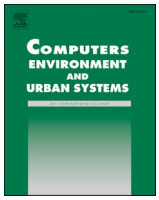
Computers, Environment and Urban Systems 112 (2024) 102141

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus



Learning the rational choice perspective: A reinforcement learning approach to simulating offender behaviours in criminological agent-based models☆

Sedar Olmez [a,b,*], Dan Birks [b,c], Alison Heppenstall [d], Jiaqi Ge [a]

[a] School of Geography, University of Leeds, Seminary St, Woodhouse, Leeds LS2 9JT, United Kingdom
[b] The Alan Turing Institute, 2QR, John Dodson House, 96 Euston Rd, London NW1 2DB, United Kingdom
[c] School of Law, University of Leeds, Belle Vue Rd, Woodhouse, Leeds LS2 9JT, United Kingdom
[d] School of Social & Political Sciences, University of Glasgow, Adam Smith Building, Bute Gardens, Glasgow G12 8RT, United Kingdom

ARTICLE INFO

Keywords:
Agent-based model
Reinforcement learning
Environmental criminology
Rational choice perspective
Decision-making

ABSTRACT

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 article 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.

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, Bar-Yam, & Morales, 2020; Sutton & Barto, 2018; Wiering & Van Otterlo, 2012).

RL algorithms can learn advanced problems to solve using neural networks (Islam, Chen, & Jin, 2019), such as playing complex games and defeating human players (Justesen, Bontrager, Togelius, & Risi, 2020). In health research, RL was used to develop treatment plans for patients (Jalalimanesh, Shahabi Haghighi, Ahmadi, & Soltani, 2017). In social science, researchers 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 article 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

☆ This document is the result of research funded by the Economic and Social Research Council (ESRC), grant numbers: ES/P000401/1 and ES/R007918/1, UK Prevention Research Partnership (UKPRP) MR/S037578/2, Medical Research Council MC_UU_00022/5 and Scottish Government Chief Scientist Office SPHSU20

* Corresponding author at: School of Geography, University of Leeds, Seminary St, Woodhouse, Leeds LS2 9JT, United Kingdom.
E-mail address: solmez@turing.ac.uk (S. Olmez).

https://doi.org/10.1016/j.compenvurbsys.2024.102141
Received 16 August 2022; Received in revised form 7 March 2024; Accepted 11 June 2024
Available online 27 June 2024

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, Guerette, & Bowers, 2014; Park & Buckley, 2016).

Rational choice perspective (RCP) (CORNISH and CLARKE 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, Townsley, & Stewart, 2012; Clarke, 1997a; Cornish & Clarke, 2003; Hayward, 2007; Wortley, 2001). RCP underpins situational crime prevention intervention (SCPI) (CLARKE 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, Heppenstall, & See, 2010) or probability of offence (Birks et al., 2012)) which define conditions where offences take place.

Additionally, recent advancements in environmental criminology, particularly the focus on Crime Prevention Through Environmental Design (CPTED), have expanded the perspectives on crime prevention strategies. Cozens (2013) provides valuable insights into CPTED strategies, emphasizing the role of environmental design in preventing crime. While historically rooted in SCPI, acknowledging the contemporary emphasis on CPTED adds a nuanced dimension to the discussion on crime prevention.

There are many studies where environmental criminologists adopt ABMs (Birks et al., 2012; Bosse & Gerritsen, 2008; Gerritsen, 2015; Gialopsos & Carter, 2014; Groff, 2007; Joubert, Saprykin, Chokani, & Abhari, 2022; Malleson et al., 2010; Malleson, See, Evans, & Heppenstall, 2012). Most models propose traditional condition-action rules to respond to situations preventing agents from learning and adapting to change. This article 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 and CLARKE Cornish & Clarke, 1987).

Using this model, in this article, 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 article begins by providing a brief literature review of related work 2 of ABMs in environmental criminology. Methodology 3 section describes the model logic, including conceptualised theories. Results 4 section describes a series of experiments run using the model, and the outcomes of those experiments are presented. Lastly, discussion and conclusion 5 of findings, drawbacks and contributions to environmental criminology are presented.

2. Related work

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, Crooks, See, & Batty, 2012). Most criminological applications of ABMs have been applied within the field of environmental criminology due to their spatial modelling capabilities (Groff, Johnson, & Thornton, 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, 2018; Wiering & Van Otterlo, 2012). This article presents RL as a means to contribute to behavioural decision-making in these models. The following section describes some prominent applications of ABM in environmental criminology.

2.1. ABMs in environmental criminology

According to Johnson and 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 and CLARKE 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.

A central tenet of the Rational Choice Perspective (RCP) is the concept of bounded rationality. Bounded rationality posits that decision-making is constrained based on the knowledge that an individual possesses; thus, risk/reward is relative to what is known. This perspective aligns with the RL component of our study, engaging the idea that an individual’s ability to make rational decisions can be altered as they learn from past/prior experience. In other words, an evaluation of risks/rewards becomes more accurate as rational decision-making becomes more informed.

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.

Significant early contributions to agent-based crime simulation were made by Gutiérrez, Orozco-Aguirre, and Landassuri-Moreno (2013), who developed a framework to analyse the interrelated social and individual-level factors of crime events and opportunities. Their model was supported by established criminological literature and validated against macro-level crime patterns, though it faced limitations in replicating the nuanced individual decision-making processes of offenders. Troitzsch (2017) discussed the dynamics of Extortion Racket Systems (ERSs) using ABM, offering insights into criminal system parameters that influence normative behaviour. However, their model’s ability to generalise beyond the specific context of Southern Italy’s provinces may be limited. In the realm of organised crime, Nardin, Székely, and Andrighetto (2017) introduced GLODERS-S, a simulator for protection racketeering groups that adopted an event-based approach. While innovative, the model’s configurability might pose challenges in accurately capturing the complex, emergent behaviour of criminal networks. Lastly, Devia and Weber (2013) presented an agent-based model

for generating artificial street-crime data, which proved beneficial in evaluating policing strategies. The model was validated in fictitious and real cities, although the simplifications necessary for such simulations may overlook significant social complexities.

Bosse and 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, Wasek, and Franz (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 and CLARKE 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, Evans, & Jenkins, 2009). Some models have addressed challenges of testing criminological theory through ABM, these include (Birks et al., 2012; Bosse & Gerritsen, 2008; Bosse, Gerritsen, Hoogendoorn, Jaffry, & Treur, 2011; Caskey et al., 2018).

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, Rajagopalan, & Miikkulainen, 2010). Consequently, a gap in behavioural representation in crime models has emerged. Cornelius, Lynch, Modelling and Gore 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; Malleson et al., 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 article 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 novel 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.

3. Agent-based model (Methodology)

In this section, our proposed ABM is described. Including the chosen RL algorithm proximal policy optimisation (PPO) (Schulman, Wolski, Dhariwal, Radford, & Klimov, 2017).

Our ABM simulates dynamics of residential burglary events. This particular offence was chosen as examples of burglary models are vast (Birks et al., 2012; Malleson et al., 2010, 2012). Furthermore, several studies have looked at spatio-temporal properties of residential burglary (Johnson et al., 2007; Zhang & McCord, 2014; Zhang & Song, 2014); thus, the features of residential burglary that may allow appropriate model validation are well documented.

Fig. 1 depicts the model and its key components: offender agents, static targets, interventions, nodes, routine activity nodes and two spatial localities. These components were inspired by the following articles (Birks et al., 2012; Malleson et al., 2010; Park & Buckley, 2016).

3.1. Model environment

The model applies situational interventions (CLARKE 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 defined as some abstract property protection by increased guardianship (such as CCTV) which increases the risk of victimising the property.

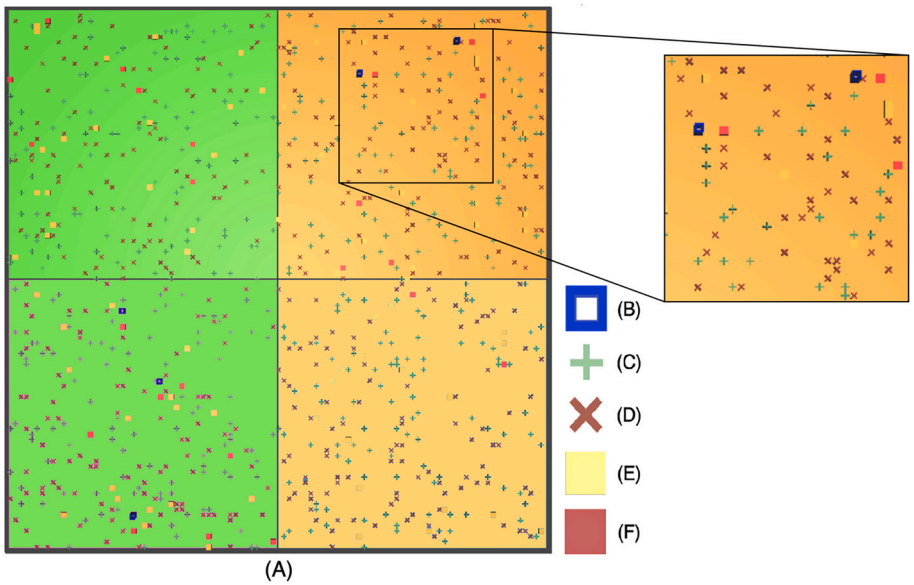


Fig. 1. Example model environment, where (A) is the environment on a 100 × 100 grid which includes five offender agents (B), 100 targets (C) and interventions (D) at each of the two spatial localities, 100 nodes (E) of which, 23 are routine activity nodes (F), where each offender agent has five assigned nodes (the same node can be assigned to two or more offender agents).

3.1.1. Interventions

Interventions in the model represent physical protection of a property and are implemented as static objects in the environment (Fig. 1, 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 Clarke, 1980).

3.1.2. Targets

Targets (represent residential properties) are static, like interventions, they are randomly distributed (Fig. 1, C). The number of targets per locality can be specified. In keeping with RCP (CORNISH and CLARKE 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 1 for each offender agent. A formal example of the logic behind *Reward*, *Risk* and *Effort* can be found in Appendix A.1.

$$Target\_Attractiveness(T_i) = T_i(Reward) - (T_i(Effort) + T_i(Risk)) \quad (1)$$

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, Brantingham, & Taylor, 2006; Clarke, 1997b).

3.1.3. Nodes

In order to represent a rudimentary transport network where agents can navigate their environment, the model utilises navigation nodes (following (Birks et al., 2012; Park & Buckley, 2016)). Nodes are randomly distributed across the environment in unoccupied *cells* (Fig. 1, 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 Fig. 1 A, we see 100 nodes distributed, where 23 of these are routine activity nodes.

3.1.4. Routine activity nodes

Routine activity nodes were inspired by RAT (Cohen & Felson, 1979), these are a subset of navigation nodes (Fig. 1, 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).

3.2. Offender agents

Agents within our model are known as offenders, Fig. 1, B. These agents navigate the environment, perceive surroundings and make decisions influenced by spatial stimuli inferred by RL (sub-section 3.3). 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 A).

3.2.1. 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 Fig. 2. A formal definition of movement can be found in a.2

3.2.2. 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 A.3 for a formal definition of the described process.

