Modelling and Validation of Social Influences in Human Security Behaviour

Peter James Carmichael

School of Computing

Newcastle University

This thesis is submitted for the degree of

Doctor of Philosophy

September 2019

I would like to dedicate this thesis to my late grandfather Alexander Carmichael. Without his passion for learning or his motivation for inspiring me to challenge myself I do not think I would have ever started a PhD. A special mention to my partner Abigail, who has consistently provided me with a voice of reason and reassurance in times of success and failure. She has kept me level headed throughout my studies and encouraged me to work to the best of my ability. Thank you Abi! Finally, I would like to thank my parents and sister, Neil, Julia and Laura for being my academic cheerleaders during my 8 years of study time. Thank you for always believing in me!

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Peter James Carmichael

September 2019

Acknowledgements

I would like to acknowledge Charles my supervisor for the constant support and many hours he has invested working with me. Without his patience throughout all of my ideas (some good, many bad), I would not be in the position I am to submit the thesis, I am truly grateful. I'd also like to acknowledge Thomas for his continued support and ensuring that I become as rigorous as possible to produce work to the best of my ability. I would like to thank Sean, John and Luca for the many conversations and useful feedback they have provided over the years. Finally, I would like to thank my examiners Jason Steggles and Sasa Radomirovic for their time and effort towards the final stages of my PhD journey.

Abstract

A major challenge for organisations is effectively implementing security policies where employees have a choice to comply. For instance, an organisational requirement such as *no unauthorised personnel in restricted areas* may have a policy stating authorised persons must wear identification badges and, therefore, places the onus on employees to make a choice and have the responsibility of wearing their badge. Existing literature in psychology and human factors in security tells us that a person's compliance behaviour for a security policy depends on their *compliance attitude*. For the organisation, there exists uncertainty for quantifying compliance attitudes of employees towards security policies where those employees have a choice to comply. Quantifying the compliance attitudes would allow an organisation to further establish its current risk environment. In the case of not wearing identification badges due to poor compliance attitudes, it would be challenging to identify unauthorised personnel and the organisational requirement would not be met. A person's compliance behaviour depends on their compliance attitude which itself depends on the compliance behaviour of themselves and of others. For example, the compliance behaviour from top-level management can influence others to be non compliant. *This thesis poses the question how could one quantify compliance attitudes for security policies in an organisation where people can observe other people's compliance behaviour.* This thesis contributes the following: 1) modelling of social influences in Coloured Petri Nets 2) a rule based model to represent agents that observe the actions of other agents 3) the application of machine learning to identify hidden compliance attitudes 4) a user study with behavioural interventions towards security policy compliance 5) a simulation tool for assessing how compliance attitudes evolve amongst agents 6) validation of the simulation tool by comparison to the empirical data from the user study. Overall, we believe that this thesis provides a holistic approach towards social influences over compliance attitudes and the simulation tool paves the way towards accurately assessing compliance attitudes for security policies.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Roughly two thirds of medium/large organisations in the UK experienced some form of cyber breach in 2018 [3]. In 2014, Coca-Cola had a number of laptops stolen which led to 74,000 employees, suppliers and contractors having personal data compromised due to an unencrypted laptop [10]. Organisations classify these cyber breaches as security incidents, where there may be financial and operational loss or loss of reputation. Security policies are often deployed by organisations to ensure some high level requirements are achieved. This is to ensure security incidents are avoided or handled in the most appropriate way should they occur.

One of the major challenges for organisations is ensuring security policies where employees have a choice are effectively implemented, and ensuring employees have a positive security attitude [91]. Employees unintentionally not following a security policy can create a risk environment which does not align with the organisation's beliefs [112]. The exact wording of these security policies may change between organisations. For example, a security policy at the National Health Service (NHS) is as follows:

> *"In order to improve security all staff are required to wear a photo identification badge in a visible position at all times during working hours. In order to ensure security, every member of staff should be prepared to challenge individuals without identification badges where it is safe to do so."* [1] [1]

[1] The notion of challenging others or policies requiring interaction extends to others such as use of USB devices, tailgating, password management and so on.

A high level requirement for this policy would most likely be *ensuring that no unauthorised personnel are permitted in restricted areas*. For the organisation, the challenge of whether or not this type of policy is being effectively implemented is complex. People move around in many different locations and areas and interact with a range of other people. Without constant surveillance it is reasonable to assume that we cannot reliably assess or manage the number of people complying with the policy. That is, are they a) wearing their identification badge and b) challenging others who are not wearing their identification badges. For the organisation, it is a major challenge to achieve the high level requirement. However, a different type of policy might be capable of achieving the high level requirement.

Security policies placing the onus on the employee to make a choice may expect proactive behaviour in the form of social interaction. Two or more people making some form of exchange is classed as social interaction. For example, the previously mentioned NHS security policy expects employees to challenge others not wearing identification badges [1]. The challenge behaviour is a form of exchange occurring and will require a minimum of two people. Choosing not to comply for any policy could lead to a security incident. In the NHS, a lack of compliance for the challenging policy could allow an unauthorised person to access a restricted area. This could lead to the theft of hardware, which would have comparisons with the aforementioned Coca-Cola breach [10].

For an organisation, these security policies expecting social interaction and a choice are extremely complex to assess and determine how many individuals have a positive attitude and would be compliant with the security policy. CCTV surveillance is one example for measuring compliance for the NHS challenging security policy; however, this would come at an extremely high financial cost. Often, organisations will use surveys to assess how their employees currently perceive security policies but this can be time consuming and provides only a snapshot. For example, a survey to assess security awareness provided a holistic view for how employees viewed security in their organisation [89]. To counter the idea that a poor culture of perceptions and attitudes around security exists, we often see campaigns deployed to change people's perceptions towards security [11].

A person presented with a choice can often be influenced to make a different choice. In 2010, the UK government created The Behavioural Insights Team [2]. Commonly referred to as the Nudge Unit, they influence public policy. One of their projects increased the uptake

of registrations for organ donor cards by targeting people with personalised messages when they completed their tax return online [104]. The social influence mechanisms they used made use of MINDSPACE which is a framework of social influences [44].

For security policies where there is a choice that is the responsibility of the employee, these choices can be *do I wear my ID badge?* or *do I challenge someone tailgating?* and so on. Like any choice, for certain people they can be *nudged*. In the context of security, Zhu *et al.* demonstrated that by mutually exchanging information, they could utilise reciprocity norms and by providing different types of information, they were able to increase the amount of personal information that people reciprocated [117]. As a nudging mechanism, reciprocity along with others are well established, most notably from the behavioural scientist Robert Cialdini [33]. The social influences from Cialdini form a base for the previously mentioned MINDSPACE framework [44].

For a particular person, their beliefs and attitudes towards security policies will dictate whether or not they are compliant. This is often referred to as the Theory of Planned Behaviour [8]. For a security policy, the attitudes towards complying with the policy may be influenced by the benefit of compliance, the cost of compliance and the cost of non compliance. This holistic assessment of consequences for compliance or non-compliance can shape a person's belief and subsequently their compliance behaviour [63]. The Theory of Planned Behaviour has been investigated numerous times when assessing an individual's decisions to comply with security policies [56, 76, 79]. With such a vast array of attitudes and beliefs, people must be considered heterogeneous. We introduce the term compliance attitude which we use to capture a person's beliefs, attitudes and perceptions towards complying with security policies.

We often see poor security decisions as a result of phishing, where social influences are common (e.g. an email impersonating a bank is a form of social influence) [41]. Phishing attacks are one example of how a social influence can nudge a person's decision. On the physical side, where people interact face to face, it has been shown that top level management can significantly increase policy compliance attitudes at lower levels of the employee hierarchy [62]. Influencing people's compliance attitudes is not limited to social influences. Other factors such as the surrounding physical space and any tools/capabilities may affect a person's compliance attitude. A physical space which creates or removes the

opportunity to change a person's compliance attitude could have a major impact. Additionally, a tool/capability such as physically wearing a lanyard of a bright colour to display a badge may have an influence over a person's compliance attitude. Therefore, it is not just a social influence which impacts a person's security decision.

When comparing top level management as a social influence we can see similarities with the *Messenger* effect from MINDSPACE [44]. The social influences from MINDSPACE, such as *Messenger* or *Norms* could participate in influencing the compliance attitudes for the NHS security policy:

- ***Messenger:*** If individuals in top level management stopped wearing their badge or did not challenge those not complying, this could influence others to be less compliant.
- ***Norms:*** If a number of people are not complying, then it may become the norm behaviour and others would adopt a lack of compliance.

Continuing with the NHS policy, a social engineer could exploit influence techniques to avoid being challenged [87]. Furthermore, the behaviours of others may unintentionally influence the compliance attitude of more people. The *Messenger* effect from MINDSPACE describes the authoritative influence that people have over other peoples choices [44]. These social influences from MINDSPACE exist and change a person's behaviour. Furthermore, a social engineer will exploit these social influences to achieve their goal. We declare our first remark based on our observations so far.

Remark 1 (Informs)**.** *The compliance attitude of people informs their behaviour.*

Social influences also assist with good security decisions. For the *Messenger* effect, top-level management could actively challenge employees, re-enforcing the message that security policies need to be followed [62]. *Norms* could be exploited as an incentive not to lose out on something for lack of security policy compliance, thereby creating a subjective norm for those not complying [70].

The compliance attitude governs a person's behaviour towards security policies [8]. Social influences nudge a person when they are presented with a choice [44]. If the behaviours of others can influence a person's compliance attitudes, then the behaviour must have been observed [62]. These observed behaviours then act as a social influence and change a person's

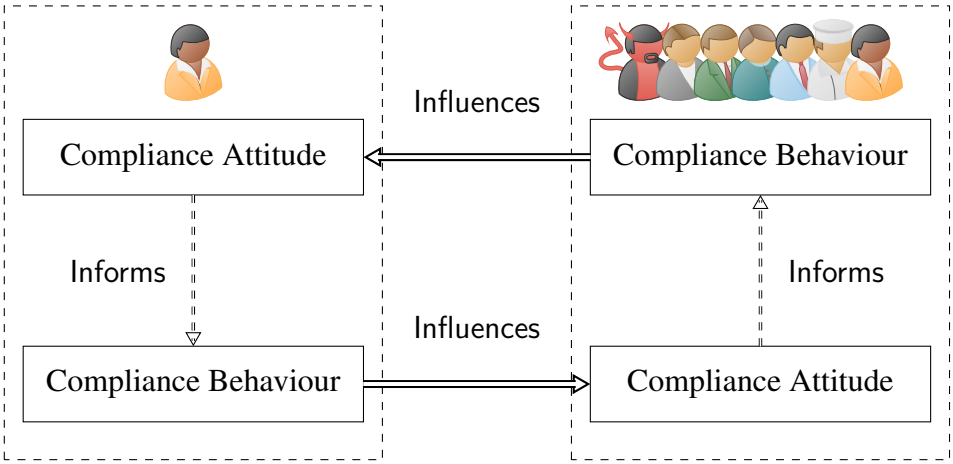


Figure 1.1 Relationship between compliance attitude and compliance behaviour.

compliance attitude, which then changes their behaviour, which then changes the compliance attitude of themselves and others. A social influence, therefore has the potential to propagate around a group of people. We declare the following remark based on the discussions so far.

Remark 2 (Influence)**.** *The compliance behaviour of people can be observed and is a social influence.*

Figure 1.1 provides an overview of Remark 1 by considering the dependencies between a person's compliance behaviour (i.e. how they behave towards a security policy) and their compliance attitude. Additionally, Figure 1.1 includes the relationship between the compliance attitude and the compliance behaviours of others.

When designing and managing any security policy, an organisation will consider risk management processes, as this is crucial to designing a solid security architecture [46]. Often, standards such as the ISO Information Security Management Systems standard are adopted to enhance certainty that any policies are being implemented in the most effective way [**?**]. However, an organisation does not know how a social influence impacts employee's compliance attitudes and beliefs towards security policies. Should certain people behave in a particular manner, i.e. not comply with a policy, this could propagate to others and create a general norm of non-compliance attitude amongst employees. For the organisation, they do not have all the pieces of the puzzle, and where the risk framework identifies a need for a policy, in practice it is not effectively implemented due to the social influences over employees' compliance attitudes.

If an organisation cannot assess the impact a social influence has on employee's compliance attitudes towards security policies, they cannot guarantee that employees will always comply with the security policy. A lack of compliance without any malicious intent may place an employee in a state of being or becoming an unintentional insider threat [53]. If it is the case that they are unaware of their potential damage capability, it begs the question, have they been influenced towards this current security attitude? Attaining this information would allow the organisation to fully understand if policies are being effectively implemented.

Some objectives for an organisation when designing and enforcing security policies are to a) reduce the number of security incidents and b) improve how staff respond to security incidents. One metric to understand this would be gathering the compliance attitudes of all members of staff for security policies. Regardless of what the employee's attitudes are, an organisation would want to know how a social influence in the organisation impacts the global compliance attitude for a particular policy[2]. Knowledge of how compliance behaviours of others and social influences impact a person's compliance attitude would provide an organisation a clear understanding of the compliance levels towards their security policies where employees have a choice, which leads us to our final remark:

Remark 3 (Compliance Attitude Prediction)**.** *It is useful to predict compliance attitudes evolving in an organisation.*

Assessing the compliance attitudes of a group of people is a complex problem as many people will have different compliance attitudes (i.e they are heterogeneous). We have established that a person's compliance attitude can be influenced by other peoples' compliance behaviour which subsequently impacts their own behaviour and the attitudes of others. Nevertheless, for the organisation, they must understand the implications of how compliance attitudes evolve towards security policies, which leads us to the aim of this thesis:

> ***Aim****: To provide a foundation for an organisation to build tools and methodologies to predict and analyse the impact social influences have towards compliance attitudes for security policies.*

[2] A global compliance attitude refers to the compliance attitudes of all employees for an organisation towards a security policy.

1.1 Research Questions

In this thesis we identify six research questions. We consider a research question to be something that we have identified from the current state of literature from psychology, behavioural economics, risk management and cyber security. Chapter 2 provides the background and related work for this thesis and supports each of the research questions that we address. Based on Remarks 1 and 2 we have the following research questions:

Research Question 1 (Propagating Effect)**.** *How can we quantify propagating social influences impacting compliance attitudes?*

Research Question 2 (Cyclic Observational Behaviour)**.** *How do we model observations of compliance behaviour?*

Research Question 3 (Security Behaviour Profiling)**.** *Can we accurately learn compliance attitudes from public traces of compliance behaviours?*

Research Questions 1, 2 and 3 use Remarks 1 and 2 as a foundation. Each question addresses a slightly different area of the same overarching problem: that social influences impact compliance attitudes. For instance, Research Question 3 will address the accuracy by which compliance attitudes can be learned based on traces of behaviour where social influences are present. Given Remark 3 which focuses on assessing and predicting compliance attitudes, we have two research questions:

Research Question 4 (User Study)**.** *How can we investigate and measure the effect of behavioural interventions such as social influences on compliance behaviour?*

Research Question 5 (Application)**.** *Can we design a tool to allow for the prediction and impact of compliance behaviours towards compliance attitudes?*

We offer two approaches for assessing compliance attitudes: a user study where we collect real world data; and a tool development approach where we use simulation as a basis for assessing evolving compliance attitudes. Addressing the research questions so far leads us to our final research question, which is:

Research Question 6 (Validation)**.** *How can we ensure that our methods and processes to address the research questions are valid?*

The confidence that can be placed in our methods and processes is of great importance, as an accurate and validated method or tool would indicate the trust that an organisation can place in such deliverables.

1.2 Contributions

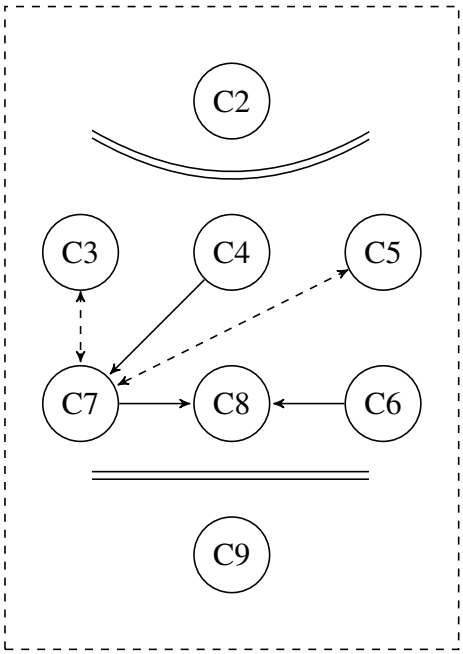
We provide six contributions in this thesis. They are 1) modelling of social influences in Coloured Petri Nets 2) the design of a rule based model to represent agents that observe the actions of other agents 3) the application of machine learning via decision trees to identify hidden compliance attitudes 4) a user study with behavioural interventions towards security policy compliance where participants are asked to swipe their smart card 5) a simulation tool for assessing how compliance attitudes evolve amongst agents 6) validation of the simulation tool by comparison to the empirical data from the user study.

The tool presented in this thesis, PCASP (Predicting Compliance Attitudes for Security Policies), allows an organisation to predict and analyse the impact of social influences towards compliance attitudes for security policies. We addressed the aim by providing a holistic approach and by identifying and answering appropriate research questions that we have perceived from existing literature. In particular, PCASP is a proof of concept whereby we demonstrated the core principles outlined in Figure 1.1.

1.2.1 Thesis Structure

Figure 1.2 provides an overview of the relationships between chapters. Chapters 2 and 9 support and reflect on the work carried out in this thesis. In each chapter we provide an overview for the purpose of the chapter and the research question it is addressing. We also provide a conclusion for each chapter and, if applicable, a section about validation of the models and methods used.

- **Chapter 2:** Background research. In this chapter we highlight the concepts and work which provide the foundation for Chapters 3 to 8.
- **Chapter 3:** Related work. In this chapter we highlight the concepts and work which are similar to those in Chapters 3 to 8.



Chapter 3: Propagation of social influences in Coloured Petri Nets [24].

Chapter 4: Building a model for observations of agent behaviour [23].

Chapter 5: Identifying compliance attitudes using decision trees [23].

Chapter 6: Assessing the impact of behavioural interventions on compliance behaviour [25].

Chapter 7: A simulation tool to predict compliance attitudes towards security policies [26].

Chapter 8: Validation of the simulation tool against empirical data from user study.

Figure 1.2 Relationships between chapters. Dashed line indicates a discussion between these chapters. Filled line indicates building contributions between chapters. Chapter 2 serves as the support for background literature for all other chapters. Chapter 9 is the conclusion and discusses the contributions of this thesis.

- **Chapter 4:** We model influence tokens using Coloured Petri nets to assess the impact an adversary can have with different types of behaviour change effects. The work is motivated through a tailgating example where the adversary has different amounts of finite influence tokens. This work targets Research Question 1 by addressing the notion of influence propagation through which social influences disperse across human agents [24].

- **Chapter 5:** We introduce a rule based model to capture influencing behaviour between agents. We focus on the model development to express agents that are capable of observing the behaviours of other agents. The focus here is on Research Question 2 whereby we design the model to allow for cyclic observational behaviour [23].

- **Chapter 6:** We assess the accuracy of Decision Trees over synthetic traces of agent behaviour. The focus of the Decision Trees is towards identifying hidden markers for compliance attitudes based on the publicly available information, i.e. traces of observed behaviours [23].

- **Chapter 7:** A user study to show the impact of interventions on human behaviour. The user study is split into three groups which are the Control, Discrete, and Continuous groups. We show how the broken-window theory negatively reduces the compliance attitude of swiping a smart card for subjects. This user study also addresses Research Question 4 for which we collected a data set containing what we believe to be social influences from the study leader and social influences amongst participants [25].

- **Chapter 8:** We introduce PCASP (Predicting Compliance Attitudes for Security Policies), a tool allowing for the rule based model to be easily utilised and providing an interface for performing simulation and assessing the compliance attitude of agents for a given policy. Additionally, we demonstrate the application of PCASP with a running example to illustrate how one might mitigate against poor compliance attitudes amongst agents. We envision that future versions of the tool would enable organisations to make more informed security policy decisions about employee behaviour, such as the best intervention to use. The tool provides a contribution to Research Question 5 for usable software that an organisation could use to quantify the impact of social influences [26].

- **Chapter 9:** We validate PCASP via comparisons of simulated data versus the empirical data from Chapter 6. We address Research Question 6 with this final contribution.

- **Chapter 10:** The thesis concludes with reflections about the research questions and how we addressed the aim. We then discuss the many avenues this area of work could follow.

1.2.2 Publications

Some chapters in this thesis are formed from four publications. To be clear, I (Peter) am first author on the papers and the technical contributions come from myself. The role of Charles Morisset and Thomas Groß was from a supervisory and editorial perspective.

- Carmichael, P., Morisset, C., and Groß, T. (2018a). Interventions over smart card swiping behaviour. In *STAST (Socio-Technical Aspects of Security and Trust)*. In publication

- Carmichael, P., Morisset, C., and Groß, T. (2018b). Simulating influencing human behaviour in security. In *STAST (Socio-Technical Aspects of Security and Trust)*. In publication

- Carmichael, P. and Morisset, C. (2017). Learning decision trees from synthetic data models for human security behaviour. In *International Conference on Software Engineering and Formal Methods*, pages 56–71. Springer

- Carmichael, P., Morisset, C., and Groß, T. (2016). Influence tokens: analysing adversarial behaviour change in coloured petri nets. In *Proceedings of the 6th Workshop on Socio-Technical Aspects in Security and Trust*, pages 29–40. ACM

Chapter 2

Background

In this chapter we provide two sections, they are the background and the related work. They do overlap, however, the background focuses on providing the supporting literature whereas the related work provides research that is focusing on addressing similar problems to that of people's compliance attitudes impacted by social influences.

2.1 Social Influences

2.1.1 Cialdini's Seven Influence Principles

There are seven well established social influences, most notably from the behavioural scientist Robert Cialdini [33]. Cialdini is noted for demonstrating compliance psychology [33, 32]. The seven influences as part of his response to how people comply are Reciprocity, Commitment and Consistency, Social Proof, Authority, Liking, Scarcity and Unity. Cialdini based his first six effects on observations he made towards persuasions and influencing as an undercover worker at used car dealerships, fund-raising organizations and telemarketing firms[1]. These social influences have been popularised and adopted by the use of the MINDSPACE framework which we discuss in Subsection 2.1.3.

[1]The Unity principle was only recently added in 2016 [32].

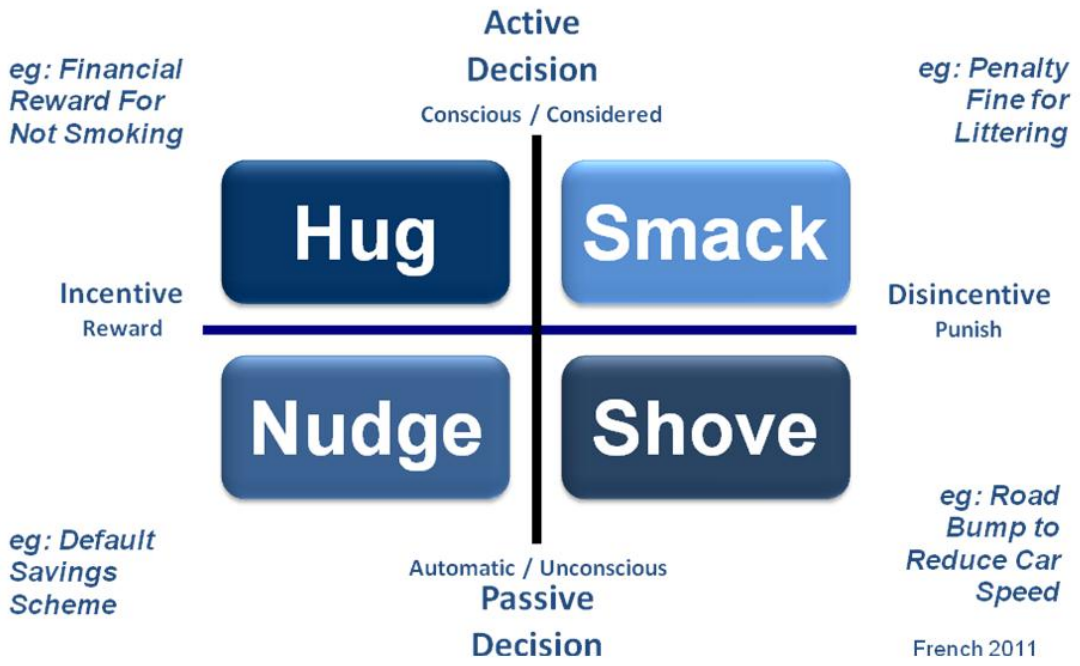


Figure 2.1 The value cost exchange matrix, taken from [49]

2.1.2 Nudging

Thaler and Sunstein are credited for impacting the direction of the MINDSPACE from the book *Nudge: Improving Decisions about Health, Wealth, and Happiness* which heavily focused on Choice Architecture, where choices can be presented to people in different ways in order to influence their final decision [106]. A nudge is a concept in behavioural science which elicits some re-enforcement for a decision. A topic of debate is when a nudge becomes shove, Figure 2.1 illustrates the a value cost exchange matrix showing the relationship between decisions and incentives for hugs, smacks, nudges and shoves.

From Figure 2.1 we can see that a nudge can impact a person's passive decision making [49]. The choices people/agents have in this thesis, such as challenging others for wearing their badge or choosing to tailgate behind someone can be seen as both active and passive decisions. From the exchange matrix in Figure 2.1, we can see why a nudge or influence over a choice can cause a change in a person's behaviour. In comparison to our work, the influence affects the compliance attitude which impacts the compliance behaviour.

When discussing passive decision and active decision making, comparisons can be made to the dual process theory which itself is likely to stem back to the work by William

James in the late 1800's where he proposed two different kinds of thinking, associative and true reasoning [65]. There does exist a number of interpretations for dual process theory and it became popular culture when the book *Thinking fast and slow* was released by Kahneman [69]. The core message of the book is System 1 and System 2 modes of thinking:

- **System 1:** Fast, automatic, frequent, emotional, stereotypical, unconscious. For example, reading text on a billboard will most likely by fast, automatic and perhaps unconscious.
- **System 2:** Slow, effortful, infrequent, logical, calculating, conscious. For example, count the number of times the letter *A* appears in a sentence will often be slow and effortful.

A choice to comply with a security policy can be either System 1 or System 2. It would depend on many factors such as the person, choice to make, environment etc. With this in mind, it's the case that sometimes a person's compliance attitude will be system 1 driven and sometimes it will be system 2 driven.

2.1.3 MINDSPACE: A Framework of Social Influences

In this thesis we adopt some of the social influences in the MINDSPACE framework. The UK government launched the Behavioural Insights Team known as the *Nudge* unit in 2010 [2]. The team used MINDSPACE as their foundation for influencing public policies. For example, they increased the uptake of people registering to be an organ donor when completing their car tax registration online [104].

MINDSPACE is an acronym listing nine different influencing effects. They are: **m**essenger, **i**ncentives, **n**orms, **d**efaults, **s**alience, **p**riming, **a**ffect, **c**ommitment, and **e**go [44]. According to the MINDSPACE framework, the nine effects are classified as[2]:

- Messenger: We are heavily influenced by who communicates information
- Incentives: Our responses to incentives are shaped by predictable mental short cuts such as strongly avoiding losses.

[2] According to the descriptions in the white paper for MINDSPACE [44]

- Norms: We are strongly influenced by what others do.
- Defaults: We *go with the flow* of pre-set options.
- Salience: Our attention is drawn to what is novel and seems relevant to us.
- Priming: Our acts are often influenced by sub-conscious cues.
- Affect: Our emotional associations can powerfully shape our actions.
- Commitment: We seek to be consistent with our public promises, and reciprocate acts.
- Ego: We act in ways that make us feel better about ourselves.

In this thesis we mainly use the *Messenger*, *Norms* and *Priming* effects in our work. The motivation and reasoning for using these different effects is due to the contrasting differences that they offer. For example, the *Messenger* effect has strong connections to a person's belief about another person whereas the *Priming* effect relies on some sub-conscious belief that previously existed when a trigger is engaged. Chapter 4 classifies effects as impacting agents differently. For example, the use of *Priming* demonstrates the impact a propagating social influence can have on the effect of agents' compliance attitudes.

Whilst we do adopt the use of the MINDSPACE framework, it is important to note that Cialdini's body of work is the main inspiration and direction for the framework. For instance, the *Authority* influence principle from Cialdini states that people in perceived positions of power have an ability to change our choices and has a strong connection with the *Messenger* effect.

We now discuss two of the seven influences which are seen regularly in this thesis, they are Authority and Social Norms:

Authority: People follow the lead of credible knowledgeable experts. The credibility is subjective to how a person perceives others based on some criteria such as their role or personality. In social engineering, we often see the authority principle exploited by an adversary as they pose as a worker in uniform to gain the trust of employees [87].

Social Norms: Cialdini and Trost defined Social Norms as *rules and standards that are understood by members of a group, and that guide and/or construct social behaviour without the force of laws* [34]. Whether or not a security policy is a rule, standard or law is a discussion we do not have in this thesis, however, it is clear that a security policy could develop its own social norms. Consider a security policy to stop tailgating, if nobody challenges and there are no consequences for tailgating or permitting it, then a social norm where everyone tailgates and nobody challenges may evolve.

2.2 COM-B Model

The COM-B (Capabilities, Opportunities, Motivation - Behaviour) model Section 2.3 conceptualizes long-term behaviour change. Figure 2.3 illustrates the behaviour change elements COM-B covers. Capabilities, Opportunities and Motivation together can influence a persons behaviour, which is similar to the concept of compliance attitude in this work. In Chapter 8 we utilise a behavioural intervention in a running example to demonstrate the impact of restructuring the physical infrastructure that agents operate in. A behavioural intervention in the context of human behaviour is an event causing a change of behaviour. There are many behaviour change theories to consider when looking at behavioural interventions [83].

We use the COM-B Model in Chapter 7 as inspiration for the impact analysis that we perform when using the simulation PCASP to impact the compliance attitudes of agents.

2.3 Understanding the Compliance Attitude of a Person

A persons behaviour is typically governed by the choices they make and can be directly related to their beliefs and attitudes surrounding the environment they currently occupy. The likelihood of people's behaviour varies in systematic ways has been widely investigated in psychology. We but name a few well-known models to support this point, which apply to different levels of abstraction.

COM-B. The COM-B model [82] (short for: Capability, Opportunity, Motivation – Behaviour) conceptualizes long-term behaviour change and has been related to how influencers change such behaviour.

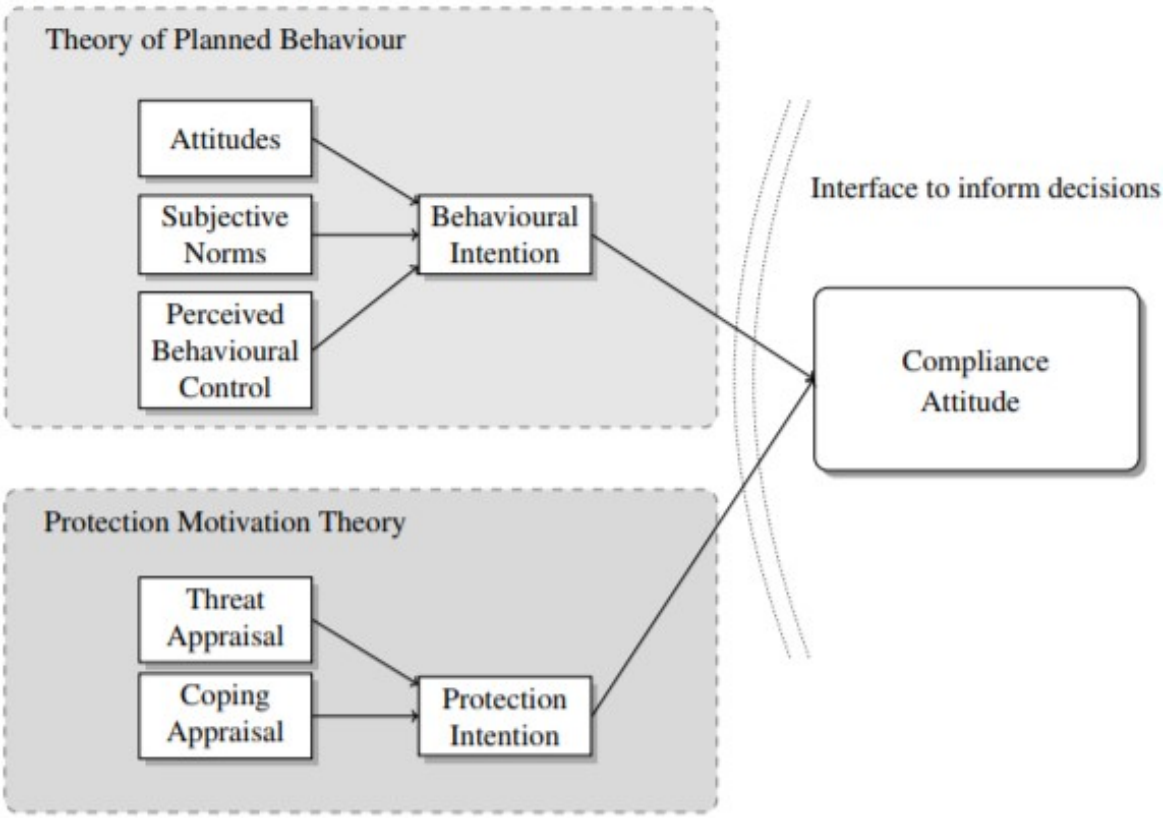


Figure 2.2 Interpretation of how compliance attitudes can be formed.

TPB. The Theory of Planned Behaviour (TPB) [8] governs more short term behaviour, especially how conscious and planned behaviour emerges out of a person's attitude, subjective norm, and perceived behavioural control (PBC). Therein, PBC covers the person's beliefs in the efficacy of their behaviour. These predictors impact the person's intention to act and finally the behaviour itself.

BDI. The Belief-Desires-Intention (BDI) [19] model relates reasoning about beliefs of others to our own actions.

PMT. The Protection Motivation Theory (PMT) [17] considers to what extent fear-inducing messages, such as "You will be fired, if you don't follow the policy" will induce behaviour change through a person's threat and coping appraisal.

