainability (Switzerland)*. 195, 217

Harrison, G.W. (1994). Expected utility theory and the experimentalists. In *Experimental economics*, 43–73, Springer. 31

Hasan, M.A., Frame, D.J., Chapman, R. & Archie, K.M. (2021). Costs and emissions: Comparing electric and petrol-powered cars in New Zealand. *Transportation Research Part D: Transport and Environment*, **90**, 102671. 211, 212

Hawick, K.A., Scogings, C.J. & James, H.A. (2008). Defensive spiral emergence in a predator-prey model. *Complexity International*. 59

Hawkins, T.R., Gausen, O.M. & Strømman, A.H. (2012). Environmental impacts of hybrid and electric vehicles-a review. 209

Hawkins, T.R., Singh, B., Majeau-Bettez, G. & Strømman, A.H. (2013). R E S E A R C H A N D A N A LY S I S Comparative Environmental Life Cycle Assessment of Conventional and Electric Vehicles. *Wiley Online Library*, **17**, 53–64. 209

Hawryszkiewycz, I.T. (2009). Modeling complex adaptive systems. *Lecture Notes in Business Information Processing*, **20 LNBIP**, 458–468. 18

Hayward, K. (2007). Situational crime prevention and its discontents: Rational choice theory versus the 'culture of now'. *Social Policy and Administration*. 92

Hayward, K. (2017). Situational crime prevention and its discontents: Rational choice theory versus the 'culture of now'. In *Crime Opportunity Theories*, 323–341, Routledge. 94

He, J., Li, X., Yao, Y., Hong, Y. & Jinbao, Z. (2018a). Mining transition rules of cellular automata for simulating urban expansion by using the deep learning techniques. *International Journal of Geographical Information Science*, **32**, 2076–2097. 5

He, Z., Dong, J. & Yu, L. (2018b). An agent-based model for investigating the impact of distorted supply–demand information on China's resale housing market. *Journal of Computational Science*, **25**, 1–15. 44

Heckbert, S., Baynes, T. & Reeson, A. (2010). Agent-based modeling in ecological economics. 211

Hemelrijk, C. (2013). Simulating Complexity of Animal Social Behaviour. 581–615. 14

Heppenstall, A., Evans, A. & Birkin, M. (2006). Using hybrid agent- based systems to model spatially- influenced retail markets. *JASSS*. 140, 219

Heppenstall, A.J., Crooks, A.T., See, L.M. & Batty, M. (2012). *Agent-based models of geographical systems*. Springer Netherlands. 6, 93, 211

Highway Code, U. (2022). Speed limits - GOV.UK. 212

Hillier, B. (2007). *Space is the machine: a configurational theory of architecture*. Space Syntax. 104

Ho, N. (2020). How AI Can Help Build Resiliency for Small Businesses in a Global Economic Crisis. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 3606–3606. 142

Houston, A.I., McNamara, J.M. & others (1999). *Models of adaptive behaviour: an approach based on state*. Cambridge University Press. 16

Huston, M., DeAngelis, D. & Post, W. (1988). New Computer Models Unify Ecological Theory. *BioScience*, **38**, 682–691. 186, 210

Huynh, H.X., Trung Do, K., Nguyen, K.D., Truc Thi Pham, P., Vo, T.H. & Minh Truong, T. (2019). Dissolved oxygen simulation of catfish pond with cellular automata. *Proceedings of 2019 11th International Conference on Knowledge and Systems Engineering, KSE 2019*. 4

Hwang, M., Cho, S. & Lee, K.M. (2011). Virtual fish: An agent-based model of group hunting behavior using reinforcement learning. *Simulation Modelling Practice and Theory*, **19**, 1487–1497. 25

Iea (2020). Electricity Market Report - December 2020 – Analysis - IEA. Tech. rep., IEA. 212

Imam, I. & Kerschberg, L. (1997). Adaptive Intelligent Agents. *Journal of Intelligent Information Systems 1997 9:3*, **9**, 211–213. 141

Inman, P. (2022). What is the Bank of England doing in bid to stabilise UK economy? — Bank of England — The Guardian. 162

Innovation, I. (2020). New Level 3 Autonomous Vehicles Hitting the Road in 2020 - IEEE Innovation at Work. 233

Iskander, N., Simoni, A., Alonso, E. & Peter, M. (2020). Reinforcement Learning Agents for Ubisoft's Roller Champions. *arXiv preprint arXiv:2012.06031*. 40

Islam, M., Chen, G. & Jin, S. (2019). An Overview of Neural Network. *http://www.sciencepublishinggroup.com*, **5**, 7. 91, 106, 108

Izquierdo, S.S., Izquierdo, L.R. & Gotts, N.M. (2008). Reinforcement learning dynamics in social dilemmas. *JASSS*, **11**. 5

J. Lynch, D. & Adam, K. (2022). Bank of England intervenes to stabilize UK finances after Liz Truss budget - The Washington Post. 162

Jalalimanesh, A., Shahabi Haghighi, H., Ahmadi, A. & Soltani, M. (2017). Simulation-based optimization of radiotherapy: Agent-based modeling and reinforcement learning. *Mathematics and Computers in Simulation*. 6, 76, 91, 143

Janssen, M., van der Voort, H. & van Veenstra, A.F. (2015). Failure of large transformation projects from the viewpoint of complex adaptive systems: Management principles for dealing with project dynamics. *Information Systems Frontiers*, **17**, 15–29. 16

Jipp, M. (2007). *Situation Adaptation: Information Acquisition, Human Behavior and its Determining Abilities*. Ph.D. thesis, Universität zu Köln. 121

Johnson, E.J. & Goldstein, D. (2003). Do defaults save lives? 37

Johnson, S.D. & Groff, E.R. (2014). Strengthening Theoretical Testing in Criminology Using Agent-based Modeling. *The Journal of research in crime and delinquency*, **51**, 509–525. 6, 91, 93, 95, 125

Johnson, S.D., Bernasco, W., Bowers, K.J., Elffers, H., Ratcliffe, J., Rengert, G. & Townsley, M. (2007). Space-time patterns of risk: A cross national assessment of residential burglary victimization. *Journal of Quantitative Criminology*. 98

Johnson, S.D., Guerette, R.T. & Bowers, K. (2014). Crime displacement: what we know, what we don't know, and what it means for crime reduction. *Journal of Experimental Criminology*. 6, 91, 94

Jones, L. (2007). Barbara Jones. *BMJ*, **335**, 519.4. 188, 204

Jordan, R., Birkin, M. & Evans, A. (2011). Agent-Based Simulation Modelling of Housing Choice and Urban Regeneration Policy. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, **6532 LNAI**, 152–166. 6, 140

Jordan, R., Birkin, M. & Evans, A. (2012). Agent-based modelling of residential mobility, housing choice and regeneration. *Agent-Based Models of Geographical Systems*, 511–524. 6, 140

Joubert, C.J., Saprykin, A., Chokani, N. & Abhari, R.S. (2022). Large-scale agent-based modelling of street robbery using graphical processing units and reinforcement learning. *Computers, Environment and Urban Systems*, **94**, 101757. 5, 92, 96, 109, 127

Joyce, K.E., Laurienti, P.J. & Hayasaka, S. (2012). Complexity in a brain-inspired agent-based model. *Neural Networks*, **33**, 275–290. 16

Juliani, A., Berges, V.P., Vckay, E., Gao, Y., Henry, H., Mattar, M. & Lange, D. (2018). Unity: A general platform for intelligent agents. *arXiv preprint arXiv:1809.02627*. xii, xvii, 4, 50, 53, 54, 60, 64, 65, 76, 81, 89, 110, 111, 155, 192, 215

Justesen, N., Bontrager, P., Togelius, J. & Risi, S. (2017). Deep Learning for Video Game Playing. *arXiv preprint arXiv:1708.07902*. 40

Justesen, N., Bontrager, P., Togelius, J. & Risi, S. (2020). Deep learning for video game playing. *IEEE Transactions on Games*, **12**, 1–20. 91

Kaelbling, L.P., Littman, M.L. & Moore, A.W. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*. 23, 108, 142

Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *American economic review*, **93**, 1449–1475. 31

Kahneman, D. & Tversky, A. (2013). Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I*, 99–127, World Scientific. 30

Kakade, S., Kakade, S. & Langford, J. (2002). Approximately Optimal Approximate Reinforcement Learning. *IN PROC. 19TH INTERNATIONAL CONFERENCE ON MACHINE LEARNING*, 267–274. 51

Kamo, K.I., Fukui, K., Ito, Y., Nakayama, T. & Katanoda, K. (2022). How much can screening reduce colorectal cancer mortality in Japan? Scenario-based estimation by microsimulation. *Japanese Journal of Clinical Oncology*, **52**, 221–226. 4