3.2.3. 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

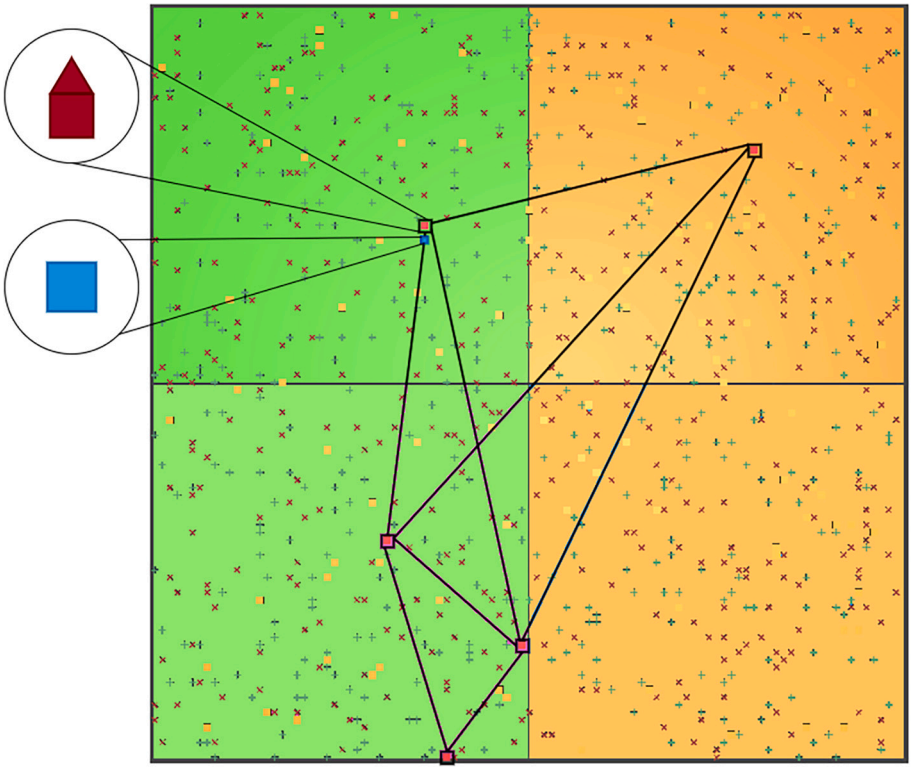


Fig. 2. Example Model Environment: a single offender agent A navigating from its home node to a routine activity node i where $RAN_i \in Offender_A(RAN) - Offender_A(RAN_H)$.

in which offences occur and vice-versa. Post-training, offender agents apply learned knowledge about the environment and make decisions. In Fig. 3, offender agents have ten individual sensors. A formal definition of this process can be found in a.4

3.2.4. 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).

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, Gordon, Iglesias, & Semeshenko, 2010) - with some offenders offending a lot more than others.

3.3. RL for decision-making

RL agents learn by interacting with their dynamic environment. At each timestep, agents perceive some state of the environment, apply an action; this causes the environment to transition into a new state. A reward signal (i.e., target attractiveness) evaluates the quality of each transition, and the agents have to maximise their cumulative reward during interaction (Buşoniu, Babuška, & De Schutter, 2010; Islam et al., 2019; Kaelbling, Littman, & Moore, 1996; Sutton & Barto, 2018; Wooldridge, 2020). Many examples of RL applied to produce behaviourally realistic agents exist; some of these can be found in (Dang, 2020; Joubert et al., 2022; Liu et al., 2020; Sert et al., 2020).

RL is split into two parts, training and testing. During training (Fig. 4, A), agents are initialised, some state s of the environment is observed, and an initial action is taken. Objective function (Formula B.2) 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 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, Paschalidis, & Cassandras, 2021; Schulman et al., 2017). These underpinning formulae can be found in Appendix B, Formulas B.1 and B.2. A simple illustration of these formulae is found below.

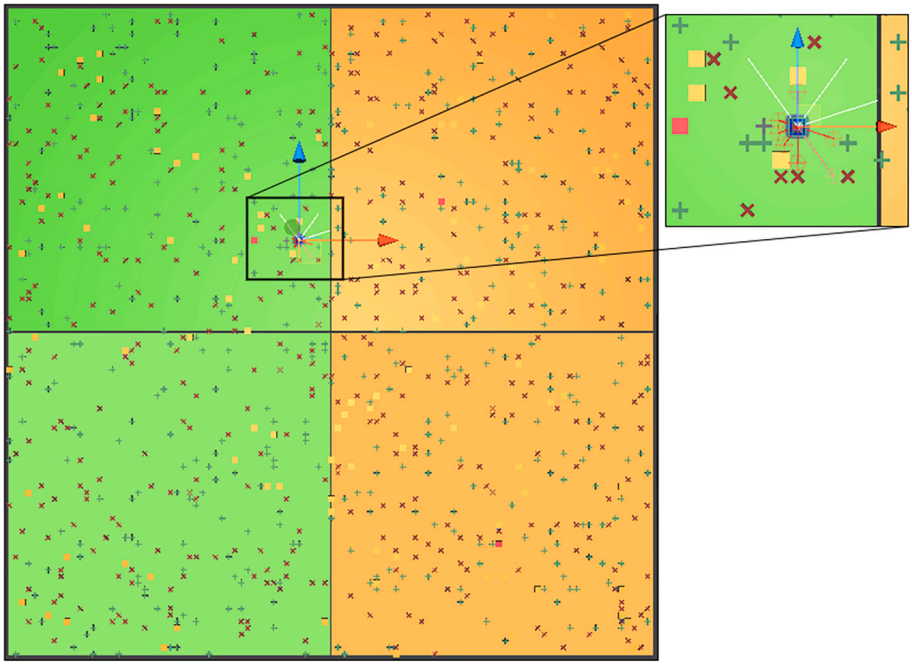


Fig. 3. Example Model Environment: offender agent perception, a sensor can be either red or white. The former means object identified, and the latter means no object. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

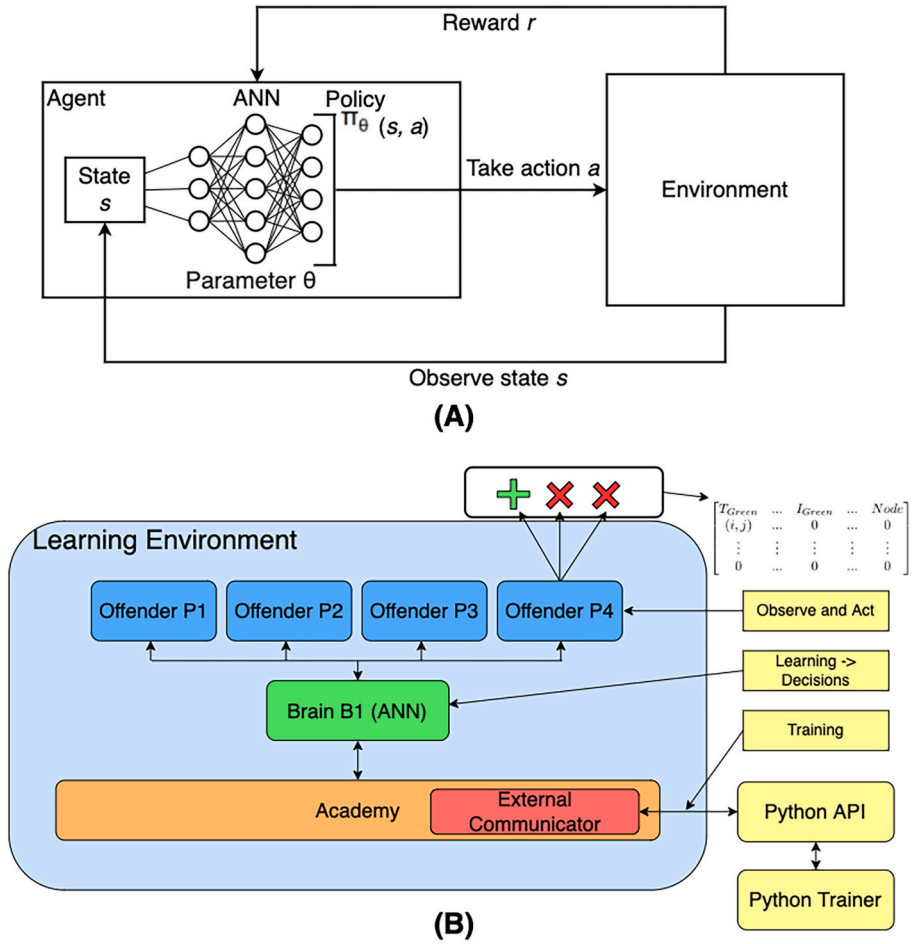


Fig. 4. Block diagrams, where (A): basic reinforcement learning ANN training architecture. (B): RL life cycle in ml-agents package (Juliani et al., 2018).

Empirical expectation $\widehat{\mathbb{E}}_t$ 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 less likely for the current than the old policy. $\widehat{\mathbb{A}}_t$ is the advantage estimate; the higher the value, the better the agent's current actions are than the actions it started with. Thus, Formula B.1, also known as the conservative policy iteration, tries to prevent substantial policy updates, which can cause unstable training outcomes. The L^{CLIP} Formula B.2, known as the clipped surrogate objective, clips the distribution change of the policy ratio between 0.8 and 1.2, where $\varepsilon = 0.2$ (Schulman et al., 2017). It takes the minimum of the current policy ratio $r_t(\theta)\widehat{A}_t$ 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, Levine, Moritz, Jordan, & Abbeel, 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 function mapping an agent's observations to the best action an agent can take in a given state (Fig. C.1). Each offender agent makes observations given localised information (sub-section 3.2.3). 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, 2018), Fig. 4, B.

PPO was chosen for this article as it was proven to outperform its rivals in many tested environments; these algorithms include 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). Similarly, it approaches optimal policy structure (Vanvuchelen, Gijsbrechts, & Boute, 2020) and utilises efficient data utilisation and good parallelism (Chu, 2018). Furthermore, ml-agents (Juliani et al., 2018) encourages simplicity in implementing these algorithms, which subsequently promotes reproducibility and explainability for model behaviours. Lastly, its presence in literature is vast with over 8032 citations (Google Scholar) as of 03/08/2022.

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 article 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 (Cornish & Clarke, 2017; CORNISH and CLARKE Cornish & Clarke, 1987; Clarke & Cornish, 1985) 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 3 are setup as treatment (left) and buffer (right) areas.

Experimental conditions are outlined in Table 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 repeated simulations** were operated for each experiment condition.
- In total, **500 episodes** per experiment condition were analysed.
- Model environment is made up of **100 × 100 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 Clarke, 1980; Clarke, 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.

Output data (Table 2) from each experiment condition (Table 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 Fig. 5 and summary statistics in Table 3 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:

- 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.

Table 1
Model experiment parameters for three conditions.

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

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

Table 2
Model output data, type and description.

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.