Figure 2.2 is an interpretation of how a person's compliance attitude can be forged based on some of the aforementioned models around human behaviour. The compliance attitude as defined here is a culmination of different behaviour models and it is the compliance

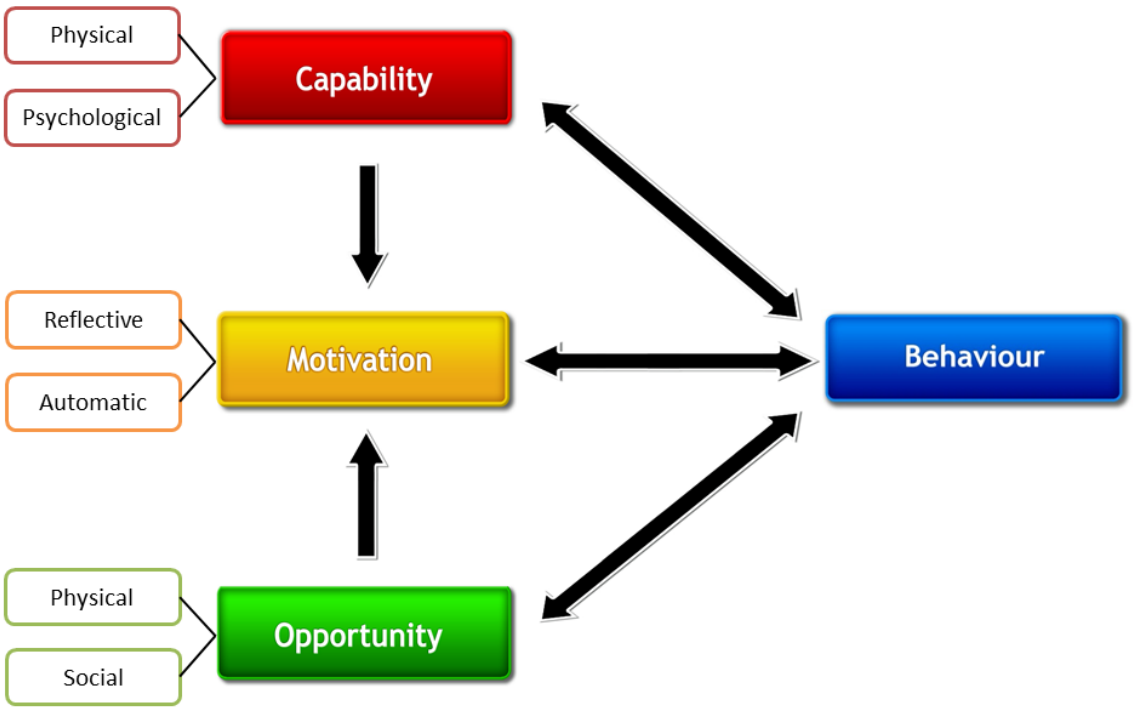


Figure 2.3 The relationship between behaviour change elements from COM-B. Figure taken from [82].

attitude that will drive a person's compliance behaviour. From Figure 2.2, which has been created from reading literature in the Theory of Planned Behaviour and Protection Motivation Theory, we see different elements which collectively form a behaviour model for compliance attitude. Assessing and addressing any change to one or more of these elements may have a resulting impact on a person's compliance attitude. We discuss this further in Chapter 8 when assessing the complexities of the simulation model and how it can be broken down into the component level, i.e. an engine to deal solely with an agents compliance attitude.

2.3.1 Variables Impacting Compliance Attitudes

Sommestad *et al.* performed a systematic review of quantitative studies towards security policy compliance. They identified thirteen different studies assessing variables impacting a person's compliance attitude [101]. Some of the variables they identified towards compliance attitudes as having at least a small effect size are threat appraisal, information security

awareness and source competency.

Threat Appraisal: The threat appraisal process encompasses the severity and vulnerability of a particular situation. Where the severity is the level of harm that comes from the thought of performing a behaviour and the vulnerability is the probability that a person will experience some harm [97]. Herath and Rao demonstrate that the perception about the severity of security breaches and the response costs of having to deal with a breach have a significant impact on a persons intentions to comply with a security policy [58].

Information Security Awareness: The work in information security awareness builds on the theory of planned behaviour. In general, it is a holistic overview of how people perceive security and includes but is not limited to to safety, rewards, intrinsic benefits, intrinsic costs and and vulnerabilities. These elements form the attitudes towards benefits towards compliance, the cost of compliance and the cost of non-compliance [21]. Bulgurcu *et al.* assessed the impact information security awareness has on compliance attitudes and found that a more well informed information security awareness campaign has a positive impact on compliance attitudes [21].

Source Competency: The notion of source competency is the perceived ability that individuals place in the source of information. For example, if a person respects their superiors then it's likely that a message from the competent source would be received and accepted [85]. The example of source competency relates to the *Messenger* effect from MINDSPACE as a person who is not trusted/respected will struggle to influence a persons decision. The findings from Johnston and Warkentin towards source competency are consistent with what we would expect and suggest that positive relationships between source credibility and attitudes and behavioural intent impacts positively how employees perceive organisational policies [68][3].

[3]We would expect that behaviours demonstrating the *Messenger* effect have an impact on a persons behaviour.

2.4 Decision Trees

In this thesis we learn Decision Trees in Chapter 6. Typically, decision tree learning uses decision trees as a predictive model to take some observations of data, say an agent's compliance behaviour and to forecast some conclusions about that data, say their compliance attitude.

Decision trees are often used in data mining and there are two main types which are classification tree analysis and regression tree analysis. The two are often used together and is know as CART (Classification And Regression Tree) analysis and it was first introduced in 1984 [20]. Our focus is on classification trees analysis where the predicted outcome is a discrete class on which the data belongs. Unlike regression tree analysis where the predicted outcome is a real number such as a person's age or the expected shift of shares in a company.

In this thesis any implementation of a classification tree analysis uses the CART algorithm which is similar to another algorithm known as C4.5. The implementation is done in python and uses the scikit-learn module. The trees that we produce are binary trees where each decision point has only two options until a discrete value from a pre-determined class is forecasted as the prediction.

The research question for the chapter featuring decision trees is:

> *Can we accurately learn compliance attitudes from public traces of compliance behaviours?*

The motivation for doing this is based on the usefulness of acquiring agents' compliance attitudes. If we can accurately assess compliance attitudes from traces then we can establish the current risk environment that an organisation is operating in. Having a wider knowledge of the risk environment allows for the organisation to apply its risk appetite and make fully informed decisions regarding the effectiveness of security policies.

Often, Decision Trees are used in Fraud Detection for classification [80]. There are similarities in the approach that we adopt here. Instead of identifying a perpetrator for a crime, we are identifying the compliance attitude of agents for how they perceive their compliance behaviour towards security policies.

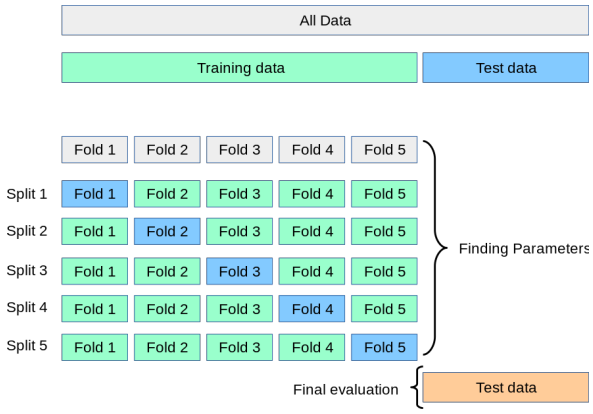


Figure 2.4 Figure representing how data is split when performing cross-validation. Figure taken from [92].

2.4.1 Cross-Validation

In Chapter 6, we use cross-validation as our technique for assessing the accuracy of decision trees. In cross-validation, we typically split a data set D into k mutually exclusive subsets, which are often referred to as folds $\{D_1, D_2, .., D_k\}$ of usually equal size. A decision tree is then trained and tested for each data set in D where it is trained on $\frac{D_x}{D_{xt}}$ and tested on D_{xt} where D_{xt} refers to the test part of the data.

A general rule of thumb for splitting training and test data is 80% as training data and 20% for test data. This is performed each time for the different folds as described above ensuring that the data is generalised as the decision trees will be exposed to all parts of the data for training and testing.

Figure 2.4 is taken from the scikit, which is the implementation we use in Chapter 6 to perform our cross-validation. The figure illustrates that with a number of different folds, there is test data that is taken out each time, allowing the process to be generalised across all of the data. For example, in Split 1 and Split 2, the test data is Fold 1 and Fold 2, the remaining folds in each split are for training the models, in our case, the decision trees.

2.5 Julia and SysModels

In this thesis we create a tool called PCASP which we will discuss later in Chapter 8. The purpose of PCASP is to allow a security practitioner to enter a set of rules describing agent

behaviour with regards to a security policy. The tool builds a model given the rules and performs some simple analysis. In it's current state the tool is manually mapped from the rules outlined to the implementation code.

The tool is implemented in a language called Julia and makes use of a package called SysModels. Julia is a high level programming language which has been designed for computational analysis. The package SysModels is for creating Systems Models [**?**]. In our use case the Systems Models is the set of behaviour rules defining how agents interact with each other.

2.6 Petri Nets

We make use of Coloured Petri nets in Chapter 3. We provide an overview of Petri nets in general here. The use of Petri nets allows us to model the propagation of social influences. The research question it helps to address is:

> *How can we quantify propagating social influences impacting compliance attitudes?*

A Petri net or a place/transition system is a modelling language to describe distributed systems. A typical Petri net contains places, transitions and arcs. Arcs connect places and transitions and can go from place to transition and transition to place. An arc cannot go from place to place or from transition to transition [88].

Tokens: In a Petri net, tokens are distributed around the places and a particular distribution of tokens is known as a marking. A marking is essentially a state of the Petri net and from any given marking, there is a finite number of reachable markings [96] Tokens can migrate if a transition is fired. In order for a transition to to fire/execute, it must be enabled. A transition is enabled if there is a sufficient number of tokens in the input places for a transition. If multiple transitions are enabled, then the Petri behaves in a non-deterministic manner.

We illustrate Petri nets in Chapter 3 with Figure 4.2 and provide the formal rules for how the extended version of Coloured Petri nets are adopted [66]. We utilise the notion of influence

tokens to express propagating social influences that exploited by an adversary to increase the adversary's likelihood of entering an organisation via tailgating.

We don't discuss Petri nets any further here as we provide formal definitions and more descriptions in Chapter 4.

2.7 Multi Agent Systems

We make use of a Multi Agent System in Chapter 4 to model agents that are capable of observing the actions of other agents. It assists with addressing Research Question 2:

How do we model observations of compliance behaviour?

In the literature, a Multi Agent System is defined as a set of agents that are interacting with an environment and or each other to achieve some common or individual goals [115]. Often, it is the case that an agent is said to be cognitive and has a reasoning process in order to perform actions or behaviours. This reasoning process is strongly dependent on the attitudes of the agent, where the attitudes impacts the choices an agent makes [110].

We utilise the concept of Multi Agent Systems in Chapters 3, 4, 5 and 6 where we use different models to address the research questions. The agents in our systems are heterogeneous and do not have a common or individualised goal, they just behave according to their current context and any observations they have made of other agents compliance behaviour.

Figure 2.5 is a holistic view of how we see social influences interacting between agents. We express the notion of a defender that positively influences agents and an adversary that negatively influences agents. We use Figure 2.5 as a base and build on it within the chapters to create an observational based system where agents interact with each other and can observe each others behaviour. The defender is explored in Section 8 where we demonstrate through impact analysis the success a defender can have with different behavioural interventions. The adversary is considered in Section 4 where we use Coloured Petri nets to illustrate influence tokens, where influence tokens can be used by an adversary in this context to increase the likelihood of tailgating into an organisation.

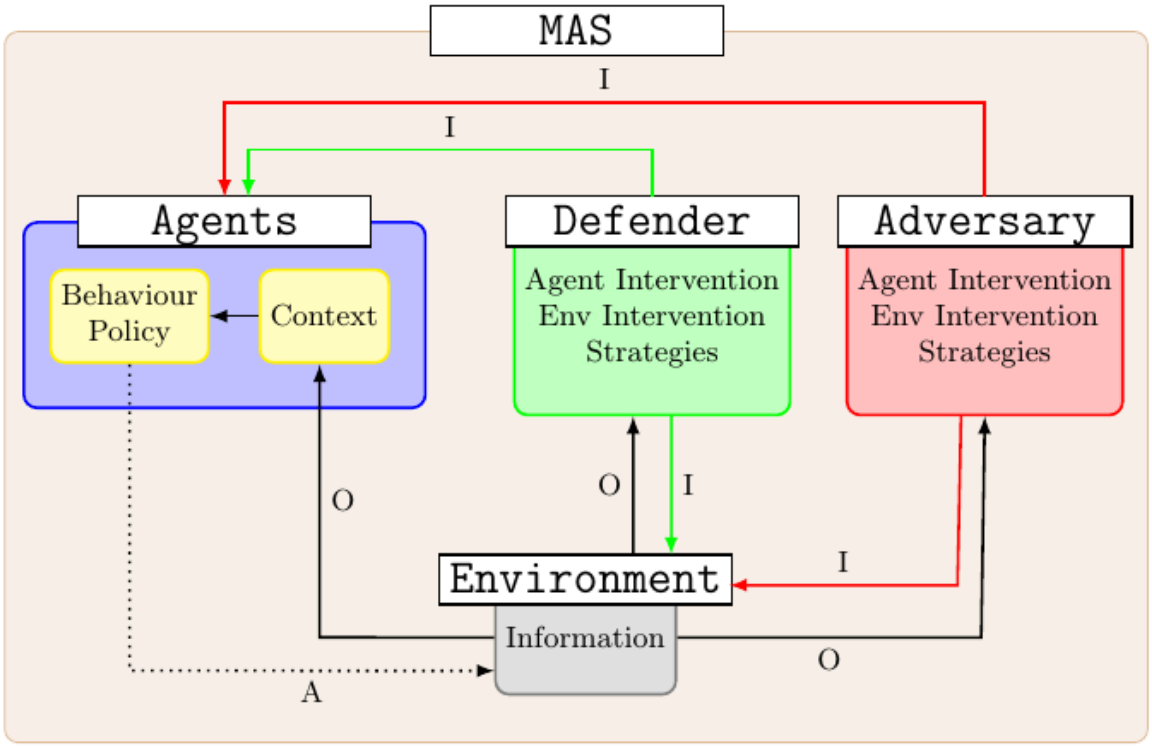


Figure 2.5 Holistic view of a Multi Agent System for social influences - A: Actions, O: Observations, I: Interventions, Green Arrow: Positive Influence, Red Arrow: Negative Influence

2.8 Statistical Tests

In Chapter 6 and Chapter 8 we use a range of statistical tests which we introduce in the chapter and also provide the full list here with a short description.

2.8.1 Planned Comparisons

Planned comparisons are not a statistical test but it is important and applicable to mention it as this point. We make use of planned comparisons in our work to ensure we focus on sensible and pre-determined comparisons in our work. Essentially, we specify before an experiment takes place, what it is that we are looking for in the data we have collected.

The requirements for performing planned comparisons ensures we have pre-specified what it is we want to compare and the statistical tests of how we are going to compare them.

For example, in Chapter 7 we pre-specified that we are looking to compare the total number of smart card swipes and that we would use an Analysis of Variance Testing.

In order to prove that we have performed planned comparisons and not just made the story up after we have collected the data, we commit a document to the OSF (Center for Open Science) website before the experiment takes place outlining how the experiment will run and the planned comparisons that we will do after the data has been collected.

2.8.2 Analysis of Variance

An analysis of variance (ANOVA) is a set of statistical methods to analyse and approximate the differences between group means of some samples. We used a one-way omnibus ANOVA which entails comparing the means of all the different groups and assessing if there is a difference. In our case, we use the ANOVA in both Chapter 6 and Chapter 8:

Chapter 7: The three groups we measure and compare against are Control, Discrete+ and Continuous−. The ANOVA will not tell us which groups are statistically significant, just that at least two of the groups are.

Chapter 9: In this chapter we used a one-way omnibus ANOVA which entails comparing the means of all the different experiment conditions against simulation conditions and assessing if there is a significant difference in the means. In total we perform six ANOVAS in Chapter 8.

We provide an overview of how a one-way omnibus ANOVA is performed to assist the reader with understanding how the results in the chapters are calculated.

An ANOVA is a statistical test for comparisons of mean values. The *one way* refers to the test only having one independent variable where the independent variable is something that is fixed and consistent for that group condition. For example, the experiment in Chapter 7 has the independent variable of the Discrete+ intervention and the Continuous− intervention. The dependent variable is a measurement that is dependent on some condition. In the experiment one dependent variable is the number of total swipes. A null hypothesis for any ANOVA is that there is no significant difference amongst any of the groups. The alternative hypothesis assumes that there is at least one significant different among the groups. In our

case, the comparisons are the Control vs Discrete+ and the Control vs Continuous−. In Chapter 7, we do not compare Continuous− vs Discrete+.

Table 2.1 Example Statistics: Total Swipes

Control	Discrete+	Continuous−
10	15	6
11	14	4
10	17	9
11	11	5
5	15	6
13	12	6

Consider Table 2.1 which provides some example descriptive statistics for the experiment in Chapter 7. The dependent variable being measured is the total number of swipes. Assuming the experiment has 18 participants where 6 participants take part in only one group then the result is Table 2.1 which is the raw data to be examined. We now demonstrate each step of the one-way ANOVA:

Step 1: Calculate the mean values of each group:

$$\bar{Y}_1 = \frac{1}{6}\sum Y_{\text{Control}i} = \frac{10+11+10+11+5+13}{6} = 10 \tag{2.1}$$

$$\bar{Y}_2 = \frac{1}{6}\sum Y_{\text{Discrete+}i} = \frac{17+14+17+13+16+13}{6} = 15 \tag{2.2}$$

$$\bar{Y}_3 = \frac{1}{6}\sum Y_{\text{Continuous-}i} = \frac{6+4+10+5+3+2}{6} = 5 \tag{2.3}$$

Step 2: Calculate the overall mean of the groups where a is the number of groups:

$$\bar{Y} = \frac{\sum \bar{Y}_i}{a} = \frac{\bar{Y}_1+\bar{Y}_2+\bar{Y}_3}{a} = \frac{10+14+5}{3} = 10 \tag{2.4}$$

Step 3: Calculate the between-groups sum of squared differences where n is the of data values per group:

$$S_B = n(\bar{Y}_1-\bar{Y})^2 + n(\bar{Y}_2-\bar{Y})^2 + n(\bar{Y}_3-\bar{Y})^2 \tag{2.5}$$

$$= 6(10-10)^2 + 6(15-10)^2 + 6(5-10)^2 = 50 \tag{2.6}$$

Step 4: Calculate the between-groups degrees of freedom:

$$f_b = |\bar{Y}| - 1 = 3 - 1 = 2 \tag{2.7}$$

Step 5: Calculate the between-group square mean value:

$$MS_B = \frac{S_B}{f_b} = \frac{50}{2} = 25 \tag{2.8}$$

Step 6: Calculate the within-group sum of squares. We do this calculation by centering the data in each group, see Table 2.2:

Table 2.2 Example Statistics: Within Groups Preparation

Control	Discrete+	Continuous−
$(10-10)^2 = 0$	$(17-15)^2 = 4$	$(6-5)^2 = 1$
$(11-10)^2 = 1$	$(14-15)^2 = 1$	$(4-5)^2 = 1$
$(10-10)^2 = 0$	$(17-15)^2 = 4$	$(10-5)^2 = 25$
$(11-10)^2 = 1$	$(13-15)^2 = 4$	$(5-5)^2 = 0$
$(5-10)^2 = 25$	$(16-15)^2 = 1$	$(3-5)^2 = 4$
$(13-10)^2 = 9$	$(13-15)^2 = 4$	$(2-5)^2 = 9$

$$S_W = 0+1+0+1+25+9+4+1+4+4+1+4+1+1+25+0+4+9 \tag{2.9}$$

$$= 94 \tag{2.10}$$

Step 7: Calculate the within-groups degrees of freedom:

$$f_W = a(n-1) = 3(6-1) = 15 \tag{2.11}$$

Step 8: Calculate the withing group mean square value:

$$MS_W = \frac{S_W}{f_W} = \frac{94}{15} \approx 6.3 \tag{2.12}$$

Step 9: Calculate the F-ratio:

$$F - ratio = \frac{MS_B}{MS_W} \approx \frac{25}{6.3} \approx 4 \tag{2.13}$$

For an ANOVA, the result is testing if there is a significance level, typically at the 5% significance level. In order for that to be true, the F-ratio must be greater than the F-critical value which can be calculated using an F-distribution table as a lookup or via many different tools. For the purposes of this thesis we don't provide the formula for calculating the F-critical value but we do provide the value which is $F - crit(2,15) = 3.86$ at 5% significance level. As the F-ratio (4) is greater than the F-crit value (3.68) then the results are significant at the 5% significance level. We would then fail to accept the null hypothesis that the expected values in the three groups do not differ and that there is a difference in the mean values amongst them. We therefore accept the alternate hypothesis that the three groups differ. We use this method to compare means in Chapter 7.

2.8.3 Effect Sizes

We only use effect sizes in Chapter 6. An effect size is a quantitative measure of an observation made about some metric, such as the correlation between two variables. A rule of thumb for effect sizes is that the larger the absolute value of an effect size then the stronger the effect is. We calculate the Hedges'g effect size which tells us the difference between two groups [57]. The value of the effect size provides an indicator for the magnitude of the effect. Typically, small, medium and large effects are referred to when the effect size is .2, .5 and .8 respectively [36].

As Hedges'g only permits the comparisons between two groups we perform planned comparisons. Planned comparisons are pre-specified before the results are collected to avoid *looking* for results that may be interesting if the data does not quite add up. The planned comparisons we perform in Chapter 6 are Control vs Discrete+ and Control vs Continuous−. The calculation for Hedges'g effect size is the following:

Step 1: Calculate the pooled standard deviation s*:

$$s^* = \sqrt{\frac{(n_1-1)s_1^2 + (n_1-1)s_2^2}{n_1+n_2-2}} \tag{2.14}$$

Step 2: Calculate the effect size:

$$g = \frac{\bar{x}_1 - \bar{x}_2}{s^*} \tag{2.15}$$

For the Control vs Discrete+ effect size we have the following for:

$$s^*_{\text{Discrete+}} = \sqrt{\frac{(6-1)4.84 + (6-1)7.18}{6+6-2}} \tag{2.16}$$

$$s^*_{\text{Discrete+}} = 3.16 \tag{2.17}$$

$$g_{\text{Discrete+}} = \frac{15-10}{3.16} \tag{2.18}$$

$$g_{\text{Discrete+}} = 1.58 \tag{2.19}$$

The value for the Discrete+ effect size when comparing to Control group is 1.58 which would be considered a large effect.

We would then perform the same calculation for the Continuous− group against the Control group. Note that we would not compare Continuous− and Control.

2.8.4 Confidence Intervals

The use of confidence intervals are only used in Chapter 6. A confidence interval is an interval estimate we compute to outline the range of values that an observed value could take for some confidence level. Where the confidence level is typically 95%, which is what we use in this chapter [59]. The confidence interval we provide is over the effect sizes that we observe.

2.8.5 Chi Squared Goodness of Fit test

The use of a Chi squared goodness of fit test is only used in Chapter 8. The Chi squared goodness of fit test is used to determine how significantly different some observed data (for

us, simulation data) is when compared to the expected data (for us, empirical data). For a Chi Squared Goodness of Fit test there are four assumption criteria that must be met:

Assumption 1 (Categorical Variable)**.** *One categorical variable (i.e., the variable can be dichotomous, nominal or ordinal). Examples of dichotomous variables include gender (2 groups: male or female), treatment type (2 groups: medication or no medication), educational level (2 groups: undergraduate or postgraduate) and religious (2 groups: yes or no).*

Assumption 2 (Independence of Observations)**.** *We should have independence of observations, which means that there is no relationship between any of the cases (e.g., participants).*

Assumption 3 (Mutually Exclusive)**.** *The groups of the categorical variable must be mutually exclusive.*

Assumption 4 (Expected Frequencies)**.** *There must be at least 5 expected frequencies in each group of your categorical variable.*

The assumptions were taken from laerd statistics, which is a website composed of comprehensive definitions and techniques to perform statistical techniques [4].

Chapter 3

Related Work

The TREsPASS project was launched consisting of many research projects into *socio-technical* security by means of risk estimation and predictive assessment [6]. The socio aspect focused on human behaviour in security. In this section we focus in on the work commissioned by the TREsPASS project and any other sources that are relevant.

3.1 Formal Modelling

Formal methods when applied to human behaviour was demonstrated by the analysis of insider threat. Probst *et al.* discussed the inclusion of explicit *socio* aspects and demonstrated proof of attacks towards this insider threat by use of a running example consisting of stealing a baker's birthday cake [93]. The socio-technical aspect is taken to the next step where automated verification is performed to produce a framework for insider threat analysis [71].

The formalism we adopt describes the semantics of a Multi Agent System in Chapter 4 where agents can observe the actions of other agents. To the best of our knowledge, we have not come across another system which models the observations of actions for the purposes of social influences impacting compliance attitudes. We then take the modelling process one step further and describe the syntax for the simulation tool PCASP which uses the semantics of the model from Chapter 4. One possible next step of the work for the semantics would be automated verification to produce a framework for social influences analysis.

Lenzini *et al.* addresses whether a security policy together with physical access controls protects from socio-technical threats [78]. The approach they take deals with reasoning

about an organisations security when enforces policies. It is an extension of previous work where they defined a framework for automatic security analysis of socio-technical physical systems [77]. The automatic security analysis has links with the work we present in Chapter 4 as a natural next step of defining the semantics of the rule based model would for assessing social influences.

3.2 Security Aware Organisational Culture

One intuition for a person whose compliance attitude falls under the broad definition of being security aware is that they are able to assess their organisations current risk environment. Another intuition is that they are fully aware of the costs, benefits and risks of complying with security policies [15].

Two possibilities, which is not an exhaustive list, for a person that is security aware is that they a) know about and comply with security policies and b) know an adversary may attempt to socially engineer them. For option b), if an employee is at risk of being targeted by social engineering, should the employee be as vigilant towards countering such a threat? A question such as this raises an aspect of accountability. An option to defend against this would be to design a work environment supporting the employee to not succumb to a social engineer [108].

For this thesis, the next stage of security awareness is to discuss influencing security awareness. Bullee *et al.* demonstrated that by the use of a priming social influence as an intervention they could decrease the percentage of people that would hand over their keys on request at an organisation [22]. In Chapter 7 we demonstrate interventions by social influences or restricting the observations of agents and the impact this has towards their compliance attitudes.

Evaluating behaviour with regards to security policies has been noted to be similar to conducting performance appraisals [111]. Where a performance appraisal analyses behaviour with regards to some targets or outcomes, a *compliance appraisal* would analyse behaviour with regards to the security policies. In essence, it would be a qualitative/quantitative overview of how much an employee is compliant with security policies. In this method,

it would heavily rely upon the current compliance attitude of the employee and may only provide a snapshot of how they behave.

3.3 User Studies Influencing Compliance Behaviour

In Chapter 7 we collect a data set which is the observation of compliance behaviour towards a security policy where participants in the study must swipe their smart card on entry and exit of a room. We designed the study using the Cyber Security Room at Newcastle University. We provide an overview of related work to the study that we performed.

Groß and Coopamootoo demonstrated that strong cognitive depletion impacted a person's ability to choose a strong password which indicated that cognitive effort is required to choose strong passwords. If a rule in a security policy is to *choose a strong password* and person is influenced to perform a task that will cognitively deplete them, then chooses a weak password for a different task, they have been socially influenced. This approach provides a wider scope for how the simulation tool PCASP could be enhanced, by providing cognitive effort towards tasks and the impact this has on password selection.

Compliance behaviour for security policies can also be present for information security policies. One body of work considered that theory-based training achieved positive results and was practical to deploy when dealing with information security policies [64]. In this thesis, the security policies we deal with require people to perform some physical behaviour and would not usually be information security policies. Nevertheless, the theory-based training is one possibility to assist with improving the compliance attitudes of people for security policies where they have a choice. Extending the concept of theory-based training to the application of social influences is something that we do not consider in this thesis, however, it would provide useful material to the holistic approach we assess, as training self-awareness could reduce the negative impact social influences can have on compliance attitudes.

3.4 Nudging in Security

The majority of the work in nudging towards information security falls under the RISCS which is Research Institute in Science of Cyber Security in the UK [5]. A central part of this thesis is focused around the influencing of people's compliance attitudes. In general, we are discussing influences that come from other people and where the providers and receivers of the influence can be unaware. Of course, in the context where an adversary and or defender is present, then self-awareness of influencing behaviour would be true.

Nudging for positive security was demonstrated by Turland et al. where they present a first version application promoting the choice of secure wireless network options by use of colours (green for good, red for bad), and the application targeted users who are unfamiliar with the wireless networks available in their area. The purpose was to influence users to make a secure choice when selecting which wireless network to connect to [107]. The research was taken to the next step in the form of validation. The notion of influence power is introduced by Yevseyeva et al. to characterize the extent to which an influencer can influence decision makers and they illustrated their approach using data from the controlled experiment on techniques to influence which public wireless network users select [116]. In this thesis we don't explore the notion of influence power, however, we acknowledge that this would be potential avenue for a research question in regards to simulation and validation.

3.5 Simulation Tools for Human Security Behaviour

In this thesis we build the simulation tool PCASP catering for the prediction of compliance attitudes towards security policies. The tool makes use of different behavioural elements that is built from the semantics of the rule based model in Chapter 4.

One part of the PCASP uses compliance attitudes to inform agent's actions. A similar approach was developed by Kothari et al. where they presented *DASH*, an agent-based simulation framework that supports the dual-process model of cognition, reactive planning, modeling of human deficiencies (e.g., fatigue, frustration), and multi-agent interactions [73]. The approach of *DASH* focused towards password policies and how agents responded to it and did not consider any observations of agent actions in their framework. Nevertheless,

the approach for the cognition modelling certainly has similarities towards the compliance attitude driven actions in PCASP.

Chapter 4

Influence Tokens

This chapter address Research Question 1 by focusing on the modelling of social influences where propagation occurs. A social influence is when a person's attitudes or behaviours are affected by others [33]. As a reminder the research question is:

Research Question 1 (Cyclic Observational Behaviour)**.** *How do we model observations of compliance behaviour?*

To address the research question, this chapter considers the impact and success an adversary could have using different social influences. We use the MINDSPACE framework to inform us about different social influences as it provides a collection of different social influences [44]. For a more comprehensive overview of MINDSPACE, please see Chapter 2. The MINDPSACE framework is not a complete guide to all social influences, however, it does provide enough richness when considering the complexity of the problem. When we refer to complexity we mean that it is computationally intensive to solve.

The notion of propagation or *ripple* for social influences has similarities to how malware propagates. For the malware, it embeds itself, makes a copy and distributes itself to local neighbours [99, 54]. For a social influence, it changes the compliance behaviour of a person and subsequently, that person can influence other people that surround them resulting in a social influence that has propagated amongst a group of people.

In this chapter we use Coloured Petri nets to model propagation of social influences. We do this because Coloured Petri nets allows for prescriptive behaviour, such as the decision to comply with a security policy or not. The benefits of this prescriptive behaviour allow us to

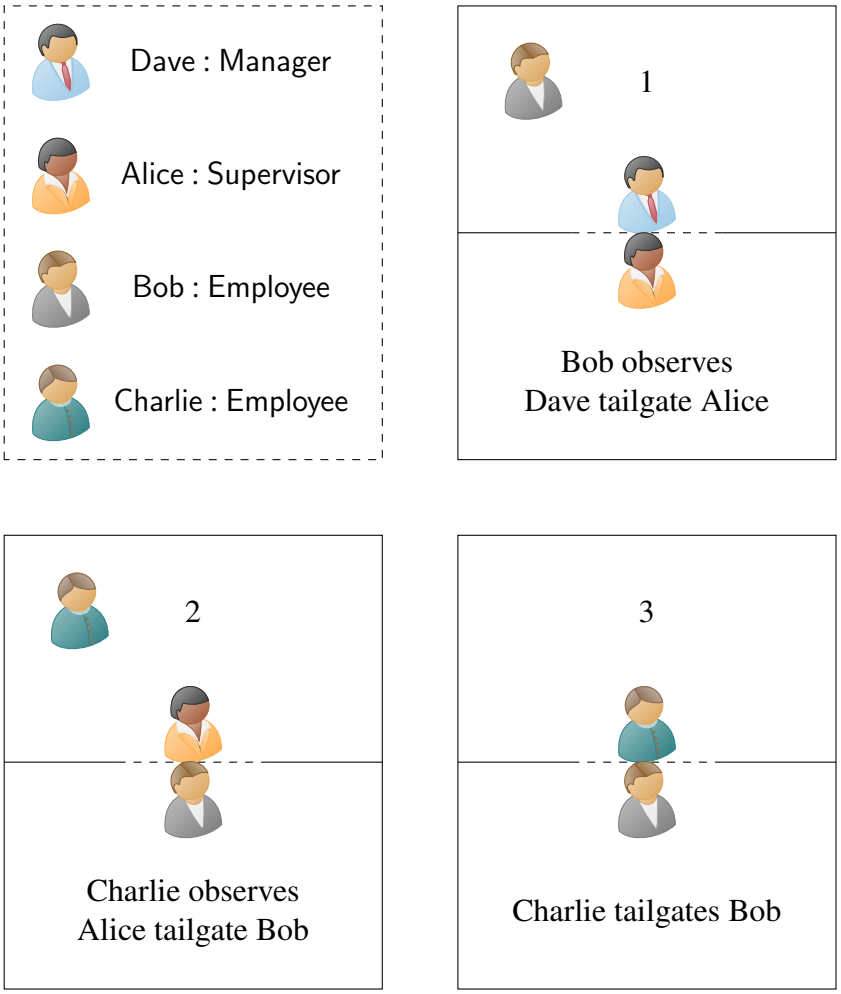


Figure 4.1 Example of Propagating influence. Initially, Bob tailgates Alice which leads to Charlie tailgating Bob.

understand the impact an adversary can have towards a security policy based on a given set of behaviours. One hope of future work in this area would be to understand how to contain social influences by addressing the opportunities for a given set of behaviours.

We do not consider models that express propagation such as the modelling of infectious diseases. Our reasoning is based on the assumptions that these models make are generally tailored towards large populations. Additional assumptions tend to generalise the rate of physical contact between individuals in sub groups [42]. Our interests are within the medium-large organisations where the populations level can be small and people may only make virtual contact (e.g. by phone or email).

To give an intuition of the problem, consider the following example. An adversary influences the compliance attitude of one person at an organisation to be more likely to allow tailgating into the main office building. The behaviour of this one person relaxes the compliance attitude of many at the organisation. The global compliance attitude descends as employees regularly don't challenge anyone tailgating. An adversary can easily tailgate into the building without being challenged to take equipment. We use Figure 4.1 as an abstract view of the problem of propagation between compliance behaviours of employees and the influences this can have on others. In Figure 4.1 the behaviour of Dave tailgating Alice influences the behaviour of Bob and as a result, Charlie, who was originally not present ends up tailgating Bob.

Chapter Overview: This chapter is mainly built from a previous publication at the STAST (Socio-Technical Aspects in Security and Trust) workshop in 2016 [24]. The paper is a more condensed form of this chapter. The chapter features in addition to the paper a discussion around the introduction of a concept called Influence Tokens and a more concise introduction to Petri nets in general.

4.1 Influence Tokens

A token in a Petri net is representation of some unique element which has its own set of rules. For instance, a person/agent is a token that can move between locations and carry or distribute other tokens. These other tokens are Influence Tokens which is a term we define to capture a special set of tokens that will be the focus of this chapter. An influence token is a representation of a social influence. Recall the MINDSPACE framework, a framework of social influences. An influence token can take the form of these social influences, in this chapter they are:

1. **Priming** a target can change their compliance attitude, such that at a later point in time, their compliance behaviour can be exploited. As a reminder, *priming* is to prepare someone for some behaviour in the future. It is often associated with being subconscious driven behaviour, i.e. it occurs automatically [45].

2. Use of the **Messenger** influence to exploit the current compliance attitude of a target. A person susceptible to the *Messenger* influence believes the adversary has a right to enforce a form of obedience [84].

A person can carry an influence token and distribute it to others. Depending on the policy for the influence token, it can exhibit many different behaviours. For example, a priming token can propagate and influence many people. Whereas, a messenger token relies on the provider of the token to be perceived as the relevant authority. The impact of an influence token can strongly depend on the context and the compliance attitude of the person receiving the token.

4.1.1 The Modelling of Influence Tokens

The notion of influence tokens indicates a distributed system where many people can carry an influence token. In this context, a distributed system refers to many people interacting with each other.

A Coloured Petri net is a mathematical technique for modelling a system that is distributed [66]. We adopt it to formally represent the notion of influence tokens. It allows for a number of different types of objects to be represented. In our case, human behaviour requires us to not only model the influence, but the person as well.