Kang, D.O., Bae, J.W., Lee, C., Jung, J.Y. & Paik, E. (2019). Data Assimilation Technique for Social Agent-Based Simulation by Using Reinforcement Learning. *Proceedings of the 2018 IEEE/ACM 22nd International Symposium on Distributed Simulation and Real Time Applications, DS-RT 2018*, 220–221. 144

Kangur, A., Jager, W., Verbrugge, R. & Bockarjova, M. (2017). An agent-based model for diffusion of electric vehicles. *Journal of Environmental Psychology*, **52**, 166–182. 49, 211

Katoch, S., Chauhan, S.S. & Kumar, V. (2021). A review on genetic algorithm: past, present, and future. *Multimedia Tools and Applications*, **80**, 8091–8126. 23, 24

Kayhan, B.M. & Yildiz, G. (2021). Reinforcement learning applications to machine scheduling problems: a comprehensive literature review. *Journal of Intelligent Manufacturing*, **34**, 905–929. 58

Kazakov, R., Howick, S. & Morton, A. (2021). Managing complex adaptive systems: A resource/agent qualitative modelling perspective. *European Journal of Operational Research*, **290**, 386–400. 15, 18

Kennedy, J., Gorell, R., Crinson, L., Wheeler, A. & Elliott, M. (2005). 'Psychological' traffic calming Prepared for Traffic Management Division, Department for Transport. *20splentyforus.nationbuilder.com*. 185

Kennedy, W.G. (2012). Modelling human behaviour in agent-based models. In *Agent-Based Models of Geographical Systems*, 167–179, Springer Netherlands. 2, 15

Kerr, C.C., Stuart, R.M., Mistry, D., Abeysuriya, R.G., Rosenfeld, K., Hart, G.R., Núñez, R.C., Cohen, J.A., Selvaraj, P., Hagedorn, B. & others (2021). Covasim: an agent-based model of COVID-19 dynamics and interventions. *PLOS Computational Biology*, **17**, e1009149. 1

Ketkar, N. & Ketkar, N. (2017). Introduction to keras. *Deep learning with python: a hands-on introduction*, 97–111. 28, 29

Khan, N.A. & Habib, M.A. (2021). Microsimulation of mobility assignment within an activity-based travel demand forecasting model. *https://doi.org/10.1080/23249935.2021.1983664*. 4

Kibble, T. & Berkshire, F. (2004). *Classical mechanics*. World Scientific Publishing Company. 234

Kim, T. & Lee, J.H. (2019). Effects of Hyper-Parameters for Deep Reinforcement Learning in Robotic Motion Mimicry: A Preliminary Study. In *2019 16th International Conference on Ubiquitous Robots, UR 2019*, 228–235, Institute of Electrical and Electronics Engineers Inc. 64

Kitajima, M. & Toyota, M. (2012). Simulating navigation behaviour based on the architecture model Model Human Processor with Real-Time Constraints (MHP/RT). *Behaviour & Information Technology*, **31**, 41–58. 35

Kitto, K.J. (2006). *Modelling and generating complex emergent behaviour*. Flinders University, School of Chemistry, Physics and Earth Sciences. 18

Kloeden, C., Ponte, G. & McLean, A. (1997). Travelling Speed and the Risk of Crash Involvement on Rural Roads. *Ponte*. 183

Kloeden, C., Ponte, G. & McLean, J. (2001). *Travelling speed and risk of crash involvement on rural roads*. Transport Safety Bureau. 183

Kober, J. & Peters, J. (2013). Introduction to Deep Reinforcement Learning. *Neural Networks*, **45**, 185–204. 28

Koster, C. & De Jong, K. (2009). Reward hacking in multi-agent learning. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*, 637–644. 25

Kothari, V., Blythe, J., Smith, S. & Koppel, R. (2014). Agent-based modeling of user circumvention of security. In *ACM International Conference Proceeding Series*. 140, 219

Koutn'k, J., Cuccu, G., Schmidhuber, J. & Gomez, F. (2013). Autonomous driving in urban environments using deep reinforcement learning. In *International conference on artificial neural networks*, 322–329. 25

Kuhnle, A., Schaarschmidt, M. & Fricke, K. (2017). Tensorforce: a TensorFlow library for applied reinforcement learning. Web page. 145

Kullback, S. & Leibler, R.A. (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics*. 52

Kvassay, M., Krammer, P., Hluch\'y, L. & Schneider, B. (2017). Causal analysis of an agent-based model of human behaviour. *Complexity*, **2017**. 34

Laird, J.E., Newell, A. & Rosenbloom, P.S. (1987). Soar: An architecture for general intelligence. *Artificial Intelligence*, **33**, 1–64. 36

Langton, N., Polhill, J.G. & Gotts, N. (2005). Social influence in a multi-layered model of decision-making. *Environmental Modelling & Software*, **20**, 1159–1168. 33

Langton, N., Polhill, J.G. & Gotts, N. (2007). BRIDGE: A multi-layered model of decision making in emergency response. *Environmental Modelling & Software*, **22**, 8. 33

Lapan, M. (2018). Deep Reinforcement Learning Hands-On. Apply modern RL methods, with deep Q-networks, value iteration, policy gradients, TRPO, AlphaGo zero and more. 23, 27

Lecun, Y., Bengio, Y. & Hinton, G. (2015). Deep learning. 19

Lee, J. & Kockelman, K.M. (2019). Energy implications of self-driving vehicles. In *98th Annual Meeting of the Transportation Research Board in Washington, DC*. 230, 232, 233

Lee, J.D., McGehee, D.V., Brown, T.L. & Reyes, M.L. (2002). Collision warning timing, driver distraction, and driver response to imminent rear-end collisions in a high-fidelity driving simulator. *Human Factors*, **44**, 314–334. 187

Leibo, J.Z., Zambaldi, V., Lanctot, M., Marecki, J., Graepel, T. & Lillicrap, T. (2017a). Multi-agent reinforcement learning in sequential social dilemmas. In *Proceedings of the International Conference on Machine Learning*, 3053–3062. 25, 26

Leibo, J.Z., Zambaldi, V., Lanctot, M., Marecki, J., Graepel, T. & Lillicrap, T. (2017b). Virtual wildlife conservation model based on deep reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, 2329–2338. 25

Lelei, D.E.K. & McCalla, G. (2019). How many times should a pedagogical agent simulation model be run? *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, **11625 LNAI**, 182–193. 151

Letmathe, P. & Suares, M. (2017). A consumer-oriented total cost of ownership model for different vehicle types in Germany. *Transportation Research Part D: Transport and Environment*, **57**, 314–335. 212

Levy, L., Santhakumaran, D. & Whitecross, R. (2014). *What Works to Reduce Crime?: A Summary of the Evidence*. Scottish Government, Social Research. 128

Lewis, R.L., Dabney, W., Hessel, M., Van Hasselt, H. & Silver, D. (2020). A Survey of Deep Reinforcement Learning: Classic, Deep, and Actor-critic. *arXiv preprint arXiv:1909.06560*. 20, 28

Li, S.E., Zheng, Y., Li, K., Wu, Y., Hedrick, J.K., Gao, F. & Zhang, H. (2017). Dynamical Modeling and Distributed Control of Connected and Automated Vehicles: Challenges and Opportunities. *IEEE Intelligent Transportation Systems Magazine*, **9**, 46–58. 233

Licato, J., Sun, R. & Bringsjord, S. (2014). Structural representation and reasoning in a hybrid cognitive architecture. *IEEE Access*, **2**, 1190–1203. 35

Lilly, C. (2021). Electric vehicle market statistics 2021 - How many electric cars in UK ? 220

Lin, M. & Yao, Y. (2018). Simulation of water pollution accident based on cellular automata. *ACM International Conference Proceeding Series*, 270–274. 4

Linden, R. (2007). Situational crime prevention: Its role in comprehensive prevention initiatives. *IPC Review*, **1**, 139–159. 92

Littman, M.L. (2015). Reinforcement learning improves behaviour from evaluative feedback. *Nature 2015 521:7553*, **521**, 445–451. 97, 141

Liu, S., See, K.C., Ngiam, K.Y., Celi, L.A., Sun, X. & Feng, M. (2020). Reinforcement Learning for Clinical Decision Support in Critical Care: Comprehensive Review. *J Med Internet Res 2020;22(7):e18477 https://www.jmir.org/2020/7/e18477*, **22**, e18477. 6, 109, 143

Lockwood, P.L. & Klein-Flügge, M.C. (2021). Computational modelling of social cognition and behaviour—a reinforcement learning primer. *Social Cognitive and Affective Neuroscience*, **16**, 761–771. 14, 18, 97