4.1. Experiment 1 - uniform distribution of rewards (1)

In this experiment, *Rewards* at each target is uniform set to 1, therefore, $0 \leq Target\_Attractiveness \leq 1$ pre-intervention and $-1 \leq Target\_Attractiveness \leq 1$ 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 (Fig. 6, A). However, once interventions are introduced and *Risk* increases (treatment area), offences concentrate in the buffer area (Fig. 6, B). Small pockets of offences in the treatment area still occur. In Fig. C.2a 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 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 Fig. C.2a, B.

These results alone do not prove offender agents behave in ways that

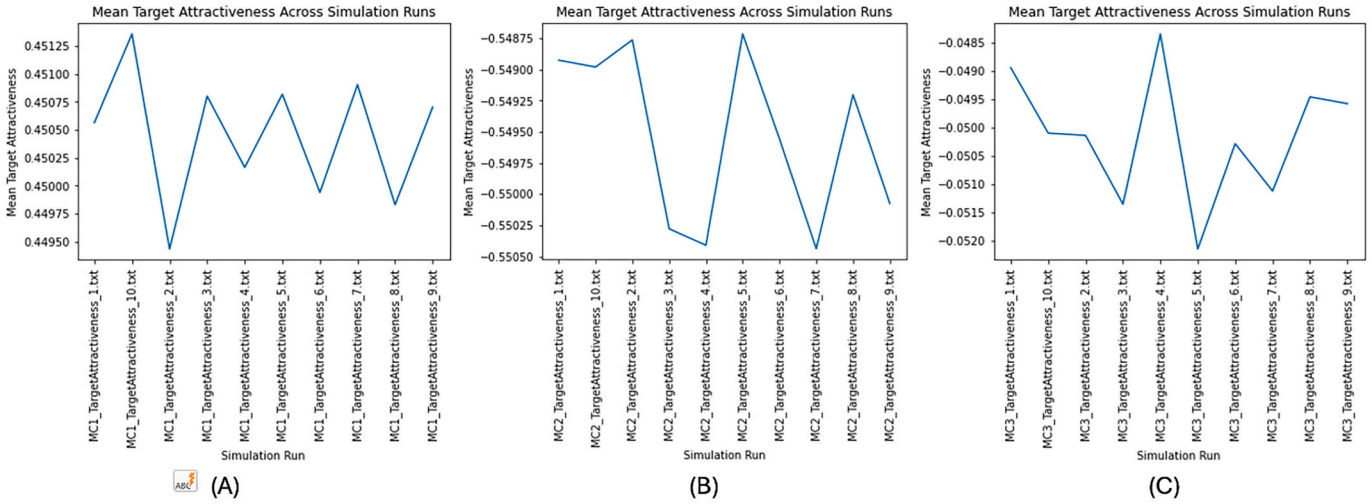


Fig. 5. Variation of mean across 10 batch runs, where (A)-(C) are Experiments 1 to 3 respectively.

Table 3
Statistics of variability across simulation runs per experiment condition.

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.

would be characterised by RCP. The target attractiveness at each target over time must be quantified and compared to validate behaviours. In Fig. 6, D we observe a concentration of positive target attractiveness pre-intervention. If we compare these patterns to Fig. 6, A, offences clustered at the centre and decreased as we branch out. When comparing Figs. 6, 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 (Figs. 6, 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.

These results show offender agents successfully learned to offend at targets where rewards outweighed risk and effort, evidenced in Fig. 7. This is true for both pre and post-intervention; however, post-intervention accumulated rewards drastically dropped as high rewarding opportunities decreased.

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, Braithwaite, & Johnson, 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. Fig. 6, 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).

4.2. *Experiment 2 - uniform distribution of rewards (0)*

In this experiment, rewards are set to 0, therefore, $-1 \leq$ *Target_Attractiveness* ≤ 0 pre-intervention and $-2 \leq$ *Target_Attractiveness* ≤ 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 (Figs. 8, 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 Fig. C.2b. 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 Fig. C.2b, 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.

Figs. 8, 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 Figs. 8, 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 Fig. 9 indicates that the accumulated reward is near zero for every offender agent with an average total cumulative reward of −0.38.

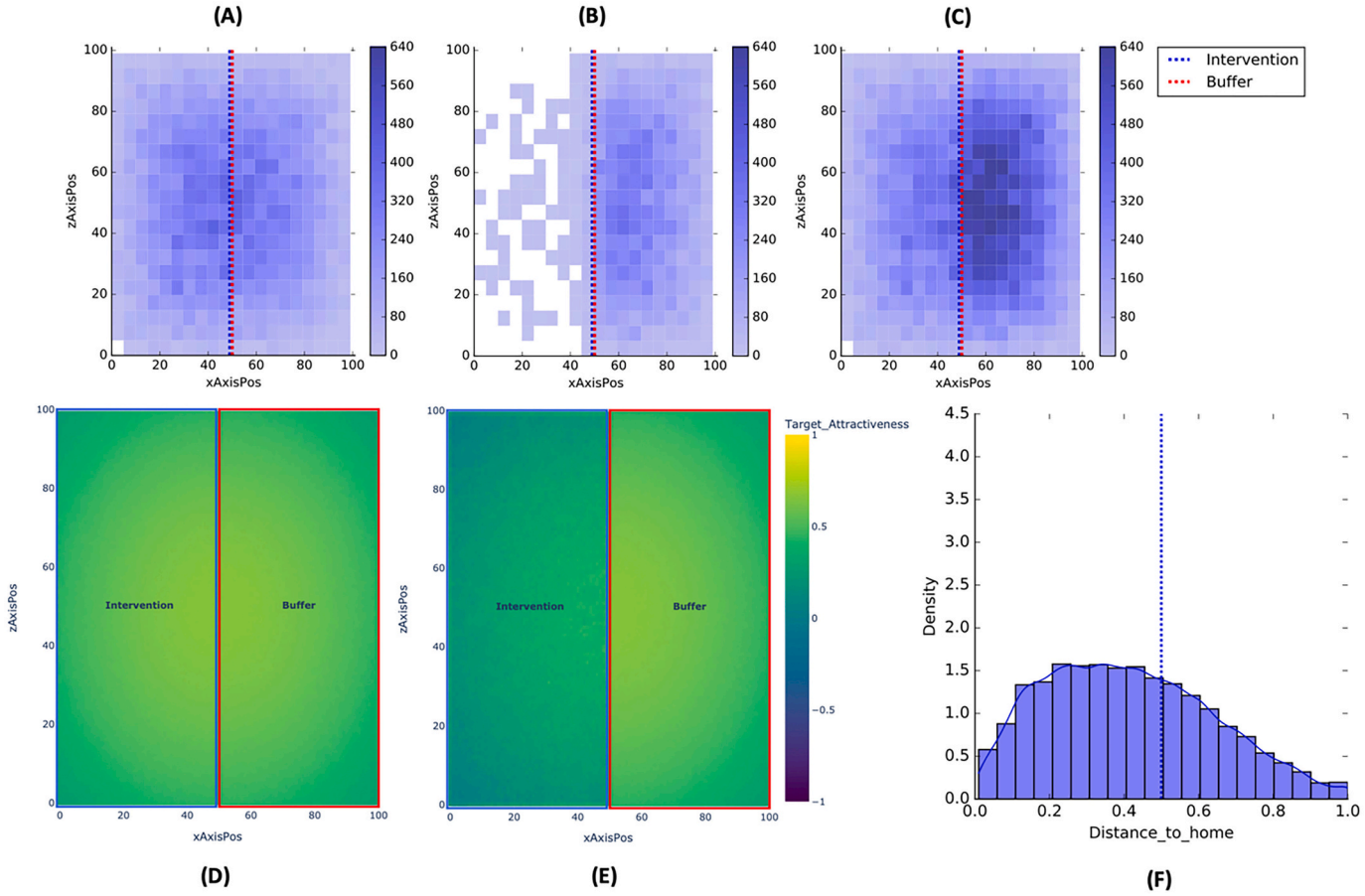


Fig. 6. Distributions of offences, average target attractiveness and distance between home and offence locations, where (A-C): offences across targets pre, post and pre-post merged. (D-E): target attractiveness pre and post-intervention, respectively. (F): distance between home and offence locations (Intervention = Treatment area).

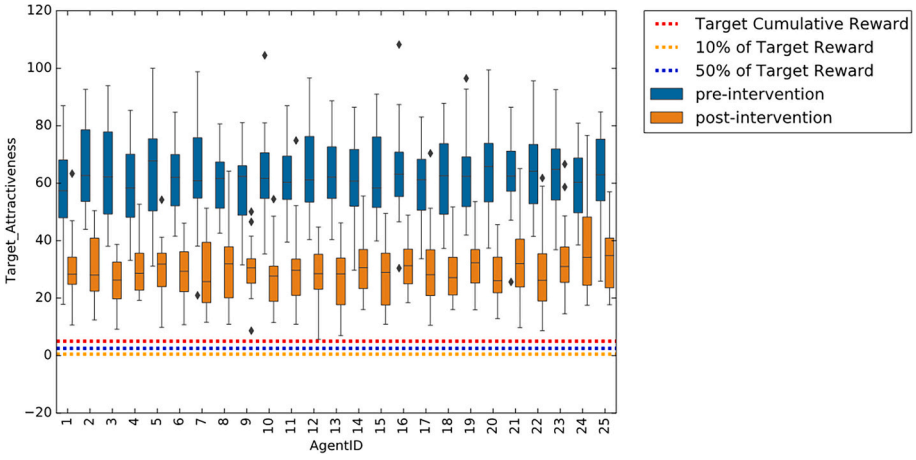


Fig. 7. The cumulative reward distribution for each offender agent pre-post intervention across all episodes.