To express our modelling of Influence Tokens in this chapter we use a running example based on tailgating:

Scenario: *In a building where employees must swipe a smart card to gain entry, we assume people can tailgate to get inside. We assume the default behaviour of employees is to challenge tailgaters and to not allow them entry. A social engineer wears maintenance clothing (Messenger), can try and tailgate, as another measure they can email (Priming) ahead and interact with different departments to notify them that some maintenance work will take place. We make a final assumption that employees are also under pressure to make a deadline and are prone to becoming stressed out which impacts their decision making.*

The goal of the adversary is to gain access to the building. The adversary has two measures of influence they can deploy; the maintenance clothing (Messenger) and the emails (Priming). The dynamic behaviour of the employees is outside of the adversaries control, however, it affects the state of the employees and in turn, can impact if the social engineer is successful. The adversary in this sense is a risk based opportunist as they have no knowledge about the compliance attitude of the person they are tailgating, this is hidden information.

We have chosen to use Messenger and Priming in this chapter as they are the social influences illustrated in the scenario. During the course of this thesis we address different social influences and whilst we are prescriptive about what those social influences are, it's important to ensure they align with the narrative. Without the descriptors in the narrative, we have no grounding for justifying why a social influence is appearing in our higher level discussions and analysis.

4.2 Modelling Influence Tokens with Coloured Petri nets

In this section we provide the definitions for Coloured Petri nets, adapt the verification techniques from Coloured Petri nets towards our problem and then we introduce the running example that we use throughout this chapter. It is important to note that Coloured Petri nets make use of tokens as a system evolves, which are not necessarily influence tokens. We also provide an overview of Petri nets in Chapter 2.

4.2.1 Coloured Petri Nets

A Coloured Petri net is an extension of Petri nets which can be used to analyse a distributed system providing verification for properties and simulations to identify behaviour trends. Intuitively speaking, a Petri net is a system where some tokens flow from places to transitions, and from transitions to places. The main goal of a Petri net is to analyse if, given an initial distribution of tokens (which is often referred to as a marking), a specific distribution of tokens can eventually be reached. A Coloured Petri net is a Petri net where tokens can be of different colours, which allows for a greater range of modelling concepts to be taken in to account. Different definitions of Coloured Petri nets exist in the literature, and in order to avoid confusion, we repeat the notion of Coloured Petri nets we use here, inspired from [66].

Definition 1 (Coloured Petri net)**.** *A Coloured Petri net (Coloured Petri net) is a tuple* $ColouredPetrinet = (\Sigma, P, T, A, C, N)$ *where:*

1. $\Sigma = \{\mathscr{T}_1, \ldots, \mathscr{T}_n\}$ *is a finite set of finite and non-empty colour sets, where each colour set represents a specific type of tokens. In the following, we write* $\mathscr{T} = \mathscr{T}_1 \cup \cdots \cup \mathscr{T}_n$ *for the set of all possible tokens.*
2. P *is a finite set of places.*
3. T *is a finite set of transitions.*
4. A *is a finite set of arcs.*
5. $C : P \to \Sigma$ *is a colour function and maps a place to its assigned colour set.*
6. $N : A \to P \times T \cup T \times P$ *is the node function.*

From the node function in Definition 1, we know that a place in a Coloured Petri net can only be connected to a transition, and a transition can only be connected to a place. We see this in Figure 4.2 which illustrates a Coloured Petri net with only two colours used for places where tokens can reside. A token beginning in a red place must end in a red place. A token beginning in a blue place must end in a blue place.

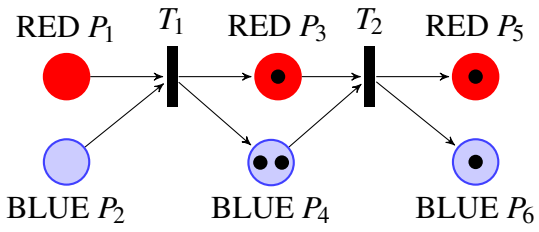


Figure 4.2 A Petri net showing places connected to transitions. A marking here would take the form of $\{0,0,1,2,1,1\}$ which indicates the distribution of tokens for the places $P_1, P_2, P_3, P_4, P_5, P_6$.

In Figure 4.2, places are depicted with circles, transitions as long narrow rectangles and arcs as unidirectional to show connections between places to transitions or transitions to places.

4.2.2 Motivation for Using Petri Nets

Petri nets are powerful tools to analyse distributed systems. Petri nets provides a user with a graphical notation which helps to illustrate any processes which are being modelled. Furthermore, Petri nets generally have tool support allowing for modelling and analysis.

In this chapter we make use of CPNtools which provides a user with the ability to edit, simulate and analyse Coloured Petri nets. The tool provides model checking through state based methods[67]. We make use of the tool to explore our running example to express influence tokens.

4.2.3 Behaviour of Coloured Petri nets

We now cover the behaviour of a Coloured Petri net, where we define how transitions execute. At a high level view, the arcs directed into transitions are the pre-conditions for the transition to execute, where a pre-condition indicates that a token must exist in the place that the arc originates from. A post condition of the same transition is all places where the transition is directly connected via an arc, it indicates where tokens will go once the transition has executed. This sequence allows for tokens to migrate throughout a Coloured Petri net. In order to express Coloured Petri nets, we first recall the notion of multi-sets.

Definition 2 (Multi-sets)**.** *A multi-set m, over a non-empty and finite set S, is a function* $m : S \to \mathbb{N}$ *where the integer value is non-negative and represents the number of appearances of the element s in the multi set m.*

A marking is a distribution of tokens over a Coloured Petri net, and represents the state that the Coloured Petri net is currently in. A token can only be in a place, not a transition. The transition serves to migrate tokens between places via the arcs.

Definition 3 (Marking)**.** *A marking is a function* $M : P \to m(\mathscr{T})$ *such that:*

$$\forall p \in P \, \forall \tau \in \mathscr{T} \; M(p,\tau) > 0 \Rightarrow \tau \in C(p)$$

where we used the curried notation for the function M, i.e., $M(p,t)$ refers to $m(t)$ where $m = M(p)$. In addition, when no confusion can arise, we write $M(p)$ for the number of all tokens in place p, i.e., $\sum_{\tau \in C(p)} M(p,\tau)$.

A Coloured Petri net will have an initial marking M_0, which is formed by distributing the tokens over the net. The main semantic aspect of a Coloured Petri net is that, in order for a transition to be enabled, there should be enough tokens in the places that are connected to that transition. In particular, if a place is connected n times to a transition, then there must be n tokens in that place for that transition to be enabled. For the sake of exposition, we write $in(p,t)$ for the number of arcs from the place p to the transition t, and $out(t,p)$ for the number of arcs from the transition t to the place p. More formally, for any place p and any transition t:

$$in(p,t) = |\{a \in A \mid N(a) = (p,t)\}|$$
$$out(t,p) = |\{a \in A \mid N(a) = (t,p)\}|$$

Definition 4. *A transition t is said to be enabled in a marking M, and in which case we write $M \vdash t$, if, and only if:*[1]

$$\forall p \in P\ (\exists a \in A\ N(a) = (p,t))) \Rightarrow M(p) \geq in(p,t)$$

Once a transition is enabled, it is capable of occurring. In its simplest form the occurrence of a transition changes the current marking M to M'. Intuitively speaking, the value of M' is calculated by moving across the transition the tokens that were present in the places connected to the transition. More formally, we say that the marking M' can be obtained from the marking M with the transition t, and in this case we write $M \rightsquigarrow_t M'$ if, and only if the transition t is enabled:

$$M'(p) = \begin{cases} M(p) - in(p,t) & \exists a \in A\ N(a) = (t,p) \wedge M \vdash t \\ M(p) + out(t,p) & \exists a \in A\ N(a) = (p,t) \wedge M \vdash t \\ M(p) & \text{otherwise.} \end{cases}$$

Finally, the reachability graph from a marking M indicates all the markings that can be eventually reached from M by triggering enabled transitions. More formally, given a marking

[1] For the sake of exposition, we use here a simplified version for the notion of enabled transition, which is enough to present our model for influence tokens. We refer to [66] for further details about the semantics of Coloured Petri nets.

M, we define inductively the reachability graph $R_n(M)$ as follows:

$$R_0(M) = \{M\}$$
$$R_{n+1}(M) = \{M' \mid \exists M'' \in R_n(M)\ \exists t \in T\ M'' \vdash t \wedge M'' \rightsquigarrow_t M'\}$$

In the following, we write $R(M)$ to express the unbounded reachability graph of a given marking M, i.e., $R(M) = \bigcup_n R_n(M)$.

4.2.4 Coloured Petri nets Verification

From a practical perspective, if a state of the system or marking in our case exists which is of interest, then we want to verify that it can exist. An example of something of interest would be a specific marking indicating that the model is in an bad state.

From a resilience aspect, a Coloured Petri net should not violate any rules set out. We can do this by deadlock detection to identify how and if a Coloured Petri net deadlocks. As we are modelling people as agents, a property of interest is that we have no duplication. For example, if the model begins with n agents, at each marking, n agents will always be present.

The reachability property states that given an initial marking M, is a marking M' is part of the reachability graph of a Coloured Petri net. The appeal of this is apparent as it may be the case that we want to deduce the current state of a set of agents compared to the adversary to represent some form of distance between the two.

The boundedness property states that a place in a Coloured Petri net is *k-bounded* if it can contain no more than *k-tokens*, this is a property of interest as we do not want to begin with one adversary and somehow finish with two, unless we consider a case where a cognitive agent becomes an insider and an adversary is created within the net. Unless we consider the case where specific behaviour traits can cause agents to become insider threats, for example, the specific composition of an agent leads their compliance attitude to not only become non-compliant but malicious [90].

We define the upper and lower bounds of a Coloured Petri net, where given an initial marking it refers to the number of tokens that exist for a given marking.

Definition 5. *Let a place $p \in P$, a non-negative integer $n_1, n_2 \in \mathbb{N}$ and all the reachable markings from M_0 be $R(M_0)$, such that:*

- *Upper Bounds:* $\forall M \in R(M_0) : M(p) \leq n_1$
- *Lower Bounds:* $\forall M \in R(M_0) : M(p) \geq n_2$

For our models, $n_1 \equiv n_2$ would indicate that the number of tokens does not change in any Coloured Petri net. Whether or not this consistency check should hold in our models depends on the tokens that we are considering. For instance, the type of agent tokens should always remain constant, however, the number of influence tokens may change as an influence propagates through agents.

4.3 Influence Model

A Coloured Petri net is capable of having places which are bespoke to holding specific types of coloured tokens. For example, given a red token, it could only exist in a place designated to hold red tokens[2]. A place can hold tokens of different colours if it is assigned to do so. In this section we introduce the core concepts of how we use Coloured Petri nets to represent social influences, in particular the focus on defining our different types of coloured tokens.

The modelling approach expresses three core aspects which are the behaviour of influence tokens, the adversary and the agents. The influence tokens are assigned to the adversary as a special set of tokens. The adversary must have the capability to distribute influence tokens to agents by a set of behaviours, which are captured through transitions. The agents have the capability to interact with the model, other agents and the adversary.

We pre-define the set of behaviours for influence tokens, the adversary and agents. The influence model described in this chapter captures the three core aspects of our modelling approach. Firstly, the priming and messenger influence tokens are assigned to the adversary. Secondly, the adversary gas the capability to influence with the priming or the messenger token. Thirdly, the agents interact with each other and the adversary. By addressing the three core concepts for our modelling approach, the behaviour of influence tokens can be illustrated as propagating between agents.

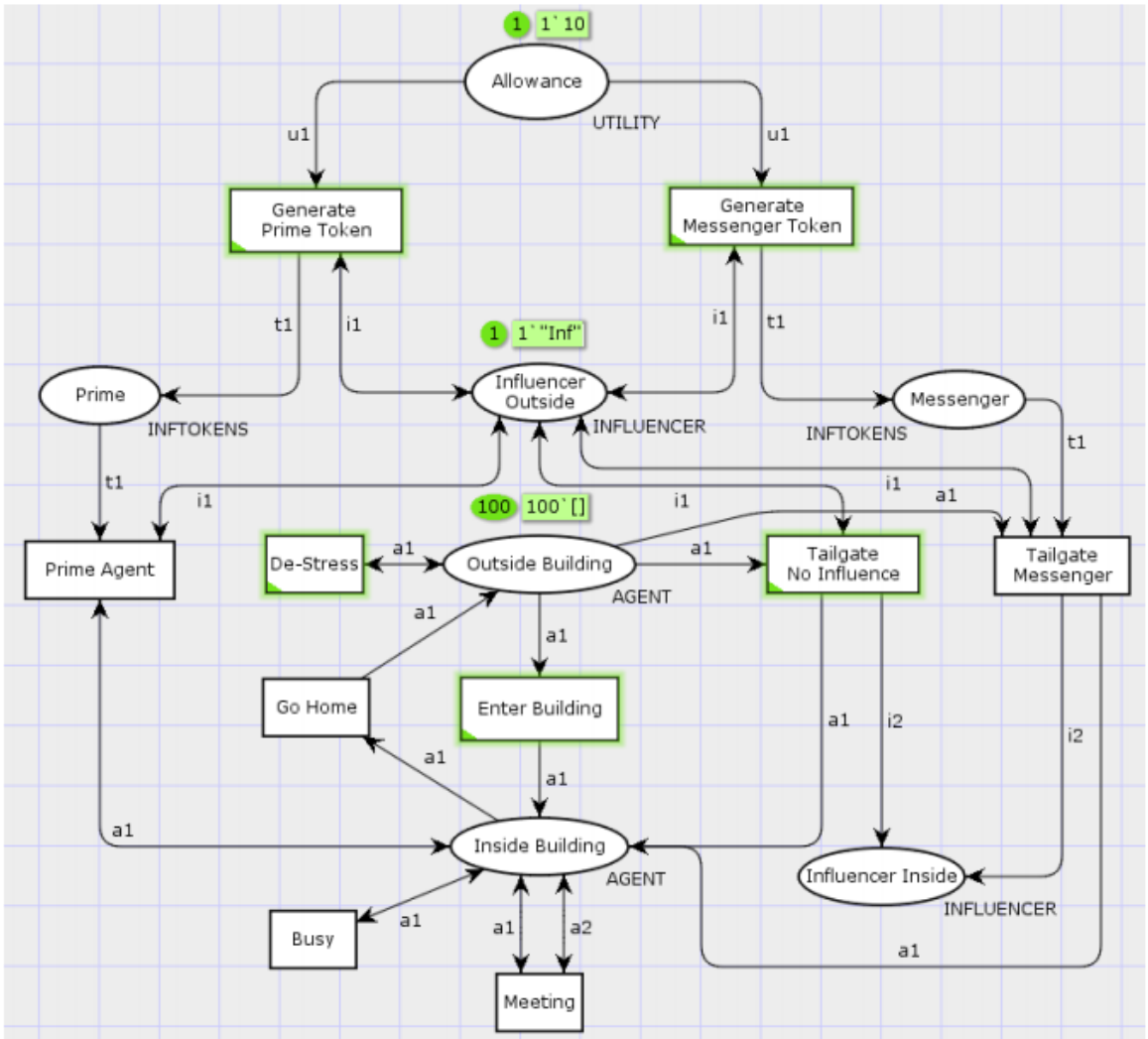


Figure 4.3 Running Example: Adversary working towards tailgating with a finite number of influence tokens available.

4.3.1 Coloured Petri net Influence Model Definitions

Applying our model to the Coloured Petri net, we now explain our meaning of colour sets, places, transitions, arcs etc. As a reminder of the notation for the Coloured Petri net we provide our implementation of the Coloured Petri net with regards to the following:

1. Σ is a finite set of colour sets.
2. P is a finite set of places
3. T is a finite set of transitions
4. A is a finite set of arcs
5. $C : P \rightarrow \Sigma$ is a colour function
6. $N : A \rightarrow P \times T \cup T \times P$ is the node function;

For each part of the Coloured Petri net, we define our implementation. With regards to the tailgating running example where agents represent people, the social engineer is an influencer that has a pre-defined finite set of influence tokens. We illustrate the running example in Figure 4.3[3].

1: Colour sets is a term used to capture different types of tokens, the set of colour sets $\Sigma = \{\lambda, \gamma, A_\lambda, \Phi, U, Env\}$ where:

- λ is the set of behaviour change elements.
- γ is the set of influence tokens and $\gamma \subseteq \lambda$.
- A_λ is the set of agents and $A_\lambda \subseteq \mathbb{P}(\lambda)$.
- Φ is the set of influencers.
- U is the set of utility values available to influencers.
- Env is the set of environments.

Table 4.1 Mapping of Places

Type	Example Scenario	Text Description
Place	Influencer Outside	InfOut
Place	Influencer Inside	InfIn
Place	Prime	Prime
Place	Messenger	Messenger
Place	Agent Outside	AgentOut
Place	Agent Inside	AgentIn
Transition	Tailgate Messenger	TailgateM

2, 3 & 4: To assist the reader, we map places and transitions in the example figure to the places described in this section in Table 4.1. Places P within Coloured Petri nets can store tokens, where the tokens can have a data value attached to them. An example place $\mathsf{InfOut} \in \mathsf{P}$ is a place within the net and refers to a physical location. For other places such as $\mathsf{Prime} \in \mathsf{P}$ and $\mathsf{Messenger} \in \mathsf{P}$ which is of type γ, the place still refers to a physical location but in relation to the place of the influencer. In the InfOut place, the adversary has access to tokens in both Prime & Messenger. Once the token representing the adversary changes place to $\mathsf{InfIn} \in \mathsf{P}$ then access to these social influence tokens is no longer possible.

The set of places must always be finite. Transitions are a finite set and capture how the models markings change, a transition has an input, output, some internal function and can be constrained with a guard to ensure it can only be enabled once the guard is satisfied. An example transition is $\mathsf{TailgateM} \in \mathsf{T}$ which is connected by arcs to many places. The transition TailgateM has multiple places connected to, the input places are $\mathsf{InfOut} \in \mathsf{P}$, $\mathsf{AgentOut} \in \mathsf{P}$ and $\mathsf{Messenger} \in \mathsf{P}$. The output places are InfOut, AgentIn and InfIn. A token must therefore, be in the three input places in order for the transition TailgateM to be enabled, assuming the guard is satisfied.

5: The colour function C maps a place to its respective colour set. Therefore, for all tokens in a place p, the token must have a colour set which belongs to $C(p)$. For example, $C(\mathsf{InfOut}) = \Phi$ and indicates that any token in InfOut must be of the colour set Φ, that is it

[2]We don't use *red* tokens, this is just provided as an example.

[3]The arcs here are bidirectional, this is breaching the rules previously set out, however, we only do this for illustrative purposes.

must be an influencer.

6: The node function maps each arc into a pair where the pair is of type $\mathsf{P} \times \mathsf{T} \cup \mathsf{T} \times \mathsf{P}$ where the first element is the source node of the arc and the second element is the destination node. Several arcs can exist between the same ordered pair of nodes. In the running example the places $\mathsf{InfOut} \in \mathsf{P}$ and $\mathsf{InfIn} \in \mathsf{P}$ and transition $\mathsf{TailgateM} \in \mathsf{T}$ are matched by many nodes. Consider the arcs which connect them $i_1, i_2 \in \mathsf{A}$, where:

$$N(i_1) = ((\mathsf{InfOut}, \mathsf{TailgateM}), (\top, \bot)) \tag{4.1}$$

$$N(i_2) = ((\mathsf{InfOut}, \mathsf{TailgateM}), (\top, \mathsf{InfIn})) \tag{4.2}$$

The use of $\top$ refers to the same transition already used and $\bot$ is for the same place used. The two arcs i_1 and i_2 come from the same place but i_1 returns to InfOut and i_2 goes to InfIn. This allows for a transition to send tokens to either place but also to control when the token goes to a specific place. If the adversary fails at tailgating then they should not end up inside the building, similarly if they succeed they should not remain outside. An arc should only be unidirectional, however, we breach this convention for simplicity. An arc which is bidirectional simply refers to two unidirectional arcs.

4.3.2 Model Behaviour

Implementing the example into a Coloured Petri net provides us with the colour sets outlined in Section 4.3.1 where we have agents and each agent is a set of behaviour change elements. An influencer with some finite value can generate influence tokens to distribute amongst agents and use tailgating to achieve their goal and gain access to the building.

Figure 4.3 illustrates the example implemented in CPNTools, which is the tool used to implement the running example [67]. The ovals refer to the places, the rectangles are for transitions and the arcs connect places to transitions. In Figure 4.3 the colour sets are captured by INFTOKENS for γ, AGENT for A_γ, UTILITY for U and INFLUENCER for Φ. This was from the restrictions CPNTools provided for naming colour sets. In Figure 4.3, each place is assigned its respective colour set which can be identified near the place, such as the place *Influencer Outside* which is captured by INFLUENCER. The transitions mostly

have an internal algorithm which deals with the input arcs and produces the appropriate output arcs. The transition *Tailgate Agent* takes tokens of a colour set (a_1, i_1) which refers to one agent and one influencer respectively. It can output (a_2, i_2, i_3) where (i_2, i_3) are both influencers and refers to whether or not the adversary as an influencer managed to tailgate into the building. The transition TailgateM has an internal function that takes an agent, an influencer and outputs an agent, an influencer and another influencer. For clarity we shorten TailgateM to $\mathsf{TA_{Msg}}$ as the transition requires the use of the messenger token. Note that in the transition TA an influence token is required, however, we do not include it in the function definition as the token is only required to enable and allow the transition to occur, it does not capture the internal function which is the following:

$$TA_{Msg} : A_\gamma \times \Phi \rightarrow A_\gamma \times \Phi \times \Phi$$

$$TA_{Msg}(a,i) = \begin{cases} (a, i, \emptyset) & \mathsf{Stressed} \in a \vee \mathsf{Primed} \in a \\ (a \cup \mathsf{Aware}, \emptyset, i) & \text{otherwise} \end{cases}$$

The return of the function $TA_{Messenger}$ will always return one empty set, where the empty set refers to the success or failure of the adversary, the second element in the tuple indicates a success if it is not an empty set and the same is true for the third element. The adversary also has the option to tailgate without using the messenger token:

$$TA_{Noinfluence} : A_\gamma \times \Phi \rightarrow A_\gamma \times \Phi \times \Phi$$

$$TA_{Noinfluence}(a,i) = \begin{cases} (a, \emptyset, i) & \mathsf{Aware} \in a \\ (a, i, \emptyset) & (\mathsf{Stressed} \in a \vee \mathsf{Primed} \in a) \wedge \mathsf{Aware} \notin a \\ (a \cup \mathsf{Aware}, \emptyset, i) & \text{otherwise} \end{cases}$$

The initial marking of the net is captured by the distribution of the tokens over the places. In this example the initial marking is as follows:

- $Allowance = 1'1000$;

- $InfluencerOutside = 1'10$;
- $AgentOutside = 100'()$;

The notation above is in the format $y'x$ where y refers to the number of tokens for that Colour type, e.g. there is one influencer and x refers to the initial value. For $Allowance$ and $InfluencerOutside$, the values are integers so they must take a numerical format. For the place $AgentOutside$ it is a list, and in CPNtools that is represented by $()$ to show an empty list of elements, in this case strings.

Where the initial marking indicates the influencers allowance is of one thousand. There is only one influencer who is the adversary, indicating that the adversary has ten attempts at tailgating and there are one hundred agents who are outside of the building and are all subject to no behaviour change elements or influence tokens.

Transitions that can occur without the adversary still impact the success rate, even though the adversary has a small impact towards this. An agent that is carrying a *Primed* token, can duplicate the token and allow another agent to carry a *Primed* token. This allows the token to propagate throughout the net, which creates the uncertainty for the adversary, as it is not clear when tailgating, if the agent being tailgated is carrying the *Primed* token.

4.3.3 Token Types & Internal Transition Guards

In this model we exploit two types of functionality that is offered by CPNtools. They are token types and internal transition guards.

Token Types

A type token in CPNtools can be assigned to many different types. In the example we run, we exploit three types which are; integer, string, list. We use the adversaries allowance as a type integer. We use the influence tokens as type strings. Finally, we capture each agent as a list of strings which collects influence tokens as they traverse the Coloured Petri net.

Internal Transition Guards

We use internal guards to capture token behaviour during a transition. An internal guard in CPNtools can permit or deny tokens to progress along from an outgoing arc of a transition

to a place. Therefore, the earlier definition for how a marking updates is true, however, an internal guard can prevent the marking from being updated.

We use these internal guards at decision points in the Coloured Petri net. For example, the decision to permit the adversary to tailgate is dependent upon an agent have the correct influence tokens. The guard handles this query and if the tailgate is successful, the adversary will progress to a new place or they will return along an outward arc from the transition to their starting place. For simplicity, figure 4.3 does not show all of the arcs connecting a transition to a place and vice versa but they are there.

One intuition that a normal Petri net user would have about the CPN in figure 4.3, is that the number of tokens increases as transitions occur. However, this is not the case as the guards prevent an increase of tokens based the queries they check and the results each guard produces.

4.4 Analysis

In this section we provide our analysis where we consider six test case experiments to identify the success rate of the adversary under different conditions. At the end of this section we provide a discussion to summarise our findings.

4.4.1 Analysis: Verification

The two properties we wish to satisfy are the reachability and boundedness property. The use of CPNtools allows us to perform these verification checks to identify if the properties are satisfied. The reachability property to ensure that the adversary can reach their goal, as in a marking $M(InfIn) > 0$ and the boundedness property of the following:

$UpperBounds : \forall M \in R(M) : M(InfIn) \leq 1$
$LowerBounds : \forall M \in R(M) : M(InfIn) \geq 0$

For each test case experiment we ran the verification checks to ensure that the properties were met. The outputs of verification confirmed that the lower and upper bounds were met and that the place was reachable. We consider the same verification for agents and influence

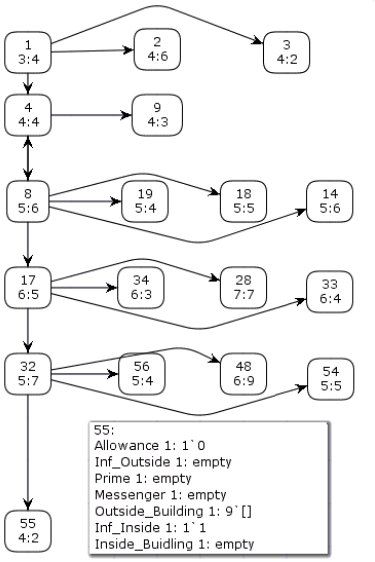


Figure 4.4 Running Example: Marking showing the adversary inside the building.

tokens to ensure that the same properties are true with regards to each type of token in the net. An agent in the place AgIn $\in$ P refers to the physical location of an agent being inside the building. The upper and lower bounds are captured by the following, assuming there are multiple places for agents to transition to and where n is the number of agents:
$UpperBounds : \forall M \in R(M) : M(AgIn) \leq n$
$LowerBounds : \forall M \in R(M) : M(AgIn) \geq 0$

One of the benefits of using CPNtools is the ability to calculate model checking properties. We have discussed the use of the upper and lower bounds for a marking where we want to ensure that a specific place can be reached by a specific token. We can use CPNtools to manually verify this property. We can generate a state space tree illustrating all of the reachable markings from our initial marking. We then step through manually the tree following the markings of interest until we reach a marking showing the desired result.

Figure 4.4 demonstrates the verification of CPNtools. We have constructed a subset of the reachability graph from the initial marking in the running example. We have reduced the number of agents and the number of attempts the adversary has to increase performance of producing the reachability graph.

Each box within Figure 4.4 contains the three numbers. The first number is the number of the node. The two numbers at the bottom of the node are the number of predecessors for that node and the number of successors that have been calculated. In the case of the last box which is $\frac{55}{4:2}$, it is node 55 with 4 predecessors and 2 successors. If we expand the marking of node 55 it will display the marking which shows the distribution of tokens across the Coloured Petri net. Node 55 in Figure 4.4 states that the place Inf_Inside contains one token. That one token refers to the adversary who has successfully entered the building.

4.4.2 Influence Tokens and Propagation: Test Case Experiments (1-4)

In order to quantitatively analyse the performance of the adversary in the running example, we compare the success of the adversary reaching their goal based on four test cases. We consider a range of agents within the net and a range of utility values for the adversary. When we calculate the success rate of the adversary, we measure it between 0 and 1 where 1 is a guaranteed success for the adversary and 0 is the adversary failing. For each trace in the simulation, the adversary will either fail or succeed meaning we calculate a success rate based on the following:

$$SuccessRate = \frac{successes}{runs} \tag{4.3}$$

We are only considering the two influence tokens of messenger and priming, the first four test case experiments are the following:

1. An adversary with no allowance and therefore, no influence tokens and must try to tailgate without any influencing:

2. Considering the messenger token where the adversary does not invest into the priming token:

3. Considering the priming token where the adversary does not invest into the messenger token:

4. Influence token propagation, the likelihood that the adversary can tailgate after allowing the priming token to propagate.

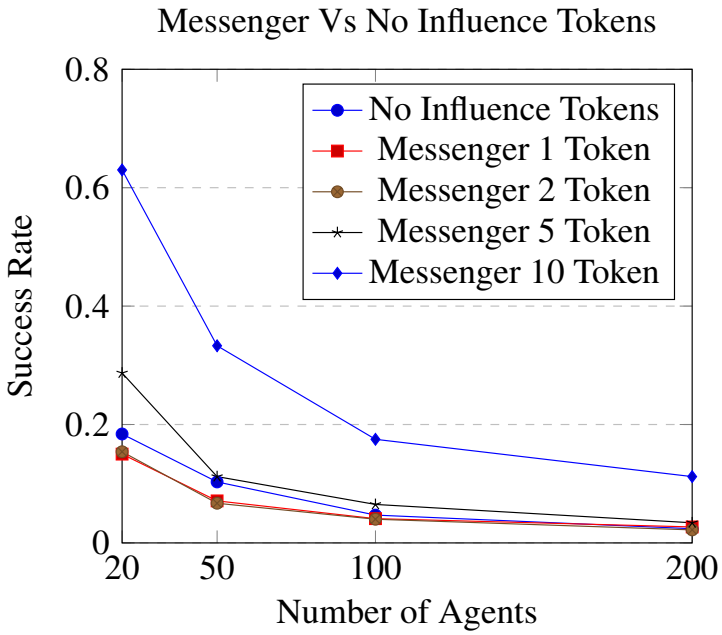


Figure 4.5 The Messenger token used in the running example compared to using no influence tokens. The difference in the impact of the number of tokens is clear when ten *Messenger* tokens are used.

We are only considering these two influence tokens within this context. The inclusion of the current two influences is to show the comparisons that two influence tokens can have within our model, where the priming token can propagate from agent to agent and the messenger token, where it is more likely for someone to be convinced by the one delivering the content.

The adversary was allocated ten attempts to tailgate into the building before stopping the attack. The following experiments were simulated under that constraint, to ensure that the adversary does not keep tailgating until they succeed.

Experiment 1 - No Influence Tokens: In Figure 4.5 and 4.6 the use of no influence tokens showed that the adversary in this context of tailgating has a greater likelihood of achieving their goal when fewer agents exist. This is mainly down to the constraint that the adversary can only tailgate successfully if an agent is stressed or primed. As the priming token is not used in this experiment, the adversary must wait until an agent is stressed to gain entry. As this is a behaviour change element out of their control, the adversary must risks uncertainty by not knowing the compliance attitude of the agent.

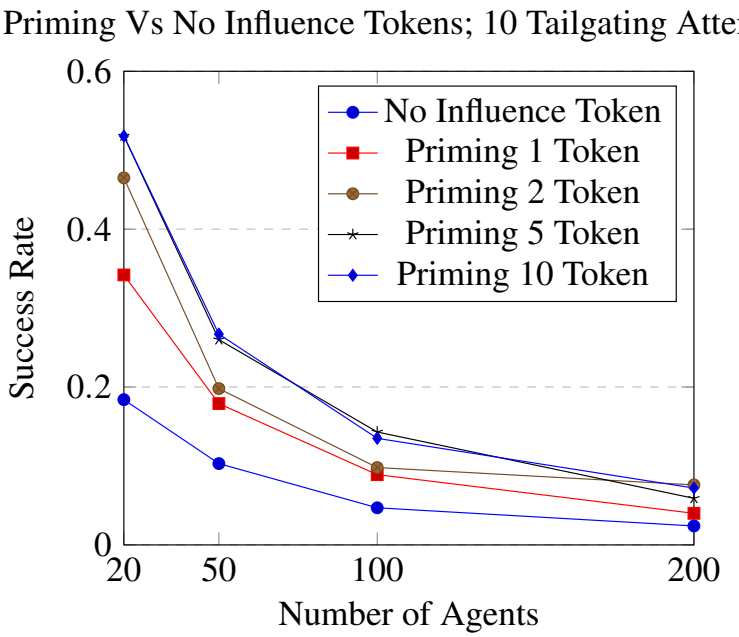


Figure 4.6 The priming token used in the running example compared to using no influence tokens. Without the priming token being given the opportunity to propagate throughout the agents, the strength of the token becomes limited where using five tokens is no different than using ten.

Experiment 2 - Messenger Token Figure 4.5 shows the use of the messenger token. Where the adversary is still restricted to ten attempts of tailgating. The use of the messenger token over twenty agents with a utility value of ten tokens yields the greatest likelihood for the adversary with a value of .63. The messenger token initially reduces the success rate when compared against no influence tokens when only one messenger token is used, indicating that investing a small amount into this influence token would be counter intuitive in this context.

Experiment 3 - Priming Token: The use of the priming token requires that an agent is primed, then at a later point a transition occurs which triggers the prime, in this case the trigger is the tailgating. Figure 4.6 shows the results of using priming based on a utility value of deploying a certain number of priming tokens. The prime improves the success rate of the adversary but does not continue to scale it consistently for more primes. Using five or ten primes has very little difference on the outcome an adversary will have in this scenario. The point of the prime is to pass an influence token into place for later use where the token can propagate amongst agents. The obvious weakness in this experiment is in the purpose

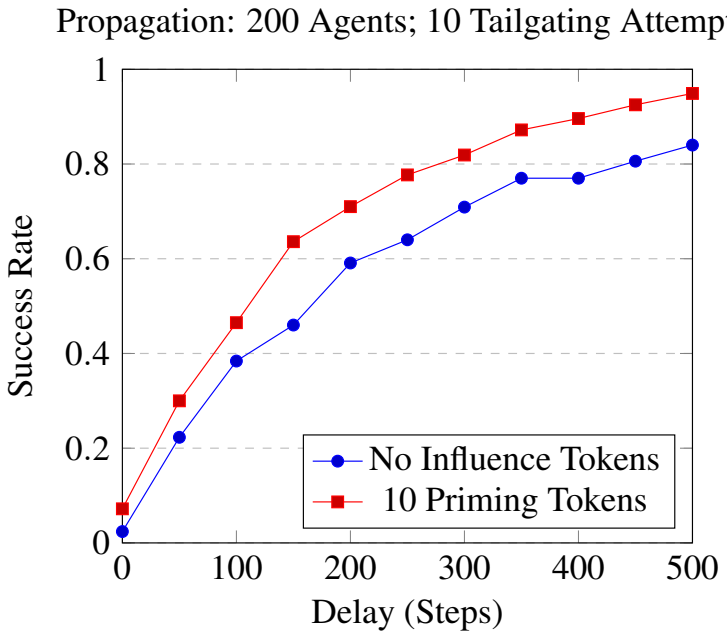


Figure 4.7 The priming token used in the running example compared to using no influence tokens with ten attempt at tailgating. This shows how ten attempts can negate the impact of the priming token. Because the adversary has so many attempts, the increase in success rate because of priming is almost insignificant

of the prime, which is to propagate throughout the net. By not delaying the adversaries attacks, the prime has no chance to circulate amongst the agents and therefore, the adversary is sometimes tailgating with no influence tokens. This creates a clear strength for the initial use of the messenger token.