LondonAssembly (2021). *Electric Vehicle Infrastructure — London City Hall*. 2021-10-13. 209

Lopes, G.C., Ferreira, M., Da Silva Simoes, A. & Colombini, E.L. (2018). Intelligent control of a quadrotor with proximal policy optimization reinforcement learning. In *Proceedings - 15th Latin American Robotics Symposium, 6th Brazilian*

Robotics Symposium and 9th Workshop on Robotics in Education, LARS/SBR/WRE 2018. 49, 54, 76

Lorscheid, I. (2014). Learning agents for human complex systems. *Proceedings - IEEE 38th Annual International Computers, Software and Applications Conference Workshops, COMPSACW 2014*, 432–437. 8

Loukis, E., Kyriakou, N. & Maragoudakis, M. (2020). Using Government Data and Machine Learning for Predicting Firms' Vulnerability to Economic Crisis. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, **12219 LNCS**, 345–358. 142

Lowe, R., Wu, Y., Tamar, A., Harb, J., Abbeel, P. & Mordatch, I. (2017). The limitations of deep reinforcement learning in cooperative multi-agent systems. *arXiv preprint arXiv:1705.08926*. 26

Lowes, A., Kudenko, D., Tuyls, K., Bull, P., Pickett, A. & MacDonald, B. (2017). Multi-Agent Deep Reinforcement Learning for Complex Cooperative and Competitive Environments. *arXiv preprint arXiv:1703.05407*. 27

Luce, R.D. & Raiffa, H. (1989). *Games and decisions: Introduction and critical survey*. Courier Corporation. 37

Luhnow, D. & Colchester, M. (2022). U.K.'s Central Banker Struggles With Inflation, a Financial Crisis and His Own Government - WSJ. 162

Luo, L., Zhou, S., Cai, W., Low, M.Y.H., Tian, F., Wang, Y., Xiao, X. & Chen, D. (2008). Agent-based human behavior modeling for crowd simulation. In *Computer Animation and Virtual Worlds*, vol. 19, 271–281. 48

Luong, M.T., Sutskever, I. & Le, Q.V. (2015). Deep reinforcement learning for machine translation. *arXiv preprint arXiv:1509.00685*. 24

Lynch, M.F., Sun, R. & Wilson, N. (2011). CLARION as a cognitive framework for intelligent virtual agents. In *International Workshop on Intelligent Virtual Agents*, 460–461. 35

Maes, P. (1987). Concepts and experiments in computational reflection. *ACM Sigplan Notices*, **22**, 147–155. 32

Maghdid, H.S. & Ghafoor, K.Z. (2020). A Smartphone Enabled Approach to Manage COVID-19 Lockdown and Economic Crisis. *SN Computer Science*, **1**, 1–9. 142

Mahmoud, M.A., Ahmad, M.S., Ahmad, A., Mohd Yusoff, M.Z. & Mustapha, A. (2012). Norms detection and assimilation in multi-agent systems: a conceptual approach. In *Knowledge Technology: Third Knowledge Technology Week, KTW 2011, Kajang, Malaysia, July 18-22, 2011. Revised Selected Papers*, 226–233. 31, 34

Malin, F., Norros, I. & Innamaa, S. (2019). Accident risk of road and weather conditions on different road types. *Accident Analysis and Prevention*. 207

Malleson, N. & Andresen, M.A. (2016). Exploring the impact of ambient population measures on London crime hotspots. *Journal of Criminal Justice*. 41

Malleson, N., Evans, A. & Jenkins, T. (2009). An Agent-Based Model of Burglary:. *https://doi.org/10.1068/b35071*, **36**, 1103–1123. 96

Malleson, N., Heppenstall, A. & See, L. (2010). Crime reduction through simulation: An agent-based model of burglary. *Computers, Environment and Urban Systems*. 1, 41, 92, 96, 97, 100, 104, 210, 219

Malleson, N., See, L., Evans, A. & Heppenstall, A. (2012). Implementing comprehensive offender behaviour in a realistic agent-based model of burglary. *SIMULATION*, **88**, 50–71. 34, 43, 92, 97

Malleson, N., Heppenstall, A., See, L. & Evans, A. (2013). Using an agent-based crime simulation to predict the effects of urban regeneration on individual household burglary risk. *Environment and Planning B: Planning and Design*. 49

Malyshev, N.A. (2015). The Importance of Regulatory Policy. *SSRN Electronic Journal*. 142

Manson, S.M. (2006). Bounded rationality in agent-based models: Experiments with evolutionary programs. In *International Journal of Geographical Information Science*, vol. 20, 991–1012. 125

Maqbool, S.D., Imthias Ahamed, T.P. & Malik, N.H. (2011). Analysis of adaptability of Reinforcement Learning approach. In *2011 IEEE 14th International Multitopic Conference*, 45–49. 8, 65, 76

Markkula, G., Benderius, O., Wolff, K. & Wahde, M. (2012). A review of near-collision driver behavior models. In *Human Factors*, vol. 54, 1117–1143, SAGE PublicationsSage CA: Los Angeles, CA. 187

Martens, M. (1997). Deliverable D1 The Effects of Road Design on Speed Behaviour: A Literature Review Public MASTER PROJECT FUNDED BY THE EUROPEAN COMMISSION UNDER THE TRANSPORT RTD PROGRAMME OF THE 4th FRAMEWORK PROGRAMME The Effects of Road Design on Speed Behaviour: A Literature Review. Tech. rep., Master. 184

Martin, C., Schmitt, N. & Westerhoff, F. (2022). HOUSING MARKETS, EXPECTATION FORMATION AND INTEREST RATES. *Macroeconomic Dynamics*, **26**, 491–532. 142, 156, 162

Martín-Guerrero, J., Olivas, E., Martínez-Sober, M., Serrano-López, A., Magdalena, R. & Gómez-Sanchís, J. (2008). Use of Reinforcement Learning in Two Real Applications. 191–204. 111

Martinez, M.T. (2016). An Overview of Google's Machine Intelligence Software TensorFlow. Tech. rep., Sandia National Lab.(SNL-NM), Albuquerque, NM (United States). 4

Mataric, M. (2002). A review of multi-agent reinforcement learning: Independent vs. cooperative agents. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, **32**, 257–262. 27

Matthews, R.B., Gilbert, N.G., Roach, A., Polhill, J.G. & Gotts, N.M. (2007). Agent-based land-use models: A review of applications. 30

Maycock, G., Brocklebank, P. & Hall, R. (1999). Road layout design standards and driver behaviour. *Proceedings of the Institution of Civil Engineers: Transport*. 183, 199

McLane, A.J., Semeniuk, C., McDermid, G.J. & Marceau, D.J. (2011). The role of agent-based models in wildlife ecology and management. *Ecological Modelling*, **222**, 1544–1556. 186, 211

Melzer, B.T. (2010). Mortgage Debt Overhang: Reduced Investment by Homeowners with Negative Equity. 157

Miao, X. & Xu, D. (2007). MAS-based model framework of enterprise GDSS. In Q. Wang, G. Chen, G. Yan, X. Zhang, X.L. Jianguo, L. Huang & B. Wu, eds., *PROCEEDINGS OF THE 2007 CONFERENCE ON SYSTEMS SCIENCE, MANAGEMENT SCIENCE AND SYSTEM DYNAMICS: SUSTAINABLE DEVELOPMENT AND COMPLEX SYSTEMS, VOLS 1-10*, 1609–1615. 30

Midgley, D., Marks, R. & Kunchamwar, D. (2007). Building and assurance of agent-based models: An example and challenge to the field. *Journal of Business Research*, **60**, 884–893. 3

Miller, J. & Yu, L. (2008). On initial segment complexity and degrees of randomness. *Transactions of the American Mathematical Society*, **360**, 3193–3210. 3

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G. & et al. (2013). Human-level control through deep reinforcement learning. *Nature*, **518**, 529–533. 40

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S. & Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*. 4

Mnih, V., Badia, A.P., Mirza, L., Graves, A., Harley, T., Lillicrap, T.P., Silver, D. & Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. In *33rd International Conference on Machine Learning, ICML 2016*, 1–19. 51, 111

Moglia, M., Podkalicka, A. & McGregor, J. (2018). An Agent-Based Model of Residential Energy Efficiency Adoption. *Journal of Artificial Societies and Social Simulation*, **21**, 3. 35

Mohandas, B.K., Liscano, R. & Yang, O.W.W. (2009). Vehicle traffic congestion management in vehicular ad-hoc networks. In *2009 IEEE 34th Conference on Local Computer Networks*, 655–660. 200