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 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, Paich, & Rao, 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 article, 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 novel situations (previously unseen situations) which is an essential aspect of human cognition (Jipp,

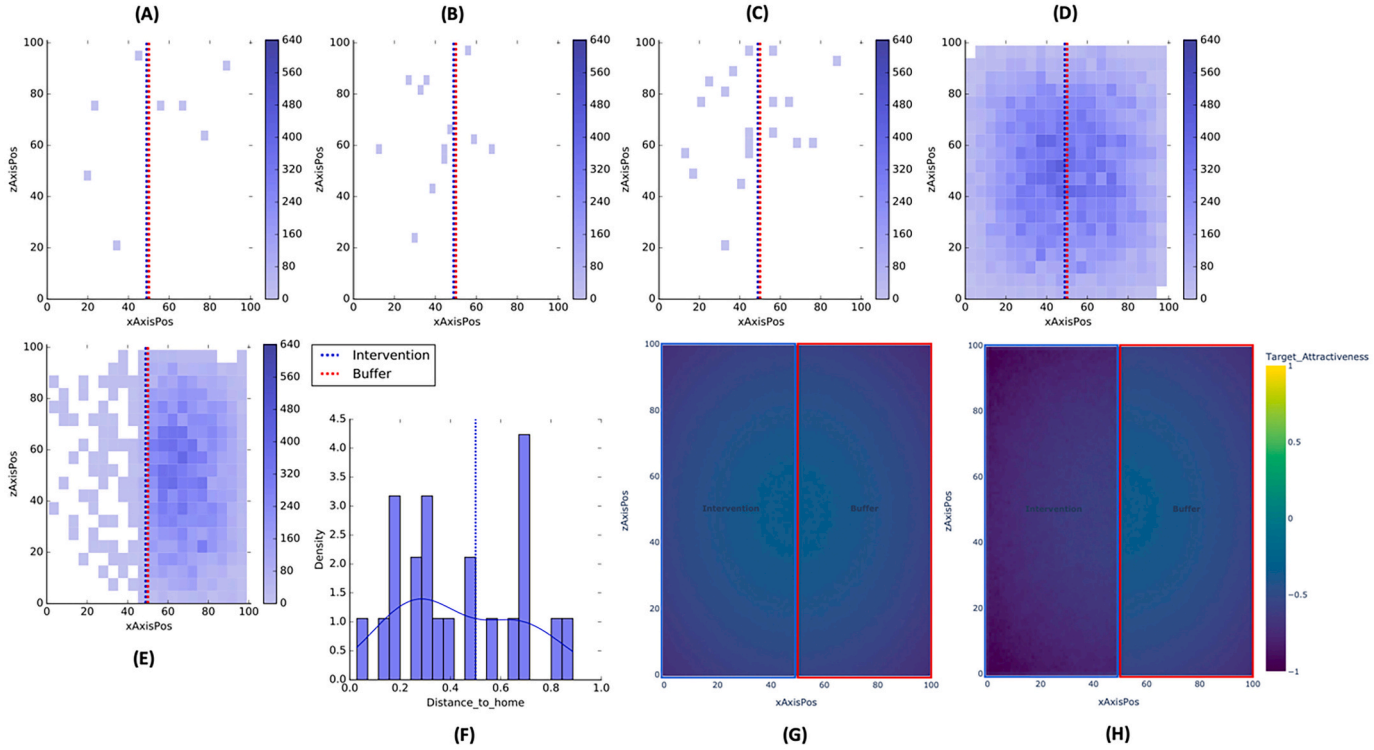


Fig. 8. Distributions of offence and not to offend target locations, the distance between home and offence locations and average Target_Attractiveness, where (A-C): offences across targets pre, post and pre-post intervention merged. (D-E): no offence decisions across targets pre and post-intervention, respectively. (F): distance between home and offence locations. (G-H): Target_Attractiveness pre and post-intervention, respectively (Intervention = Treatment area).

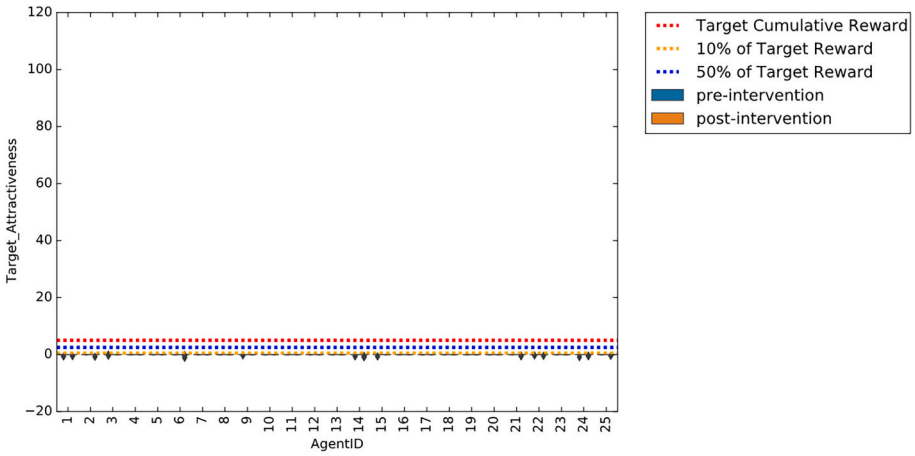


Fig. 9. The cumulative reward distribution for each offender agent pre-post intervention across all episodes.

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.

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

In the final experiment, the reward at each target is randomly distributed between 0 and 1 inclusive. Therefore, $-1 \leq$ *Target_Attractiveness* ≤ 1 pre-intervention and $-2 \leq$ *Target_Attractiveness* ≤ 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.

Figs. 10, A-C show that spatial patterns are similar to those in MC1 Figs. 6, 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 Fig. 10, 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 Fig. 10, D, focusing on more rewarding surrounding locations.

For reference, the average target attractiveness across episodes,

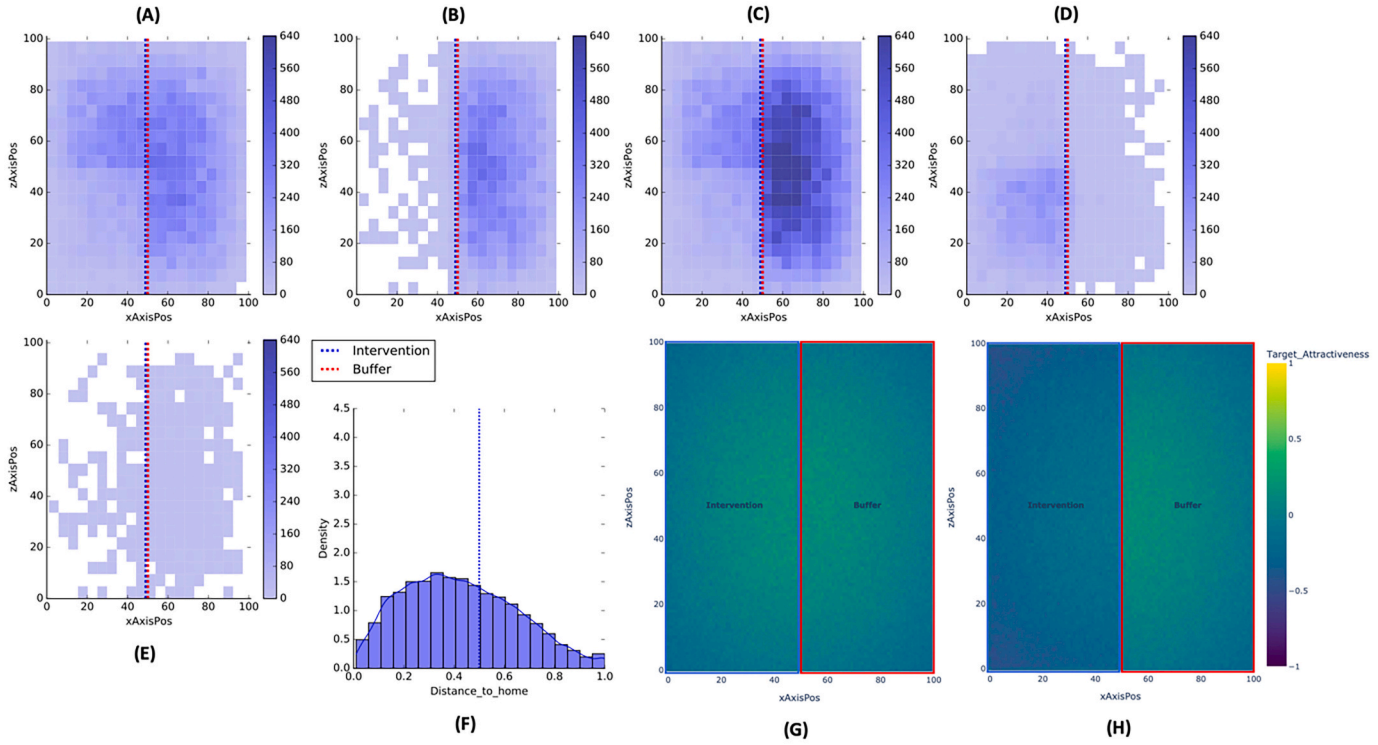


Fig. 10. Distributions of offence and not to offend target locations, the distance between home and offence locations and average Target_Attractiveness, where (A-C): offences across targets pre, post and pre-post intervention merged. (D-E): no offence decisions across targets pre and post-intervention, respectively. (F): distance between home and offence locations. (G-H): Target_Attractiveness pre and post-intervention respectively (Intervention = Treatment area).

Figs. 10, G-H when compared with Figs. 10, 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 Fig. 11 compared to MC1, Fig. 7.

These results show offender agents have acquired an average total cumulative reward greater than TCR pre-intervention Fig. 11. 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 Fig. C.2c, A. The proportion of offences per offender agent in the treatment area pre-intervention is dissimilar to post-intervention Fig. C.2c, B. Some offender agents offend substantially more post-intervention than they did pre-intervention and vice-versa. When SCPIs are adopted, offender agent decision-making becomes more heterogeneous.

These results demonstrate that RL as a behavioural framework can incorporate intelligent 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

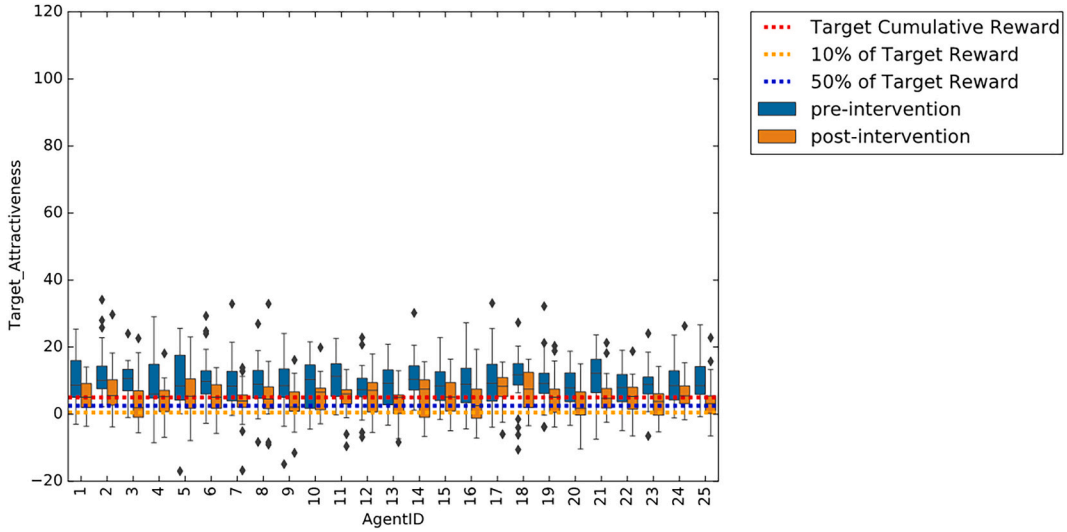


Fig. 11. The cumulative reward distribution for each offender agent pre-post intervention across all episodes.

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 agents living closer to or in the treatment area commit fewer offences than those living elsewhere, as observed in Table 4. Consequently, some offender agents can be "spatially better suited" to offending than others, evidenced by their net gains in Fig. 11 and proportion of offences in Fig. C.2c. Some offender agents consistently maintained high crime levels in the treatment area across all post-intervention episodes Figs. C.2c, B.

Overall, JTC (Rengert, 2002) is positively skewed, where offender agents learnt to minimise effort by offending closer to home Figs. 6, 8, 10, F. Crime Concentration (Farrell, 2015; Weisburd et al., 1993) is clustered to areas maintaining greater target attractiveness such as the centre and buffer areas Figs. 6, A-C, 6, D-E, 10, A-C and 10, G-H. When no rewards exist, decision to not offend is vast and clustered in the centre Figs. 8, D-E (similar empirical patterns of SCPI found in (Eck & Clarke, 2019)).