When simulating the Coloured Petri net the questions asked so far refer to simulating from the initial marking and identifying the chances for a specific result to occur over a number of simulation runs.

Constrained transitions for a number of steps can open up a new range of attacks where we consider propagation. By propagation we mean the transfer and duplication of behaviour change elements within the net. The priming token in the running example is a propagating token. In the example, once an agent is primed, the transition *Meeting* can cause the priming token to duplicate and be passed from agent to agent.

If the influencer waits after priming an agent, they can allow the token to propagate throughout the net and perhaps improve their success rate.

Experiment 4 - Priming Token Propagation (Part 1): Figure 4.7 shows the success rate of the adversary after delaying their attack when two hundred agents are in the initial marking. As the initial marking of this net includes agents who are not subject to any behaviour change element then the more the net is allowed to run, the more likely it is that the distribution of agents are subject some behaviour change that is outside of the adversaries control. Therefore, the use of no influence tokens and just waiting for some time increases the success rate. The use of ten priming tokens shows a clear difference against the use of no influence tokens as the success rate after 500 steps when using the priming tokens is 94.9% compared to 84% when using no influences.

Experiment 4 - Priming Token Propagation (Part 2): Figure 4.8 shows the success rate of the adversary after delaying their attack when one hundred agents are in the initial marking. Initially, a wait of one hundred steps yields very little difference but as the adversary delays more and more, the success rate improves significantly when using the priming token. It shows the propagation that an influence token has throughout a system over time. We only considered 1000 simulation runs here and at a delay of 800 the token has already propagated enough to provide a 63% success rate for the adversary.

4.4.3 Combination of Influence Tokens (Test Case Experiments 5 & 6)

We have considered influence tokens independently, the separation between messenger and priming has shown that sometimes it is better to use one token over another, or that investing more in an influence token yields no further reward for an influencer. Furthermore, we have considered the propagation of behaviour change elements in a system, one which considered no influence and out of bounds for control of an influencer and the other where priming was introduced and left to grow and propagate amongst agents.

Now, we will consider the combination of influence tokens, allowing the influencer to have a range of priming and messenger tokens at their disposal to quantify the strength of exploiting these tokens. We deploy two further experiments:

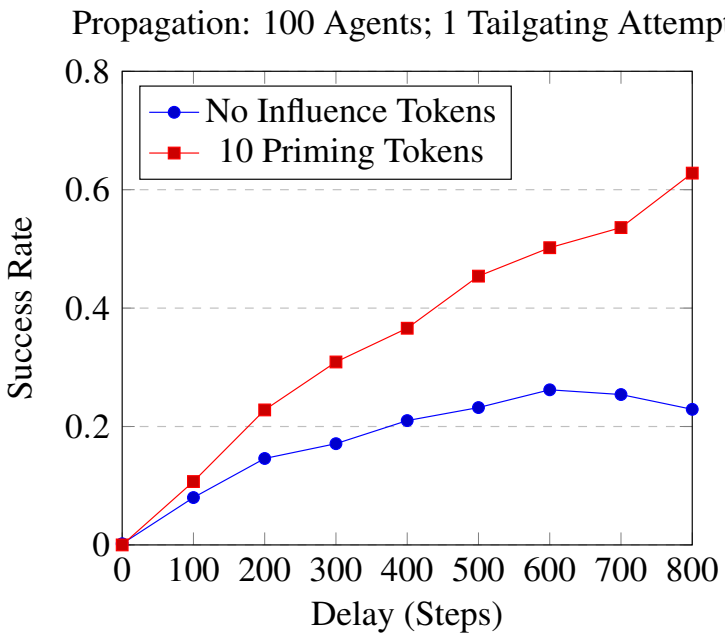


Figure 4.8 The priming token used in the running example compared to using no influence tokens with only one attempt at tailgating. This shows the true power of the *Priming* token when only one attempt at tailgating is provided. The success rate is proportional to the delay, where, the longer the delay, the better the success rate.

5. This experiment will consider the combination of five priming and five messenger tokens with no delay to identify the impact they have with a maximum of ten attempts over a range of agents.

6. The last experiment will consider the same combination of influence tokens as experiment 5, however, their will be a delay on the use of the messenger token to allow the priming token to propagate where one hundred agents are present.

Experiment 5: The purpose of this experiment is to identify the impact of a combination of influence tokens towards the success rate of an adversary. Figure 4.9 shows that the combination of tokens yields a strong output for the lack of physical presence required by the adversary. The combination only considers five attempts at tailgating where those five attempts use the messenger token. Although the investment is in ten influence tokens, the number of actual attempts to tailgate is reduced. As this is just a small example, we can see the usefulness of combining influence tokens to reduce the overall effort for an adversary. Looking back at Experiment 3, investing in more than five priming tokens where there is no

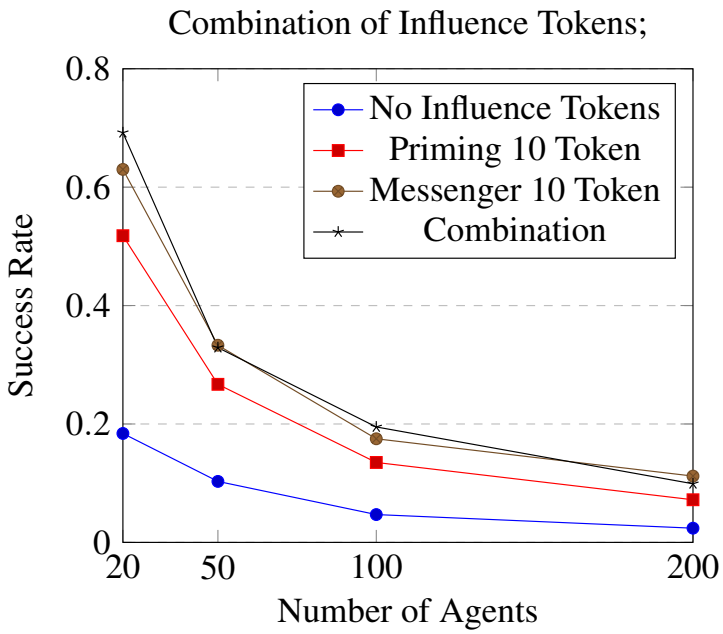


Figure 4.9 The comparison of a combination of influence tokens with no delay against singular and no influence tokens. The combination provides a slight increase to the success rate of the adversary, especially when fewer agents are in the net.

delay did not yield a significant gain for the success rate.

Experiment 6: Figure 4.10 shows the impact of combining the influence tokens with a delay and only considering five tailgating attempts. It compares against using five priming tokens with five attempts at tailgating without any messenger tokens. The results show that the priming tokens impact the most as the use of messenger when coupled with priming and a delay does not provide a strong impact.

4.4.4 Test Case Experiments - Conclusion

The six experiments show the varying impact of influence tokens and how it might be better at one point to use one token over another. For example, if the adversary is short on time, then the best option would be to use the messenger token. If the adversary can wait, allowing a priming token to propagate throughout the system would yield a greater success rate than the messenger token. These experiments show the impact of influence tokens, the strength of the messenger token and the propagation of the priming token. The adversary may have used

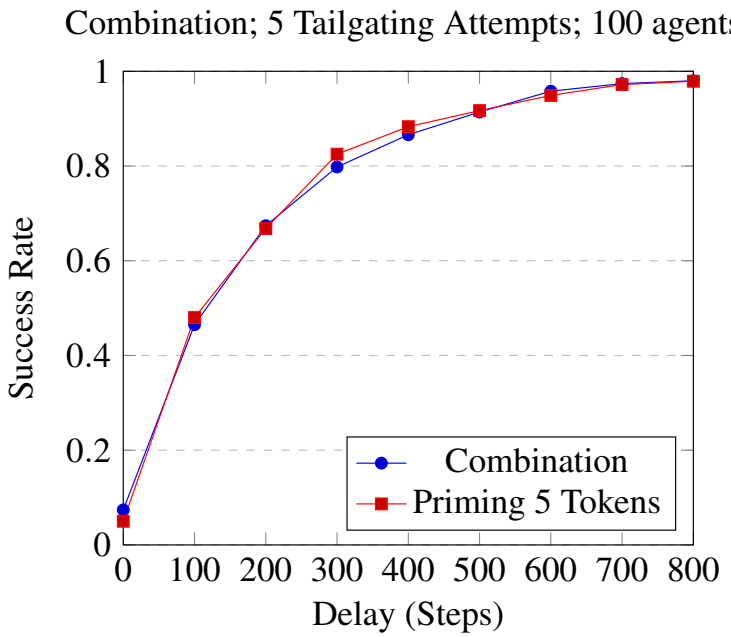


Figure 4.10 Combination of influence tokens when a delay is included, this shows the strength of the *Priming* token and that with so many attempts, the adversary has a high success rate, provided that they have delayed the attack for certain amount of time.

a priming token, but there is no guarantee that the agent they are tailgating is a direct result of the prime that they distributed. This creates an element of risk for adversary, where they are unsure about the compliance attitude of the agent.

These preliminary results from the experiment show that influence tokens alter the success rate a considerable amount compared to using no influence tokens. It now opens the path to allow us to consider other influence tokens. For example, the use of an *Incentive* token, which intuitively, would be a token that an agent holds for some time until they choose to accept or deny the incentive, where the output would be some positive or negative repercussions. A contextual example of an incentive would be blackmail. Another example of an influence which could be a token is *Reciprocity* where both adversary and agent mutually exchange information, with the hope of the adversary getting the required information to further or complete their attack [117].

In regards to the computation time, we managed to simulate the largest experiment, which consisted of two hundred agents, ten influence tokens, ten tailgating attempts, simulated for one thousand steps and the simulation ran one thousand times. The time for computation was on average thirty five seconds on a windows machine with 8GB of RAM and an i7 processor.

4.5 Influence Tokens: Validation

When we refer to Petri nets in this section we also refer to Coloured Petri nets, as an unravelled Coloured Petri net is a Petri net.

4.5.1 Are Influence Tokens Subject to Diminishing Returns?

The modelling of influence tokens in this chapter has made certain assumptions about how the behaviour of the tokens are formulated. For example, both the priming and messenger instantiations as influence tokens had a consistent impact on an agent's behaviour. However, we know from literature in psychology that influences can have diminishing returns, such as the impact of incentives towards consumer behaviour for purchasing products [102].

4.5.2 Petri nets Allow for Reachability and Deadlock Detection

When analysing any Petri net, properties about that Petri net can be quickly established which are reachability and deadlock detection:

Reachability: We described reachability in Section 4.2.3 where we defined a reachability graph that given an initial marking, the possible markings that can be reached. For the application of influence tokens used by an adversary or a defender for that matter, reachability allows for specific properties to be assessed. For example, assessment of whether or not all agents are non-compliant towards a security policy would be a property of interest.

Deadlock Detection: When describing agent behaviour reflects artefacts of human behaviour, we would not expect to see a deadlock where agents cannot move or perform actions. The validation a user

4.6 Influence Tokens: Conclusion

An adversary using influence tokens can improve their success rate through the impact of an influence token, by exploiting the features of a specific influence token such as priming, they can allow the token to propagate throughout the system to further improve the success

rate. The key components of this chapter are (1) the notion of influence tokens which can propagate, alter and impact a system, (2) the modelling of a social engineering scenario in Coloured Petri nets where influence tokens are present, (3) the verification of the model to ensure that properties such as reachability are satisfied, and (4) an analysis to consider a wide variety of test cases allowing for an in depth look at influencing techniques.

The success rate of the adversary depends upon the influence tokens used, furthermore, depending on when the adversary wishes to get inside the building, it may be better to use the messenger token over priming. Of course, if a delay is possible, then priming becomes the best choice if only one token is being used. These preliminary results provide us with the foundation to consider other influence tokens and the impact those would have within more complex scenarios.

For addressing Research Question 1, we have demonstrated that if the quantification of the social influence is known, we can analyse its impact when considering compliance attitudes, and furthermore, it would allow for the prediction of compliance attitudes evolving. Currently, we have not validated this approach, as we do not know if the quantification used in this chapter reflects the accuracy of social influences in the real world.

The key takeaway message from this chapter for the reader should be that social influences can be captured through modelling techniques. The results themselves provide an illustration of how propagation can improve the adversary's success rate for entering the building, however, assessing and understanding the necessary building blocks for a social influence model is the foundation that this chapter provides.

The successes of this chapter reside in it's focus on the representing the social influences. By implementing them as their own colour, the analysis of propagation for the social influences was intuitive and straightforward to perform. The model itself in CPNTools was easy to execute and we recommend any repeat experiments to make use of the tool when it comes to simulation and validation.

This chapter could be considered to be a stand alone chapter in the thesis as the modelling aspects shift later on in the thesis. However, it is important to acknowledge that without this piece of work we would not have the insights to build a more refined model focusing on agent behaviour.

A limitation of this chapter is that we did not identify any optimal costings for an adversary. For example, it would be interesting to understand how many tokens an adversary needs to distribute before it makes no difference for propagation. Should they distribute one more token it does not improve their success rate. Answering a question such as the optimal number of influence tokens would be a great contribution to this chapter.

Any insights towards understanding validation of this method is an area that we did not improve on in this chapter. Improving confidence in a model that is designed to be faithful toward human behaviour is crucial to ensuring it has any applicability in the real world. Unfortunately, this chapter did not provide us any insight into validation techniques. Whilst it was never the aim to validate within this chapter, we did hope that ideas or opportunities to validate would appear, for example using real world data in the chapter.

The use of Coloured Petri nets allows for the behaviour of influence tokens to be demonstrated but it does not force the use of Coloured Petri nets for influence tokens. One could use other tools of implementation such as PRISM which is a probabilistic model checker to model a Markov Decision Process or a Markov Chain [75]. However, the limitation of the tool tends to be around the number of agents/people that can be represented. In a Markov Decision Process or Markov Chain adding additional agents can be exponential and lead to state explosion.

This choice to use Coloured Petri nets came from the appeal of the simulations offered by the tool. The ability to formally represent our problem and run simulations consisting of many agents can demonstrate the power of social influences.

We did include some validation, however, even a large number of tokens in CPNtools would struggle to render a reachability graph. This is a limitation with our approach but we are confident validating for a small number of tokens would provide confidence in the model even when the number of tokens is increased and we can't validate due to limited processing power. This needs to be done cautiously and we feel that the example chosen was applicable to provide that confidence through validation.

The next stages of this work should consider the optimality of an adversary and the validation. By establishing how much investment an adversary requires we can begin to understand what sort of defence mechanisms work. In parallel, improving the confidence in the model through user validation is crucial to ensuring that this work can gain some traction.

Chapter 5

Building a Rule Based Model

This chapter addresses Research Question 2 by focusing on building a model which allows for observations of compliance behaviours:

Research Question 2 (Cyclic Observational Behaviour)**.** *How do we model observations of compliance behaviour?*

If we consider breaking this research question down we can identify four key elements for observing behaviour:

1. People must have some sort of context/beliefs in order to perform a behaviour [52]. We have defined this as the compliance attitude.
2. There is the actual behaviour that a person exhibits.
3. We know that behaviours can be observed by other people.
4. The observation of other peoples' behaviour could cause a person's compliance attitude to change [33, 44, 43].

The research question is not trivial to address, as the people performing behaviours and those observing are not static. They move and the neighbours of one person can easily change as that person's location changes. By dissecting the research question into four categories we can then begin to address each category on its own. This allows for greater refinement should the literature in the area change and we need to redesign a model.

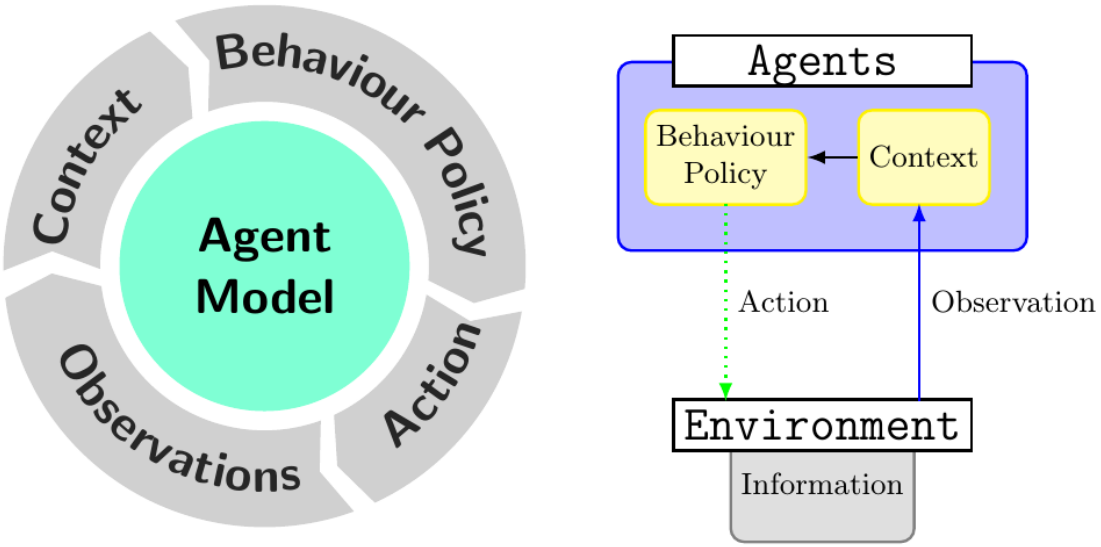
A model to express social influences towards compliance attitudes will touch on the notion of behaviour change. Michie demonstrated that there are over 83 different types of behaviour change models [83]. Unfortunately, these models offer no guidelines for dealing with the complexity of different influences or interventions. In general, they tend to be focused at the individual level. If we consider the COM-B Model, which is discussed in Chapter 2, then applying it to our research question does not take into account a person's location, who they interact with or other conflicting interventions [82]. Furthermore, the compliance attitudes of individuals is a variable which may be different for the many individuals. Assessing the composition of these different attitudes with regards to their behaviours and social influences towards others is something a behaviour change model does not cater for.

When we consider interactions that have complexity such as many people interacting with each other, individuals moving locations and capturing social influences then a direct comparison to the literature is the notion of Multi-Agent Systems [115]. A Multi Agent System demonstrates a number of individually driven agents that are capable of interacting with each other, sometimes this is to achieve a common goal. Often it's the case that a Multi Agent System is deterministic and the output of a Multi Agent System is strictly related to the initial input. In Section 2.7 we provide some background research to Multi Agent Systems.

As with any system that comprises of agents or people, one starting point is to ask the relevant questions. For this chapter, there are three we ask:

- What is required to build a system for human security behaviour?
- How do we formally describe how a human behaves and is socially influenced?
- What are the properties of such a system?

Chapter Overview: To the best of our knowledge there is no model that currently deals with the complexity of social influences for many people and the effects this has on their compliance attitudes. Therefore, this chapter addresses each of the questions in turn. In section 5.1 we introduce the building blocks of our model. In Section 5.2, we describe the formalisms of our model, we make use of operational semantics to describe our interactions. We extend this description to cover adversarial behaviour in Section 5.3. In Section 5.4, we address the properties of the system and what we can measure. The work reflects some of the material from the publication in 2017 at DataMods: From Data to Models and Back [23].



(a) Ordering of agent interactions

(b) Agent Interaction with Environment

Figure 5.1 Multi Agent System: Overview of agent interaction.

5.1 Building a Cyclic Observational Agent Based Model

A socio-technical system is often used to describe a physical system where people are described as agents [77]. The model we present is classified as a socio-technical system. The building blocks for our model require that people can interact with each other and are capable of executing actions in a representation of a physical environment.

5.1.1 Model Requirements

From here on in and for this chapter, we refer to an agent or agents when discussing human behaviour. An agent within the model has a context, which is their internal state or their compliance attitude. A similar concept is explained in the belief-desire-intention model, where a belief shares similarities to a context [52]. Figure 5.1 provides an overview of how we define the agents interactions. An agents context informs the actions they can execute and actions influence the context of other agents. Traditionally, these are concepts from Multi Agent System's (Multi Agent Systems) and ABM (Agent Based Modelling). For our model we have the following:

- **Action** - Agents have the capability to execute actions which may or may not include other agents.
- **Observation** - Agents can observe other agents actions and perceive what has happened.
- **Context** - Often referred to as beliefs, agents have internal metrics which guide their actions [52].
- **Behaviour change** - An agent's behaviour changes due to an influence of the observations they have witnessed.
- **Location** - Agents have a physical presence and can observe the actions of others who are nearby.

Actions: An agent, which is similar to a process, has a series of actions they execute. Once an agent executes an action, they still remain as the same agent [86]. Certain actions, much like human behaviour can occur simultaneously. For example, an agent can both move location and interact with another agent within one executable action. Depending upon the context and location of an agent, specific actions for an agent may be restricted.

Observation: Actions performed by agents can be observed by other agents. An observation is stored as part of the agent. Therefore, an agent builds up a unique trace of observed actions as a system transitions.

Context: In its simplest form, the context of an agent drives the action they execute. These actions can be observed which can modify the context of themselves or other agents. An agent's context and possible actions that have been modified is said to have had its behaviour changed. Decisions to enforce or circumvent security policies are individual to each agent. Typically, attitudes towards policies can be impacted by personality, past experiences and a productivity trade-off, to name a few [14]. In this chapter we focus in on compliance attitudes by means of altering the context of an agent.

Behaviour Change: To influence an agents behaviour is to change it. An influence relies on a psychological effect, such as the use of *Messenger* or *Social Norms* effects. We know that humans are influenced with varying degrees of success. Therefore, not all agents will have their behaviour changed due to an influence. We see this regularly in self influence towards efficacy when people who have achieved more are able to motivate themselves more than those who have achieved less [12]. It is common in phishing emails or social engineering attacks to use an influence to effect a change of behaviour [87]. For an agent, a behaviour change may result in different choices for actions when the opportunity for those actions is presented.

Location: If an agent represents a person, then an observation of an agent action is one person witnessing another do something. In order for an agent to observe, it must have a physical location. An agent can't observe another agents action when line of sight is not present, unless aided by a third party device such as a camera. Some security policies rely upon physical locations. For example, tailgating relies on locations are connected by some entry system such as a door, corridor and so on. Furthermore, employees express unique behaviours for moving between multiple locations. In recent work, it has been shown that malicious insider behaviour can be detected using historical data from a building access control system [31]. The data allows for suitable models to be learned surrounding movement behaviour. Using such techniques, Hidden Markov Models have been successfully used to predict with up to 92% accuracy, the next movement of someone given some historical data [51]. The context of such work could allow for future models that we build to consider using real data with predicted movements to have a better understanding of agent's interactions between locations.

5.2 Agent Based Model for Human Security Behaviour

In this section we introduce the formalism for our Rule Based Model. To motivate it, we use a tailgating example taken from previous work [23]. This is a similar to the example used in Chapter 4.

Scenario 1. *Agents arrive at the back of the workplace reception and there are two possibilities. Firstly, if nobody is at the front of the reception, the agent must progress to the front of the reception. Secondly, and dependent upon the agents security preference (usable or secure), if less than five agents are at the front of reception and about to enter, and the newly arrived agent is usable, they will attempt to tailgate. A perfectly secure agent will never attempt to tailgate. If an agent is being tailgated, they can either permit or deny the action, where a permit would allow both agents into the main building, a deny would force the tailgater to the front of the reception. A perfectly usable agent will always permit tailgating, a secure agent will deny tailgating.*

We now introduce the model incrementally and build up the rules to demonstrate how agents interact with each other.

5.2.1 Location Based Agents

Agents exist in a physical space and can move between locations. As an agent moves between locations, they are associated with a context.

Definition 6 (Location Based Agents)**.** *Given a set of agents A, a set of locations L, and a set of contexts C, we define the set of* location agents *as* $LA = C \times L$*. We define a state of location based agents LBA as a set of pairs* $\{(a_1, ls_1), \ldots, (a_n, ls_n)\}$ *where* $a_i \in A$ *and* $ls_i \in LA$

Definition 7 (Location based Actions)**.** *Given an agent a and two locations* l *and* l'*, a location based action is defined as* $\mathsf{m}(a, l, l')$*, indicating that a moved from* l *to* l'*. A location based action does not modify the context of the agent.*

In general, there could be many different ways to capture the concrete set of location based actions in a system, for instance by going through the actual logs of a smart card system. For the sake of simplicity, we consider here a set of links $Link \subseteq L \times L$ where (l_1, l_2) indicates a physical link between the location l_1 and l_2. Intuitively speaking, any agent can move from a location to another as long as there is a link between them. We characterise this with the following rule:

$$\frac{(l, l') \in Link}{(a, (c, l)) \xrightarrow{\mathsf{m}(a,l,l')} (a, (c, l'))} \tag{5.1}$$

The action 5.1 is defined as an inference rule at the atomic level expressing that one agent moves from one location to another, provided the two locations are connected. In general, we write $(a_1,(c_1,l_1)) \mid \ldots (a_n,(c_n,l_n))$ for a set of location based agents. We utilise Rule 5.1 in Rule 5.2 to express how an agent can move with respect to all other agents.

$$\frac{(a,(c,l)) \xrightarrow{\mathsf{m}(a,l,l')} (a,(c,l'))}{(a,(c,l)) \mid LBA \xrightarrow{\mathsf{m}(a,l,l')} (a,(c,l')) \mid LBA} \tag{5.2}$$

Let us consider that a context of an agent refers to their security preference, where a preference is simply *usable* or *secure*. To be *usable* is to permit a breach of a security policy, to be *secure* is to deny a breach of a security policy. Whilst it is not the case that an agent is *usable* or just *secure*, for now, we use these polar opposites to express our model. In the tailgating scenario, a breach of the policy would be an agent permitting tailgating.

We introduce two actions, $\mathsf{tgp}(a_1,a_2,l,l')$ and $\mathsf{tgd}(a_1,a_2,l,l')$, indicating that a_1 tailgated a_2, and that a_2 denied a tailgate from a_1, respectively. Intuitively speaking, a *usable* agent is permitted as they tailgate a *usable* agent. A *usable* agent is denied as they tailgate a *secure* agent. Formally:

$$\frac{(a_1,(c_1,l)) \xrightarrow{\mathsf{m}(a_1,l,l')} (a_1,(c_1,l')) \quad (a_2,(c_2,l)) \xrightarrow{\mathsf{m}(a_2,l,l')} (a_2,(c_2,l')) \quad c_1 = c_2 = usable}{(a_1,(c_1,l)),(a_2,(c_2,l)) \mid LBA \xrightarrow{\mathsf{tgp}(a_1,a_2,l,l')} (a_1,(c_1,l')),(a_2,(c_2,l')) \mid LBA} \tag{5.3}$$

$$\frac{(a_2,(c_2,l)) \xrightarrow{\mathsf{m}(a_2,l,l')} (a_2,(c_2,l')) \quad c_1 = usable \quad c_2 = secure}{(a_1,(c_1,l)),(a_2,(c_2,l)) \mid LBA \xrightarrow{\mathsf{tgd}(a_1,a_2,l,l')} (a_1,(c_1,l)),(a_2,(c_2,l')) \mid LBA} \tag{5.4}$$

The rules introduced so far explain how agents with a context move between locations. They can either just move, or they can tailgate and be permitted or denied. A permit moves them into the tailgated location, the denied leaves an agent in the same location before the action occurred.

5.2.2 Observing Agents

In an organisation, when an action happens, people may notice. In our example, this translates to agents observing the compliance behaviour of other agents. Much like in the workplace, if someone is challenged for tailgating, people within close proximity will notice. One report recorded that users security sensitivity for other peoples security behaviour was prominent in a working environment [39]. We know that people have security awareness in the workplace, particularly when policies are regularly followed [39].

We began with Location Based Agents, where the interactions of agents are restricted by physical locations and the security preference of each agent. We extend this notion and introduce Observing Agents:

Definition 8 (Observing Agents)**.** *We define* $SOA = C \times L \times \mathbb{P}(\Theta)$ *as the set of observing agents states where C is the set of contexts, L is the set of Locations and* $\Theta \subseteq A \times Act_\theta$*. We introduce a set of observable actions* Act_θ*, where any* $act \in Act_\theta$ *can be observed by an agent.*

Definition 9 (System State)**.** *Given a set of agents A, a set of observing agents state SOA, the set of states are* $\Delta = \{\delta_1, \delta_2, ..., \delta_n\}$*, a state* δ *of the system is a set of pairs* $\{(a_1, s_1), \ldots, (a_n, s_n)\}$ *where* $a_i \in A$ *and* $s_i \in SOA$ *and:*

$$\forall a \in A, \exists s \in SOA, (a,s) \in \delta \wedge ((a,s) \in \delta \wedge (a,s') \in \delta \implies s = s')$$

The above states that for any agent $(a,s) \in \delta$*, there will not be another agent* $(a,s') \in \delta$ *where the agent id a is the same and the state of the observing agent s is different. Essentially, we don't want to duplicate an agent with the same identifier.*

Let us consider that $Act_\theta = \{permit, deny\}$, which refers to the permitting or denying of tailgating. An observing agent during the course of their interactions, may accumulate observations of other agents permitting or denying tailgating. As it is currently defined, an agent can only observe and store one observation for each agent. At the atomic level, an inference rule for an observation would take the form:

$$\frac{}{(a,(c,l,\Theta_a)) \xrightarrow{\mathsf{obs}(\theta)} (a,(c,l,\Theta'_a \cup \theta))} \tag{5.5}$$

Intuitively, an agent a_1 with the action $\mathsf{obs}(a_2, permit)$ would indicate that they have observed a_2 permitting tailgating.

For an agent observing a particular act, such as tailgating, we must consider all agents in the observable area. We provide the following definition:

$$loc : SOA \rightarrow LA$$

$$loc(c,l,\theta_a) = (c,l)$$

Therefore, the observable actions for tailgating are as follows:

$$\frac{\begin{array}{c}(a_1,(c_1,l,\Theta_{a_1})) \xrightarrow{\mathsf{obs}(a_2,permit)} (a_1,(c_1,l,\Theta_{a_1} \cup (a_2,permit))) \\ \forall (a',(c',l,\Theta'_a)) \in \delta \Rightarrow (a',(c',l,\Theta'_a)) \xrightarrow{\mathsf{obs}(a_2,permit)} (a',(c',l,\Theta'_a \cup (a_2,permit)))\end{array}}{\begin{array}{c}(a_1,(c_1,l)),(a_2,(c_2,l))|loc(SOA) \xrightarrow{\mathsf{tgp}(a_1,a_2,l,l')} \\ (a_1,(c_1,l')),(a_2,(c_2,l'))|loc(SOA)\end{array}} \\ \hline (a_1,(c_1,l,\theta_{a_1})),(a_2,(c_2,l,\theta_{a_2}))|\delta \xrightarrow{\mathsf{tgp}(a_1,a_2,l,l')} (a_1,(c_1,l',\theta_{a_1})),(a_2,(c_2,l',\theta_{a_2}))|\delta' \tag{5.6}$$

$$\frac{\begin{array}{c}(a_1,(c_1,l,\Theta_{a_1})) \xrightarrow{\mathsf{obs}(a_2,deny)} (a_1,(c_1,l,\Theta_{a_1} \cup (a_2,deny)) \\ \forall (a',(c',l,\Theta'_a)) \in \delta \Rightarrow (a',(c',l,\Theta'_a)) \xrightarrow{\mathsf{obs}(a_2,deny)} (a',(c',l,\Theta'_a \cup (a_2,deny)))\end{array}}{\begin{array}{c}(a_1,(c_1,l)),(a_2,(c_2,l))|loc(SOA) \xrightarrow{\mathsf{tgd}(a_1,a_2,l,l')} \\ (a_1,(c_1,l)),(a_2,(c_2,l'))|loc(SOA)\end{array}} \\ \hline (a_1,(c_1,l,\theta_{a_1})),(a_2,(c_2,l,\theta_{a_2}))|\delta \xrightarrow{\mathsf{tgd}(a_1,a_2,l,l')} (a_1,(c_1,l,\theta_{a_1})),(a_2,(c_2,l',\theta_{a_2}))|\delta' \tag{5.7}$$

The rule in Equation 5.6 consists of three layers. The bottom layer states how the two agents a_1 and a_2 are engaged in the previously defined action $\mathsf{tgp}(a_1,a_2,l,l',permit)$ which states that a_1 tailgated a_2, so a_2 is the agent who permitted the action. The middle layer describes how this occurs with respect to all agents. Finally, the top layer describes how agents who are in the same location as the action will observe the action and update their set of observable actions. This is then repeated for the rule in Equation 5.7, however, it is for the action of denying tailgating.

5.2.3 Behaviour Change Agents

The concept of behaviour change as a body of research contains many different models. Not all of the models are fit for our purpose. However, the COM-B model splits behaviour change into three elements; Capabilities, Opportunities and Motivation [82]. In this chapter, we focus on the aspect of Motivation, which can be changed by influencing effects. Such effects as *Messenger* and *Social Norms* are of interest to us [44]. The former relates to those people/agents we perceive to be in a position of authority, the latter is all about those people around us in our immediate vicinity [33]. For more background on these topics look at Chapter 2 where discuss them in more detail.

In our model, a behaviour change would be a change of context. A *secure* agent can become *usable* and vice versa. The following rule captures behaviour change for security preferences:

$$\frac{c \neq c'}{(a,(c,l,\Theta_a)) \xrightarrow{\mathsf{bchange}(c')} (a,(c',l,\{\}))} \tag{5.8}$$

In the previous chapter we used messenger and priming which was targeted towards an adversary exploiting agents that were susceptible to an influence. In this chapter, the scenario allows for agents to enter multiple times and this provides a notion of social norms whereby agents become culturally defined with how to behave. This is different to priming in the previous chapter where the adversary pro-actively sought to prime agents for tailgating. To the best of our knowledge, for the influencing effects *Messenger* and *Social Norms* there does not exist a strategy to quantify formally these effects. Unsurprisingly, an effect is unique to each agent. Nevertheless, we provide a definition to demonstrate how we have interpreted it:

Definition 10 (Influencing Agents)**.** *The set $IA \subseteq A \times A$ captures Influencing Agents, where any $(a,a') \in IA$ indicates that a' can influence a and $a \neq a'$.*

The influencing agents is for the purpose of defining inference rules for the messenger effect. The following rules capture this:

$$\frac{(a,(c,l,\Theta_a)) \xrightarrow{\mathsf{bchange}(usable)} (a,(c',l,\Theta_a)) \ (\exists(a',permit) \in \Theta_a \wedge (a,a') \in IA)}{(a,(c,l,\Theta_a))|\delta \xrightarrow{\mathsf{messP}(usable)} (a,(c',l,\Theta_a))|\delta} \tag{5.9}$$

$$\frac{(a,(c,l,\Theta_a)) \xrightarrow{\mathsf{bchange}(secure)} (a,(c',l,\Theta_a)) \ (\exists(a',deny) \in \Theta_a \wedge (a,a') \in IA)}{(a,(c,l,\Theta_a))|\delta \xrightarrow{\mathsf{messD}(secure)} (a,(c',l,\Theta_a))|\delta} \tag{5.10}$$

For social norms, we care about the number of agents that have been observed for a particular action. Given a set of observations, we can establish how many agents have been observed performing a particular action. This is then a different interpretation of how an agent can be influenced due to the observations they have made. Instead of it relying on the agent observed performing the action, the influence is the number of times the action has been observed, regardless of the agent who performed it.

5.3 The Adversary

Defining an adversary in this rule based model is done by isolating one agent to take the role of the adversary. We use adv as the agent id to specify that the agent takes on the role of adversary.