Mongin, P. (1998). Expected utility theory. 31

Moravčík, E. & Jaśkiewicz, M. (2018). Boosting car safety in the EU. In *11th International Science and Technical Conference Automotive Safety, AUTOMOTIVE SAFETY 2018*, 1–5, Institute of Electrical and Electronics Engineers Inc. 184

Morescalchi, A., van Veldhuizen, S., Voogt, B. & Vogt, B. (2018). Negative home equity and job mobility. *Data-Driven Policy Impact Evaluation: How Access to Microdata is Transforming Policy Design*, 183–202. 157

Moriarty, P. & Wang, S.J. (2017). Can Electric Vehicles Deliver Energy and Carbon Reductions? *Energy Procedia*, **105**, 2983–2988. 222, 231

Munro, M. (2018). House price inflation in the news: a critical discourse analysis of newspaper coverage in the UK. *https://doi.org/10.1080/02673037.2017.1421911*, **33**, 1085–1105. 157

Murphy, A. (2020). Reported road casualties in Great Britain: 2019 annual report. Tech. rep., Department for Transport. 183, 205

Nadal, J.P., Gordon, M.B., Iglesias, J.R. & Semeshenko, V. (2010). Modelling the individual and collective dynamics of the propensity to offend. *European Journal of Applied Mathematics*, **21**, 421–440. 108

Napoletano, M., Guerci, E. & Hanaki, N. (2018). Recent advances in financial networks and agent-based model validation. 2

Narayan, P.K. & Narayan, S. (2011). The Importance of Real and Nominal Shocks on the UK Housing Market. *SSRN Electronic Journal*. 158

Nawrocki, P. & Sniezynski, B. (2018). Adaptive Service Management in Mobile Cloud Computing by Means of Supervised and Reinforcement Learning. *Journal of Network and Systems Management*. 54

Newell, A. (1994). *Unified theories of cognition*. Harvard University Press. 36

Nik, P.A., Jusoh, M.A., Shaari, A.H. & Sarmdi, T. (2016). Predicting the probability of financial crisis in emerging countries using an early warning system: Artificial neural network. *Journal of Economic Cooperation and Development*, **37**, 25–40. 142

Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology*, **53**, 139–154. 4, 14, 93, 127

Niv, Y., Joel, D., Meilijson, I. & Ruppin, E. (2002). Evolution of Reinforcement Learning in Uncertain Environments: A Simple Explanation for Complex Foraging Behaviors. *Adaptive Behavior*. 5

Noland, R.B. & Quddus, M.A. (2005). Congestion and safety: A spatial analysis of London. *Transportation Research Part A: Policy and Practice*, **39**, 737–754. 189, 204

Nolfi, S. (2004). Behaviour as a Complex Adaptive System: On the Role of Self-Organization in the Development of Individual and Collective Behaviour. *Complexus*, **2**, 195–203. 17

Nordfjærn, T., Jørgensen, S.H. & Rundmo, T. (2010). An investigation of driver attitudes and behaviour in rural and urban areas in Norway. *Safety Science*, **48**, 348–356. 188

Norwegian, T. (2009). A Heterodox Economic Analysis of the Housing Market Structure in Budapest Using Neural Network Classification. *Journal of Real Estate Literature*. 143

Nunes, T. (2019). Safe Reinforcement Learning Applications. 57

Oakland, T. & Harrison, P.L. (2008). Adaptive Behaviors and Skills: An Introduction. *Adaptive Behavior Assessment System-II*, 1–20. 15

O'Doherty, J.P., Lee, S.W. & McNamee, D. (2015). The structure of reinforcement-learning mechanisms in the human brain. *Current Opinion in Behavioral Sciences*, **1**, 94–100. 4

O'Donnell, C.J. & Connor, D.H. (1996). Predicting the severity of motor vehicle accident injuries using models of ordered multiple choice. *Accident Analysis and Prevention*. 206

Oei, L.H. & Polak, P. (1992). Effect of automatic warning and surveillance on speed and accidents: results of an evaluation study in four provinces. Tech. rep., Transportation Research Board. 185

Office for National Statistics (ONS) (2024). Housing affordability in England and Wales: 2023. xiv, 157

Oguibenine, B. (2011). Economic recession and mental health: an overview. *Neuropsychiatrie : Klinik, Diagnostik, Therapie und Rehabilitation : Organ der Gesellschaft Osterreichischer Nervenarzte und Psychiater*, **25 3**, 113–7. 7, 140

Oldford, R.W. (2016). Self-Calibrating Quantile–Quantile Plots. *http://dx.doi.org/10.1080/00031305.2015.1090338*, **70**, 74–90. 151

Olmez, S. (2022). SedarOlmez94/Pythonic_UK_Housing_Market_ABM_2022: UK Housing Market Model 2022. xiv, 144, 145, 151, 154, 163, 166

Olmez, S. & Heppenstall, A. (2021). Drive Cycle Data from the 3D Urban Traffic Simulator (ABM) in Unity (version 1.1.0). 234

Olmez, S., Douglas-Mann, L., Manley, E., Suchak, K., Heppenstall, A., Birks, D. & Whipp, A. (2021a). Exploring the Impact of Driver Adherence to Speed Limits and the Interdependence of Roadside Collisions in an Urban Environment: An Agent-Based Modelling Approach. *Applied Sciences*, **11**, 5336. xv, xviii, 49, 211, 213, 214, 216, 217, 218, 219

Olmez, S., Sargoni, O., Heppenstall, A., Birks, D., Whipp, A. & Manley, E. (2021b). 3D Urban Traffic Simulator (ABM) in Unity. xv, xviii, 185, 187, 190, 191, 192, 193, 194, 195, 197, 210, 212, 215, 217, 219, 221, 223, 224, 225, 227

Olmez, S., Birks, D. & Heppenstall, A. (2022a). Learning Complex Spatial Behaviours in ABM: An Experimental Observational Study. *Arxiv*. 155

Olmez, S., Thompson, J., Marfleet, E., Suchak, K., Heppenstall, A., Manley, E., Whipp, A. & Vidanaarachchi, R. (2022b). An Agent-Based Model of Heterogeneous Driver Behaviour and Its Impact on Energy Consumption and Costs in Urban Space. *Energies*, **15**. 140

Olner, D., Evans, A. & Heppenstall, A. (2015). An agent model of urban economics: Digging into emergence. *Computers, Environment and Urban Systems*. 211

Olsen, M.M. & Fraczkowski, R. (2015). Co-evolution in predator prey through reinforcement learning. *Journal of Computational Science*, **9**, 118–124. 5, 17, 76, 81

Orr, J. & Dutta, A. (2023). Multi-Agent Deep Reinforcement Learning for Multi-Robot Applications: A Survey. *Sensors (Basel, Switzerland)*, **23**. 57

Orsi, F. (2019). Centrally located yet close to nature: A prescriptive agent-based model for urban design. *Computers, Environment and Urban Systems*, **73**, 157–170. 1

O'sullivan, D., Millington, J., Perry, G. & Wainwright, J. (2012). Agent-based models-because they're worth it? In *Agent-Based Models of Geographical Systems*. 48

Outlander, M. (2021). *Running costs - Mitsubishi Outlander PHEV — Cut your costs*. 2021-09-30, https://www.mitsubishi-motors.ie/cars/outlander-phev/cost. 224

Page, S.E. (2015). What sociologists should know about complexity. *Annual Review of Sociology*, **41**, 21–41. 3

Palasz, B., Walus, K.J. & Wargula, L. (2019). The determination of the rolling resistance coefficient of a passenger vehicle with the use of roller test bench method. *MATEC Web of Conferences*, **254**, 04007. 235

Palmer, K., Tate, J.E., Wadud, Z. & Nellthorp, J. (2018). Total cost of ownership and market share for hybrid and electric vehicles in the UK, US and Japan. *Applied Energy*, **209**, 108–119. 211

Pammer, K., Freire, M., Gauld, C. & Towney, N. (2021). Keeping safe on australian roads: Overview of key determinants of risky driving, passenger injury and fatalities for indigenous populations. 183

Pantangi, S.S., Fountas, G., Sarwar, M.T., Anastasopoulos, P.C., Blatt, A., Majka, K., Pierowicz, J. & Mohan, S.B. (2019). A preliminary investigation of the effectiveness of high visibility enforcement programs using naturalistic driving study data: A grouped random parameters approach. *Analytic Methods in Accident Research*, **21**, 1–12. 188

Park, A.J. & Buckley, S. (2016). Three-Dimensional Agent-Based Model and Simulation of a Burglar's Target Selection. *Proceedings - 2015 European Intelligence and Security Informatics Conference, EISIC 2015*, 105–112. 6, 91, 100, 103

Park, D. & Ryu, D. (2021). A Machine Learning-Based Early Warning System for the Housing and Stock Markets. *IEEE Access*, **9**, 85566–85572. 143