Crime does not spatially concentrate when rewarding opportunities are non-existent Figs. 8, A and G.

Most importantly, our intelligent 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 Fig. 12 which was also observed in the following articles (Stokes & Clare, 2019; Tillyer, Wilcox, & Fissel, 2018), journey to crime is typically short (positively skewed), and lastly, non-static patterns of crime over time.

5. Discussion and conclusion

This article has introduced an Agent-Based Model (ABM) to examine the spatio-temporal dynamics of burglary, addressing behavioural limitations in existing ABMs within environmental criminology (Groff et al., 2019; Johnson & Groff, 2014). Our model supports enhanced simulation of offender behaviour and crime events, addressing the previously overlooked aspect of behavioural learning in crime dynamics (Johnson & Groff, 2014). Traditional models often relied on fixed behaviours, leading to unrealistic offender actions (Arthur, 1994; Cornelius, Lynch, Modeling, & Gore, 2024; Manson, 2006), lacking the ability to adapt to changes—a critical aspect of real-world offender decision-making (Gialopsos & Carter, 2014; Sigurdsson, Gudjonsson, & Peersen, 2008; Topalli, 2005). To address these gaps, our research objectives were:

1. Assess if RL-based offender agents' behaviours align with Rational Choice Perspective (RCP), learning to offend when rewards surpass risks and efforts.
2. Investigate RL-based offender agents' adaptation to environmental changes following crime prevention interventions.
3. Examine if the simulated crimes by RL-utilising agents reflect empirical crime patterns.

Table 4
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).

Model Condition	Buffer Area	Treatment Area	Difference(+/−)	t-test(p)[1]
1	(711, 214.11)	(574, 188.91)	− 137	2.40(0.02 $p < 0.05$)
2[2]	(0, 0.0)	(0, 0.0)	–	–
3	(669, 196.86)	(615, 155.70)	− 54	1.08(0.28 $p > 0.05$)

[1] H_0: that two independent samples have identical average (expected) values. H_a: the means of the distributions underlying the samples are unequal.
[2] No offences occurred in the buffer or treatment areas by local offenders for MC2.

By adopting a multi-agent RL framework, we integrated neurologically inspired decision-making with ABMs in environmental criminology. This integration allows offenders to learn and adapt to Situational Crime Prevention Interventions (SCPIs), aiming to achieve specific goals like wealth accumulation (Niv, 2009; Sutton & Barto, 2018). Despite critiques on large-scale RL application (Joubert et al., 2022), our experiments demonstrated that offenders could learn to select targets based on the RCP, with SCPIs significantly altering their behaviour (Eck & Clarke, 2019; CLARKE Clarke, 1980; CORNISH and CLARKE Cornish & Clarke, 1987).

Our findings confirm that an RL-based model, grounded in the principles of RAT, RCP, and CPT (Brantingham & Brantingham, 2019; Cohen & Felson, 1979; CORNISH and CLARKE Cornish & Clarke, 1987), can produce outcomes resonating with empirical crime patterns, including spatial crime concentration and journey to crime (Brantingham & Brantingham, 1995; Rengert, 2002; Weisburd et al., 1993). The model reveals that RL offender agents, with partial knowledge and employing Artificial Neural Networks (ANNs), make diverse decisions reflecting RCP, generating crime patterns observed in empirical burglary studies (Eck & Liu, 2004; Piquero & Rengert, 2006; Short et al., 2011; Vandeviver, Neutens, van Daele, Geurts, & Vander Beken, 2015).

Traditional ABMs in this field often lack the dynamic learning and adaptation mechanisms our model incorporates. Our agents demonstrated adaptability to spatial changes and developed behaviours that align with the dynamic introduction of SCPIs, adjusting their actions based on reward availability (Bernasco & Luykx, 2003; Eck & Clarke, 2019; Levy, Santhakumaran, & Whitecross, 2014; Short et al., 2011).

Our experiments suggest that reducing opportunities (making target attractiveness ≤ 0) is more effective in lowering crime rates than SCPIs alone, as observed in Table 5. Despite computational limitations (Ding & Dong, 2020; Faghri, 2021; Farkas, Kertesz, & Lovas, 2020), this research paves the way for future studies on SCPIs' impacts on real-world offending patterns, including crime displacement and diffusion of benefits (GUERETTE and BOWERS Guerette & Bowers, 2009; Barr & Pease, 1990; Wortley, 2016).

In conclusion, our model demonstrates that RL can produce agents whose behaviours are both theoretically sound and empirically valid. This approach opens new avenues for research and provides insights into crime dynamics, supporting environmental criminologists and crime reduction efforts.

6. Open-source model access

The agent-based model can be found at the following source with documentation https://zenodo.org/records/6722701, accessed on 01/08/2022. The datasets and jupyter notebook used for analyses in this article can be found at the following link: https://figshare.com/articles/dataset/Agent-Based_Reinforcement_Learning_Model_of_Burglary_ARLMB_datasets_for_article_Learning_the_Rational_Choice_Perspective_A_Reinforcement_Learning_Approach_to_Simulating_Offender_Behaviours_in_Criminological_Agent-Based_Models/20418735, accessed on 02/08/2022.

CRediT authorship contribution statement

Sedar Olmez: Writing – original draft, Validation, Software, Project administration, Methodology, Investigation, Data curation, Conceptualization, Writing – review & editing. **Dan Birks:** Project administration, Supervision, Writing – review & editing. **Alison Heppenstall:** Funding acquisition, Project administration, Supervision, Writing – review & editing. **Jiaqi Ge:** Supervision, Writing – review & editing.

Data availability

I have attached open-source model code and data in the article and all software access is open and free.

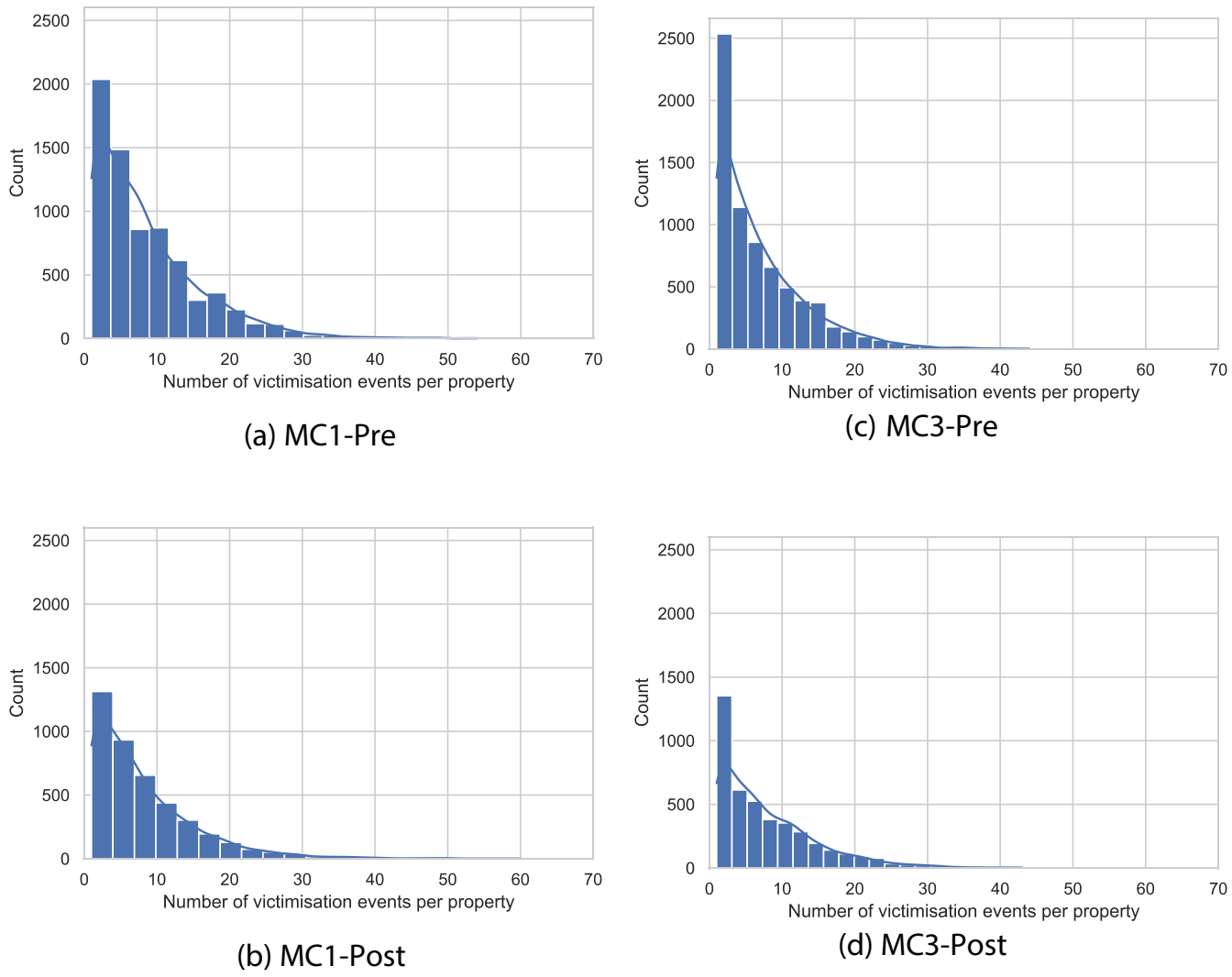


Fig. 12. Number of victimisation events per residential property ($n = 177{,}129$, bins $= 20$) pre-post intervention episodes for experiment conditions one and three.

Table 5
The (mean, std) of the number of offences committed for each experiment condition across pre-post intervention episodes.

Experiment Condition	Pre-Intervention	Post-Intervention	Difference(+ / −)	t-test(p)[1]
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	-[2]
3	(2078.08, 104.08)	(1282.80, 180.58)	− 795.28	19.02 ($p < 0.00$)

[1] H_0: that two independent samples have identical average (expected) values. H_a: the means of the distributions underlying the samples are unequal.
[2] t-test could not be applied to condition two as the number of offences was small.

Appendix A. Formal definitions

A.1. The logic behind risk, effort and reward

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 Fig. 13, $T_A(Risk) = 4/8$ 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.

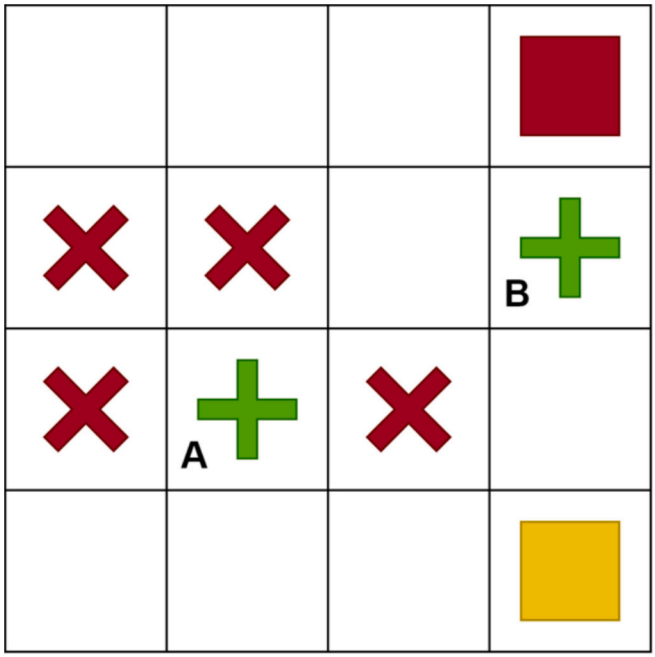


Fig. 13. An Example Scenario, where 4 Interventions, 2 Targets, 1 Node and Routine Activity Node.