5.3.1 Influencing Adversary

Where as the previous rules defined were general rules for all agents, for the adversary, we define specific rules. We define an influencing adversary that has the power to change another agents context:

$$\frac{l_{adv} = l_1 \wedge \mathsf{adv} \neq a}{(\mathsf{adv},(c,l_{adv},\Theta_{\mathsf{adv}}))(a,(c_1,l_1,\Theta_a))|\delta \xrightarrow{\mathsf{inf}(\mathsf{usable})} (\mathsf{adv},(c,l_{\mathsf{adv}},\Theta_{\mathsf{adv}}))(a,(\mathsf{usable},l,\Theta_a))|\delta} \tag{5.11}$$

For Rule 5.11, it states that an adversary agent labelled as adv can influence another agents context to be usable providing both the adversary and agent are in the same location. In this chapter we don't explore the adversary beyond the assessment of properties, however, for future work it would be useful to understand how the influence tokens methodology from Chapter 4 aligns with the model defined in this chapter.

5.4 Rule Based Model: Properties

Traditionally, a security property is often associated with a set of protocol actions. For example, Perfect Forward Secrecy is a security property commonly associated with communication protocols, such as the TLS protocol [38]. In access control, the properties of a model are related to the policies, such that an access request to protected resources or a physical area is evaluated with regards to that policy [37]. An associated property for access control would be to ensure that only authorised access requests are permitted.

In a socio-technical system, a physical security property is associated with a breach of a policy. Whilst our model is not a socio-technical system a physical security property for a policy breach would be an unauthorised agent entering the building (In the example of tailgating). This can occur in many different settings, to list a few:

1. An unauthorised employee tailgates an authorised employee.
2. An adversary tailgates an authorised employee.
3. An adversary authorises themselves and enters - In our scenario this would not occur as the adversary has no capability to authorise themselves. In reality this could be achieved by taking/cloning a smart card to enter an access controlled building.

5.4.1 Formalisation of Security Properties

Establishing the validity of a security property can sometimes be captured as a reachability problem. We wish to know, given some initial state, if another state where a property is always true or eventually true can be reached where the conditions for that property have been satisfied or maintained. For example, will a state exist where all agents would comply with a security policy (i.e. it is a secure state).

In the model, a state represents all of the internal states of all agents. These internal states include the context, location and observations of agents. Table 5.1 is an example of how a state is defined for agents.

Definition 11 (Transition Relation)**.** *A transition* $\tau = (\delta, \delta')$ *where* $\delta \in \Delta$ *and* $\delta' \in \Delta$ *indicates that a state change from* δ *to* δ' *can occur, a transition relation is therefore, defined as* $T \subseteq \delta \times \delta$.

Table 5.1 Rule Based Model: Internal States of agents

A	δ_0
alice	$(\mathsf{secure}, \mathsf{inside}, \emptyset)$
bob	$(\mathsf{usable}, \mathsf{outside}, \emptyset)$
.	.
.	.
eve	$(\mathsf{secure}, \mathsf{hall}, \emptyset)$

Table 5.2 Rule Based Model: Transition of internal states

A	δ_0	$\xrightarrow{\tau}$	δ_1	$\xrightarrow{\tau'}$	δ_2
alice	$(\mathsf{secure}, \mathsf{inside}, \emptyset)$		$(\mathsf{secure}, \mathsf{inside}, \emptyset)$		$(\mathsf{secure}, \mathsf{hall}, \emptyset)$
bob	$(\mathsf{usable}, \mathsf{outside}, \emptyset)$		$(\mathsf{usable}, \mathsf{inside}, \emptyset)$		$(\mathsf{usable}, \mathsf{hall}, (\mathsf{alice}, \mathsf{p}))$
.	.	.	.	.	.
.	.	.	.	.	.
eve	$(\mathsf{secure}, \mathsf{hall}, \emptyset)$		$(\mathsf{secure}, \mathsf{hall}, \emptyset)$		$(\mathsf{usable}, \mathsf{hall}, (\mathsf{alice}, \mathsf{p}))$

Table 5.2 is an example of state transition, where in this example, the agents alice and bob eventually move into the location hall and in τ', we see that the observation $(\mathsf{alice}, \mathsf{p})$ was acquired for $\mathsf{alice}, \mathsf{bob}$ and eve. The observation $\mathsf{alice}, \mathsf{p}$ refers to the agent alice permitting tailgating, and as a result, the agent eve has their context changed to usable.

As part of our design we have elected to include actions, observations and behaviour changes as one transition. We do this to avoid any obvious problems such as the ordering of transitions, for example, should an observation occur before or after an action which increases the complexity of the model.

Definition 12 (Model). *A model $M = (\Delta, T, \delta_0, IA, L)$ where $\delta_0 \in \Delta$ is the initial state of the system.*

The model dictates how the compliance attitude (context) of the agents and their behaviour evolves. Let us consider some examples of how a system can evolve dependent upon the constraints placed upon it. We consider three different models that are slightly differentiated by the configuration of the influencing agents. Let us consider a set of fixed locations.

$$\mathsf{L} = \{(\mathsf{outside}, \mathsf{inside}), (\mathsf{inside}, \mathsf{outside}), (\mathsf{inside}, \mathsf{hall}), (\mathsf{hall}, \mathsf{inside})\} \tag{5.12}$$

Table 5.3 Cyclic Influencing - Agents locked in a loop influencing each other.

A	δ_0	$\xrightarrow{\tau_1}$	δ_1	$\xrightarrow{\tau_2}$	δ_2
alice	(secure, hall, $\emptyset$)		(secure, hall, $\emptyset$)		(secure, hall, (eve, p))
bob	(secure, outside, $\emptyset$)		(secure, inside, $\emptyset$)		(usable, hall, (eve, tgp))
eve	(usable, inside, $\emptyset$)		(usable, inside, $\emptyset$)		(usable, hall, $\emptyset$)
$\xrightarrow{\tau_3}$	δ_3	$\xrightarrow{\tau_4}$	δ_4	$\xrightarrow{\tau_5}$	δ_5
	(secure, inside, $\emptyset$)		(secure, inside, $\emptyset$)		(secure, hall, $\emptyset$)
	(usable, hall, $\emptyset$)		(usable, hall, $\emptyset$)		(usable, hall, $\emptyset$)
	(secure, hall, $\emptyset$)		(usable, inside, $\emptyset$)		(secure, inside, (alice, tgd))
$\xrightarrow{\tau_6}$	δ_6	$\xrightarrow{\tau_7}$	δ_7	$\xrightarrow{\tau_8}$	δ_8
	(secure, inside, $\emptyset$)		(secure, inside, $\emptyset$)		(usable, inside, $\emptyset$)
	(usable, hall, $\emptyset$)		(usable, hall, $\emptyset$)		(usable, inside, $\emptyset$)
	(secure, inside, $\emptyset$)		(secure, inside, $\emptyset$)		(secure, hall, $\emptyset$)
$\xrightarrow{\tau_9}$	δ_9	$\xrightarrow{\tau_{10}}$	δ_{10}		
	(secure, inside, $\emptyset$)		(usable, hall, (bob, tgp))		
	(usable, inside, $\emptyset$)		(usable, hall, $\emptyset$)		
	(secure, hall, $\emptyset$)		(secure, hall, (bob, p))		

Example 1: Cyclic Influencing

Consider the set of influencing agents in this example to be:

$$\mathsf{IA}_1 = \{(\mathsf{bob}, \mathsf{eve}), (\mathsf{eve}, \mathsf{alice}), (\mathsf{alice}, \mathsf{bob})\}$$

The current set IA_1 states that eve, alice and bob can influence bob, alice and eve respectively. Given this set of influencers and set of rules governing how agents tailgate and challenge each other then the collective compliance attitudes of the agents can be cyclic if the correct transitions occur. Table 5.3, which was computed manually demonstrates how even agents can influence each other. For the purposes of this example, we have introduced a new rule for agents that usable who convince others to act insecurely and tailgate behind them[1].

[1]This is why bob can tailgate eve in τ_2

Table 5.4 Influence Propagation - Agent alice influencing eve by propagation.

A	δ_0	$\xrightarrow{\tau_1}$	δ_1	$\xrightarrow{\tau_2}$	δ_2
alice	$(\mathsf{usable}, \mathsf{inside}, \emptyset)$		$(\mathsf{usable}, \mathsf{inside}, \emptyset)$		$(\mathsf{usable}, \mathsf{hall}, \emptyset)$
bob	$(\mathsf{secure}, \mathsf{inside}, \emptyset)$		$(\mathsf{secure}, \mathsf{hall}, (\mathsf{alice}, \mathsf{p}))$		$(\mathsf{usable}, \mathsf{hall}, \emptyset$
eve	$(\mathsf{secure}, \mathsf{outside}, \emptyset)$		$(\mathsf{secure}, \mathsf{outside}, \emptyset)$		$(\mathsf{secure}, \mathsf{inside}, \emptyset)$
$\xrightarrow{\tau_3}$	δ_3	$\xrightarrow{\tau_4}$	δ_4		
	$(\mathsf{usable}, \mathsf{hall}, \emptyset)$		$(\mathsf{usable}, \mathsf{hall}, \emptyset)$		
	$(\mathsf{usable}, \mathsf{inside}, \emptyset)$		$(\mathsf{usable}, \mathsf{hall}, \emptyset)$		
	$(\mathsf{secure}, \mathsf{inside}, \emptyset)$		$(\mathsf{usable}, \mathsf{hall}, (\mathsf{bob}, \mathsf{p}))$		

Example 2: Influence by Propagation

Consider the set of influencing agents in this example to be:

$$\mathsf{IA}_2 = \{(\mathsf{bob}, \mathsf{alice}), (\mathsf{eve}, \mathsf{bob})\}$$

The current set IA_2 states that alice and bob can influence bob and eve respectively. By this configuration of influencers, the agent alice can influence the agent eve by propagation. Table 5.4 provides an example to demonstrate how this influence can propagate from the agent alice.

5.4.2 Example 3: Influence by Observation

Consider the set of influencing agents in this example to be:

$$\mathsf{IA}_3 = \{(\mathsf{bob}, \mathsf{alice}), (\mathsf{eve}, \mathsf{alice})\}$$

The current set IA_3 states that alice can influence both bob and eve. By this configuration of influencers, the agent alice can influence by observation the agents bob and eve. Table 5.5 provides an example to demonstrate how this influence can be observed from the agent alice.

Table 5.5 Influence by Observation - Agent alice influencing eve by observation.

OA_0	δ_0	$\xrightarrow{\tau_1}$	δ_1
alice	$(\mathsf{secure}, \mathsf{inside}, \emptyset)$		$(\mathsf{secure}, \mathsf{hall}, \emptyset)$
bob	$(\mathsf{usable}, \mathsf{inside}, \emptyset)$		$(\mathsf{secure}, \mathsf{inside}, (\mathsf{alice}, \mathsf{tgd}))$
eve	$(\mathsf{usable}, \mathsf{inside}, \emptyset)$		$(\mathsf{secure}, \mathsf{inside}, (\mathsf{alice}, \mathsf{d}))$

5.4.3 Computation Tree Logic

A transition of $\mathsf{tgp} \in T$ where $\mathsf{tgp} = (\delta, \delta')$ is an example of how a model can change state. We have defined these as rules and expressed them as $\delta \xrightarrow{\mathsf{tgp}} \delta'$.

Given a Model M, we use CTL (Computation Tree Logic) to define our security properties [55]. In CTL we define ϕ which is a state formula. These formulas in logic form are represented by *Backus-Naur* form:

$$\begin{aligned} \phi ::= & \bot | \top | p | (\neg\phi) | (\phi \wedge \phi) | (\phi \vee \phi) | (\phi \Rightarrow \phi) | (\phi \Leftrightarrow \phi) \\ & | AX\phi | EX\phi | AF\phi | EF\phi | AG\phi | A[\phi U \phi] | E[\phi U \phi] \end{aligned} \tag{5.13}$$

The logical operators in CTL are the usual ones of $\neg, \vee, \wedge, \Rightarrow$ *and* $\Leftrightarrow$. Additionally, CTL formulas make use of boolean constants true and false.

The temporal quantifier operators are the following when evaluating all paths:

- $A\phi$ - All: The condition ϕ has to hold on all paths starting from the current state.
- $E\phi$ - Exists: There exists a minimum of one path starting from the current state where ϕ holds.

In CTL, it is useful to assess a specific path, in which case we use the following:

- $X\phi$ - Next: The condition ϕ has to hold at the next state.
- $G\phi$ - Globally: The condition ϕ has to on the entire subsequent path.
- $F\phi$ - Finally: At some point on the subsequent path, the condition ϕ has to hold.
- $\phi U \psi$ - Until: ϕ has to hold at least until at some state ψ is true.

- $\phi W \psi$ - Weak Until: ϕ has to hold until ψ holds.

Given these definitions, we construct security properties for our example.

Property 1. ***Weak Unauthorised Entry:*** *"Is it possible from an initial state $\delta_0 \in \Delta$, such that another state can be reached where an unauthorised employee has tailgated an authorised employee?" This property is expressed as the following:*

$$P_1 ::= F[\exists(a,(c,l,\Theta)).(a,(c,l,\Theta)) \wedge ((a,\mathsf{p}) \in \Theta)] \tag{5.14}$$

Property 2. ***Strong Unauthorised Entry:*** *"Is it possible from an initial state $\delta_0 \in \Delta$, such that another state can be reached where an adversary has tailgated an authorised employee?" This property is expressed as the following:*

$$P_2 ::= F[\exists(a,(c,l,\Theta)).(\mathsf{adv},(c,l,\Theta)) \wedge ((\mathsf{adv},\mathsf{p}) \in \Theta)] \tag{5.15}$$

Property 3. ***Exploited Authorised Entry:*** *"Is it possible from an initial state $\delta_0 \in \Delta$, such that another state can be reached where an adversary has entered as an authorised person?" This property is expressed as the following:*

$$P_3 ::= F[\exists(a,(c,l,\Theta)).(\mathsf{adv},(c,l,\Theta)) \wedge ((\mathsf{adv},\mathsf{p}) \notin \Theta) \wedge (l = \mathsf{inside})] \tag{5.16}$$

The three properties P_1, P_2 and P_3 are focusing on the possibility that the property is violated, i.e. it returns true. One would expect that if P_2 is true then P_1 must also be true, however, it may be the case that only one agent exists that permits tailgating. If this was the case, P_1 would evaluate to false.

5.5 Rule Based Model: Validation

5.5.1 The Model Does Not Express the Ordering of Agent Actions

We do not explicitly state that the model is non-deterministic for selecting which agent will perform an action. However, in Chapters 6 and 8 we utilise the model in simulation

and use an engine known as SysModels which deals with action selection for human based agents [27].

5.5.2 The Model Does Capture Agents Observing Other Agents

The model captures agents observing other agents, which addresses the research question in this chapter. The framework of an agent that observes an action of another agent which impacts their context and a deterministic behaviour change policy dictates their actions post observation.

5.5.3 A Context is Not a True Representation of a Person's Internal State

The context in this chapter is the compliance attitude that we refer to in the rest of the thesis. We discuss in Section 2.3 that the compliance attitude is not trivial and can be impacted by many different vectors. For example, the COM-B model of behaviour change demonstrates that a person's behaviour can be changed by their interactions with others, their environment that surrounds them and their abilities to perform tasks [82].

In this chapter we defined a binary context for agents which does not reflect reality, however, it was all that was needed to express our model for an observational based multi agent system.

5.6 Rule Based Model: Conclusion

In this chapter we have compartmentalised the different elements for observing a behaviour. By defining a rule based model we have split the research question of observing behaviours into actions, observations, context and behaviour change. At the beginning of the chapter we posed three questions which were:

- What is required to build a system for human security behaviour?
- How do we formally describe how a human behaves and is socially influenced?
- What are the properties of such a system?

Firstly, for the what is required to build a system for human security behaviour we addressed the concept of compliance towards security policies. By given a state of an agent's compliance status, it informed there compliance behaviour. Secondly, we formally describe how a human behaves and is socially influenced by means of the rules outlined in this chapter. We have offered no means of validation in this chapter that it is a true reflection of reality. Instead we compartmentalised different aspects of human behaviour into actions, observations, context and a behaviour policy as depicted in Figure 5.1. Thirdly, we addressed the properties that such a system could express such as cyclic influencing where multiple contradictory social influences co-exist. We also considered influence by propagation which was a core aspect of Chapter 4.

We have addressed Research Question 2 by designing a model allowing for human agents to observe the behaviour of others. In particular, the model allows human agents to be dynamic and move location which is an important requirement for the overall aim of this thesis. This chapter will be used later on to form the basis of the model for Chapters 5 and 7. It will provide the building blocks necessary to create a model that we will use in simulations. Furthermore, we will then extend the model slightly as we implement in a tool called PCASP which provides a usable method for utilising this formal model.

The separation into the four elements of action, observation, context and behaviour change ensures that each element can be improved individually as the literature evolves. It then allows for the Multi Agent System to be composed by each element which has been designed to be as faithful as possible to the current state of the literature.

The decisions for those four elements of action, observation, context and behaviour is partly captured in the remarks in Chapter 1 of this thesis. Particularly remarks 1 and 2 which focus on compliance attitude and compliance behaviour. Our research into the state of the literature created those remarks which we have demonstrated with our modelling approach. Of course, there are areas which would require refinement, but one of the purposes of each of the chapters within this thesis is to provide proof of concepts. To the best of our knowledge there exists little research into this area and we must start have some ground to build our proof of concepts on.

A key insight of this chapter is the fundamental modelling aspects that were chosen. Those aspects are action, observation, context and behaviour policy to capture an agent's

nature. The relationship between action and observation is crucial to ensuring that an agent's context can be influenced by their behaviour policy.

The manual traces defined earlier in this chapter demonstrated that two conflicting influences could co-exist in the same system. Whilst it may be possible for these to both exist in one agent, we did not explore that option of uncertainty.

The successes of this chapter remain in those modelling aspects where we can clearly demonstrate how an agent's behaviour can be influenced by the observations they make of other agents. Without this relationship this chapter would not be relevant to the thesis as that was the research question we identified at the beginning of the thesis. We phrased this as *How do we model observations of compliance behaviour?* and by ensuring a model can capture observations we ensured that we could create a model that represented influencing behaviour.

A clear limitation is the extent to which the behaviour policy reflects human nature. We understand that a person does not have a policy stating that if they observe an action multiple times contrary to their belief, they will then change their belief. We know this isn't always the case as some people may use the perceived incorrect actions of others to validate their own beliefs. For example, observing a robbery would not always influence a person to steal, one would hope it would be the latter and that it would re-enforce their belief that stealing is not in their nature.

In this chapter we never addressed or managed to solve the problem of agent sequencing. For example, which agent should go first was something we did not address or solve. Even after doing the work it is still not clear how to resolve this issue.

In the future for this work, it should address correctness and completeness. Without this, any model would be lacking confidence and trust.

For the correctness, there is validation and verification. Firstly, does the model reflect the real world elements it is trying to replicate. Secondly, does the model behave in the way it's supposed too with regards to those elements. For example, we would not want to see an agent's behaviour change without any observations, unless we are considering an ambivalent agent.

For the completeness, we would like to understand if we have captured all of the necessary elements. For example, does the behaviour policy have elements which are missing? Most

likely because we use a very simplified approach in this thesis. Providing that completeness will allow us certainty to know that the elements we are representing is the complete set and nothing more.

By ensuring completeness and correctness in this manner we can have certainty that this top layer model is faithful to the real world. Furthermore, we would not risk and errors being transitioned to implementations of this model, which is likely in it's current state for the tooling and user experiments we discuss later in the thesis.

Chapter 6

Learning Decision Trees from Synthetic Human Behaviour

In the other chapters of this thesis we work on building and validating models expressing social influences towards compliance behaviour of security policies. Whilst this chapter does still build a model, the direction is focused on assessing the output/traces that the behavioural models are capable of generating. Furthermore, if we can provide a basis for understanding and successfully analysis synthetic data then we have more reliability should we ever gather real world data. The chapter addresses Research Question 3, as a reminder the research question is:

Research Question 3 (Security Behaviour Profiling)**.** *Can we accurately learn compliance attitudes from public traces of compliance behaviours?*

The hidden information being the compliance attitudes would provide motivation for agent's publicly observable behaviours with regards to security policies. The marker we target in this chapter is an agent's compliance attitude towards their compliance behaviour. Figure 6.1 demonstrates how someone monitoring their employees' behaviour could only ever observe public behaviours. Private compliance attitudes is the hidden information that we aim to identify based off the public compliance behaviours we observe.

One method of assessing the compliance attitudes of agents is to watch their behaviour. This behaviour could come in one of two forms. We could build a classifier from actual traces observed within an organisation. This can be costly and time-consuming and it can be

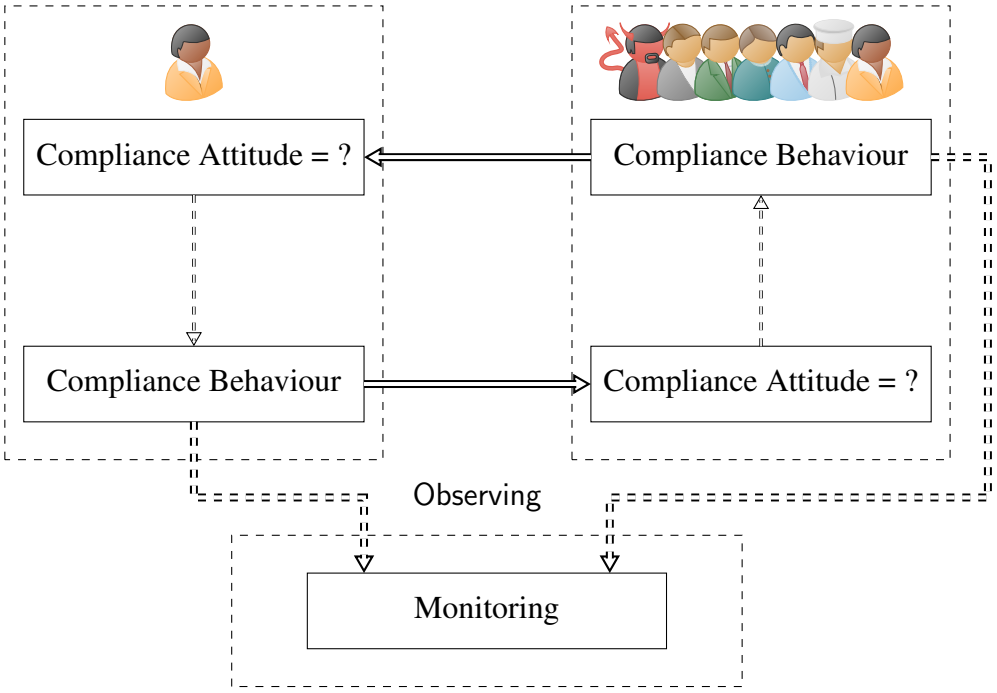


Figure 6.1 Lack of knowledge for compliance attitudes.

difficult to select the best classifier, e.g. which is the best attributes to classify on. This would require some form of surveillance and monitoring equipment to record how agents behave. In a recent study, researchers observed via CCTV people entering a building where a secure door was replaced with a turnstile. Over 1800 sequences of behaviour events were logged over one working day for approximately 600 employees and visitors where they calculated the individual cost of security placed upon employees [28]. In this chapter, we perform a methodical assessment of decision trees to predict the impact of human behaviour on the security of an organisation, by learning them from different sets of traces generated by the rule based model defined in Chapter 5. The simulation used to generate the traces allows agent behaviour to be distributed by time, that is, agent synchronisation is not deterministic.

Chapter Overview: This chapter is inspired by work carried out in the research paper submitted to DataMods: From Data to Models and Back [23]. We generate synthetic data from a simulation environment using the rule based model defined in Chapter 5. We then assess the accuracy of the decision trees models towards identifying hidden markers which

we deemed as the compliance attitude. In this chapter, the compliance attitude is considered to be binary, i.e. 0 or 1.

6.1 Building Synthetic Traces

In this chapter we make use of synthetic traces that are based on the model from Chapter 5. The previously defined rule based model which focused on actions and observations provides a base in this chapter for us to define a parameterised instantiation of said model where agents perform actions and make observations.

We don't use a real world data set for two reasons. Firstly, we could not find a data set which is suitable to represent the model we proposed. Secondly, because it is not yet clear how we could capture all of the information that we would need in order to build a reliable data-set. For example, uncertainty will be present in the compliance attitudes of people and this lack of information would find it's way into the data set.

We now provide two subsections to support building synthetic traces. The first is a discussion on the uncertainty of compliance attitudes. The second is on the methodology used to build the synthetic traces.

6.1.1 Compliance Attitude Uncertainty

Within an organisation, a global compliance attitude exists for how individuals and groups respond to security incidents. Depending on the type of security incident and those involved, it could become a security violation or it could be prevented. We hope that individuals trained to perform tasks are security aware, but we regularly find that they circumvent organisational security policies [16].

Consider working with a company for a short period of time in order to identify the compliance attitudes of employees. We could ask them, where responses from interviews have led to popular theories such as the compliance budget [15]. Of course, respondents could lie, answer honestly but not behave consistently, or even fail to acknowledge that their behaviour is insecure.

Even if survey respondents answer honestly, it does not mean that this holds for the future. A secure employee interacting with a usable (non-secure) employee may be influenced

towards usable behaviours, creating an insecure compliance attitude. Of course, this is bi-directional where secure behaviour could inform more secure behaviour.

From an organisation's perspective, they only have so many tools to establish the global compliance attitude. For example, they could interview employees, then manually observe them via CCTV recordings to establish if their compliance attitudes matches their behaviour [28]. This is of course, costly and time consuming, where we would need to manually record the exact behaviour of each employee.

To add further complexity to the uncertainty of a global compliance attitude, someone who is secure may make a judgement of error causing a security incident. For example, Zhu *et. al.* showed they could get more information from people simply by providing them with information up front, exploiting a concept known as reciprocity where people who believed they were generally secure acted in an insecure manner [117]. They were able to influence people to sacrifice more information than they would usually part with.

The global compliance attitude of a company can be changed, for example, via training employees [100]. This behaviour change impacts how people respond to security incidents, for example a recently trained employee may have more awareness for *spear phishing* emails, and is less likely to click suspicious links.

6.1.2 Implementing to Build Synthetic Traces

We manually coded rules based on the model in Chapter 5 in the language Julia. We used the Sysmodels package which provided a management tool for deciding when agents would perform actions and so on [30, 27]. The model in Chapter 5 provides no basis for ordering agent's behaviour, as such we used a library in Julia to do this for us. It does raise the question that a different library or controller for action sequences would produce different outcomes. However, this is not in the scope of the work for now and could lend itself to future work.

We added parameterisation to the agent's profiles to model a range of situations. We introduce and discuss these parameters in this chapter in 6.3.1.

In the future we would desire a tool which would automatically build a model and transform the rules defined in Chapter 5.

6.2 What Can We Learn?

An organisation observes employee behaviour and accumulates information about security incidents. Using this data, the organisation wants to learn the compliance attitudes of employees. One possible solution is machine learning.

The data generated from observing employees forms a trace, where an entry in a trace describes who did what and when. It is similar to an intrusion detection system, where the logs of what happened are the entries, a collection of logs/entries forms a trace. The problem is, given a trace of interactions, can we use machine learning techniques to correctly identify the compliance attitudes of agents?

Table 6.1 Example Trace of Agents: A collection of entries forming a number of violations and preventions for four agents. Each agent is accompanied with a known compliance attitude.

Agent	**Violations**	**Preventions**	**compliance attitudes**
Alice	4	1	*Usable*
Bob	2	3	*Secure*
Charlie	1	0	*Usable*
Dan	2	5	*Secure*

Let us consider a simple example, where an agent's compliance attitude can be *usable* or *secure*. A *usable* agent is more likely not to follow a policy, whereas a *secure* agent is more likely to follow the policy. Table 6.1 lists four agents, the number of violations and preventions for a policy and their compliance attitudes. Given this data, the Decision Tree in Figure 6.2 can be formed. It is deterministic and will resolve to a value of *usable* or *secure* dependent on the entry being evaluated. In this case, an entry is the log of violations and preventions for an agent. The decision returned is the classification for an agent's compliance attitudes.

Table 6.1 is a small sample, however, an accurate tree has been learned in python using a machine learning algorithm with 100% accuracy for the training data[1]. A decision tree offers predictive power and, given an agent with some information, we wish to predict their compliance attitude. This is complex, as the decision tree from Figure 6.2 can be easily

[1]For more of a background on decision trees, refer to Chapter 2.

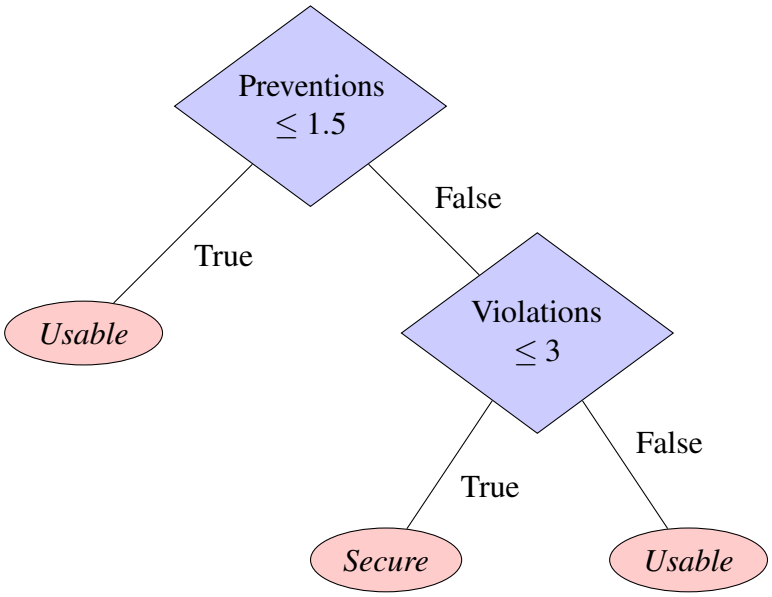


Figure 6.2 Learned Decision Tree from the data in Table 6.1. A diamond is a decision point, an oval is a decision.

fooled. For example, an agent with ten preventions and five violations will return as *usable*, even though they have acted securely for the majority of the time.

Establishing which features to consider could provide meaningful results for observing the global compliance attitude of employees in an organisation. Unfortunately, the problem is of greater complexity than what we identify here, as human behaviour is complex in itself and leads us to the previously introduced research question can we accurately learn compliance attitudes from public traces of compliance behaviours?

6.3 Multi Agent System - Simulation

We provide traces of agent's behaviour which is simulated on the semantics of the model described in Chapter 4. We provide further details of the simulation as a tool PCASP in Chapter 7. For now, it's important to know that traces of agent behaviour can be generated for the purpose of assessment via decision trees.

The use of rules for transitions are explained in Chapter 5. For clarification, we acknowledge that at times, certain rules within the system can be executed asynchronously. For example, choosing which agent will move location. In the case of two similar rules

conflicting with each other, we let the simulation use a random number generator to address this.

6.3.1 Model Parameters

We need to generate traces from a behavioural model in order to assess the accuracy of decision trees. The model and example we used is the one described in Chapter 5. As a reminder, Scenario 1 focuses on tailgating and the compliance attitudes of agents towards tailgating.

For the simulation, we reflect behaviours and attributes that we know exist. We understand that, whilst we will not have a model that truly reflects human behaviour, we at least can parameterise concepts that we know exist from the literature. Our parameters are as follows:

- p_1: Expected Arrival Rate - Agents arrive stochastically to the workplace reception, the arrival rate follows a normal distribution, agents can arrive at any point within some bounds. For example, if a start time for work is 9AM, we might expect some agents to turn up early, just before, just after, late or precisely on time. The value of this is implemented as an integer based on a probability density function where agents all arrive within one hour of each other.

- p_2: Probabilistic Decision - Assumptions have been made towards individuals as being *homo economicus*, where we make decisions based on personal gain or internal heuristics for guiding behaviour which look to maximise some reward [48]. Additionally, each day, experience is slightly different and for an agent, this could be the difference between a *secure* agent acting *usable* and vice versa, which is what we capture with our probabilistic decision, the ability for agents to act against their compliance attitudes [105]. The value for this is set at 0.8 for *secure* agents and 0.2 for *usable* agents respectively. We have no grounding for these values.

- p_3: Social Norms Influence - Social proof, where individuals assume the actions of those they have observed in order to reflect the interpreted cultural norms is apparent in many societies [95]. This is a boolean which is either true of false.

- p_4: Messenger Influence - *Authority*, is influencing by social/professional status, those we perceive to be in a position of power or responsibility can influence our behaviours [43, 33]. This is a boolean which is either true or false.

- p_5: Personality - Different personalities react differently to the same influences. We implement personality traits for agents as a string to refer to which personality, each personality is then subject to different influences:

    1. *Conscientiousness* - influenced by *Messenger*.
    2. *Agreeableness* - influenced by *Social Norms* and *Messenger*.
    3. *Extroversion* - influenced by *Social Norms* [109].

In a model, there is a distinct set of actions and observations recorded for agents. A trace generated from a model records agents moving between locations, tailgating being permitted or denied, agents successfully tailgating, agents failing to tailgate and agents observing tailgating being permitted or denied. This is all public information, the private information, such as the model parameters and agent contexts are not in a trace. A trace does not contain information such as if an agent can be influenced. The trace does not contain information such as the compliance attitude of the agent, which in this case is the hidden marker that we wish to observe from the publicly available information.

In Chapter 5 we did not discuss or introduce the notion of an expected arrival rate, a probabilistic decision and a personality. However, the implementation allowed us freedom to test these different parameters which we believe are crucial to capturing human behaviour. It does mean that the model from 5 has a lot of scope for future work by considering these potential avenues. We only provide a snapshot of what needs to be done for each parameter and it is at a very high level. Each of these parameters would require a significant amount of time and investment to replicate something that could be comparable to a real world data set.

6.3.2 Assessment Methodology

We define a model with a set of parameters such that each model contains a different set of parameters. However, the initial state of each model is identical in terms of agent context

and agent location. We then run a number of simulations for each model and generate a trace for each simulation run. A model will therefore, be associated with many traces.

A trace contains the number of entries equal to the number of agents, where an entry contains all of the features for an agent and is accompanied by the final compliance attitudes of that agent. The features of agents are the number of violations, preventions, attempts at tailgating and the number of times an agent is in close proximity when tailgating between other agents occurs.

Given all of the traces for a model, we use cross validation to construct and assess the accuracy of learning decision trees for each model. The cross validation consists of a training and testing phase, where the training is inclusive for the compliance attitudes of an agent and the testing phase is exclusive of the compliance attitudes.

A prediction from a decision tree in this chapter is either usable or secure. If we consider secure as our target value then a *true-positive* (tp) is a correct prediction for secure, *true-negative* (tn) is a correct prediction for usable. *False-positive* (fp) is an incorrect prediction for secure and *false-negative* (fn) is an incorrect prediction for usable. From these we can calculate the error rate, precision and recall:

$$err(fp,fn,tp,tn) = \frac{fp+fn}{tp+tn+fp+fn} \tag{6.1}$$

$$pr(tp,fp) = \frac{tp}{tp+fp} \quad r(tp,fn) = \frac{tp}{tp+fn} \tag{6.2}$$

The cross validation creates a number of decision trees for each parameterised model. Given a set of Decision Trees D where a set of testing traces T are present. A testing trace contains a set of entries E excluding the compliance attitudes, where a function $f : D \times E \to O$ takes a decision tree, an entry and returns an outcome O where $O = \{fp_o, fn_o, tp_o, tn_o\}$.

$$g : \mathbb{N} \times \mathbb{N} \times \mathbb{N} \times \mathbb{N} \times D \times \mathbb{P}(E) \to \mathbb{N} \times \mathbb{N} \times \mathbb{N} \times \mathbb{N} \tag{6.3}$$

$$g(i,j,k,l,d,E) = \begin{cases} g((i+1),j,k,l,d,(E \setminus e)), & \exists e \in E \text{ where } f(d,e) = fp_0 \\ g(i,(j+1),k,l,d,(E \setminus e)), & \exists e \in E \text{ where } f(d,e) = fn_0 \\ g(i,j,(k+1),l,d,(E \setminus e)), & \exists e \in E \text{ where } f(d,e) = tp_0 \\ g(,j,k,(l+1),d,(E \setminus e)), & \exists e \in E \text{ where } f(d,e) = tn_0 \\ (i,j,k,l) & \text{otherwise} \end{cases} \tag{6.4}$$

Using the function g which is a counter function we can calculate the number of different types of outcomes a decision tree produces. We can then use the function $calc$ to assess the accuracy of a decision tree:

$$calc : D \times \mathbb{P}(P) \rightarrow [0,1]$$

$$calc(d,E) = err(g(0,0,0,0,d,E)) \tag{6.5}$$

Once we can calculate the error rate for one decision tree, we can then assess the accuracy of all the decision trees generated for a particular model:

$$\mu_{error}(D,E) = \frac{\sum_{d \in D} calc(d,E)}{|D|} \tag{6.6}$$

We do the same for the precision and recall of the decision trees, however, we do not provide the notation for this as it is follows the same principles. Each model is associated with a set of decision trees. Therefore, for each model we calculate μ_{error} to identify the accuracy of decision trees for a given set of parameters.