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L. & others (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, **32**. 28

Payne, J.W., Bettman, J.R. & Johnson, E.J. (1988). Adaptive strategy selection in decision making. *Journal of experimental psychology: Learning, Memory, and Cognition*, **14**, 534. 37

Peden, M., Scurfield, R., Sleet, D., Hyder, A.A., Mathers, C., Jarawan, E., Hyder, A.A., Mohan, D. & Jarawan, E. (2004). *World report on road traffic injury prevention*. World Health Organization. 185

Peichl, A. (2016). Linking microsimulation and CGE models. *International Journal of Microsimulation*, **9**, 167–174. 4

Percy, D. (1989). The Process of Learning. *Adult Study Tactics*, 88–94. 14

Pere, P.P. (2017). The effect of pedestrianisation and bicycles on local business. Tech. rep., futureplaceleadership. 206

Petrol Prices, G. (2021). United Kingdom gasoline prices, 25-Oct-2021 — GlobalPetrolPrices.com. 228

Picascia, S. (2014). A theory driven , spatially explicit agent-based simulation to model the economic and social implications of urban regeneration. 6, 140

Piquero, A. & Rengert, G.F. (2006). Studying deterrence with active residential burglars. *http://dx.doi.org/10.1080/07418829900094211*, **16**, 451–450. 128

Places Catapult, C. (2019). Connected Places Catapult Market Forecast For Connected and Autonomous Vehicles. Tech. rep., Catapult Connecting Places. 232

Point, P. (2021). Cost of Charging an Electric Car — Pod Point. 228

Polydoros, A.S. & Nalpantidis, L. (2017). Survey of Model-Based Reinforcement Learning: Applications on Robotics. *Journal of Intelligent & Robotic Systems*, **86**, 153–173. 56

Poole, D.L. & Mackworth, A.K. (2010). *Artificial Intelligence: foundations of computational agents*. Cambridge University Press. 64

Popescu, I.V. (2014). Analysis of the Behavior of Central Banks in Setting Interest Rates. The Case of Central and Eastern European Countries. *Procedia Economics and Finance*, **15**, 1113–1121. 162

Porta, S., Crucitti, P. & Latora, V. (2006). The network analysis of urban streets: A dual approach. *Physica A: Statistical Mechanics and its Applications*. 195, 196, 217

Poyner, B. (1991). Situational crime prevention in two parking facilities. *Security Journal*, **2**, 96–101. 92

Prud'homme, R. & Koning, M. (2012). Electric vehicles: A tentative economic and environmental evaluation. *Transport Policy*, **23**, 60–69. 229

Prystawski, B., Mohnert, F., Tošić, M. & Lieder, F. (2021). Resource-rational Models of Human Goal Pursuit. *Topics in Cognitive Science*, **00**, 1–21. 8

Pusse, F. & Klusch, M. (2019). Hybrid online POMDP planning and deep reinforcement learning for safer self-driving cars. In *IEEE Intelligent Vehicles Symposium, Proceedings*, vol. 2019-June, 1013–1020, Institute of Electrical and Electronics Engineers Inc. 184

Qiu, L. & Nixon, W.A. (2008). Effects of adverse weather on traffic crashes: Systematic review and meta-analysis. *Transportation Research Record*. 207

Queeney, J., Paschalidis, I.C. & Cassandras, C.G. (2021). Generalized Proximal Policy Optimization with Sample Reuse. In *NeurIPS*, 11909–11919. 109

Quimby, A., Maycock, G., Palmer, C. & Buttress, S. (1999). The factors that influence a drivers choice of speed. *Journal of Safety Research*. 183, 199

Rahimiyan, M. & Mashhadi, H.R. (2010). An adaptive Q-Learning algorithm developed foragent-based computational modeling of electricity market. *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*. 81, 97

Ram, A. (2020). Braking energy calculation for a given drive cycle and different methods of regenerative braking. : Skill-Lync. 236

Ramchandani, P., Paich, M. & Rao, A. (2017). Incorporating learning into decision making in agent based models. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, **10423 LNAI**, 789–800. 2, 3, 8, 121

Ramey, V.A. (2016). Macroeconomic Shocks and Their Propagation. *Handbook of Macroeconomics*, **2**, 71–162. 140

Rao, A. & Georgeff, M. (1995). BDI Agents: From Theory to Practice. In *Proceedings of the First International Conference on Multi-Agent Systems (ICMAS-95)*. 30

Rawal, A., Rajagopalan, P. & Miikkulainen, R. (2010). Constructing competitive and cooperative agent behavior using coevolution. In *Proceedings of the 2010 IEEE Conference on Computational Intelligence and Games, CIG2010*, 107–114. 97

Reason, J., Manstead, A., Stephen, S., Baxter, J. & Campbell, K. (1990). Errors and violations on the roads: A real distinction? *Ergonomics*, **33**, 1315–1332. 188

Rengert, G. (2002). *The journey to crime*. 107, 118, 125, 128

Retallack, A.E. & Ostendorf, B. (2019). Current understanding of the effects of congestion on traffic accidents. *International journal of environmental research and public health*, **16**, 3400. 189, 200

Retallack, A.E. & Ostendorf, B. (2020). Relationship between traffic volume and accident frequency at intersections. *International journal of environmental research and public health*, **17**, 1393. 204, 205

Richards, S.H. & Dudek, C.L. (1986). IMPLEMENTATION OF WORK-ZONE SPEED CONTROL MEASURES. *Transportation Research Record*. 184

Rios, D.R. & Springer, D.A. (2017). Evolutionary optimization of recurrent neural network architectures for language generation. *IEEE Transactions on Evolutionary Computation*, **21**, 932–944. 24

Robins, G., Elliott, P. & Pattison, P. (2001). Network models for social selection processes. *Social networks*, **23**, 1–30. 32

Rosenfield, A., Chingcuanco, F. & Miller, E.J. (2013). Agent-based Housing Market Microsimulation for Integrated Land Use, Transportation, Environment Model System. *Procedia Computer Science*, **19**, 841–846. 141

Ruan, X., Li, Y., Zhou, X., Jin, Z. & Yin, Z. (2020). Simulation method of concrete chloride ingress with mesoscopic cellular automata. *Construction and Building Materials*, **249**, 118778. 4

Ruder, S. (2016). An overview of gradient descent optimization algorithms. 51

Russel, S. & Norvig, P. (2012). *Artificial intelligence—a modern approach 3rd Edition*. 39

Sajan, B., Mishra, V.N., Kanga, S., Meraj, G., Singh, S.K. & Kumar, P. (2022). Cellular automata-based artificial neural network model for assessing past, present, and future land use/land cover dynamics. *Agronomy*, **12**, 2772. 5

Sali, R., Adewole, S. & Akakpo, A. (2021). Feature Selection Using Reinforcement Learning. *ArXiv*, **abs/2101.09460**. 143

Salim, F.D., Loke, S.W., Rakotonirainy, A., Srinivasan, B. & Krishnaswamy, S. (2007). Collision pattern modeling and Real-Time collision detection at road intersections. In *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*, 161–166. 187

Sampson, R.J. (1988). Local friendship ties and community attachment in mass society: A multilevel systemic model. *American sociological review*, 766–779. 94

Samuel, A.L. (1959). Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, **3**, 210–229. 19

Sarlioglu, B., Morris, C.T., Han, D. & Li, S. (2017). Driving Toward Accessibility: A Review of Technological Improvements for Electric Machines, Power Electronics, and Batteries for Electric and Hybrid Vehicles. *IEEE Industry Applications Magazine*, **23**, 14–25. 209

Saunier, N., Sayed, T. & Ismail, K. (2010). Large-scale automated analysis of vehicle interactions and collisions. *Transportation Research Record*. 185

Savarimuthu, B.T.R. & Cranefield, S. (2011). Norm creation, spreading and emergence: A survey of simulation models of norms in multi-agent systems. *Multiagent and Grid Systems*, **7**, 21–54. 34

Schakner, Z. & Blumstein, D.T. (2016). Learning and conservation behavior: an introduction and overview. *Conservation Behavior*, 66–92. 14

Schelling, T.C. (1969). Models of Segregation. *American Economic Review*. 5

Schelling, T.C. (1971). Dynamic models of segregation. *The Journal of Mathematical Sociology*. 31

Schmidt, B. & Schneider, B. (2004). Agent-based modelling of human acting, deciding and behaviour-the reference model PECS. In *Proceedings of the European Simulation Multiconference*. 34

Schulman, J., Levine, S., Moritz, P., Jordan, M.I. & Abbeel, P. (2015). Trust Region Policy Optimization. *32nd International Conference on Machine Learning, ICML 2015*, **3**, 1889–1897. 3, 20, 51, 110, 111