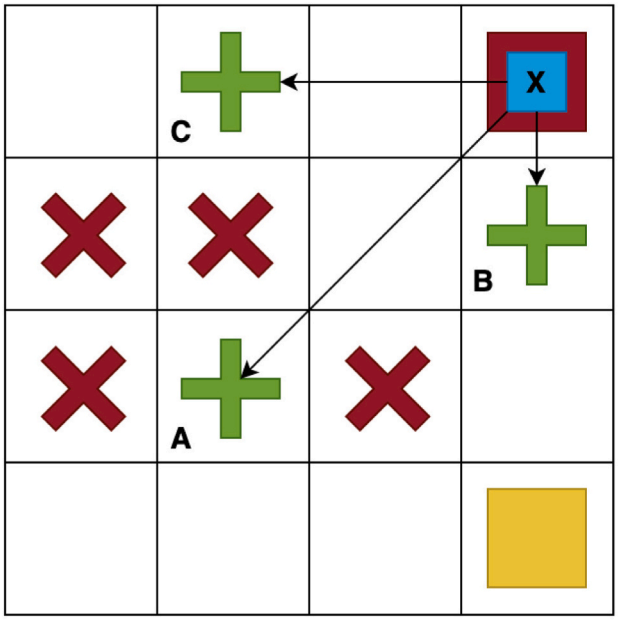


Fig. 14. An Example Scenario, where 4 Interventions, 3 Targets, 1 Node, Routine Activity Node and Offender Agent.

Given Fig. 14, 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

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

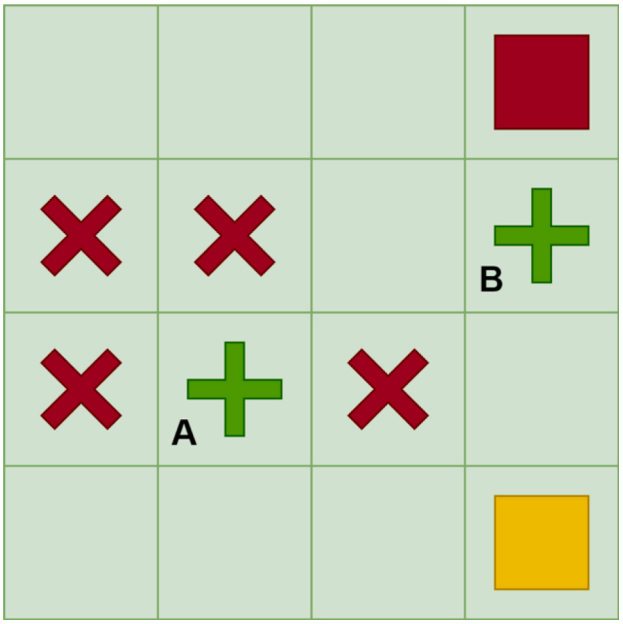


Fig. 15. An Example Scenario, where Area is Green, 2 Targets, 4 Interventions, 1 Node and Routine Activity Node. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Given Fig. 15, lets assume:

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

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.

A.2. 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.

A.3. 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

$Target\_Attractivness(T_i) < 0.$

A.4. 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):

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

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

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

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:

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

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.

Appendix B. Formulae

$$L^{CPI}(\theta) = \widehat{\mathbb{E}}_t\left[\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta old}(a_t|s_t)}\widehat{A}_t\right] = \widehat{\mathbb{E}}_t[r_t(\theta)\widehat{A}_t] \tag{B.1}$$

$$L^{CLIP}(\theta) = \widehat{\mathbb{E}}_t[min(r_t(\theta)\widehat{A}_t, clip(r_t(\theta), 1-\varepsilon, 1+\varepsilon)\widehat{A}_t)] \tag{B.2}$$

Where θ is the policy parameter, π is the policy, a and s are action and state respectively, $\widehat{\mathbb{E}}_t$ is the empirical expectations over timesteps. r_t is the probability ratio under the new and old policies, respectively. $\widehat{A}_t$ 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 B.1). The second term, $clip(r_t(\theta), 1-\varepsilon, 1+\varepsilon)\widehat{A}_t)$, 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).

Appendix C. Supporting figures

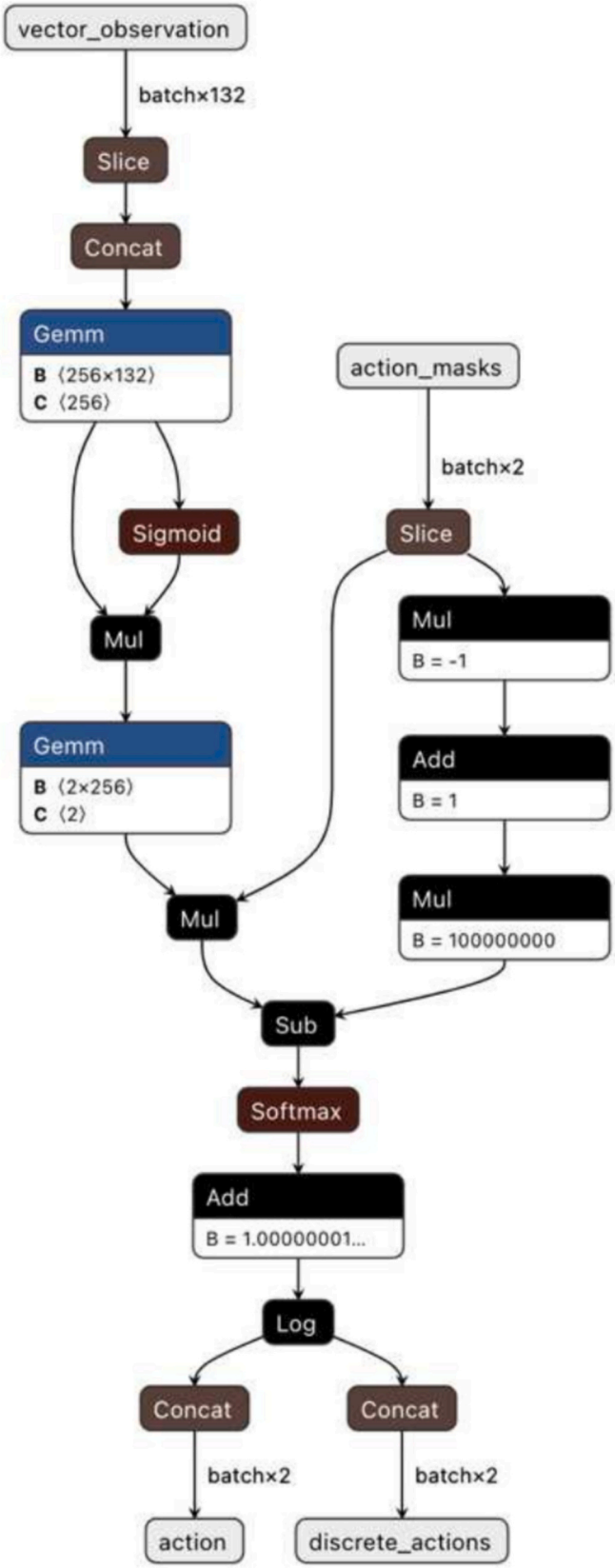


Fig. C.1. The Artificial Neural Network architecture diagram for the Offender agents.