A simplified overview of Functions 6.4 and 6.5 is that they count the number of appearances for true positive, true negative, false positive and false negative respectively.

6.4 Analysis - Case Study

In this section we discuss the use of parameterised models and make remarks surrounding the results for three different amounts of simulated agents which are fifty, one hundred and two hundred.

Table 6.2 100 Agents: p_1: Expected Arrival Rate; p_2: Probabilistic Decision; p_3: Norms Influence (Social Proof); p_4: Messenger Influence; p_5: Personality; $\mu(error)$: Average error rate of a model; $pr(s)$: The precision of the model towards *secure*; $r(s)$: The recall of the model for *secure*;

p_1	p_2	p_3	p_4	p_5	**Model**	$\mu(error)$	$\sigma(error)$	$pr(s)$	$r(s)$
					m_1	0.255	0.067	0.659	0.830
✓					m_2	0.001	0.002	1.000	0.997
✓	✓				m_3	0.234	0.028	0.697	0.712
✓		✓			m_4	0.073	0.019	0.963	0.953
✓			✓		m_5	0.160	0.024	0.884	0.898
✓		✓		✓	m_6	0.094	0.018	0.928	0.938
✓			✓	✓	m_7	0.114	0.016	0.904	0.910
✓	✓	✓			m_8	0.271	0.024	0.724	0.731
✓	✓		✓		m_9	0.367	0.031	0.634	0.624
✓		✓	✓	✓	m_{10}	0.027	0.012	0.975	0.969
✓	✓	✓	✓	✓	m_{11}	0.277	0.028	0.675	0.675

The number of possible parameterised models is 2^5, however, we only consider eleven of these thirty two. The expected arrival rate is included in the majority of the parameterised models, as we do not consider too many models where all agents always arrive at the exact same time, of course this could happen, but it is very unlikely. The personality parameter is dependent upon a behaviour change parameter being present, therefore, it does not add to a model if *Social Norms* and/or the *Messenger* parameters are not included. These are the motivations for considering only eleven parameterised models.

We used the *Julia* programming language to implement our case study and made use of the SysModels package [30, 27]. We generated synthetic data on a Toshiba laptop with a 2.4 GHz i5 processor and 8GB RAM. To generate the data for eleven models with two hundred agents it took twenty two minutes which is roughly two minutes per model. Each model is generated with ten traces each starting from an identical initial state for each model.

For the analysis we performed three test cases and used fifty, one hundred and two hundred agents. Table 6.2 is the results for the one hundred agents, the results for fifty and two hundred agents in Tables 6.3 and 6.4 respectively.

For each test case, we calculated the average error rate, the standard deviation, the precision and the recall of each parameterised model, where Table 6.2 shows the parameters for each model. We now make remarks regarding the results we have obtained.

Remark 4. *The average error rate for model m_1 is significantly more accurate with fifty than one hundred or two hundred agents.*

With regards to Remark 4, as the expected arrival time is not set, all agents arrive at the same time. The majority of agents don't ever permit, deny or attempt to tailgate, therefore, a decision tree will make inaccurate predictions for some agents, particularly when more than fifty agents are used.

Remark 5. *If the probabilistic parameter is set, then the average error rate significantly increases. In particular, it impacts more than both the Messenger and Social Norms parameters.*

The use of the probabilistic parameter significantly increases the average error rate of the decision trees. Due to the uncertainty of agent behaviour, i.e. secure agents acting usable and vice versa, a secure agent could have always behaved as usable. A classifier model would always conclude that they are usable when they are in fact secure. Whilst Remark 5 is not surprising, the impact of uncertain behaviour against social influences is a useful result for a security practitioner. In the real world, some people will always be secure or usable, some hover between the two and some are slightly more secure or slightly more usable, some insight towards these numbers would allow us to calculate the impact of agents towards a model.

Remark 6. *The Messenger influence has a slightly more of an impact to the error rate, precision and recall of a model than Social Norms. It is true for all four of the test cases. They both impact the error rate, precision and recall of every model.*

The influences themselves differ in how they are implemented. The *Messenger* relies on an agent observing a behaviour of another agent that they consider to be an authoritative figure. The social norms is a cumulative influence, where the number of observations of a particular action can trigger the compliance attitudes of an agent to change. For Remark 6, the interest is that they are not probabilistic behaviours, they are private behaviours. We have defined very simple rules for our influences. We wish to know if decision trees are capable of generating rules to deal with these simple behaviour changes. Given the data for our number of agents. We can see a slight improvement when 200 agents are present. However, the decision trees still perform poorly for these basic implementations of influences.

Remark 7. *On average, the models for two hundred agents are more accurate than the models for fifty or one hundred agents.*

A trend emerged for the accuracy of models as we increased the number of agents. Whilst some of the models were more accurate for 50 agents, in general Remark 7 holds, in particular for the complex models where influences and probabilistic decisions are present. This is due to an increase number of entries to train decision trees, improving its accuracy.

Overall, we can see that with some basic aspects of human behaviour such as an uncertainty of decisions between secure and usable agents, the decision trees perform poorly. Even more so that we are just considering the polar opposites for compliance attitudes. Whilst the influencing effects implemented are relatively simple, we believe as they increase in complexity, i.e. become heterogeneous for each agent, this would reduce the accuracy of decision trees even more. On another note, and mainly due to processing limitations, it's not clear if the accuracy can be improved by generating thousands of traces to use in the cross validation analysis.

Table 6.3 50 Agents; See Table 6.2 for column definitions.

Model	$\mu(error)$	$\sigma(error)$	$pr(s)$	$r(s)$
m_1	0.070	0.037	0.896	0.943
m_2	0.018	0.010	0.974	0.980
m_3	0.279	0.035	0.658	0.642
m_4	0.050	0.025	0.947	0.950
m_5	0.162	0.029	0.853	0.867
m_6	0.031	0.018	0.955	0.977
m_7	0.091	0.030	0.893	0.937
m_8	0.266	0.039	0.701	0.686
m_9	0.365	0.051	0.694	0.656
m_{10}	0.017	0.012	0.976	0.986
m_{11}	0.325	0.056	0.622	0.581

6.5 Discussion

The decision trees constructed in this chapter was on a boolean case for compliance attitudes. An agent that is usable would in general, act differently to an agent that is secure. We

Table 6.4 200 Agents; See Table 6.2 for column definitions.

Model	$\mu(error)$	$\sigma(error)$	$pr(s)$	$r(s)$
m_1	0.201	0.135	0.861	0.912
m_2	0.006	0.003	0.996	0.998
m_3	0.170	0.013	0.910	0.888
m_4	0.014	0.006	0.993	0.993
m_5	0.050	0.009	0.976	0.972
m_6	0.024	0.007	0.984	0.990
m_7	0.047	0.011	0.976	0.973
m_8	0.140	0.021	0.933	0.912
m_9	0.277	0.029	0.833	0.812
m_{10}	0.040	0.008	0.975	0.980
m_{11}	0.161	0.016	0.920	0.892

acknowledge that this is not a complete model representing human behaviour, instead it is characterising elements that we see in the real world.

Figure 6.3 is an overview of the rate when comparing fifty to two hundred agents. Intuitively, one would expect that the more data points a ML algorithm has, the more accurate it should become. In general, this does tend to be the case, however, the parameters for m_1 show that a decision tree for fifty agents is more accurate than the decision trees for one hundred or two hundred agents. We suspect that this is due to the clash of more agents appearing at once and queuing as a result of the parameter p_1 which is the expected arrival rate of agents.

6.6 Validation

In this chapter we utilised cross validation as a method to assess the accuracy of decision trees for the traces generated by simulation. The traces used could have been collected by real world data, however, collecting the data for compliance attitudes would have been a much more complex process.

The decision trees used in this Chapter is a model built on a trace output from the simulation of a synthetic data set. In Chapter 7, we collect a data set which focuses on just the compliance behaviour, which itself is complex and does not include the compliance attitudes.

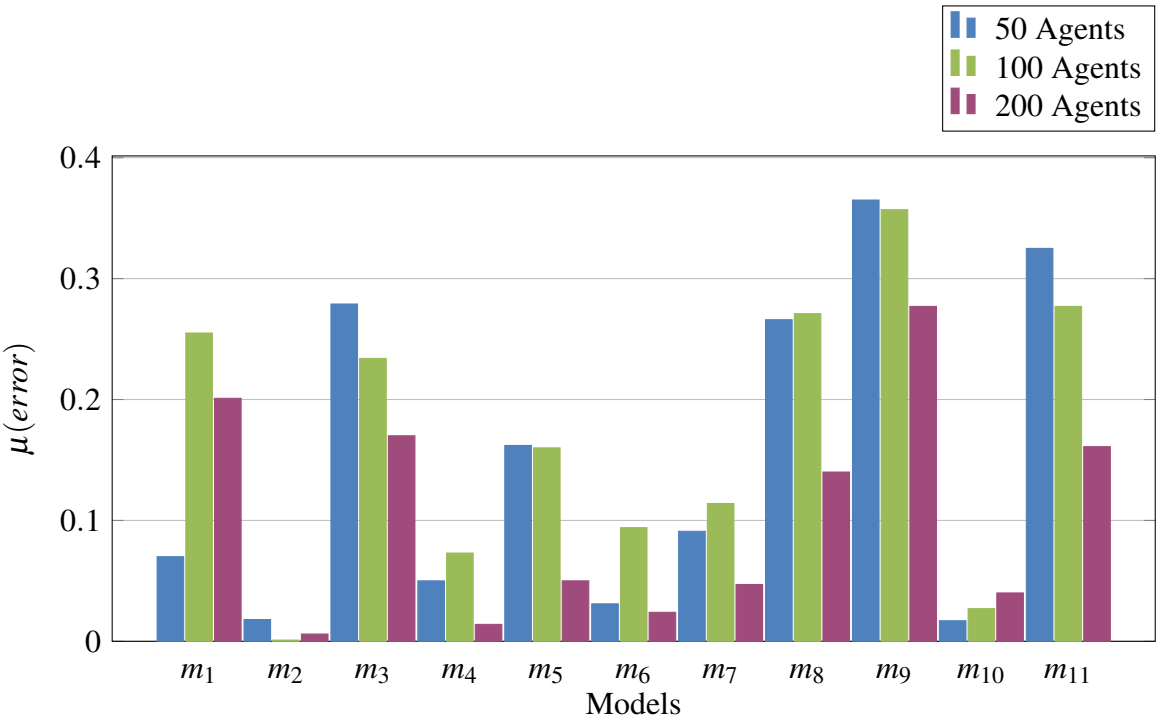


Figure 6.3 Overview of the mean error rate for 50,100 and 200 agents.

However, if we knew the compliance attitudes of the participants in the user study then we could apply the machine learning techniques to assess how accurate decision trees are for real data as opposed to synthetic data.

In this chapter we have presented a model which is based on the model from Chapter 5. The model from Chapter 5 is one step away from reality meaning its not clear how at the present time how accurate or truly reflective the model is of human behaviour. Adding another model on top of this is to be puts the DT model two steps away from reality. Any effects or results that we are seeing could be amplified by errors in the first model and errors in the second model. The work in this thesis and chapter is a starting point and for any confidence in the models to be established, future work must iterate over and improve the current models.

This stacked model approach of a first layer consisting of the model from Chapter 5 and the decision tree model as the second layer creates a set of four possibilities. Firstly, both models in both layers are a true reflection of reality, in which case further work is required to increase confidence. Secondly, the first layer is accurate and the second layer is flawed, meaning that machine learning applied to human behaviour is no further forward. Thirdly, the

original model is incorrect and the decision trees are correct which would require a complete overhaul of both models to ensure accuracy of reality is truly reflected. Fourthly, both models are incorrect which would require a completely new approach to build a foundation in this area.

6.7 Chapter Conclusion

This chapter focused on addressing Research Question 3 and that of identifying the compliance attitudes from publicly observable behaviours. With very simple model traces to analyse we saw a relatively high success rate but as the complexity of the traces increased, the accuracy of the decision trees decreased. This was particularly apparent when dealing with randomised behaviour.

It may be the case that the decision trees are biased and lend themselves towards latching onto specific values [113]. Even with identifying a binary compliance attitude we had a mean error rate of nearly 40% on one of the decision trees. This is very close to the flip of a coin when considering just *usable* or *secure* behaviour.

For an organisation, this method offers little in guarantees towards identifying compliance attitudes behind agents behaviours for security policy compliance. Nevertheless, it does open up an avenue for future work where we can assess at what point does a decision tree become accurate if any, and what behaviour observations are necessary to identify compliance attitudes.

A key insight from this chapter is the relationship between complex traces and the accuracy of decisions trees. As the complexity of the traces increased, the accuracy of the decision trees decreased. A question arises as to what data would be required in order to maintain the accuracy of the decision trees.

The success of this chapter is how the different parameters impacted the complexity of each trace. A parameter such as expected arrival time had enough of an impact to change agent's observations which subsequently impacts how they are influenced. As a comparison to reality, if a person arrives after or before a security policy breach then they will not witness it and as such, will not be influenced by the breach and any subsequent actions.

One of the parameters not discussed was the number of days for observations. In the simulations, they would run for five working days, so there would be at least five opportunities for tailgating to occur. Had we have extended the parameter for number of days to be ten or one hundred it's not clear whether or not this would have improved the accuracy of the decision trees. Nevertheless it is something to note for future work. Is there a point where enough data will ensure accurate decision trees.

Some of the models were more accurate for 50 agents than the larger populations. It is most likely caused by the lack of a probabilistic decision and the expected arrival time not forming a queue of people to enter. Without a queue of people waiting to enter it's likely that tailgate happens less in the simulations when there are fewer agents. However, we do not know this for certain and more work is required in this area to understand if it is the case.

A clear limitation of this work is the lack of evidence for smaller groups. In the later Chapters we discuss experiments consisting of 4-6 participants which means it is unreasonable to assume we can offer some comparison. If we performed a study to simulate for smaller groups it may offer insights into key influencers within the system. Furthermore, a range of participants would demonstrate if an influencer strength decreases or increases with regards to a given set of agents.

Chapter 7

Social Influences Towards Compliance Behaviour

This chapter address Research Question 4, as a reminder the research question is the following:

Research Question 4 (User Study)**.** *How can we investigate and measure the effect of behavioural interventions such as social influences on compliance behaviour?*

To address Research Question 4 we consider that a social influence of the *Messenger* effect or the broken-window effect has an impact on people's compliance behaviour for security policies.

We conducted a study with three conditions in a between-subjects design on a sample of $N = 54$ university students owning a university-issued access-control smart card. In all three groups, participants were asked to complete a set of Capture the Flag challenges in our Cyber Security Room where they had no Internet access. Outside of the room, the participants could access the Internet. Thereby, participants were compelled to leave and enter the room, without disclosing the experiment purpose.

We asked participants to swipe their smart card on entering and exiting the Cyber Security Room. The Control group had no intervention. The Discrete+ experiment group was exposed to a *Messenger* influence [44]. The Continuous− experiment group was exposed to a an *untidy* Cyber Security Room, which was inspired by the broken-window Theory [114]. The broken-window theory states that visible signs of crime or anti-social behaviour encourages

further crime and anti-social behaviour. We use this theory as a basis to decrease compliance behaviour.

We measured swiping behaviour on entry and exit, computing total swipes and swipe rate ratio as key metrics subjected to an Analysis of Variance. We found a statistically significant large effect of the Continuous− intervention, that is, the broken-window effect impacting the total swipe rate negatively, Hedges' $g = -1.04$, 95% CI $[-1.78, -0.28]$.

While having observed a negligible effect size between the Control and Discrete+ groups, we acknowledge that the swipe rate was high for both of these groups. From the effect size of the Control group and Continuous− group, we offer evidence towards an untidy area and lack of compliance for security policies.

One of the social influences from MINDSPACE is the *Messenger* effect which is strongly related to Cialdini's *Authority* principle. The Messenger effect states that a person in authority can influence the decisions that other people make [44]. Social influences feature regularly when discussing social engineering and it is commonplace to hear that wearing a uniform will get you past the front door [87]. Social influences are one example of a behavioural interventions [81].

The broken-window theory states that an area left in disrepair can increase the crime rate in the surrounding area [114]. The theory goes on to discuss how the perceptions from people that nobody is taking care of the area also contributes to an increase in crime [114]. To the best of our knowledge, no one has taken this concept and applied it to the compliance of security policies. The broken-window theory is an example of one behavioural intervention.

An organisational requirement such as *no unauthorised personnel in the building* can be implemented with a security policy where employees must swipe their smart cards to access certain rooms and help to record where they are, where they are going and where they have been. The information can be used for assessing who could be responsible for permitting unauthorised personnel entry should a breach occur. Depending on the compliance intentions of employees, they may not always comply with the policy [15]. We pose the question can behavioural interventions can change their behaviour for how they interact with swiping their smart card? Which leads us to the aim for this chapter:

Aim: ***We investigate the effect of socio-environmental interventions on smart card swiping behaviour.***

Chapter Overview: This chapter addresses Research Question 4 and offers a user study where social influences impact compliance behaviour. The chapter is formed from a paper at Socio-Technical Aspects of Security and Trust 2018, which at the time of thesis submission, is currently in the process of publication [25]. This data set also offers up speculative work which we carry out in Chapter 9 towards the social influences within each group that participated in the study. The data set was collected via a user study at Newcastle University with 54 participants.

7.1 Preliminaries

In this section we provide a brief overview of the methods and metrics used in this chapter to familiarise the reader.

7.1.1 Analysis of Variance

An analysis of variance (ANOVA) is a set of statistical methods to analyse and approximate the differences between group means of some samples. In this chapter we used a one-way omnibus ANOVA which entails comparing the means of all the different groups and assessing if there is a difference. In our case, the three groups are Control, Discrete+ and Continuous−. The ANOVA will not tell us which groups are statistically significant, just that at least two of the groups are.

7.1.2 Effect Sizes

An effect size is a quantitative measure of an observation made about some metric, such as the correlation between two variables. We calculate the Hedges'g effect size which tells us the difference between two groups [57].

As Hedges'g only permits the comparisons between two groups we perform planned comparisons. Planned comparisons are pre-specified before the results are collected to avoid *looking* for results that may be interesting if the data does not quite add up. The planned comparisons we perform in this chapter is Control vs Discrete+ and Control vs Continuous−.

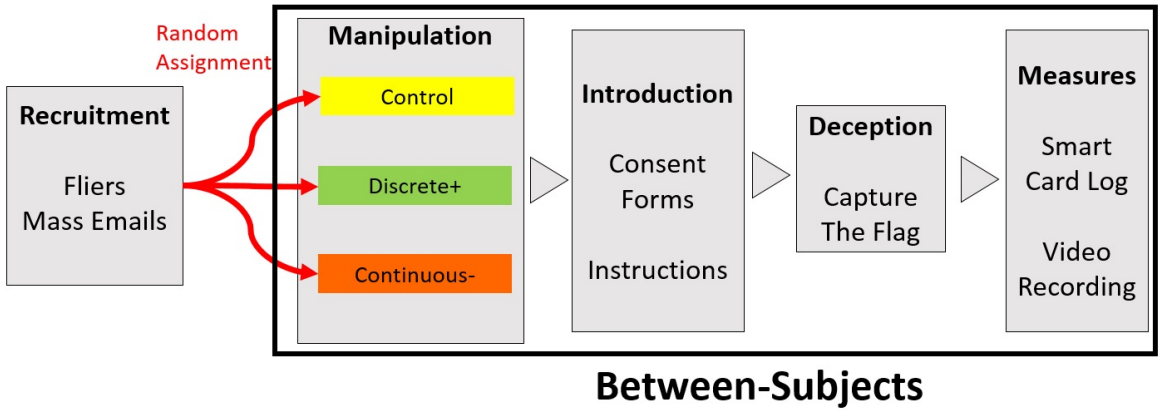


Figure 7.1 Study Structure: An overview of what study participants were exposed too.

7.1.3 Confidence Intervals

A confidence interval is an interval estimate we compute to outline the range of values that an observed value could take for some confidence level. The confidence level is typically 95%, which is what we use in this chapter [59]. The confidence interval we provide is over the effect sizes that we observe.

7.1.4 The Broken-Window Theory

When crime is quantified, the environment by which a person lives or resides in can impact the level of crime for that area [18]. Social disorganization theory focuses on the relationship between neighbourhood structure, social control, and crime [74]. An intervention in a neighbourhood structure could change the resulting crime rate.

One study explored the possibility that academics are *messy*. By creating an environment where *mess* was present, they found that people were more likely to litter [94]. They state that when academics observed that others are violating the social norm of keeping the room clean then the probability of littering increases by roughly 40% [94]. This work is an example interpretation of the broken window theory where the researchers adapted it for the workplace.

7.1.5 Social Influences - Messenger

When considering COM-B the Behaviour Change Framework which is discussed in more depth in Chapter 2, the motivational element can be targeted by social influences [81]. The term *social influence* refers to the ways in which a person's beliefs, attitudes and subsequent actions change as a result of social interactions with people [50]. One example is that normative beliefs regarding the expectations of colleagues has an impact on employee security behaviours. Not only the expectations of others, but also the perceived behaviour of others was found to contribute in employee attitudes towards security policy compliance [58]. Consider it a form of exchange between two or more individuals.

MINDSPACE is a behaviour change framework that lists many different types of social influences and is discussed more in depth in Chapter 2 [44]. One social influence they describe and one that is repeated throughout this thesis is the *Messenger* effect which is derived by the work carried out by Cialdini [33]. The *Messenger* effect states that:

"A person in authority or one that is trusted has an influence over the decisions that other people make" [44].

The effect is not described as having a positive or negative impact, just that it can alter the decision a person makes.

In the context of security, Zhu *et al.* demonstrated by mutually exchanging information, they could utilise another type of influence which is reciprocity. By providing different types of information, they were able to increase the amount of personal information that people reciprocated [117]. As an influence mechanism, reciprocity along with others are well established, most notably from the behavioural scientist Cialdini [33]. We often see poor security decisions in Phishing, where social influences are common (e.g. an email impersonating a bank is a form of social influence) [41]. Phishing attacks are one example of how a social influence can nudge a persons decision. On the physical side, where people interact face to face, it has been shown that top level management can significantly increase policy compliance rates at lower levels [62]. When comparing top level management as a social influence we can see similarities with the *Messenger* effect from MINDSPACE [44].

Table 7.1 Operationalisation of Study: Interventions on Smart Card Swiping Behaviour

	Groups	Intervention	Instrument
IV: Condition	Control	None	None
	Discrete+	*Messenger* [44]	Reminder
	Continuous−	broken-window [94]	Messy Room
DV: Swipe	All	Visual	Video Camera
		Log	Adafruit PN532 Reader

The social influences can target on an individual basis, such as the study by Zhu *et al.* on reciprocity [117]. In the context of influencing a group, one can consider that social conformity could play a vital role [9]. It has been shown that social conformity can be dependent on different cultures as they vary in the amount to which they nurture conformity [35]. Should an influence be targeted at the group, it can very well depend on which participants respond to it, as they could be key to creating conformity amongst others.

A social influence over a persons attitudes can often come from interactions and observations of others [34]. As such, it may be the case that the social influence propagates in a group and a particular behaviour establishes itself as the social norm [24].

7.2 Chapter Aims

We provide sub research questions and the hypotheses that form this chapter:

Sub Research Question 1 (Messenger Effect)**.** *To what extent does a Messenger effect have on the swipe rate of participants?*

Table 7.1 provides an overview of the operationalisation for this research question. As an independent variable (IV), we have selected the use of the *Messenger* effect [44].

Sub Research Question 2 (Broken-window Effect)**.** *To what extend does the broken-window effect have an effect on the swipe rate of participants?*

Table 7.1 also provides the info for the second research question. As the IV for this case, we consider a messy ('untidy') room as the intervention.

For both sub research questions, we intend to compare them against a Control group which has no experimental treatment. For all three groups (Control, Discrete+ and Continuous−),

we will measure the dependent variable (DV) by use of a video camera and smart card logging system. Using both of these metrics, we will generate an event log for all participants.

Hypotheses on Total Number of Swipes The overall null hypothesis for this experiment is $H_{\mathsf{T},0}$ *: There is no mean difference in the total number of swipes for participants exposed to an intervention.*

We have subordinate null hypotheses for each condition:

$H_{\mathsf{T},0,\mathsf{D}+}$**:** *The* Discrete+ *intervention does not impact the mean total number of swipes.*

$H_{\mathsf{T},0,\mathsf{C}-}$**:** *: The* Continuous− *intervention does not impact the mean total number of swipes.*

We have these as alternative hypotheses.

$H_{\mathsf{T},1,\mathsf{D}+}$**:** *The* Discrete+ *intervention (Messenger effect) impacts the total number of swipes.*

$H_{T,1,\mathsf{C}-}$**:** *A* Continuous− *messy ('untidy') room (broken-window effect) impacts the total number of swipes.*

Hypotheses on Swipe Rate Ratio The overall null hypothesis for this experiment is $H_{\mathsf{R},0}$ *: There is no mean difference in the swipe rate ratio for participants exposed to an intervention.*

The subordinate null hypotheses for each condition are:

$H_{\mathsf{R},0,\mathsf{D}+}$**:** *The* Discrete+ *intervention does not impact the mean swipe rate ratio.*

$H_{\mathsf{R},0,\mathsf{C}-}$**:** *The* Continuous− *intervention does not impact the mean swipe rate ratio.*

$H_{\mathsf{R},1,\mathsf{D}+}$**:** *The* Discrete+ *intervention (Messenger effect) impacts the swipe rate ratio.*

$H_{R,1,\mathsf{C}-}$**:** *A* Continuous− *messy ('untidy') room (broken-window effect) impacts the swipe rate ratio.*

7.3 Method

For reproducibility and scientific integrity, the study has been registered and updated with statistical analysis procedures before proceeding with any tasks at the Open Science Framework (OSF) [1]. Tables for descriptive statistics, ANOVA results and effect size graphs were generated from the data by R. We also provide a copy of the pre-registration in Appendix A.

[1] https://osf.io/3jsc7/

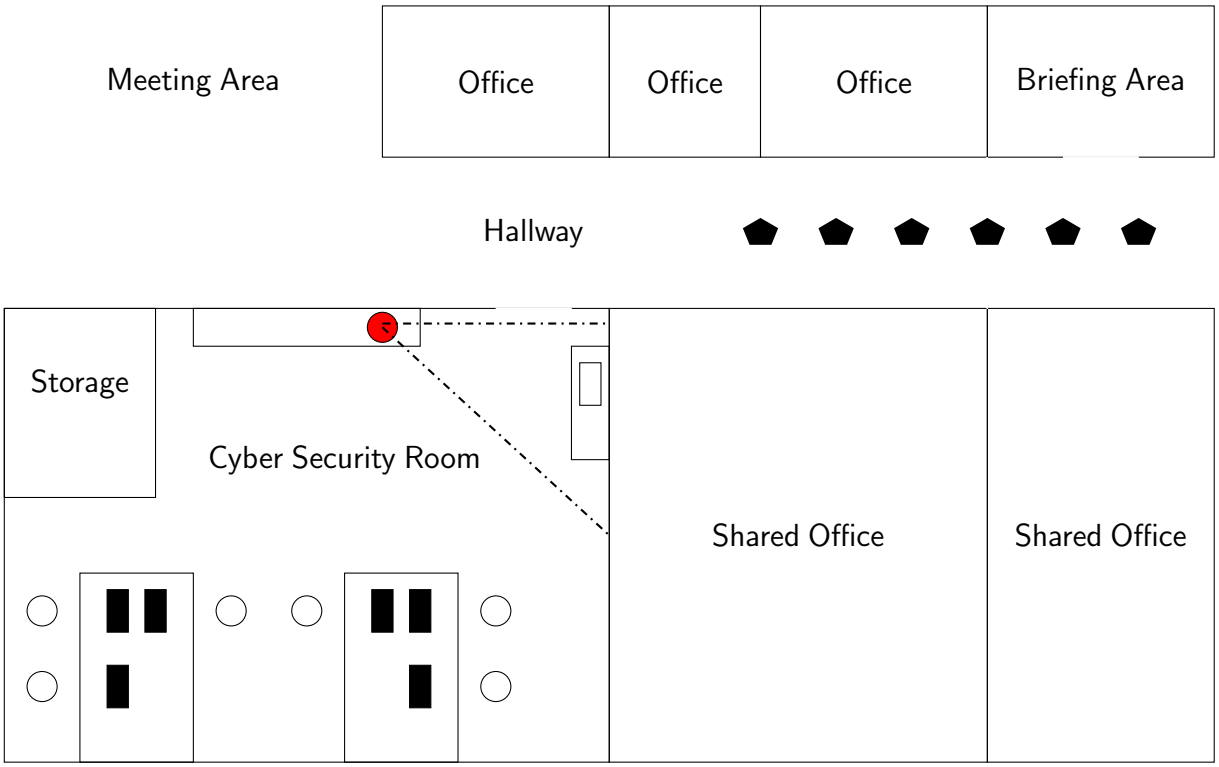


Figure 7.2 Floor Plan of the office space used for the study.
Raspberry Pis; Desktops (Challenges); Camera; Smart Card Reader;

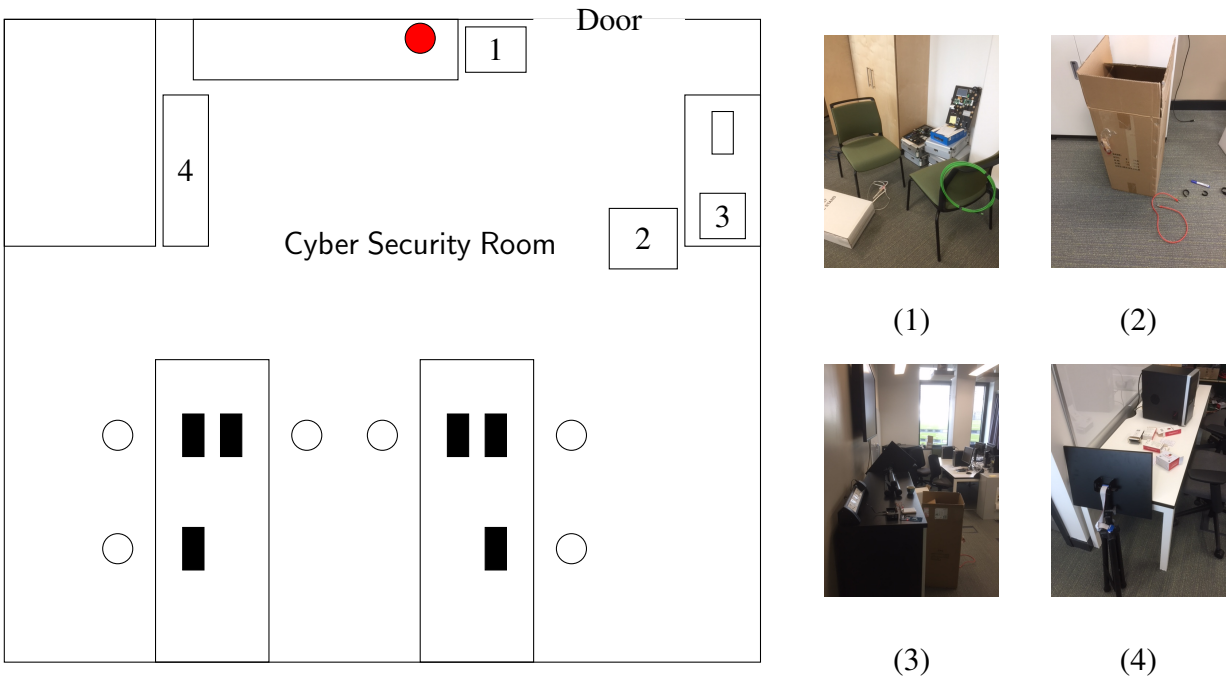


Figure 7.3 Larger view of the Cyber Security Room and the messy items that were placed around the room.

7.3.1 Experiment Setup

The experiment is setup as a between-subjects experiment, with one IV (condition), with three levels (Control, Discrete+ and Continuous−).

We ran all groups as morning, afternoon or evening sessions randomly distributed over conditions. At no point did we raise awareness to the existence of any other group types. The study itself had to be conducted with physical presence as we required participants to engage in a physical behaviour, that of swiping their smart card. An overview of the study as experienced by participants is provided in Figure 7.1.

7.3.2 Sampling

The survey population was students of Newcastle University. We chose this population as all students are issued a smart card during induction, which we use in this experiment. In the study, only undergraduate and postgraduate students were recruited. No members of staff took part in this study. The participants were recruited through an emailing list with in the University, flyers and announcements in lectures and seminars. Overall, we classify the sampling process as convenience sampling.

The participants for all groups were recruited as one big cohort under the offer of a Cyber Security Study.

7.3.3 Grouping & Assignment

Participants are randomly assigned to three groups:

- Control: Our first group is our baseline where participants perform a set of tasks over fifty five minutes. No intervention is in place here. The room is set in a defined ‘tidy’ state.

- Discrete+: The first experiment group is exposed to a discrete intervention. By discrete we mean at one time point the group is exposed to some intervention. The purpose of this intervention is to positively influence group behaviour around swipe compliance. The room is kept in a defined ‘tidy’ state. We anticipate that this will increase participants swipe levels.

- Continuous−: The second experiment group is exposed to a continuous intervention present in the *offline* Cyber Security Room. The intervention is specifically to set the room in a defined 'untidy' state. The purpose of this intervention is to negatively influence group behaviour around swipe compliance. We anticipate that this will decrease participants swipe levels.

In each group we will ensure a short break is in place at the 25 minute mark. At this point we will gather all of the participants in the Cyber Security room and ask them how they are getting on. The instructor will swipe their smart card as they enter and say *"How is everyone finding the tasks?"*. The instructor will then swipe their smart card then leave the room. Each groups session time will be fifty five minutes ($\pm$20s).

7.3.4 Experimental Environment

The experiment was conducted in a purpose-built environment for cyber-security capture-the-flag tasks. We offer a floor plan of the environment in Fig. 7.2.

In the experiment, we designated *offline* and *online* areas. The *offline* area had no Internet connectivity, which would be considered a high-security clean-room, which in this study is the Cyber Security Room. The *online* area had Raspberry Pis provided to allow for Internet connectivity. Subjects must swipe their student smart card as they enter and exit the *offline* area.

Offline Area: The *offline* area includes a Cyber Security Room and contains sensitive experimental equipment. During the building's construction, the room was not fitted with a smart card reader. We make the following changes to this room:

- Using a Raspberry Pi we design and program an Adafruit PN532 RFID Card Reader to work with Newcastle University Smart Cards.
- We fit a Video Camera to record the experiment. The camera is inside the offline room, points at the door, and observes the smart card reader.