Schulman, J., Wolski, F., Dhariwal, P., Radford, A. & Klimov, O. (2017). Proximal Policy Optimization Algorithms. 19, 49, 50, 51, 52, 53, 54, 98, 109, 110, 111, 135

Schulze, J., Müller, B., Groeneveld, J. & Grimm, V. (2017). Agent-based modelling of social-ecological systems: Achievements, challenges, and a way forward. *JASSS*. 154

Secchi, D. (2015). A case for agent-based models in organizational behavior and team research. *Team Performance Management*, **21**, 37–50. 1, 140

Sellén, M. (2022). How much does a Car Weigh? [Average by Car Model & Type]. 213

Sert, E., Bar-Yam, Y. & Morales, A.J. (2020). Segregation dynamics with reinforcement learning and agent based modeling. *Scientific Reports*. 5, 49, 53, 91, 97, 108, 219

Shafiei, E., Leaver, J., Production, B.D.J.o.C. & 2017, u. (2017). Cost-effectiveness analysis of inducing green vehicles to achieve deep reductions in greenhouse gas emissions in New Zealand. *Elsevier*. 228

Shankar, Z.P., Gaikwad, M.M., Bhor, V. & Jain, N. (2020). HOUSING PRICE PREDICTION USING MACHINE LEARNING. 143

Shefer, D. (1994). Congestion, air pollution, and road fatalities in urban areas. *Accident Analysis & Prevention*, **26**, 501–509. 189, 200

Short, M.B., D'Orsogna, M.R., Pasour, V.B., Tita, G.E., Brantingham, P.J., Bertozzi, A.L. & Chayes, L.B. (2011). A STATISTICAL MODEL OF CRIMINAL BEHAVIOR. *https://doi.org/10.1142/S0218202508003029*, **18**, 1249–1267. 128

Sierla, S., Ihasalo, H. & Vyatkin, V. (2022). A Review of Reinforcement Learning Applications to Control of Heating, Ventilation and Air Conditioning Systems. *Energies*. 56

Sierzchula, W., Bakker, S., Maat, K. & Van Wee, B. (2012). The competitive environment of electric vehicles: An analysis of prototype and production models. *Environmental Innovation and Societal Transitions*, **2**, 49–65. 209

Sigurdsson, J.F., Gudjonsson, G.H. & Peersen, M. (2008). Differences in the cognitive ability and personality of desisters and re-offenders: A prospective study among young offenders. *https://doi.org/10.1080/10683160108401781*, **7**, 33–43. 6, 127

Silva, P.C.L., Batista, P.V.C., Lima, H.S., Alves, M.A., Guimarães, F.G. & Silva, R.C.P. (2020). COVID-ABS: An agent-based model of COVID-19 epidemic to simulate health and economic effects of social distancing interventions. *Chaos, Solitons & Fractals*, **139**, 110088. 1

Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D. & Riedmiller, M. (2014). Deterministic Policy Gradient Algorithms. In E.P. Xing & T. Jebara, eds., *Proceedings of the 31st International Conference on Machine Learning*, vol. 32 of *Proceedings of Machine Learning Research*, 387–395, PMLR, Bejing, China. 162

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T.P., Simonyan, K. & Hassabis, D. (2017). Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm. *ArXiv*, **abs/1712.01815**. 5, 24

Simon, H.A. (1955). A behavioral model of rational choice. *The quarterly journal of economics*, **69**, 99–118. 31, 37

Simon, H.A. (1957). Models of man; social and rational. 37

Simon, H.A. (1962). The architecture of complexity. *Proceedings of the American philosophical society*, **106**, 467–482. 3

Singh, S.P., Kearns, M.J. & Mansour, Y. (2004). Emergence of complex behaviour from simple rules in a multi-agent reinforcement learning model. *International Journal of Human-Computer Studies*, **61**, 69–101. 27

Sledge, I.J. & Principe, J.C. (2017). Balancing exploration and exploitation in reinforcement learning using a value of information criterion. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2816–2820. 162

Son, Y.J., Kim, S., Xi, H. & Mungle, S. (2013). An extended BDI model for human behaviors: decision-making, learning, interactions, and applications. In *2013 Winter Simulations Conference (WSC)*, 401–411. 32

Spatharis, C., Blekas, K., Bastas, A., Kravaris, T. & Vouros, G.A. (2019). Collaborative multiagent reinforcement learning schemes for air traffic management. In *10th International Conference on Information, Intelligence, Systems and Applications, IISA 2019*. 49, 54, 76, 143

Spectrum, N.G.I. & 2016, u. (2016). Can you program ethics into a self-driving car? *ieeexplore.ieee.org*. 184

Squazzoni, F. (2012). *Agent-Based Computational Sociology*. John Wiley & Sons. 1, 211

Sternberg, R.J. & Gastel, J. (1989). Coping with novelty in human intelligence: An empirical investigation. *Intelligence*, **13**, 187–197. 121

Stevenson, R.J. & Forsythe, L. (1998). *The stolen goods market in New South Wales: An interview study with imprisoned burglars*. NSW Bureau Crime Statistics and Research. 94

Stokes, N. & Clare, J. (2019). Preventing near-repeat residential burglary through cocooning: post hoc evaluation of a targeted police-led pilot intervention. *Security Journal*, **32**, 45–62. 125, 129

Su, M., Liu, Z., Chen, Y., Li, G. & Shou, Z. (2017). Communication-Based Multi-Agent Reinforcement Learning. *IEEE Transactions on Cybernetics*, **48**, 2795–2807. 26

Subramanian, A., Chitlangia, S. & Baths, V. (2022). Reinforcement learning and its connections with neuroscience and psychology. *Neural Networks*, **145**, 271–287. 4

Sullivan, J.F. & Atchison, G.J. (1978). Predator-prey behaviour of fathead minnows, Pimephales promelas and largemouth bass, Micropterus salmoides in a model ecosystem. *Journal of Fish Biology*, **13**, 249–253. 59

Sun, C. & Wang, S. (2008). Modeling adaptive behaviors on growing social networks. *Proceedings - 4th International Conference on Natural Computation, ICNC 2008*, **1**, 465–469. 15

Sun, R. (2006). The CLARION cognitive architecture: Extending cognitive modeling to social simulation. *Cognition and multi-agent interaction*, 79–99. 38

Sun, R. (2007). The importance of cognitive architectures: An analysis based on CLARION. *Journal of Experimental and Theoretical Artificial Intelligence*. 35

Survey, O. (2021). OS Open Roads. xiv, 197

Sutton, R. & Barto, A. (2018a). *Reinforcement learning: An introduction*. MIT Press. 3

Sutton, R.S. & Barto, A.G. (2018b). *Reinforcement Learning: An Introduction, Second Edition*. 14, 18, 19, 23, 24, 27, 28, 39, 40, 49, 51, 53, 91, 93, 108, 111, 127, 142, 154, 155, 160, 162, 177

Suzuki, M., Nakatani, R. & Nishikawa, I. (2014). A Mechanism Design of Solar Power Trading by Autonomous Agent based on Reinforcement Learning. *Frontiers in Artificial Intelligence and Applications*, **262**, 392–401. 143

Szita, I. & Lorincz, A. (2006). Learning tetris using the noisy cross-entropy method. *Neural Computation*, **18**, 2936–2941. 51, 111

Szumska, E.M., Świętokrzyska, P., Frej, D., Szumska, E. & Grabski, P. (2020). Analysis of the Causes of Vehicle Accidents in Poland. *LOGI-Scientific Journal on Transport and Logistics*, **11**. 183

Talebpour, A. & Mahmassani, H.S. (2016). Influence of connected and autonomous vehicles on traffic flow stability and throughput. *Transportation Research Part C: Emerging Technologies*, **71**, 143–163. 233

Tan, X., Zeng, Y., Gu, B., Wang, Y. & Xu, B. (2018). Scenario Analysis of Urban Road Transportation Energy Demand and GHG Emissions in China—A Case Study for Chongqing. *Sustainability 2018, Vol. 10, Page 2033*, **10**, 2033. 209

Tang, W. & Bennett, D.A. (2010). Agent-based modeling of animal movement: A review. 140

Tang, Y., Zhu, J., Wu, D. & Chen, Y. (2017). Multi-agent reinforcement learning in water resource management. *IEEE Access*, **5**, 15916–15923. 25

Taylor, M.H., Stone, P., Littman, M.L. & Liu, Y. (2009). Challenges of Real-World Reinforcement Learning. In *International Conference on Machine Learning and Applications*, 1–6. 28

Taylor, R.B. & Gottfredson, S. (2015). Environmental Design, Crime, and Prevention: An Examination of Community Dynamics. *https://doi.org/10.1086/449128*, **8**, 387–416. 95