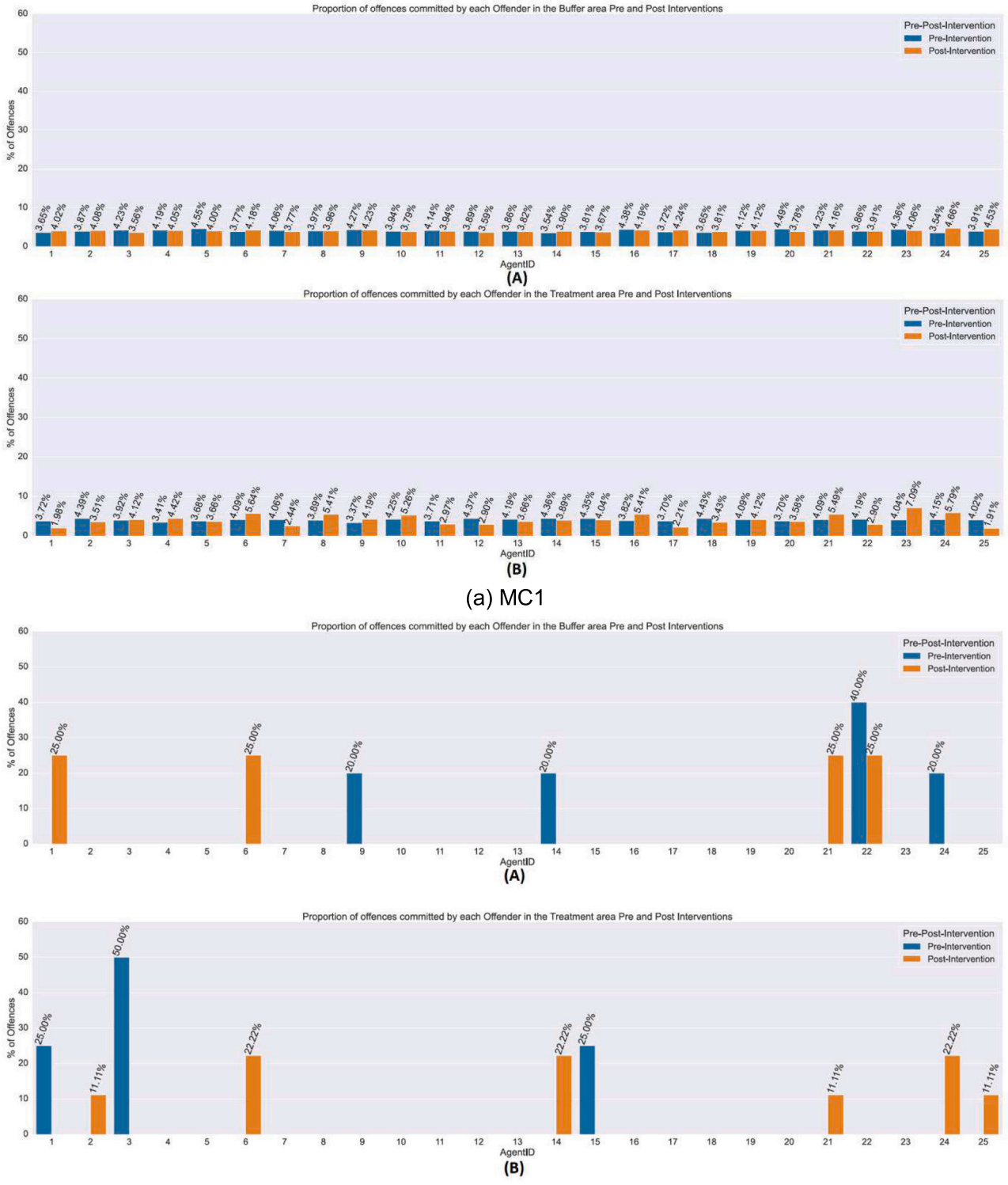
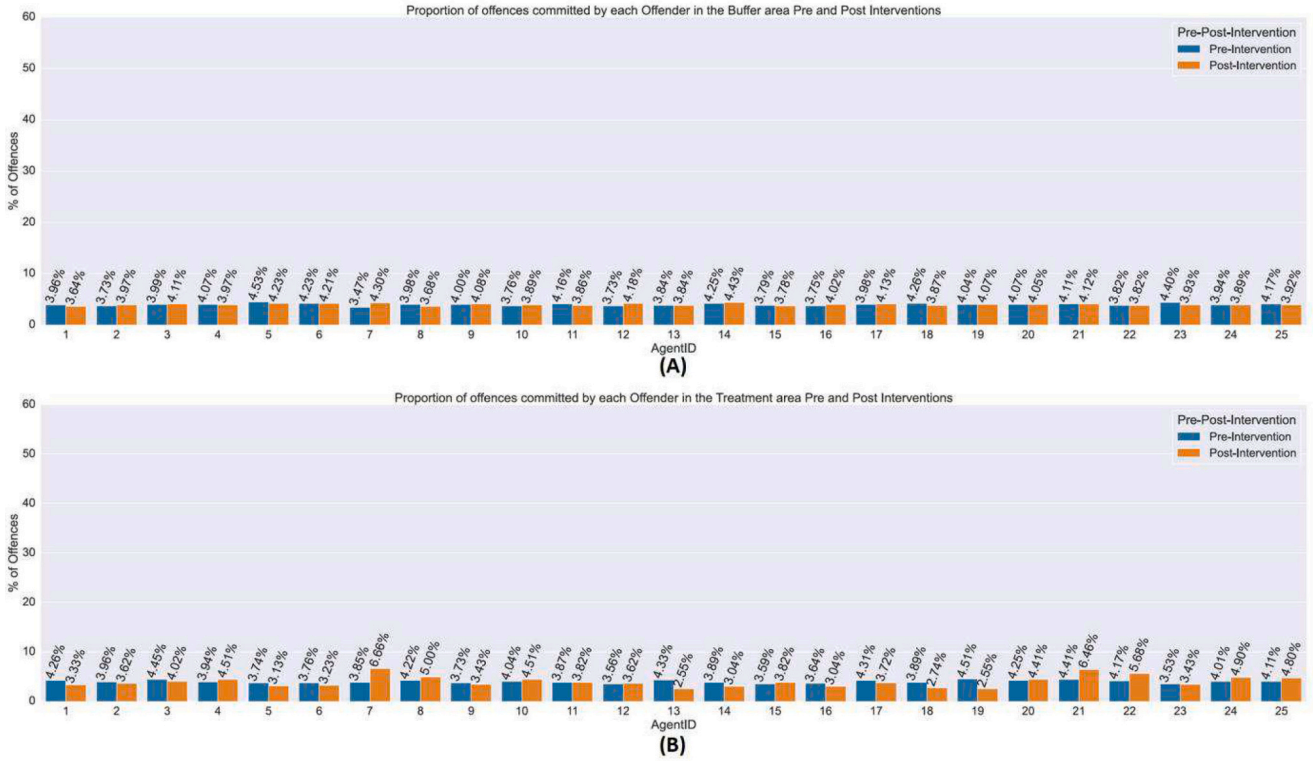


Fig. C.2. The proportion of offences per model condition committed by each offender agent across all simulations pre-post interventions in the Buffer (A) and Treatment (B) areas.



(c) MC 3

Fig. C.2. (*continued*).

Appendix D. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compenvurbsys.2024.102141.

References

Arthur, W. B. (1994). Inductive reasoning and bounded rationality: The El Farol problem. *The American Economic Review, 84*(2), 406–411.

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.

Barr, R., & Pease, K. (1990). Crime placement, displacement, and deflection. *Crime and Justice.* https://doi.org/10.1086/449167

Baudains, P., Braithwaite, A., & Johnson, S. D. (2013). Target choice during extreme events: A discrete spatial choice model of the 2011 London riots. *Criminology.* https://doi.org/10.1111/1745-9125.12004

Bernasco, W., & Luykx, F. (2003). Effects of attractiveness, opportunity and accessibility to burglars on residential burglary rates of urban neighborhoods. *Criminology, 41*, 981–1002. https://doi.org/10.1111/J.1745-9125.2003.TB01011.X. URL: https://onlinelibrary.wiley.com/doi/full/10.1111/j.1745-9125.2003.tb01011.x.

Birks, D., Townsley, M., & Stewart, A. (2012). Generative explanations of crime: Using simulation to test criminological theory. *Criminology.* https://doi.org/10.1111/j.1745-9125.2011.00258.x

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, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC* (pp. 1129–1136).

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.* https://doi.org/10.3233/WIA-2011-0212

Brantingham, P., & Brantingham, P. (1995). Criminality of place: Crime generators and CrimeAttractors. *European Journal on Criminal Policy and Research, 13*, 5–26.

Brantingham, P. L., & Brantingham, P. J. (2019). Environment, routine, and situation: Toward a pattern theory of crime. In *Routine activity and rational choice* (pp. 259–294). Routledge. https://doi.org/10.4324/9781315128788-12.

Brantingham, P. L., Brantingham, P. J., & Taylor, W. (2006). *Situational Crime Prevention as a Key Component in Embedded Crime Prevention. 47* pp. 271–292). URL: https://utpjournals.press/doi/10.3138/cjccj.47.2.271 https://doi.org/10.3138/CJCCJ.47.2.271.

Buşoniu, L., Babuška, R., & De Schutter, B. (2010). Multi-agent reinforcement learning: An overview. Studies in Computational Intelligence. *Innovations in multi-agent systems and applications-1*, 183–221.

Caskey, T. R., Wasek, J. S., & Franz, A. Y. (2018). Deter and protect: Crime modeling with multi-agent learning. *Complex & Intelligent Systems.* https://doi.org/10.1007/s40747-017-0062-8

Chu, X. (2018). Policy optimization with penalized point probability distance: An alternative to proximal policy optimization. *ArXiv*, 1–11 (abs/1807.00442).

Clarke, R. V. (1997a). Situational Crime Prevention: Successful Case Studies. *Harrow and Heston* (2, pp. 1–47). http://www.popcenter.org/library/reading/pdfs/scp2_intro.pdf.

Clarke, R. V. (1997b). *Situational crime prevention: Successful case studies.*

Clarke, R. V., & Cornish, D. B. (1985). Modeling Offenders' decisions: A framework for research and policy. *Crime and Justice, 6*, 147–185. http://www.jstor.org/stable/1147498.

Clarke, R. V., & Weisburd, D. (1994). Diffusion of crime control benefits: Observations on the reverse of displacement. *Crime Prevention Studies, 2*(1), 165–184.

Clarke, R. V. G. (1980). "Situational" crime prevention: Theory and practice. *The British Journal of Criminology, 20*, 136–147. URL: https://doi.org/10.1093/oxfordjournals.bjc.a047153.

Cohen, L. E., & Felson, M. (1979). Social change and crime rate trends: A routine activity approach. *American Sociological Review.* https://doi.org/10.2307/2094589

Cornelius, C. V. M., Lynch, C. J., Modeling, V., & Gore, R. (2024). Aging out of crime: exploring the relationship between age and crime with agent based modeling. Ads '17. In *Proceedings of the Agent-Directed Simulation Symposium.*

Cornish, D., & Clarke, R. (2003). A reply to Wortley's critique of situational crime prevention. *Crime Prevention Studies, 16*, 41–96.

Cornish, D. B., & Clarke, R. V. (1987). Understanding crime displacement: An application of rational choice theory. *Criminology.* https://doi.org/10.1111/j.1745-9125.1987.tb00826.x

Cornish, D. B., & Clarke, R. V. (2017). *The reasoning criminal: Rational choice perspectives on offending.* https://doi.org/10.4324/9781315134482

Cozens, P. (2013). Crime prevention through environmental design. In *Environmental criminology and crime analysis* (pp. 175–199). Willan.

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.

Dang, Q. V. (2020). *Reinforcement learning in stock trading. Advances in intelligent systems and computing 1121 AISC* (pp. 311–322). URL: https://link.springer.com/chapter/10.1007/978-3-030-38364-0_28. doi:10.1007/978-3-030-38364-0{_}28/FIGURES/10.

Devia, N., & Weber, R. (2013). Generating crime data using agent-based simulation. *Computers, Environment and Urban Systems, 42*, 26–41. https://doi.org/10.1016/j.compenvurbsys.2013.09.001

Ding, Z., & Dong, H. (2020). Challenges of reinforcement learning. Deep reinforcement learning: Fundamentals. *Research and Application*, 249–272. URL: https://link.springer.com/chapter/10.1007/978-981-15-4095-0_7, doi:10.1007/978-981-15-4095-0{_}7/FIGURES/6.

Eck, J. E., & Clarke, R. V. (2019). Situational crime prevention: Theory, Practice and evidence. In *Handbooks of Sociology and Social Research* (pp. 355–376). URL: https://link.springer.com/chapter/10.1007/978-3-030-20779-3_18. doi:10.1007/978-3-030-20779-3{_}18/FIGURES/3.

Eck, J. E., & Liu, L. (2004). *Routine activity theory in a RA/CA crime simulation*. Nashville, TN: American Society of Criminology.

Epstein, J. M., & Axtell, R. (1997). Artificial societies and generative social science. *Artificial Life and Robotics, 1*, 33–34. https://doi.org/10.1007/bf02471109

Faghri, F. (2021). Training efficiency and robustness in deep learning. *ArXiv, (abs/2112.01423).*, 1–149.

Farkas, A., Kertesz, G., & Lovas, R. (2020). Parallel and distributed training of deep neural networks: A brief overview. In *INES 2020 - IEEE 24th international conference on intelligent engineering systems, proceedings* (pp. 165–170). https://doi.org/10.1109/INES49302.2020.9147123

Farrell, G. (2015). Crime concentration theory. *Crime Prevention and Community Safety, 17:4*, 233–248. URL https://link.springer.com/article/10.1057/cpcs.2015.17 https://doi.org/10.1057/CPCS.2015.17.

Farrell, G., & Pease, K. (2001). *Repeat victimization. 12*. Criminal Justice Press.