Online Area: Subjects will use this area to collect supporting information to assist the completion of the Capture-The-Flag challenges in the cyber security room. This area has Raspberry Pis with a mouse, keyboard, monitor and wireless network connection.

7.3.5 Procedure

For all the groups, a mass email was sent out to all students. A potential participant responded showing interest was randomly assigned to a group (Control, Discrete+, Continuous−). A response was then sent to each potential participant with the choice of three sessions (morning, afternoon, evening). In the response, the potential participants are made fully aware that they will need a university smart card, the study will be partially filmed and that the study will take place with other participants. Participants would then respond selecting either a morning, afternoon or evening slot. We send a confirmation receipt to them informing that they are fully booked in for the group and should turn up at a specific time. One day before each session we sent a reminder confirming that the session is still going ahead.

Registration & Welcome

Upon arrival, participants were met at the entrance of the study area. Participants were provided with a random number they select from a box upon arrival. They were asked to read through documentation and sign in the appropriate places. We made it clear that they could opt out and ask any question at any point.

Once all participants had arrived and had read through the documentation and signed to agree to participate, we briefed them. We explained they would be completing Capture the Flag (CTF) challenges in the Cyber Security Room. In the Cyber Security Room, they would have no Internet connection and it was an *offline* room and are not permitted to use any device capable of a Internet connection. Should they wish to gather information on a challenge, they can leave the Cyber Security Room and make use of machines connected to the Internet in the hallway. These machines were Raspberry Pi 3's with a mouse, keyboard, monitor and connected to the University Wireless Network. We state that when entering and exiting the room, participants swipe their smart card. We provide them with an example Adafruit PN532 smart card reader and demonstrate how swiping takes place. It should be noted that all participants are familiar with swiping their smart cards at the University, albeit not on entry and exit of a room where the door is open. Participants then ask any questions and we instruct them to leave in a specific order at fifteen second intervals to enter the Cyber Security Room. The specific order refers to the number they selected and also indicates the number desk they will sit at in the Cyber Security Rooms.

Capture the Flag Tasks

Participants were provided with CTF challenges which took on a jeopardy format. A challenge in CTF that is jeopardy requires the participant to find a *flag* in whatever security puzzle is presented to them. In total there were nine challenges for them to solve. They were instructed to solve the first eight then move onto the additional challenge which was in place should a subject finish the first eight challenges. No participant successfully completed all nine challenges. Each participant was given a desktop which matched the number they originally selected upon arrival.

Each challenge consisted of an encrypted file alongside a puzzle. When the puzzle is solved it revealed a password that could be used to decrypt the encrypted file. Once decrypted the participant is rewarded with a success message. One of the challenges was called BASE64ME and users were presented with:

The password for this challenge is : dGhhdFdhc0Vhc3lUb1NvbHZlMQ ==

Once the participant identifies that this is in the data format of Base 64 and converts it they are given the following password:

thatWasEasyToSolve1

After the password is used to decrypt the file the participant is presented with a file that says congratulations and they have completed the challenge. The challenges were presented in no particular order, allowing participants freedom to choose which order they wished to solve the challenges. Albeit, we advised they should complete the first 8 challenges before moving onto the additional 9th challenge. Challenges were designed to be random amongst participants ensuring they couldn't copy the final answer.

7.3.6 Manipulations

Experiment Group One – Discrete+*:* The goal of the discrete intervention is to create a positive outlook on swiping a smart card for participants. The re-enforcement from the instructor provides that authoritative message that this is the behaviour we expect.

During the short break, once all subjects are in the clean room we will provide our discrete intervention. As the instructor, I will swipe my card as I enter and say to all participants *"Just*

a reminder to make sure you all are swiping as you enter and exit the Cyber Security Room. How is everyone finding the tasks?"

Experiment Group Two – Continuous−*:* The goal of the continuous intervention is to reduce the swiping of individuals by creating a untidy environment. The continuous intervention will consist of making the clean room *untidy*. The break at the 25 minute mark will still take place. As the instructor I will swipe my smart card as I enter and say *"How is everyone finding the tasks?"*. We offer Figure 7.3 as an overview of the messy items that were placed around the room.

7.3.7 Measurements

In all of the studies we measure the total number of swipes per participant and the swipe rate ratio per participant. Total swipes is the cumulative number of swipes for both entry and exit combined. For swipe rate ratio, a value of 1 would indicate that the participant swiped every time they entered and exited the Cyber Security Room. A swipe rate ratio of 0 would indicate that the participant never swiped when entering or exiting the Cyber Security Room.

7.3.8 Controlling Confounding Variables

Controlling Intervention Side Effects

Experiment Group One – Discrete+*:* We pre-specify the exact wording of the message used at the short break. The method for delivering the message will be consistent. The instructor will be the same for all studies and will wear very similar clothing for all studies to ensure there is no bias or unintended intervention.

Experiment Group Two – Continuous−*:* The placement of items around the cyber security room could restrict or obstruct a persons ability to swipe. As a counter measure we will place the items around the room, entered and exited to and from each desk to ensure no physical blockade has unintentionally appeared. We can confirm that this is the case and none of the items will impact a participants ability to swipe. Additionally, we will make a log of all items used and their location within the cyber security room.

Controlling Measurement Error

We deploy our own method to ensure that when reporting the final data set we are ensuring rigour and accuracy for each event that occurs. We deploy the following technique:

1. **Synchronise** the times from the video footage and smart card logs.
2. **Correlate** the visual initial swipe of a subject with the hashed smart card log for that swipe.
3. **Generate** the list of events for the timed part of the study, i.e. where subjects are completing challenges.

Synchronise: The clock on the video footage begins at 00:00 and the smart card logs time as per Greenwich Mean Time. For each experiment, the study leader swiped their smart card after the camera is switched to record mode. From this we can gather the exact timings of swipes with regards to the smart card logs. It also assisted when generating the logs for the visual logs data set, as it provides an absolute time.

Correlate: As all participants smart cards are hashed, we have no association between the names of subjects and their smart card number. To ensure that the visual logs can be generated we correlate the initial swipe of a subject with visual features. We use the following metrics to classify a subject and address conflict resolution:

- Are they wearing glasses?
- What colour top/t-shirt/coat are they wearing?
- What colour hair do they have?
- If these three features clash with another participant in that group, do they have any unique features that are easily identifiable?
- What is the hash of their smart card?

We should know the hash of each user's smart card, they also entered in a specific ordering, which is associated with a number for where they are sat. This initial swipe was

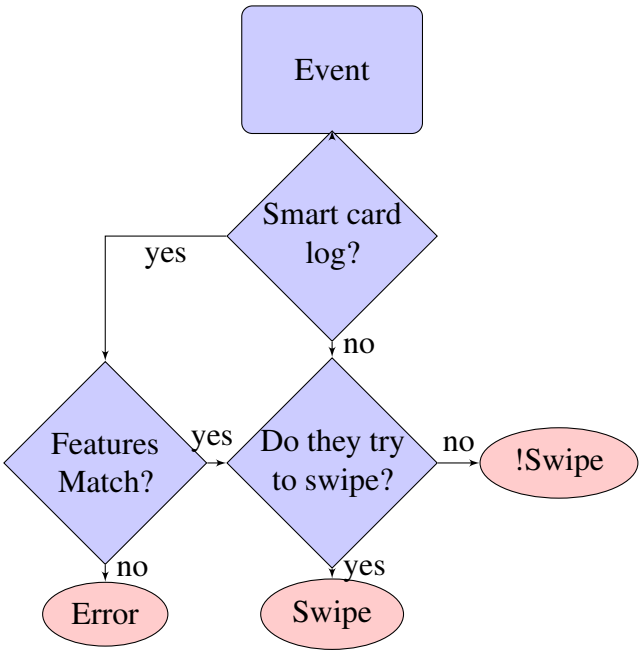


Figure 7.4 Process for recording an event, where Swipe refers to the expected behaviour of someone swiping, !Swipe is someone entering or exiting without swiping. Finally, error refers to someone swiping where either the logs are incorrect or they are using a different smart card.

before the timed section of the study began and was used to correlate their features to their smart card hash value.

Generate: In order to generate visual logs we used the video footage to gather information for each event that occurs. An event is a person entering or exiting the room.

Figure 7.4 describes the process for recording an event so that it could be logged in the visual logs data set. An outcome created a log with the timestamp and user number/smart card hash for the event. After all events were processed, the visual logs provided the ground truth for what occurred. We make the following assumptions for the data:

Assumption 1 (Multiple Swipes)**.** *If the smart card log contains identical user swipes within 1 seconds of each other, these are removed and treated as the card reader being too sensitive or the subject tried to swipe for too long of a period.*

Assumption 2 (Accurate Smart Card Logs)**.** *We assume that the smart card reader is infallible and does not produce an incorrect log when a smart card is swiped.*

Assumption 3 (Accountable Smart Card Logs)**.** *We assume that the smart card reader only produces a smart card log when a RFID card is presented. It will not create phantom logs when no RFID is present.*

7.3.9 Ethics

The study followed the institutions ethic guidelines and were approved by its ethics process. We provide the ethics and all supporting documentation for the study in Appendix B.

Informed Consent and Opt-Out Participants were informed of the requirements for each group in advance. This consisted of seventy five minutes of their time, a university smart card and that they would be filmed.

Participants received a consent form and could ask questions before, during and after the experiments. They were informed that they could withdraw at any point. All participants were able to exercise informed consent.

Deception The participants were deceived in the fact that we did not disclose the main intention of the study, which was to monitor the number of total swipes for each participant. We used a cover story of wanting to understand how users manage data when faced with a quarantined (offline) area and having access to an area of Internet connectivity (online).

Compensation Participants were reimbursed for their time spent in the experiment by means of a £10 Amazon voucher. This was given to them after the study took place. We would have honored this for participants that withdrew during the study, however, none did.

Data Protection We ensured data protection and privacy of the participants. Their smart card numbers on swipe were hashed ensuring anonymity and the video recordings are kept on one memory card in a locked room at all times. At the end of the study, this information will be deleted to ensure there is no chance of reverse engineering should the algorithms used to protect the data be compromised.

7.4 Results

Table 7.2 User Study: Participant Demographics

Gender	18-21	22 - 25	26-29	Total
Male	28	6	5	39
Female	10	4	1	15

As a general rule, statistics were computed with a significance level of $\alpha = .05$. We used an omnibus ANOVA (Analysis of Variance) with planned contrasts to assess the difference in mean number of total swipes between the three groups. Should an effect be apparent (a difference in means), we then calculate the size of the effect if a difference is present.

7.4.1 Participants

The total sample size consisted of $N = 54$ students. The Control, Discrete+ and Continuous− groups had $n_1 = 14$, $n_2 = 21$ and $n_3 = 19$ participants, respectively. Each group ran in sessions and a session would consist of 4-6 participants.

7.4.2 Outcomes of Data Preparation

Outcome of Consistency Check

We planned to exclude event observations that are inconclusive, that is, cases in which the observations from the RFID and CCTV1 sensors are contradictory and cannot be resolved with our decision algorithm. No event turned out inconclusive in our consistency check.

Exclusion Criteria Evaluation

We also planned to exclude participants who do not exit the room during the time of the study. That is, they never swipe. This occurred for two participants and they were subsequently removed from the sample.

7.4.3 Metric 1: Total Swipes

The measure total number of swipes is the number of times a participant swiped when entering and exiting the Cyber Security Room during each session.

Descriptive Statistics

We have analysed the data for univariate outliers with the Outlier Labeling Rule. We found one case with extreme values of total number of swipes. We capped the outlying value with the 5th percentile, instead of removing it. Table 7.3 shows the mean and standard deviation of the three groups in the study.

Table 7.3 Descriptive Statistics: Total Swipes

	Control	Discrete+	Continuous−
M	9.08	8.55	5.53
SD	4.48	3.75	2.29

ANOVA

Table 7.4 ANOVA Results: Total Number of Swipes (with Planned Contrasts)

	Df	Sum Sq	Mean Sq	$F-$Value	p-Value
condition	2	143.28	71.64	6.18	.004**
Discrete+ v. Control	1	0.01	0.01	0.00	.982
Continuous− v. Control	1	143.27	143.27	12.35	.001**
Residuals	49	568.28	11.60		

We conducted an omnibus Analysis of Variance (ANOVA) across conditions on the total number of swipes. We used with planned contrasts between (Discrete+ v. Control) and (Continuous− v. Control). The omnibus ANOVA shows a statistically significant difference between means on of the conditions, $F(2,49) = 6.18$, $p = .004$. We thereby reject the overall null hypothesis $H_{\mathsf{T},0}$ that the condition does not impact the mean total number of swipes.

Examining the planned contrasts, we find that the Continuous− condition ($M = 5.53, SD = 2.29$) has a statistically significant negative impact on the total number of swipes compared to

Table 7.5 Effect Sizes: Total Number of Swipes

Comparison	Hedges' g	95% CI
Continuous− v. Control	-1.04	$[-1.78, -0.28]$
Discrete+ v. Control	-0.13	$[-0.83, 0.57]$

the Control condition ($M = 9.08, SD = 4.48$), $F(1,49) = 12.35, p = .001$. We, hence, reject the subordinate null hypothesis $H_{\mathsf{T},0,\mathsf{C}-}$.

The impact of the Discrete+ condition on the total number of swipes was not statistically significant. We failed to reject the subordinate null hypothesis $H_{\mathsf{T},0,\mathsf{D}+}$.

Effect Sizes

The overall effect observed in the omnibus ANOVA was $\eta^2 = .175$ ($\omega^2 = .139$). The corresponding Cohen's f^2 (derived from the less biased ω^2) is: $f^2 = \frac{\omega^2}{1-\omega^2} = .161$. Cohen classified that as medium effect.

We calculate Hedges' g as effect size for the planned contrasts: 1) (Discrete+ v. Control) and 2) (Continuous− v. Control). We provide an overview over all effects in Table 7.5 as well as a corresponding forest plot in Fig. 7.5.

With respect to the alternative hypothesis $H_{\mathsf{T},1,\mathsf{C}-}$, we find the Continuous− condition yielded a large negative effect, Hedges' $g = -1.04$, 95% CI $[-1.78, -0.28]$.

7.4.4 Metric 2: Swipe Rate Ratio

The *Swipe Rate Ratio* is the rate at which participants successfully swiped their smart card on entering and exiting the Cyber Security Room. A swipe rate ratio of 1 indicates that the participant swiped every time they entered and exited.

Descriptive Statistics

Table 7.6 is an overview of the mean and SD for the swipe rate ratio values. The data here contained three outliers, two in the Control and one in the Discrete+ observations. These were all capped at the 5th percentile as per the Outlier Labelling rule used with Metric 1.

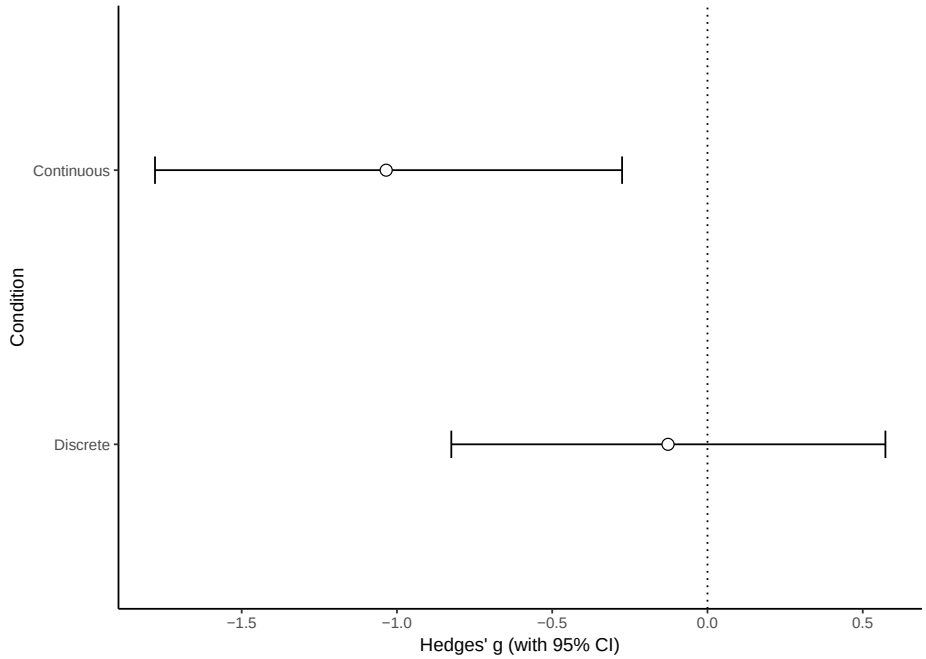


Figure 7.5 Effects of interventions on total swipes.

Table 7.6 Descriptive Statistics: Swipe Rate Ratio

	Control	Discrete+	Continuous−
M	.92	.90	.81
SD	.15	.17	.24

ANOVA

We computed an omnibus analysis of variance (ANOVA) for the swipe rate ratio. We specified planned contrasts between the conditions (Discrete+ v. Control) and (Continuous− v. Control).

The omnibus ANOVA was not statistically significant, $F(2,49) = 1.51$, $p = .231$. We thereby failed to reject the null hypothesis $H_{\mathsf{R},0}$ that the conditions impact the mean swipe rate ratio. In turn, we failed to reject the subordinate hypotheses $H_{\mathsf{R},0,\mathsf{D}+}$ and $H_{\mathsf{R},0,\mathsf{C}-}$.

Table 7.7 ANOVA Results: Swipe Rate Ratio (with Planned Contrasts)

	Df	Sum Sq	Mean Sq	F-Value	p-Value
condition	2	0.11	0.06	1.51	.231
Discrete+ v. Control	1	0.04	0.04	1.03	.316
Continuous− v. Control	1	0.07	0.07	1.99	.164
Residuals	49	1.84	0.04		

Table 7.8 Effect Sizes: Swipe Rate Ratio

Comparison	Hedges' g	95% CI
Continuous− v. Control	-0.50	$[-1.21, 0.22]$
Discrete+ v. Control	-0.12	$[-0.82, 0.58]$

Effect Sizes

The overall omnibus ANOVA yielded an effect size of $\eta^2 = .058$ and Cohen's $\omega^2 = .019$. Based on the ω^2, this constitutes a Cohen's $f^2 = .019$, less than the threshold of a small effect.

We offer parameter and interval estimation on the effects, found in Table 7.8. Fig 7.6 offers a forest plot of the same data. In the Continuous− condition, we observe an estimate of a medium (yet statistically non-significant) effect, Hedges' $g = -0.50$, 95% CI $[-1.21, 0.22]$, which asks for further examination in another experiment.

7.5 Discussion

7.5.1 The broken-window effect has an effect on how users respond to a security policy

In the Continuous− group we altered the Cyber Security Room to be *messy* with the intentions of reducing the swiping behaviour of participants. It was our interpretation of the broken-window theory [114]. It yielded a large effect size and the confidence interval, whilst being very wide suggested that we would see an effect most of the time should we repeat the study.

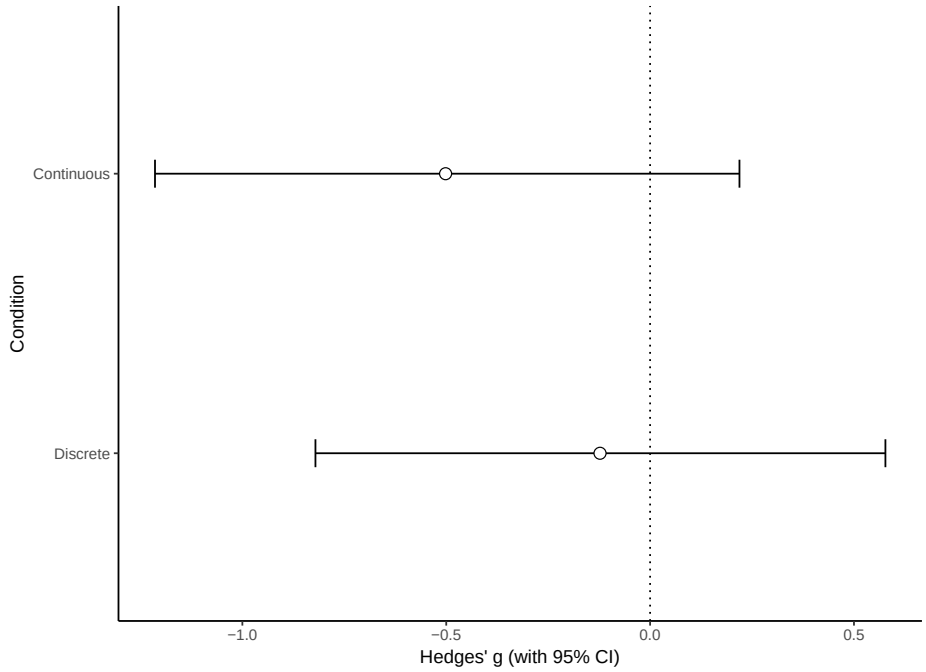


Figure 7.6 Effects of interventions on swipe rate ratio.

Clearly, a *messy* area should more likely impact the behaviour around a security policy. There are additional policies that one could consider. For instance, wearing identification badges and the impact a messy room has on participants challenging each other.

The omnibus ANOVA was not statistically significant for metric 2 (Swipe Rate Ratio) indicating that there is no clear differences in the means between the groups. However, the Hedges-g effect size indicates there is a possibility of seeing a medium effect but this is only speculation as the confidence interval does not support this.

7.5.2 The Messenger effect had no observable measurable impact on total number of swipes or the swipe rate ratio

Our interpretation of the *Messenger* effect was a reminder from the study leader at the twenty five minute mark of the Discrete+ groups [44]. In both metrics, the effect showed no report of an effect size. This does not mean that an effect does not exist, just that we did not observe one in this instance.

We do acknowledge that the swipe rate ratio is very high (0.9 and above) in both the Control and Discrete+ groups. The margin to increase the swipe rate ratio was less than

the margin to decrease it. A perfect swipe rate ratio of 100% would have given us at most a medium effect size for the *Messenger* effect had all participants always swiped. This would also have required participants to have swiped every time before the intervention was delivered.

7.5.3 Recommendations

The observations made in this study that the broken window effect does have a negative effect on how users respond to swiping a smart card. Additionally, we don't know if the broken window effect applies to other security policies. We would err on the side of caution and advise that office space mess should be minimal to ensure that there is no risk of any other security policies being negatively impacted.

7.5.4 Limitations

Generalisability

The participants were recruited from university students, limiting generalisability.

Subjects were exposed to a diversion about how they manage solving tasks when faced with a clean area and dirty area. However, it is clear that students were aware of the camera recording them. What is not clear is how much this changed their behaviour if they were not being observed.

Ecological Validity

Participants in this study were all students and were all accustomed to the principle of swiping. However, at the University, participants on a day-to-day basis swipe their smart cards to unlock doors, confirm attendance and to print documents.

The swipe card reader used for this study was an Adafruit PN532 RFID smart card reader. The ones in place at the University are supplied by an external company and look and react more professional than the makeshift reader we provided.

Statistical Power

The statistical power of the omnibus ANOVA employed in the experiment was limited at a sample size of $N = 54$. In terms of sensitivity, the ANOVA could detect large effects of Cohen's $f = .435$ at 80% power.

Future experiments testing for a medium effect size hypothesized for the swipe rate ratio (Hedges' $g = 0.5$) would need at least 128 participants to reach 80% *a priori* power under ideal conditions.

7.5.5 Reflection on Effects

We originally planned to have two effects that would be considered a pure social influence. In the end we only had the one social influence which was that of an authoritative figure trying to increase the swipe rate. We were not able to have a social influence that influence swipe rate negatively.

Our original idea for a negative social influence was to have an actor participant in the group who vocally stated that they were not going to swipe at regular intervals. Unfortunately, due to constraints and availability of people we could find no one to take up this role. Furthermore, it would have been very difficult to maintain consistency with this sort of social influence. For example, who does the actor talk too when they declare their non-compliance with the swiping policy.

If this experiment is to run again, then performing some sort of involved actor, perhaps as a discrete intervention, similar to the discrete positive in this one, it would be interesting to understand and investigate the outcome.

For the continuous intervention, which consisted of deploying our own version of the broken-window effect, it would have been interesting to have actually had a broken window. However, we did not explore this option as it would not have received much traction from building management. Our alternative to the messy room was to consider a poster that depicted some sort of anti-social behaviour towards security policies. Unfortunately, we could not source or design an appropriate poster.

7.6 Collecting Data Set: Conclusion

We offer the first study of behavioural interventions on smart card swiping behaviour in a security context. We conclude that the broken-window effect has a strong impact on how people swipe their smart card. It is an interesting observation that the mess created in a room has an impact on how users respond to a security policy. It potentially has far-reaching consequences as this may translate to how people digitally behave, e.g. selecting a password.

We did not verify the messenger effect which we expected to see an increase in swiping from participants. We do recognize that study participants in general swiped on a regular basis which may impact the chance of observing a positive effect.

The data set we have collected in this chapter is the collection of the group behaviour towards security policy compliance behaviour. We have already demonstrated that identifying hidden markers such as compliance attitude is difficult. The purpose of this data set allows us to understand if any, the relationship between observations of people and their compliance behaviours. Furthermore, the data set provides a starting point for us to validate the implementation of our Rule Based Model which we introduce in Chapter 7 and assess its accuracy in Chapter 8.

A key insight into this work is that the environmental influence had a strong impact on how people swipe their smart card. Therefore, we would need to establish what human influences have a strong impact on how people swiped their smart card. Although the significance of the data focuses on the total number of swipes as opposed to the swipe ratio, there is a clear change of behaviour. Perhaps it is the case that the broken-window effect stifles productivity and participants were not as motivated to engage in the tasks because of the messy room. Unfortunately, we have no evidence to support this theory and we would need to run the experiment again and change our surveys to understand if this was the case.

What went will with this chapter was the process from start to finish. By submitting the pre-registration to the open science framework we committed ourselves to a robust methodology. It meant we were not looking or cherry picking for any results. By pre-specifying our analysis, we could only consider that small domain. Now that the work is completed, it would be appropriate to perform that cherry picking but we would need another experiment or user study to verify what it is we seek.

A limitation with this chapter is the compliance level of the participants. By successfully engaging with the task it meant the human influence of an expert coming in at mid point to improve the swipe rate had little wiggle room. Whereas the negative effect of the messy room had a lot more to work with. On reflection, it would have made sense for us to have a human influence that was negative, for example, a senior not swiping or a user within the group being vocal about not swiping. We then could have compared the impact a positive and negative effect of the same type of delivery (by a person talking) would have on people's compliance rate.

Additionally, this experiment only consisted of 4-6 participants per run. When comparing this to the simulations where we are considering a population of up to 200 agents in some chapters, it creates a problem around comparisons. We most likely have generalised behaviours that can be drawn out of a population, whereas a small group will have little statistical power on their own. It's for that reason why we put all the data together to form a large set of observed data in the subsequent chapters. A clear limitation with this is the lack of cross influences that would have occurred had all of these groups co-existed at the same time.

7.7 Collecting Data Set: Future Work - Data Observations

From the empirical data set gathered in this chapter, we identify some final observations about the data that could inform future work.

Data Observation 1 (Behaviour Traces)**.** *The collection of the smart card logs and the video cameras allows for a set of behaviour traces to be yielded describing when people swipe or do not swipe when entering and exiting. This is the data set used in this chapter.*

Data Observation 2 (Observation Traces)**.** *Classifying events and logging the location of participants, we can document who was where when a particular event occurred. There may be a relationship between being in the Cyber Incident Room when someone swipes or does not swipe.*

Data Observation 3 (Timings)**.** *The data set documents events to the second. As such, we can understand when participants swipe within moments of each other and the effect this has on their swipe rate.*

These data observations offer a new layer of complexity for assessing the compliance behaviour of participants. In particular, do social influences occur because of observations or the timings of swipes. This area of research is one that can be delved into a lot deeper and may provide further insights towards social influences impacting compliance behaviours.

Chapter 8

Simulation Tool

This chapter address Research Question 5, as a reminder the research question is:

Research Question 5 (Application)**.** *Can we design a tool to allow for the prediction and impact of compliance behaviours towards compliance attitudes?*

Assessing security policies where people have a choice to comply is complex. Particularly when observing compliance behaviours can directly impact a person's compliance attitude. For example, a manager not complying with a policy could affect other people's compliance attitudes. For an organisation, they have no clear method to assess the impact this compliance behaviour has on other employee's compliance attitudes.

We present PCASP (Predicting Compliance Attitudes for Security Policies), a tool evaluating global compliance attitudes of human agents. PCASP simulates behaviour for security policies where social interaction is present. It is built from conclusions in literature surrounding subjective norms, attitudes and beliefs which impact an agent's context and subsequently, the decisions they make. PCASP takes as input, a list of behaviour parameters describing agent behaviour and returns the global compliance attitude from a set of traces formed through simulation. Additionally, we demonstrate the application of PCASP with a running example to illustrate how one might mitigate against poor compliance attitudes amongst agents. We envision that future versions of the tool would enable organisations to make more informed security policy decisions about employee behaviour, such as the best behavioural intervention to use.

Chapter Overview: The chapter is split as follows. In Section 8.1 we describe the high-level requirements our tool must satisfy. In Section 8.2 we present PCASP, in Section 8.3 we provide an impact analysis towards a running example and finally, we finish with a discussion on PCASP. This chapter is formed from a paper at Socio-Technical Aspects of Security and Trust 2018, which at the time of thesis submission, is currently in the process of publication [26].

8.1 Modelling Security Behaviour

In this section, we describe the high-level requirements a tool simulating social influences would need to satisfy, and we illustrate them with the example of an organization where employees must wear ID badges, and where employees should challenge employees not wearing an ID badge. In this discussion, we are interleaving observations on the example scenario with methodological considerations.

Scenario 1 (ID Badge Policy P)**.** *An organisation states the following policy:*

1. *At enrolment time, each employee is issued a personalized ID badge, with photo identification.*
2. *If an ID badge is lost or misplaced, the employee must report this loss and be issued a temporary ID badge.*
3. *Every employee must wear their ID badge visibly at all time.*
4. *Every employee should challenge and report to security persons not wearing an ID badge.*

Observation 1 (Choice to Comply)**.** *Each employee in an organization can be in a situation to decide to comply or not to comply with the currently enforced policy P.*

Employees are decision-making actors in this scenario. They can choose to comply with Policy P or choose to ignore it.

Example 1. *An employee can observe another employee not wearing an ID badge, and can decide to either report them or not. For ignoring the policy there may legitimate exceptions, such as in case of an emergency.*

Observation 2 (Likelihood of Compliance)**.** *The likelihood of an employee complying with the policy will depend on their internal traits and state.*

The likelihood of people's behaviour varies in systematic ways has been widely investigated in psychology. We but name a few well-known models to support this point, which apply to different levels of abstraction[1].

COM-B. The COM-B model [82] (short for: Capability, Opportunity, Motivation – Behaviour) conceptualizes long-term behaviour change and has been related to how influencers change such behaviour.

TPB. The Theory of Planned Behaviour (TPB) [8] governs more short term behaviour, especially how conscious and planned behaviour emerges out of a person's attitude, subjective norm, and perceived behavioural control (PBC). Therein, PBC covers the person's beliefs in the efficacy of their behaviour. These predictors impact the person's intention to act and finally the behaviour itself.

BDI. The Belief-Desires-Intention (BDI) [19] model relates reasoning about beliefs of others to our own actions.

PMT. The Protection Motivation Theory (PMT) [17] considers to what extent fear-inducing messages, such as "You will be fired, if you don't follow the policy" will induce behaviour change through a person's threat and coping appraisal.

Example 2. *An employee with a strong subjective norm in favour of security is more likely to wear an ID badge at all times and to challenge someone not wearing one than an employee with a very high belief in usability.*

Observation 3 (Influencing Behaviour)**.** *The internal state of an employee depends on the observed behaviour from other employees.*

Observations persons make have been shown to have different effects on a person's beliefs (incl. attitudes, subjective norms, perceived behavioural control, threat and coping appraisals). Similarly, the context of the current situation may affect the decision, as well.

[1] We repeat this in Chapter 2, however, we provide more detail there to support these conclusions.

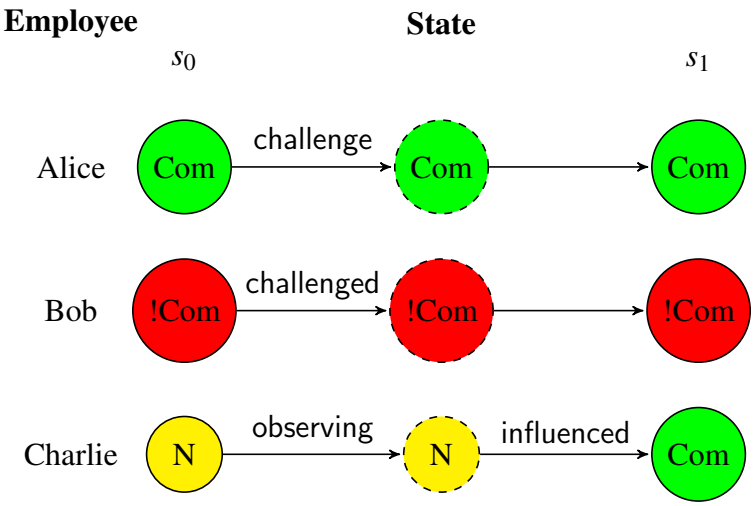


Figure 8.1 Influencing behaviour - Employee Alice challenging Bob, Charlie observing and being influenced towards a compliant attitude. Com - Compliant; !Com - Non-compliant; N-Neutral;

The aforementioned COM-B model [82] repeatedly yields evidence to that effect. Similarly, compliance psychology, such as the influencing work done by Cialdini [33] offers mechanisms by which people's behaviour is influenced by others. Such research has been surveyed and systematicized in the MINDSPACE framework [43], yielding, for instance, how *social norms* impact subjective norms and resulting behavior. Another example from compliance psychology and MINDSPACE entails how authority figures can influence others.

Example 3. *An employee who observes a person in an authoritative position challenging someone not wearing their ID badge may be more likely to wear the ID badge and challenge others.*

We make the assumption that a person observing an authoritative figure challenging a person not wearing their ID badge will be themselves, more like to challenge other people not wearing their badge. We assume this compliance from existing literature where MINDSPACE and Cialdini's work demonstrate that individuals behaviour changes based on the behaviours they have been exposed too [82, 33].

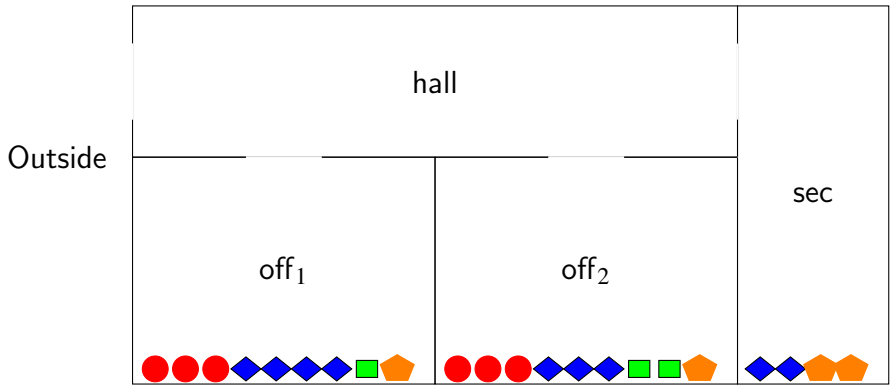


Figure 8.2 Floor plan of organisation, the agents are defined as: ● No subjective norms to security and negative attitude to compliance, ◆ Positive attitude to compliance, ■ Negative attitude to compliance, ⬟ Fear-induced protection intention and positive attitude to compliance.