Tellidou, A. & Bakirtzis, A. (2006). A Q-learning agent-based model for the analysis of the power market dynamics. *Proceedings of the 6. IASTED international conference on European power and energy systems*, **2006**, 228–233. 81

Tesfatsion, L. (2006). Agent-Based Computational Economics: A Constructive Approach to Economic Theory. Tech. Rep. 527, Society for Computational Economics. 30

Thapa, D., Jung, I.S. & Wang, G.N. (2005). Agent based decision support system using reinforcement learning under emergency circumstances. In *Lecture Notes in Computer Science*. 81

Thomasen, F. (2018). ALL-NEW FORD FIESTA ST SPECIFICATIONS PERFORMANCE AND ECONOMY. Tech. rep., Ford. xviii, 226, 227

Thompson, J., Read, G.J., Wijnands, J.S. & Salmon, P.M. (2020a). The perils of perfect performance; considering the effects of introducing autonomous vehicles on rates of car vs cyclist conflict. *https://doi.org/10.1080/00140139.2020.1739326*, **63**, 981–996. 211

Thompson, J., Stevenson, M., Wijnands, J.S., Nice, K.A., Aschwanden, G.D., Silver, J., Nieuwenhuijsen, M., Rayner, P., Schofield, R., Hariharan, R. & Morrison, C.N. (2020b). A global analysis of urban design types and road transport injury: an image processing study. *The Lancet Planetary Health*, **4**, e32–e42. 217

Thompson, J.H., Wijnands, J.S., Mavoa, S., Scully, K. & Stevenson, M.R. (2019). Evidence for the 'safety in density' effect for cyclists: validation of agent-based modelling results. *Injury Prevention*, **25**, 379–385. 219

Thompson, W.R. (1982). The Single-Sample Problem of Detection, Learning, and Prediction. *IEEE Transactions on Information Theory*, **IT-28**, 147–152. 19

Thorndike (1911). *Animal intelligence; experimental studies* . New York, The Macmillan Company, 1911. 4

Thorndike, E.L. (1927). The Law of Effect. *The American Journal of Psychology*, **39**, 212. 4

Tian, G. & Qiao, Z. (2014). Modeling urban expansion policy scenarios using an agent-based approach for Guangzhou Metropolitan Region of China. *Ecology and Society*, **19**. 45

Tieleman, S. (2022). Towards a Validation Methodology for Macroeconomic Agent-Based Models. *Computational Economics*, **60**. 2

Tillyer, M.S., Wilcox, P. & Fissel, E.R. (2018). Violence in Schools: Repeat Victimization, Low Self-Control, and the Mitigating Influence of School Efficacy. *Journal of Quantitative Criminology*, **34**, 609–632. 125, 129

Tobler, W.R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic geography*, **46**, 234–240. 148

Todd, A., Beling, P., Scherer, W. & Yang, S.Y. (2017). Agent-based financial markets: A review of the methodology and domain. *2016 IEEE Symposium Series on Computational Intelligence, SSCI 2016*. 140

Toledo, T. (2007). Driving Behaviour: Models and Challenges. *http://dx.doi.org/10.1080/01441640600823940*, **27**, 65–84. 17

Tolk, A. (2019). Limitations and Usefulness of Computer Simulations for Complex Adaptive Systems Research. 77–96. 18

Tong, C.K., On, C.K., Teo, J. & Kiring, A.M.J. (2011). Evolving Neural Controllers Using GA for Warcraft 3-Real Time Strategy Game. In *2011 Sixth International Conference on Bio-Inspired Computing: Theories and Applications*, 15–20. 24

Topalli, V. (2005). CRIMINAL EXPERTISE AND OFFENDER DECISION-MAKING: An Experimental Analysis of How Offenders and Non-Offenders Differentially Perceive Social Stimuli. *The British Journal of Criminology*, **45**, 269–295. 6, 95, 127

Torrens, P.M. (2010). Geography and computational social science. *GeoJournal*, **75**, 133–148. 3

Troitzsch, K.G. (2017). Using empirical data for designing, calibrating and validating simulation models. *Advances in Intelligent Systems and Computing*, **528**, 413–427. 17

Tse, C.B., Rodgers, T. & Niklewski, J. (2014). The 2007 financial crisis and the UK residential housing market: Did the relationship between interest rates and house prices change? *Economic Modelling*, **37**, 518–530. 157, 158

Urban, C. & Schmidt, B. (2001). PECS – Agent-Based Modelling of Human Behaviour. *Operations Research*. 43, 96

Valadkhani, A., Nguyen, J. & O'Brien, M. (2019). Asymmetric responses of house prices to changes in the mortgage interest rate: evidence from the Australian capital cities. *https://doi.org/10.1080/00036846.2019.1619026*, **51**, 5781–5792. 162

Van Houten, R., Rolider, A., Nau, P.A., Friedman, R., Becker, M., Chalodovsky, I. & Scherer, M. (1985). Large-scale reductions in speeding and accidents in Canada and Israel: a behavioral ecological perspective. *Journal of Applied Behavior Analysis*. 206

Vandeviver, C., Neutens, T., van Daele, S., Geurts, D. & Vander Beken, T. (2015). A discrete spatial choice model of burglary target selection at the house-level. *Applied Geography*, **64**, 24–34. 128

Vickrey, W. (1968). Automobile Accidents, Tort Law, Externalities, and Insurance: An Economist's Critique. *Law and Contemporary Problems*. 204

Vickrey, W.S. (1969). Congestion Theory and Transport Investment. *The American Economic Review*, **59**, 251–260. 200

Vinyals, O., Babuschkin, I., Czarnecki, W.M., Mathieu, M., Dudzik, A., Chung, J., Choi, D.H., Powell, R., Ewalds, T., Georgiev, P., Oh, J., Horgan, D., Kroiss, M., Danihelka, I., Huang, A., Sifre, L., Cai, T., Agapiou, J., Jaderberg, M., Vezhnevets, A., Leblond, R., Pohlen, T., Dalibard, V., Budden, D., Sulsky, Y., Molloy, J., Paine, T., Gulcehre, C., Wang, Z., Pfaff, T., Wu, Y., Ring, R., Yogatama, D., Wünsch, D., McKinney, K., Smith, O., Schaul, T., Lillicrap, T., Kavukcuoglu, K., Hassabis, D., Apps, C. & Silver, D. (2019). Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, **575**, 350–354. 57

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra & Martin Riedmiller (2016). Playing Atari with Deep Reinforcement Learning. 142

Von Neumann, J. & Morgenstern, O. (1944). *The Theory of Games and Economic Behavior*. Princeton University Press. 19

von Neumann, J. & Morgenstern, O. (1947). *Theory of games and economic behavior*. Princeton University Press, Princeton, NJ. 31

Wager, G., Whale, J. & Braunl, T. (2016). Driving electric vehicles at highway speeds: The effect of higher driving speeds on energy consumption and driving range for electric vehicles in Australia. *Renewable and Sustainable Energy Reviews*, **63**, 158–165. 222, 231

Wakefield, A.J., Murch, S.H., Anthony, A., Linnell, J., Casson, D.M., Malik, M., Berelowitz, M., Dhillon, A.P., Thomson, M.A., Harvey, P. & et al. (1998). Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. *The Lancet*, **351**, 637–641. 111

Walsh, M.M. & Anderson, J.R. (2014). Navigating complex decision spaces: Problems and paradigms in sequential choice. *Psychological Bulletin*, **140**, 466–486. 14

Wan, J., Yuan, Y. & Wang, Q. (2017). Traffic congestion analysis: A new Perspective. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1398–1402. 200

Wang, C., Quddus, M.A. & Ison, S.G. (2009). Impact of traffic congestion on road accidents: A spatial analysis of the M25 motorway in England. *Accident Analysis and Prevention*. 189, 204

Wang, C., Quddus, M. & Ison, S. (2013). A spatio-temporal analysis of the impact of congestion on traffic safety on major roads in the UK. *Transportmetrica A: Transport Science*, **9**, 124–148. 189

Wang, J., Wen, L., Zhang, Y. & Zhang, J. (2020). Multi-Agent Deep Reinforcement Learning for Coordinated Charging of Electric Vehicles. *IEEE Access*, **8**, 164176–164187. 26

Wang, K. & Ke, Y. (2018). Public-Private Partnerships in the Electric Vehicle Charging Infrastructure in China: An Illustrative Case Study. *Advances in Civil Engineering*, **2018**. 209

Wang, S., Wu, Y. & Li, X. (2019a). An intelligent traffic control model based on multi-agent reinforcement learning. *Transportation Research Part C: Emerging Technologies*, **102**, 1–15. 25