Florence, P. S., & Zipf, G. K. (1950). Human behaviour and the principle of least effort. *The Economic Journal*. https://doi.org/10.2307/2226729

Gerritsen, C. (2015). Agent-based modelling as a research tool for criminological research. *Crime Science*. https://doi.org/10.1186/s40163-014-0014-1

Gialopsos, B. M., & Carter, J. W. (2014). *Offender Searches and Crime Events. 31* pp. 53–70). https://doi.org/10.1177/1043986214552608. URL: https://journals.sagepub.com/doi/10.1177/1043986214552608.

Groff, E. R. (2007). Simulation for theory testing and experimentation: An example using routine activity theory and street robbery. *Journal of Quantitative Criminology*. https://doi.org/10.1007/s10940-006-9021-z

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*. https://doi.org/10.1007/s10940-018-9376-y

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. https://doi.org/10.1111/j.1745-9125.2009.00177.x. URL: http://doi.wiley.com/10.1111/j.1745-9125.2009.00177.x.

Gutiérrez, O., Orozco-Aguirre, H. R., & Landassuri-Moreno, V. (2013). Agent-based simulation of crime. In *2013 12th Mexican international conference on artificial intelligence* (pp. 24–29). https://doi.org/10.1109/MICAI.2013.9

Hayward, K. (2007). Situational crime prevention and its discontents: Rational choice theory versus the 'culture of now'. *Social Policy and Administration*. https://doi.org/10.1111/j.1467-9515.2007.00550.x

Heppenstall, A. J., Crooks, A. T., See, L. M., & Batty, M. (2012). *Agent-based models of geographical systems*. Springer Netherlands. https://doi.org/10.1007/978-90-481-8927-4

Islam, M., Chen, G., & Jin, S. (2019). *An Overview of Neural Network. 5*. https://doi.org/10.11648/J.AJNNA.20190501.12. http://www.sciencepublishinggroup.com

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*. https://doi.org/10.1016/j.matcom.2016.05.008

Jipp, M. (2007). *Situation Adaptation: Information Acquisition, Human Behavior and its Determining Abilities*. Ph.D. thesis. Universität zu Köln. URL: https://madoc.bib.uni-mannheim.de/1909/.

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*. https://doi.org/10.1007/s10940-007-9025-3

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. URL: http://www.ncbi.nlm.nih.gov/pubmed/25419001 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4230953 https://doi.org/10.1177/0022427814531490.

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*. https://doi.org/10.1007/s11292-014-9209-4

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*, Article 101757. https://doi.org/10.1016/J.COMPENVURBSYS.2022.101757

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.

Justesen, N., Bontrager, P., Togelius, J., & Risi, S. (2020). Deep learning for video game playing. *IEEE Transactions on Games, 12*, 1–20. https://doi.org/10.1109/TG.2019.2896986

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*. https://doi.org/10.1613/jair.301

Levy, L., Santhakumaran, D., & Whitecross, R. (2014). *What works to reduce crime?: A summary of the evidence*. Scottish Government, Social Research.

Linden, R. (2007). Situational crime prevention: Its role in comprehensive prevention initiatives. *IPC Review, 1*, 139–159.

Littman, M. L. (2015). Reinforcement learning improves behaviour from evaluative feedback. *Nature*, 445–451, 7553 521. URL https://www.nature.com/articles/nature14540 https://doi.org/10.1038/nature14540.

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. *Journal of Medical Internet Research, 22*(7), e18477. https://www.jmir.org/2020/7/e18477 22, e18477. URL: https://www.jmir.org/2020/7/e18477 https://doi.org/10.2196/18477.

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. URL: https://academic.oup.com/scan/article/16/8/761/5813717 https://doi.org/10.1093/SCAN/NSAA040.

Malleson, N., Evans, A., & Jenkins, T. (2009). *An Agent-Based Model of Burglary. 36* pp. 1103–1123). https://doi.org/10.1068/b35071. URL: https://journals.sagepub.com/doi/abs/10.1068/b35071.

Malleson, N., Heppenstall, A., & See, L. (2010). Crime reduction through simulation: An agent-based model of burglary. *Computers, Environment and Urban Systems*. https://doi.org/10.1016/j.compenvurbsys.2009.10.005

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. URL: http://journals.sagepub.com/doi/10.1177/0037549710384124 https://doi.org/10.1177/0037549710384124.

Manson, S. M. (2006). Bounded rationality in agent-based models: Experiments with evolutionary programs. *International Journal of Geographical Information Science*, 991–1012. https://doi.org/10.1080/13658810600830566

Mnih, V., Badia, A. P., Mirza, L., Graves, A., Harley, T., Lillicrap, T. P., … Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning, in: 33rd international conference on machine learning. *ICML, 2016*, 1–19.

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. URL: https://www.cambridge.org/core/journals/european-journal-of-applied-mathematics/article/abs/modelling-the-individual-and-collective-dynamics-of-the-propensity-to-offend/874B0A9FF3BBAC7678FFF01000036F8E https://doi.org/10.1017/S0956792510000173.

Nardin, L. G., Székely, R., & Andrighetto, G. (2017). GLODERS-S: A simulator for agent-based models of criminal organisations. *Trends in Organized Crime, 20*, 85–99. https://doi.org/10.1007/S12117-016-9287-Y

Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology, 53*, 139–154. https://doi.org/10.1016/J.JMP.2008.12.005

Park, A. J., & Buckley, S. (2016). Three-dimensional agent-based model and simulation of a Burglar's target selection. In *, 2015. Proceedings - 2015 European intelligence and security informatics conference* (pp. 105–112). EISIC. https://doi.org/10.1109/EISIC.2015.39.

Piquero, A., & Rengert, G. F. (2006). *Studying deterrence with active residential burglars. 16*. https://doi.org/10.1080/07418829900094211, 451–450. URL: https://www.tandfonline.com/doi/abs/10.1080/07418829900094211.

Poyner, B. (1991). Situational crime prevention in two parking facilities. *Security Journal, 2*, 96–101.

Queeney, J., Paschalidis, I. C., & Cassandras, C. G. (2021). Generalized proximal policy optimization with sample reuse. In *NeurIPS* (pp. 11909–11919).

Rahimiyan, M., & Mashhadi, H. R. (2010). An adaptive Q-learning algorithm developed foragent-based computational modeling of electricity market. In *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*. https://doi.org/10.1109/TSMCC.2010.2044174

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* (pp. 789–800). URL: https://link.springer.com/chapter/10.1007/978-3-319-65340-2_64. doi:10.1007/978-3-319-65340-2{_}64/FIGURES/5.

Rawal, A., Rajagopalan, P., & Miikkulainen, R. (2010). Constructing competitive and cooperative agent behavior using coevolution. In *, 2010. Proceedings of the 2010 IEEE conference on computational intelligence and games* (pp. 107–114). CIG. https://doi.org/10.1109/ITW.2010.5593366.

Rengert, G. (2002). The journey to crime. URL https://books.google.com/books?hl=en&lr=&id=vayJAgAAQBAJ&oi=fnd&pg=PA109&dq=journey+to+crime&ots=7GnvZxNImZ&sig=wfsYsbLkeLh4_YB6Bzudycop4Tk.

Schulman, J., Levine, S., Moritz, P., Jordan, M. I., & Abbeel, P. (2015). Trust Region Policy Optimization. In *, 3. 32nd International Conference on Machine Learning* (pp. 1889–1897). ICML. URL http://arxiv.org/abs/1502.05477.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). *Proximal policy optimization algorithms*.

Sert, E., Bar-Yam, Y., & Morales, A. J. (2020). Segregation dynamics with reinforcement learning and agent based modeling. *Scientific Reports*. https://doi.org/10.1038/s41598-020-68447-8

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. 18* pp. 1249–1267). https://doi.org/10.1142/S0218202508003029

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. 7 pp. 33–43). https://doi.org/10.1080/10683160108401781. URL: https://www.tandfonline.com/doi/abs/10.1080/10683160108401781.

Sternberg, R. J., & Gastel, J. (1989). Coping with novelty in human intelligence: An empirical investigation. *Intelligence, 13*, 187–197. https://doi.org/10.1016/0160-2896(89)90016-0

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. URL: https://link.springer.com/article/10.1057/s41284-018-0144-3 https://doi.org/10.1057/S41284-018-0144-3/FIGURES/1.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (2nd ed.) https://doi.org/10.1016/S0140-6736(51)92942-X

Szita, I., & Lorincz, A. (2006). Learning tetris using the noisy cross-entropy method. *Neural Computation, 18*, 2936–2941. https://doi.org/10.1162/neco.2006.18.12.2936

Taylor, R. B., & Gottfredson, S. (2015). *Environmental design, crime, and prevention: An examination of community Dynamics. 8* pp. 387–416). https://doi.org/10.1086/449128. URL: https://www.journals.uchicago.edu/doi/10.1086/449128.

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. URL: https://link.springer.com/article/10.1007/s10940-017-9347-8 https://doi.org/10.1007/S10940-017-9347-8/TABLES/5.

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. URL: http://www.jstor.org/stable/23639318.

Troitzsch, K. G. (2017). Can agent-based simulation models replicate organised crime? *Trends in Organized Crime, 20*, 100–119. https://doi.org/10.1007/S12117-016-9298-8

Urban, C., & Schmidt, B. (2001). PECS – Agent-based modelling of human behaviour. *Operations Research*, 1–6.

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. https://doi.org/10.1016/J.APGEOG.2015.08.004

Vanvuchelen, N., Gijsbrechts, J., & Boute, R. (2020). Use of proximal policy optimization for the joint replenishment problem. *Computers in Industry, 119*, Article 103239. https://doi.org/10.1016/J.COMPIND.2020.103239

Wang, Z., Bapst, V., Heess, N., Mnih, V., Munos, R., Kavukcuoglu, K., & de Freitas, N. (2016). Sample efficient actor-critic with experience replay. In *5th international conference on learning representations, ICLR 2017 - conference track proceedings*. URL: http://arxiv.org/abs/1611.01224.

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* (pp. 45–70).

Wiering, M. A., & Van Otterlo, M. (2012). Reinforcement learning. *Adapt. Learn. Optim., 12*, 729.

Wong, B. B., & Candolin, U. (2015). Behavioral responses to changing environments. *Behavioral Ecology, 26*, 665–673. URL: https://academic.oup.com/beheco/article/26/3/665/233718 https://doi.org/10.1093/BEHECO/ARU183.

Wooldridge, M. (2020). *The road to conscious machines* (1st ed.). Pelican Books.

Wortley, R. (2001). A classification of techniques for controlling situational precipitators of crime. *Security Journal*. https://doi.org/10.1057/palgrave.sj.8340098

Wortley, R. (2016). Situational precipitators of crime. In *Environmental Criminology and Crime Analysis: Second Edition* (pp. 81–105). https://doi.org/10.4324/9781315709826

Zhang, H., & McCord, E. S. (2014). A spatial analysis of the impact of housing foreclosures on residential burglary. *Applied Geography, 54*, 27–34. https://doi.org/10.1016/J.APGEOG.2014.07.007

Zhang, H., & Song, W. (2014). Addressing issues of spatial spillover effects and non-stationarity in analysis of residential burglary crime. *GeoJournal, 79*, 89–102. https://doi.org/10.1007/S10708-013-9481-2/FIGURES/5. URL: https://link.springer.com/article/10.1007/s10708-013-9481-2.