Wang, S., Wu, Y. & Li, X. (2019b). Virtual water resource management model based on multi-agent reinforcement learning. In *Proceedings of the 2019 International Conference on Computational Science and Computational Intelligence*, 945–950. 25

Wang, Y. & de Silva, C.W. (2008). A machine-learning approach to multi-robot coordination. *Engineering Applications of Artificial Intelligence*, **21**, 470–484. 25

Wang, Z., Bapst, V., Heess, N., Mnih, V., Munos, R., Kavukcuoglu, K. & de Freitas, N. (2016). Sample Efficient Actor-Critic with Experience Replay. *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*. 20, 51, 111

Wang, Z., Fan, Y. & Chen, H. (2018). Deep reinforcement learning for autonomous driving. *IEEE Access*, **6**, 834–878. 25

Watkins, C.J.C.H. (1989). Learning from delayed rewards. 40

Weisburd, D., Maher, L., Sherman, L., Buerger, M., Cohn, E. & Petrosino, A. (1993). Contrasting crime general and crime specific theory: The case of hot spots of crime. In *Advances in Criminological Theory*, 45–70. 107, 116, 125, 128

Westerhoff, F. (2010). A simple agent-based financial market model: Direct interactions and comparisons of trading profits. *Nonlinear Dynamics in Economics, Finance and Social Sciences: Essays in Honour of John Barkley Rosser Jr*, 313–332. 140

White, M. (2015). Cyclical and structural change in the UK housing market. *Journal of European Real Estate Research*, **8**, 85–103. 158

Whitehead, C. & Williams, P. (2011). Causes and Consequences? Exploring the Shape and Direction of the Housing System in the UK Post the Financial Crisis. *https://doi.org/10.1080/02673037.2011.618974*, **26**, 1157–1169. 156

Wiering, M.A. & Van Otterlo, M. (2012). Reinforcement learning. *Adaptation, learning, and optimization*, **12**, 729. 91, 93

Wikström, P.O.H. (2006). Individuals, settings, and acts of crime: Situational mechanisms and the explanation of crime. *The explanation of crime: Context, mechanisms and development*, 61–107. 94

Windrum, P., Fagiolo, G. & Moneta, A. (2007). Empirical validation of agent-based models: Alternatives and prospects. *JASSS*. 17

Wong, B.B. & Candolin, U. (2015). Behavioral responses to changing environments. *Behavioral Ecology*, **26**, 665–673. 121

Wooldridge, M. (2002). Introduction to MultiAgent Systems. *Information Retrieval*. 30, 32

Wooldridge, M. (2009). *An Introduction to MultiAgent Systems*. Wiley Publishing, 2nd edn. 39

Wooldridge, M. (2020). *The road to Conscious Machines*. Pelican Books, 1st edn. 49, 108

Wooldridge, M. & Jennings, N.R. (1995). Intelligent agents: Theory and practice. *The Knowledge Engineering Review*. 32

Wortley, R. (2001). A Classification of Techniques for Controlling Situational Precipitators of Crime. *Security Journal*. 92

Wortley, R. (2016). Situational precipitators of crime. In *Environmental Criminology and Crime Analysis: Second Edition*, 81–105. 129

Xiang, X. & Foo, S. (2021). Recent Advances in Deep Reinforcement Learning Applications for Solving Partially Observable Markov Decision Processes (POMDP) Problems: Part 1 - Fundamentals and Applications in Games, Robotics and Natural Language Processing. *Mach. Learn. Knowl. Extr.*, **3**, 554–581. 56

Xing, W., Qian, Y., Guan, X., Yang, T. & Wu, H. (2020a). A novel cellular automata model integrated with deep learning for dynamic spatio-temporal land use change simulation. *Computers & Geosciences*, **137**, 104430. 5

Xing, Y., Lv, C. & Cao, D. (2020b). Driver Behavior Recognition in Driver Intention Inference Systems. *Advanced Driver Intention Inference*, 99–134. 210

Xiong Yingfei (2009). A language-based approach to model synchronization in software engineering. 153

Yaich, R., Boissier, O., Jaillon, P. & Picard, G. (2011). Social-compliance in trust management within virtual communities. In *2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, vol. 3, 322–325. 32

Yamada, K., Takano, S. & Watanabe, S. (2011). Reinforcement learning approaches for acquiring conflict avoidance behaviors in multi-agent systems. *2011 IEEE/SICE International Symposium on System Integration, SII 2011*, 679–684. 17

Yamaguchi, S., Naoki, H., Ikeda, M., Tsukada, Y., Nakano, S., Mori, I. & Ishii, S. (2018). Identification of animal behavioral strategies by inverse reinforcement learning. *PLOS Computational Biology*, **14**, e1006122. 143

Yang, W.H., Hall, S.J. & McNicol, G. (2021). Global gases. *Principles and Applications of Soil Microbiology*, 557–579. 209

Yen, C.H. & Lu, H.P. (2008). Effects of e-service quality on loyalty intention: an empirical study in online auction. *Managing Service Quality: An International Journal*, **18**, 127–146. 32

Young, J. (2004). Voodoo criminology and the numbers game. *Critical criminology*, **12**, 225–250. 94

Youssef, A., Missiry, S.E., El-gaafary, I.N., ElMosalami, J.S., Awad, K.M. & Yasser, K. (2019). Building your kingdom Imitation Learning for a Custom Gameplay Using Unity ML-agents. In *2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, 509–514. 55

Yu, T., Davis, L., Baydar, C. & Roy, R. (2008). *Evolutionary computation in practice*, vol. 88. Springer. 18, 30

Yun, T.S. & Moon, I.C. (2020). Housing Market Agent-Based Simulation with Loan-To-Value and Debt-To-Income. *2019:122:3*, **23**, 1–19. 141

Zeng, T. (2018). Learning Continuous Control through Proximal Policy Optimization for Mobile Robot Navigation. *International Conference on Future Technology and Disruptive Innovation*, 175–184. 53

Zhan, C., Wu, Z., Liu, Y., Xie, Z. & Chen, W. (2020). Housing prices prediction with deep learning: An application for the real estate market in Taiwan. *IEEE International Conference on Industrial Informatics (INDIN)*, **2020-July**, 719–724. 143

Zhan-quan, W. (2006). Reinforcement Learning Theory,Algorithms and Application. *Journal of Hebei University of Technology*. 23, 142

Zhang, C., Vinyals, O., Munos, R. & Bengio, S. (2018). A Study on Overfitting in Deep Reinforcement Learning. *ArXiv*. 54

Zhang, H. & Li, Y. (2014). Agent-based simulation of the search behavior in China's resale housing market: Evidence from Beijing. *JASSS*, **17**. 6

Zhang, H. & McCord, E.S. (2014). A spatial analysis of the impact of housing foreclosures on residential burglary. *Applied Geography*, **54**, 27–34. 98

Zhang, H. & Song, W. (2014). Addressing issues of spatial spillover effects and non-stationarity in analysis of residential burglary crime. *GeoJournal*, **79**, 89–102. 98

Zhang, Z., Li, Z., Wang, P. & Chen, Y. (2019). A multi-agent reinforcement learning approach to modeling human decision making in dynamic environments. *arXiv preprint arXiv:1910.08722*. 26

Zhdankin, V. & Sprott, J.C. (2010). Simple predator-prey swarming model. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*. 17, 59

Zhou, M. & Sisiopiku, V.P. (1997). Relationship between volume-to-capacity ratios and accident rates. *Transportation Research Record*. 204

Zhou, X., Tong, W. & Li, D. (2019). Modeling Housing Rent in the Atlanta Metropolitan Area Using Textual Information and Deep Learning. *ISPRS International Journal of Geo-Information 2019, Vol. 8, Page 349*, **8**, 349. 143

Zhou, Y., Wu, J., Long, C., Cheng, M. & Zhang, C. (2017). Performance Evaluation of Peer-to-Peer Energy Sharing Models. *Energy Procedia*, **143**, 817–822. 143

Zhou, Y., Jiang, X., Fu, C. & Liu, H. (2021). Operational factor analysis of the aggressive taxi speeders using random parameters Bayesian LASSO modeling approach. *Accident Analysis and Prevention*, **157**, 106183. 188

Zhu, H. & Wang, F. (2021). An agent-based model for simulating urban crime with improved daily routines. *Computers, Environment and Urban Systems*, **89**, 101680. 43

Zhu, K. & Zhang, T. (2021). Deep reinforcement learning based mobile robot navigation: A review. *Tsinghua Science and Technology*, **26**, 674–691. 18

Zhuge, C. & Shao, C. (2018). Agent-based modelling of purchasing, renting and investing behaviour in dynamic housing markets. *Journal of computational science*, **27**, 130–146. 1, 44