8.1.1 Modelling Challenging Behaviour

Figure 8.1 is an overview of how we define the problem. The employee Alice challenges Bob, where Charlie observes this occurring. As a result of Charlie's observation, the compliance attitude of Charlie changes to compliant. This is a problem that can quickly become computationally intensive. With many employees and many compliance attitudes, it's unclear how this particular collective security compliance attitude would evolve because of social influences.

Consider the example in Figure 8.2 where employees are distributed across locations and have different attitudes towards policy compliance. If all employees with an attitude of non-compliant were located and remained in off_1 and those with a compliant attitude remained in off_2 then these employees could never influence or interact with each other. For the organisation, a solution would be to simply target behavioural interventions toward those in off_1. However, this is not the case that employees remain in a particular isolated room will behave in a particular way, it is much more complex. Employees have different compliance attitudes therefore, different behaviours that all interact with each other.

In Observations 1, 2 and 3, we have seen (a) that actors have a choice on following the policy or not, (b) that the likelihood to engage in a certain compliance behavior has been modelled in a number of psychological frameworks, and (c) observations made of and interventions made by others influence agents.

To model Scenario 1, we consider *challenging people not wearing a badge* as response variable. Moving forward, we will abstract from individual psychological models and consider the different aspects of internal state affecting an agent's behaviour as this agent's *context*. For instance, for the Theory of Planned Behaviour (TPB), the context would model the agent's attitude, subjective norms, and perceived behavioural control as well as the resulting behavioural intent as part of the context. Consequently, in our modelling the context will either output a resulting likelihood for a consequent behaviour or a decision from {0, 1} on said behaviour.

8.2 PCASP

The purpose of PCASP is to evaluate the compliance attitude of a set of agents for a particular security policy. The tool performs simulations where a distinct set of actions allow for agents to change state. From here on in, we refer to actors/humans/people/employees as agents.

At each state, the tool evaluates for a specific action, how an agent would behave if they were in that situation. For example, knowing the compliance attitude of an agent who is in the location hall for the challenging policy is an action we can evaluate for each state. The tool evaluates by assuming that the conditions to challenge someone are true, then returns the global security compliance attitude for that action for all agents. It uses the parameters of each agent to assess whether or not for a pre-specified action, how that agent would behave. The pre-specified action could be an expected behaviour for a security policy, for example, challenge when you see a person not wearing their badge.

Figure 8.3 is an example of how the tool simulates a state change. The global security compliance attitude is looking at a particular policy which is very specific. Consider the example in Section 2 where we defined four different types of *contexts* an agent can be in when faced with the opportunity to challenge. In Figure 8.3, the values associated with the state represents the global compliance attitude of all agents. A value of 1 would indicate that all agents are either having their context impacted by their *fear compliant attitude* or they simply believe that complying with the policy is in the best interest of all parties.

Each state change occurs due to an action performed by an agent or a group of agents. In Figure 8.3 we have three actions which are chall, move and noChall. Depending on the

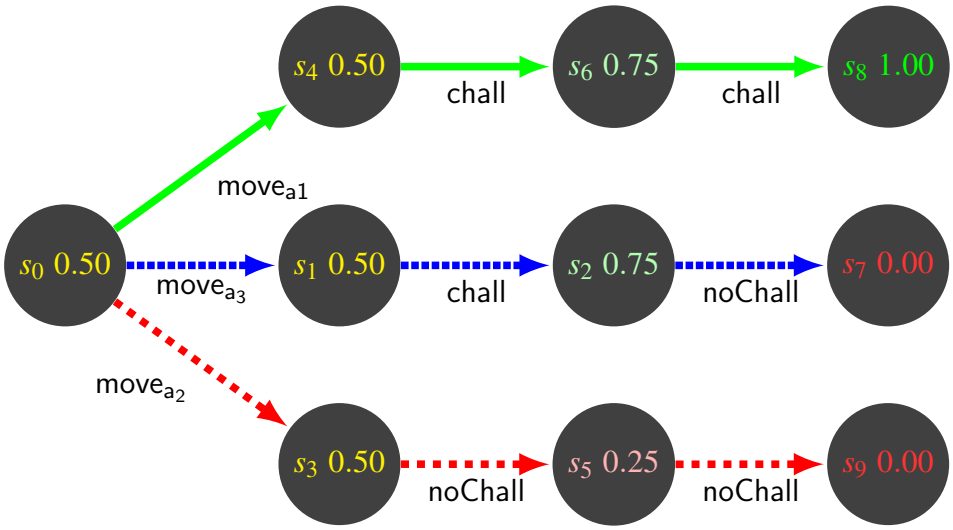


Figure 8.3 Three possible traces from an initial state s_0. Each trace calculates the *global compliance attitude* of all agents, where the red s_7 and s_9 are global compliance attitudes where all agents will act as non-compliant and the green s_8 is a global compliance attitude where all agents will act as compliant. This rate is displayed next to each state where 0 is no global compliance attitude, that is all agents have a non-compliant security attitude and 1 is a complete global compliance attitude, where all agents have a compliant attitude.

current state and the action executed, the compliance attitude is impacted accordingly.

Definition 13 (Agents)**.** *The set of agents is defined as the following:* $A \subseteq I \times C \times S \times \mathscr{P}(\Theta) \times L$ *where ID's are defined as* $I = \{i_1, \ldots, i_n\}$*, C is the set of Contexts, S is a set of statuses, the set of observations are* $\Theta \subseteq I \times Act$ *and L is the set of locations. The set Act refers to a set of actions*[2]*.*

We will discuss the attributes of agents later in this section, for now it's important to know that agents exist with these attributes. However, the observations mentioned here refer to different observations than the ones outlined in Section 8.1. Agents perform actions which change the system state. The tool automatically and randomly assigns timings to when an agent will perform an action. As such, given some initial state, a trace for one simulation may be completely different to a previous trace where the initial state was identical, as shown in Figure 8.3. In Figure 8.3 we can see the action move occurs which leads to one of three states where the global compliance attitude is identical in each state. The different states

[2]This is an extension of the definition for location based agents introduced in Chapter 5

here refer to the fact that different agents have moved and the distribution of agents across locations is now different in each of these three states. This distribution of agents can have a strong impact on how compliance attitudes are influenced.

The modelling language in PCASP is based of PRISM the Probabilistic Model Checker [75]. In PRISM, modules can be defined which express how transitions can occur. These transitions can be the state transitions in a Markov Chain or a Markov Decision Process. Each module is independently defined but can interact with other modules. The agents and later on the actions in PCASP are set out in a similar syntax to the modules. The main difference between PCASP and PRISM is any new states in PCASP are prepended to the attribute it refers too. This will become apparent later on when we cover actions. We could have followed the same style of using a right arrow ($\rightarrow$) as in PRISM but this would be quite verbose as we have labelled names for locations, contexts and attributes. Therefore, the decision to use a new line to define it was an appealing aspect from a usability perspective. It now reflects something similar to JSON or YAML for data structures.

8.2.1 Agents

An agent has a unique identifier which distinguishes them from other agents. We define the internal traits and state of an agent as a context. Agents also display their public status, such as wearing an ID badge or not, where other agents can perceive this status. Agents can observe actions executed by other agents, as such agents keep a record of these observations. At any time an agent must exist in a physical location, we associate an agent as having a location which they are occupying. Multiple agents can occupy the same location.

All agents must be defined and provided with all the relevant information. We adopt an object-oriented style description of entities in PCASP, such that an agent is defined by the list of its attributes:

```
agent
        id:a1;
        con:subjectiveNonCompliant;
        status:wearingID;
        observations:empty;
        loc:off1;
```

```
end
```

The context and status of agents contributes to the global state of the system. The context represents the internal state of an individual agent. The status is the publicly available information that other agents perceive about that agent, for instance, are they wearing their ID badge or not [29]. The contexts and statuses of the agents in this example are (As they must also be defined in the tool):

$$C : \mathsf{normsNonCompliant}, \mathsf{compliant}, \mathsf{nonCompliant}, \mathsf{fearCompliant}$$

$$S : \mathsf{wearingID}, \mathsf{carryingID}, \mathsf{noID}$$

The observations of agents are acquired as a model progresses. They are provided for an initial state of the agent, however, in this example we initialise all agents with no observations (empty). Finally, the location of where the agent is initially located must also be provided.

8.2.2 Actions

We have introduced agents which have an id, context, status, observations and a location. An action will change an agent. The id of an agent is static. The status and location can change directly because of an action. The context and observations of an agent are impacted by an action, however, we discuss this more in depth later in this section. For now, it is important to mention that actions cause observations. An action typically changes the attributes of agents and causes observations. In PCASP, we represent the change of an attribute in an action as being prefixed with new:

```
action act:
        begin agent
                attribute:val;
                newAttribute:val′;
        end agent
end
```

Agents exist in a location and can move between locations. It is an example of one such action. When we define an action, we provide the conditions required for an agent to satisfy

the action. Additionally, we provide any observations that occur as a result of the agent performing the action. In the example, a move action from off_1 to hall takes the format:

```
action offToHall:
        begin agent
                loc:off1;
                newLoc:hall;
        end agent
end
```

The action offToHall changes the agent's location attribute to a new location attribute hall. We write the following transition to express the exact meaning of the action offToHall:

$$\frac{a.loc = \mathsf{off}_1}{a \xrightarrow{\mathsf{offToHall}} a[loc = \mathsf{hall}]} \tag{8.1}$$

Not all elements of an agent need to be provided, in the case of a move action, the start and finish location may be the only requirements. To increase usability of the tool, physical links for all the locations can be given. All of the links are a defined as a set where $Link \subseteq L \times L$ and in the tool, the example for all links takes the form:

$$Link : (\mathsf{hall}, \mathsf{off}_1)(\mathsf{off}_1, \mathsf{hall})(\mathsf{hall}, \mathsf{off}_2)(\mathsf{off}_2, \mathsf{hall})(\mathsf{hall}, \mathsf{sec})(\mathsf{sec}, \mathsf{hall})$$

Note that we do not assume *Link* to be symmetrical, i.e., given two locations l and l', if (l, l') belongs to *Link*, it does not automatically imply that (l', l) also belongs to *Link*. Once given the set of links, PCASP automatically generates all actions $\mathsf{move}_{\mathsf{l}_1\mathsf{l}_2}$. We write $(i, c, s, \Theta, l)[loc = l'] = (i, c, s, \Theta, l')$ as a pointwise operation in Equation 8.1 to express that an agents location has changed. Formally, a move takes the format of a transition rule:

$$\frac{(a.loc, l) \in Link}{a \xrightarrow{\mathsf{move}_{\mathsf{l}_1\mathsf{l}_2}} a[loc = l]} \tag{8.2}$$

The context of an action can impact the choice an agent makes towards their public status. For example, an agent carrying their ID has an action to wear their ID or to have no ID.

```
action wearID:
        begin agent: //WearID
```

```
                con:fearCompliant;
                status:carryingID;
                newStatus:wearingID;
        end agent
end
```

An action can involve multiple agents, in which case they need to be all independently specified within the action. Each agent is self-contained and the requirements and observations for that action are also provided.

```
action chall:
        begin agent: //challenger
                con:compliant;
                status:wearingID;
                loc:hall;
                obs:obsChall;
        end agent
        begin agent: //challengee
                status:carryingID;
                loc:hall;
                newStatus:wearingID;
        end agent
end
```

Agents involved in the same action synchronise. We write the following: $a_1||a_2 \xrightarrow{act} a_1'||a_2'$ indicating that agents a_1 and a_2 have both been a part of the action act. Formally the action chall becomes the transition where we provide an assumption that all agents are always carrying their ID card:

$$\frac{a_1.loc = \mathsf{hall} \wedge a_1.con = \mathsf{compliant} \wedge a_2.loc = \mathsf{hall}}{a_1||a_2 \xrightarrow{\mathsf{chall}(a_1,a_2)} a_1||a_2[status = \mathsf{wearingID}]} \tag{8.3}$$

Given the action chall, the agent a_1 does not change, agent a_2 goes from not wearing an ID to wearing it.

We have defined the challenging behaviour, however, agents can perform the action of not challenging and other agents are able to observe this. The implementation for this action

is as follows:

```
action noChall:
        begin agent: //!challenging.
                con:nonCompliant;
                id:id1|id2
                loc:hall;
                obs:obsNoChall;
                newLoc:off1
        end agent
        begin agent: //agent without ID.
                loc:hall;
                status:!wearingID
        end agent
end
```

From the action noChall we can see that the ID of the agent challenging must be id_1 or id_2. It is often the case that a person/agent in a position of authority will have their actions noticed more by others [62]. It is one interpretation of the Authority principle from the work by Cialdini on social influences [34]. It has strong connections to the *Messenger* effect from the MINDSPACE framework which states *"A person in authority or one that is trusted has an influence over the decisions that other people make"* [44].

8.2.3 Observations

We previously defined the set of observations as Θ, where an observation is a record of the executed action and the ID of the agent that executed the action. Agents keep a record of observations, it is very much a personal trace for each agent's interactions with each other. An observation is the following:

```
observation obsChall:
        loc:hall;
        newObs:(a,chall);
end
```

The observation occurs for all agents satisfying the location and context provided. The set of observations for an agent include the new observation if the agent is in the correct location and has the correct context. An observation cannot occur without an action. Therefore, the observation is part of the action. We express this with regards to all agents as:

$$\frac{x \in Act \quad a \xrightarrow{x} a'}{a||A \xrightarrow{x} a'||A[\theta = (a.id, x)|a.loc]} \tag{8.4}$$

$$(i,c,s,\Theta,l)[\theta = y|l'] = \begin{cases} (i,c',s,\Theta \cup \{y\},l) & \text{if } l = l' \wedge \\ & c' = \mathsf{upd}(c,\Theta \cup \{y\}) \\ (i,c,s,\Theta,l) & \text{otherwise} \end{cases} \tag{8.5}$$

$$\mathsf{upd} : C \times \mathscr{P}(\Theta) \to C \tag{8.6}$$

Equations 8.4 and 8.5 are stating that for all agents in the same location as an action, update there own set of observations and update their context if applicable. We provide the behaviour change policy for the upd in the next subsection.

The function upd changes the context of an agent when they observe a new action. We will discuss this function more in Section 8.2.4. Again, we make use of the pointwise operation to express that an agent's observation is updated for observing an action. Given the action chall, which includes the observation obsChall, we express it as the following:

$$\frac{\mathsf{chall} \in Act \quad a_1||a_2 \xrightarrow{\mathsf{chall}} a_1||a_2}{a_1||a_2||A \xrightarrow{\mathsf{chall}} a_1||a_2||A[\theta = (a_1.id, \mathsf{chall})|a.loc]} \tag{8.7}$$

Equations 8.4 - 8.6 are the transition rules at the fundamental levels e.g. how the transition occurs without any context to an action. Equation 8.7 which is a transition is context specific as it is statically bound to the action chall and it can happen in any location. The implementation in PCASP later on will define what that location is for the running example. The transition in equation 8.7 is not neccesary but it does show how we can focus transition rules towards certain actions.

8.2.4 Behaviour Change

Once an agent has collected observations, the context of that agent can change. The work mentioned earlier by Cialdini offers mechanisms by which behaviour is influenced, where such a mechanism is authority [33]. This translates to the tool by which an agent's context can change if the agent has observed another agent who they classify as an authority perform a specific action.

A behaviour change tests the current observations against the conditions for the behaviour change to occur. A behaviour change impacts agents independently. It takes the format:

```
change: //Auth
        con:nonCompliant;
        observe:(a1,chall) in a.obs;
        newCon:compliant;
end
```

Recall the function upd which takes a context, a set of observations and returns a context. The tool expresses this function by checking behaviour change functions and updating the context accordingly. In the case of the authority behaviour change, the function upd takes the form:

$$\mathsf{upd}(c,\Theta) = \begin{cases} \mathsf{compliant} & \{(a_1,\mathsf{chall})\} \in \Theta \\ & \wedge c = \mathsf{nonCompliant} \\ c & \text{otherwise} \end{cases} \tag{8.8}$$

Given Equation 8.8, behaviour change functions can encompass a range of observations. For example, if we consider subjective norms where it is often described as a feeling of pressure to perform a behaviour, if an agent then observes many other agents performing a specific action, then their context could be influenced. It has been shown before that people may psychologically and culturally attach themselves to a group be behaving in an expected manner [7].

8.2.5 Evaluating Policy Compliance

The model expressed so far describes agents moving independently and interacting with each other over a set of behavioural elements. An agents context, status and location can change depending on the action they take or the observations they have recorded.

Definition 14 (Properties)**.** *We define the set of properties as* $P = \{p_1, \ldots, p_l\}$ *where a property refers to a set of conditions.*

The global compliance attitude for a policy is proportional to the number of agents at any given point that would be compliant with the policy should they be presented with the opportunity to comply. To evaluate the global compliance attitude we define properties which take a set of logical conditions for an agent and return all agents that meet these logical conditions. A specified property is a snapshot of a model at a given time point. As the model progresses, a snapshot of each state is taken. We use this method to evaluate the *global compliance attitude*. These snapshots form a trace describing how the global compliance attitude evolves from the initial state.

```
property strongCompliance:
        begin evalAgent:
                con:compliant|fearCompliant;
                status:wearingID
        end
end
```

A property is assessing for all agents, how many of them meet the proposed criteria. Based on the behaviour set we have defined, we know that an agent will definitely challenge when they have a context as compliant and a status as wearingID.

$$eval : \mathscr{P}(A) \rightarrow [0,1] \tag{8.9}$$

$$eval(A) = \frac{\sum_{a \in A} \mathsf{strongCompliance(a)}}{|A|} \tag{8.10}$$

$$\text{strongCompliance}(i,c,s,\Theta,l) = \begin{cases} 1 & c = \text{compliant} \wedge \\ & s = \text{wearingID} \\ 1 & c = \text{fearCompliant} \wedge \\ & s = \text{wearingID} \\ 0 & \text{otherwise} \end{cases}$$

Equations 8.9, 8.10 and 8.11 demonstrate how PCASP evaluates a behaviour. It returns the global compliance attitude for the logical conditions provided. It does this for all states in the model formed through simulation, i.e. every time an agent performs an action. We do not consider a change to the context or an observation as a state change, all state changes are due to an action.

Another property which we can evaluate is the global compliance attitude of agents who also challenge when faced with the norms pressure of others in the vicinity. We define this property as the weak compliance attitude:

```
property weakCompliance:
        begin evalAgent:
                con:compliant|fearCompliant|
                normsNonCompliant;
        end
end
```

$$\text{weakCompliance}(i,c,s,\Theta,l) = \begin{cases} 1 & c = \text{compliant} \\ 1 & c = \text{fearCompliant} \\ 1 & c = \text{normsNonCompliant} \\ 0 & \text{otherwise} \end{cases} \tag{8.11}$$

Based on the distribution of agents in Figure 8.2, the initial strong compliance attitude is 45%. Whilst there are 11/22 agents (see Figure 1 for number of agents) who are compliant or fearCompliant, we defined one to have the status as carryingID and therefore, would not satisfy the strong compliance property. The initial weak compliance attitude is 88%.

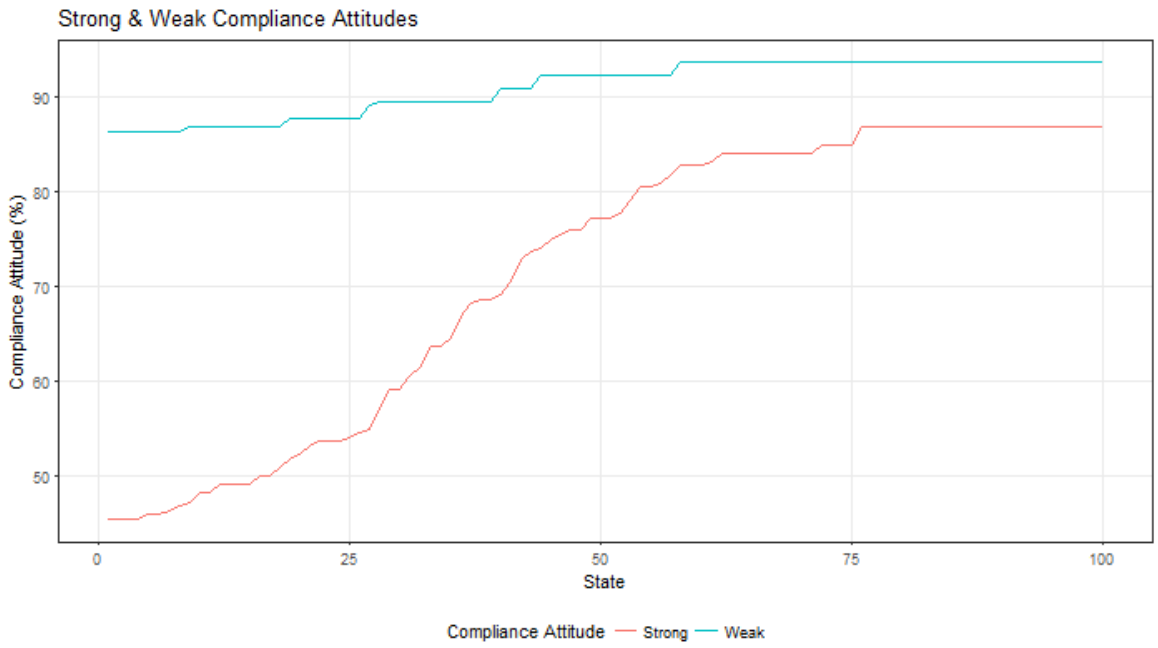


Figure 8.4 Average run of traces when comparing a weak compliance attitude VS a strong compliance attitude. The tool ran for ten traces over each property.

8.2.6 Implementation

The parameters provided to PCASP by a user is translated manually into *Julia* where the *SysModels* package is used as an engine to run the models [**?**]. The *SysModels* package allows for agent based simulation. PCASP provides a clear and concise language for a user to create and evaluate social influences. It is a platform to utilise the *SysModels* engine.

We ran the tool evaluating the two properties which are the strong and weak compliance attitudes of agents for challenging behaviour. Figure 8.4 shows the results of PCASP for the two properties. PCASP shows that from the behaviour set provided, agents will usually end up being compliant with the policy.

We used a Toshiba Portege laptop with 8GB of RAM, an Intel i5 processor and a trace of the model took approximately 3-4 seconds. The results shown are based on the average of 10 traces from the same initial state.

8.2.7 Discussion: Global Compliance Attitude

PCASP analyses the global compliance attitude of the employees by assessing and quantifying their attitudes towards particular security policies. For the organisation, we believe that this information would be relevant. However, we also see the appeal of measuring the number of security incidents.

In general, one would assume that as the global compliance attitude of employees improves, the number of security incidents decreases. However, it may be possible that a decrease in global compliance attitude increases the number of security incidents. For the organisation, this would be a major cause of concern when considering behavioural interventions.

The relationship between global compliance attitude and security incidents is one we will be addressing in future versions of PCASP as the tool. For now, we simply identify this relationship as an area of interest.

8.3 Impact Analysis

PCASP assesses the global compliance attitude of agents in a system. Given a model which contains a set of interactions and agents, a state by state value for the global compliance attitude for an action can be established. A change to the composing parts of the model can impact how the global compliance attitude evolves. Impact Analysis or Change Impact Analysis is often used to describe the effect a change has on a system.

When discussing human or agent behaviour change, we refer to the vast array of behaviour change models [83]. Mentioned earlier and in Chapter 2, the COM-B Model classifies human behaviour change as targeting towards a person's Capabilities, Opportunities or Motivation [82].

We base any changes in our model with regards to COM-B. In our Impact Analysis, we consider two behavioural change elements; 1. A change of Physical Opportunities which is the availability of actions and/or the availability of observations for an agent; 2. A change of Motivation, which is targeting the agent's context.

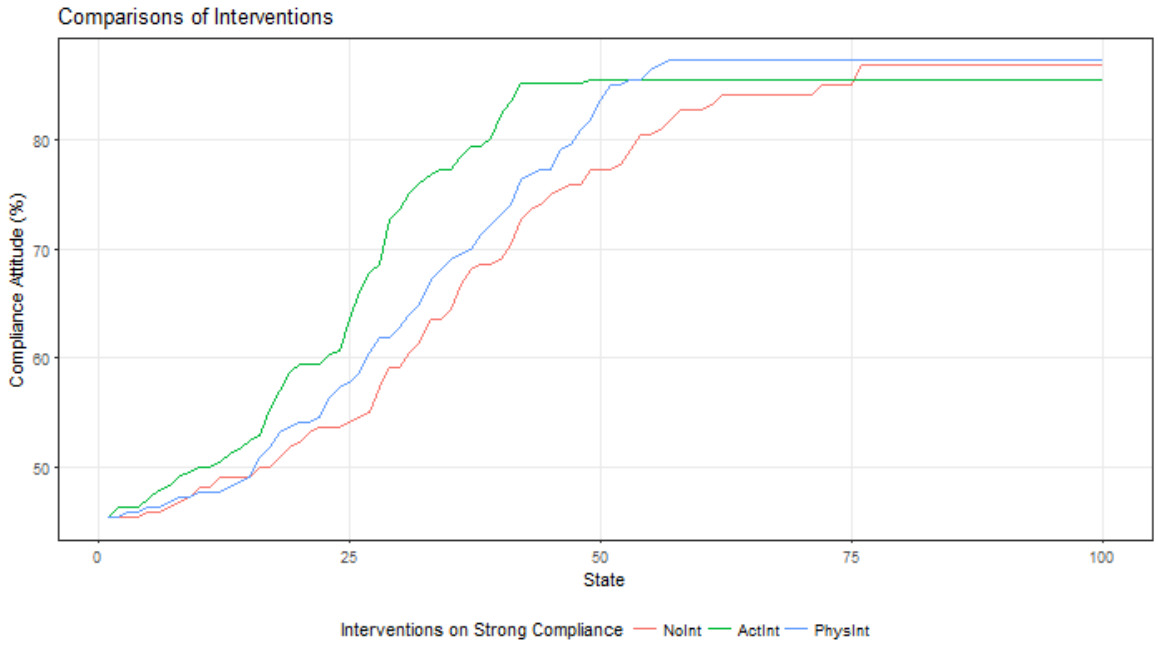


Figure 8.5 The impact analysis for two different interventions in comparison to the original behaviour of the model.

8.3.1 Impact Change Analysis

If *perfect compliance* (*all agents complying at all times*) is not being achieved, some change may impact the global compliance attitude. For example, if an organisation enforced that all employees had to attend mandatory training, they would want to assess if the training had an impact, or at which point in time is the training no longer useful. We now consider two changes or interventions in the example.

Impact Change 1 (Physical Restructure)**.** *Changing the physical structure of an organisation such that observations in specific areas become no longer possible.*

Support: The opportunities from the COM-B model discusses physically restructuring an environment as a behavioural intervention that causes behaviour change [82].

Example: By permitting a view into the hallway, employees can now observe challenging from their offices.

```
observation obsChall:
        loc:hall|off1;
```

```
        newObs:(a,chall);
end
```

Impact Change 2 (Positive Re-enforcement)**.** *People of high influence who act in a positive manner towards a policy may influence others to adopt it.*

```
action forcedChall:
        begin agent: //challenger
                id: a1|a18|a21
                status:wearingID|carryingID;
                loc:hall;
                obs:obsChall;
                newStatus:wearingID
        end agent
        begin agent: //challengee
                status:carryingID;
                loc:hall;
                newStatus:wearingID;
        end agent
end
```

Support: We know that certain people have more power than others when we consider their likelihood to influence. This power that they hold can be used to create a positive environment. We draw upon the Messenger effect from MINDSPACE as inspiration for the behavioural intervention used in the impact analysis [44].

Example: Staging an authoritative employee to challenge someone not wearing their ID badge. We still leave the action for challenging behaviour in the model, however, we define an additional challenge that forces agents of a particular ID to challenge. Consider the authoritative employees as hired actors to improve the global compliance attitude.

The results from Figure 8.5 show that both the interventions positively increased policy compliance. For a user of PCASP, measuring the relationship between the simulated changes and proposed changes in reality would ideally have some form of correlation which we don't yet know without sufficient validation. In the example in Figure 8.5 the Noint (red

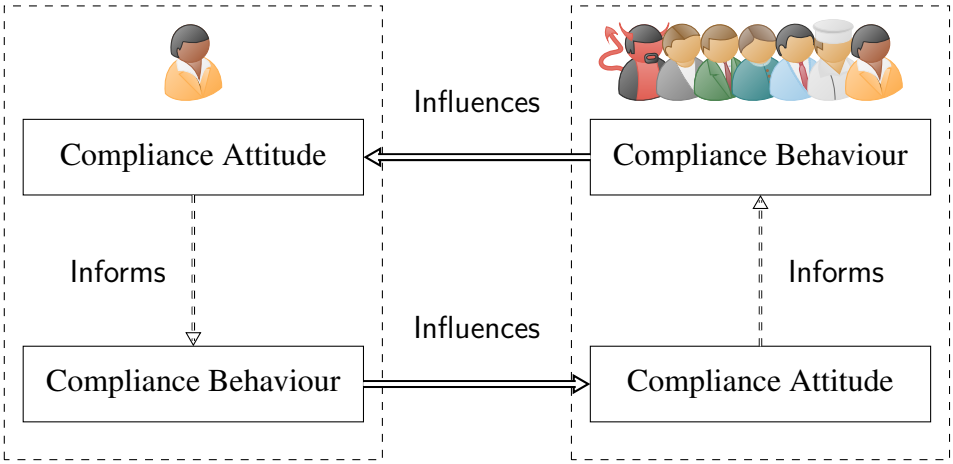


Figure 8.6 PCASP : Relationship between compliance attitude and compliance behaviour.

line) increases as the compliance attitude of agents increases due to the observations agents make of each others compliance attitudes. In this model it tends to be the case that as agen't interact the global compliance attitude always increases even without intervention.

8.4 Simulation Tool: Validation

8.4.1 PCASP is a Proof of Concept

We see PCASP as paving the way for assessing evolving compliance attitudes in an organisation. We have sought in the literature and drawn conclusions about observing compliant behaviour influencing compliance attitudes. Figure 8.6 which is the same as Figure 1.1 is a concrete example of how PCASP replicates this concept of influencing behaviour amongst agents, especially when the contexts of the agents are heterogeneous.

8.4.2 PCASP Responds to Intervention as Expected

When using PCASP to enforce a positive behaviour intervention as we did in Section 8.3, it responds as expected. That is, what we perceived to be a positive intervention produced a positive change to the compliance attitudes of agents.

An area which PCASP paves the way for is to explore perceived positive behavioural interventions which cause a negative impact towards the compliance attitude. Conversely,

does there exist a negative behavioural intervention which positively influences compliance attitudes? As a behavioural model evolved and trust/relationship between agents is integrated a scenario where poor compliance behaviour from an untrusted source improves compliance attitudes could be designed.

8.4.3 PCASP is Contextual and Does Not Generalise

At the moment, the tool provides little insight towards the power of social influences and what that means for other contexts. It's a very focused tool in that a user must design and work with one use case at a time.

Future versions of the tool would allow for estimations about the impact certain types of influences would have. For example, assessing the general impact of a *Messenger* effect would allow us to understand how effective it is across the board in a range of scenarios. A combination of PCASP and machine learning techniques would assist and support the establishments of base metrics for different social influences as more use cases are evaluated.

8.4.4 The Accuracy Between the Tool and the Formal Model

Validation techniques ensuring that the tool accurately reflects the formal model in terms of behaviour. The verification of properties associated with the model would provide certainty that it is behaving as expected. For example, a model that contains a different amount of locations during the course of a trace where the model clearly specifies fixed locations would identify that the tool or implementation is inconsistent.

8.5 Simulation Tool: Conclusion

In this work we presented PCASP which provides a platform for simulation of human agent behaviour assessing the *security culture*. The tool demonstrated its application by modifying behaviour via interventions as impact analysis. The tool PCASP differs from the implementation in Chapter 6. In Chapter 6 we hardcoded the implementation without any guidance which would lead to inconsistencies. The tool provides that platform for ensuring that models are reproducible and readable. The language used is not complex and the PCASP

translates this simple syntax into the Julia language. The syntax defined allows for a system to be modelled, simulated and analysed. These three features of the tool are the foundation of the tool.

A takeaway for this chapter is that the we established a baseline tool by which we can determine the security culture for a given set of agents. Albeit the translation between the tool and the simulation was manual, it would be possible to create a translator that takes the syntax of the code, checks for errors and performs simulation whilst returning some analysis.

Unfortunately, we don't know the accuracy of the simulation. Is it a faithful representation of the model and does it hold to the model outlined in Chapter 5. This tool takes us two and three levels away from the model. The second level is the syntax for PCASP, the third level is the simulation code. Any errors that come from the initial model may be exponential in the second and third level. We currently have not assessed this and acknowledge that it is a clear limitation of this work.

The choice to implement PCASP in the manner we did was reflective of PRISM, the probabilistic model checker. By separating agents and actions we can clearly define the system. It does have its disadvantages as a large model would make it difficult to see the global picture.

Exploring this work in the future, the focus should be on creating an automatic tool that translates the syntax of PCASP into the simulation code to ensure that many different use cases can be compared. Ideally, we should be able to build the simulation code then go back to the syntax to ensure there is correctness. Without that level of implementation, the tool will remain a manual process which takes away from its intention.

Additional future work should consider Chapter 4 and use of the Coloured Petri Nets. If PCASP can provide visualisation for connected processes/behaviours then illustrating this to the user could provide further insights. It would also assist with any model checking where users can ensure that their models behave as expected.

Chapter 9

Validating PCASP

The research question for this chapter is:

Research Question 6 (Validation)**.** *How can we ensure that our methods and processes to address the previous research questions are valid?*

We provide a brief overview of the chapter in the form of a structured abstract to inform the reader about how we address the research question.

Chapter Brief: To validate PCASP by comparing simulated data against the empirical data from Chapter 6.
Objective: In traditional statistics tests, we would aim to reject a null hypothesis. However, in this chapter, failing to reject the null hypothesis is the objective as we seek to provide assurance that PCASP has some validity when comparing against real world data. An assumption of model fitting is that the expected data and observed data have no variance. Ideally for a χ^2 test we would want to claim that the null hypothesis is true, however, we can never claim a null hypothesis is true, we can only assert with some confidence that it is true [61]. Therefore our null hypothesis will consist of comparing simulated data against real world data and the null hypothesis for this would always be that they are not significantly different.
Method: For each group in the user study, we created a model and implemented it in PCASP for $N = 54$ agents to generate the synthetic data. The synthetic data in this chapter too is referred to as the observed data and the data from the user study is referred to as the

expected data. We initialised $N = 54$ agents reflecting a high swipe ratio overall and for each model selected a unique set of these agents to participate in either the Control, Discrete+, or Continuous− simulation to generate the observed data.

Results: The observed data was simulated in PCASP and compared to the expected data. Initially, we used an ANOVA for each group to assess if there was a mean difference between the observed and the expected data sets. We then performed a Chi-Squared Goodness of Fit test for both the total number of swipes and the swipe ratio. The results for the total swipes was $\chi^2(2) = .465, p = .792$. The results for the swipe ratio was $\chi^2(2) = .716, p = .699$.

Conclusion: We conclude that PCASP was able to match with the empirical data without a significant difference, however, we know that we did not have a sample size large enough to faithfully represent a general population.

Chapter Overview: In this chapter we discuss validation techniques for PCASP. We then perform objective validation to measure the accuracy of PCASP by use of the Chi-Squared Goodness of Fit test to assess the accuracy of expected data (empirical data from the user study) against observed data (data from the PCASP simulation). Based on the results of both the Chi-Squared Goodness of Fit tests we failed to reject the null hypothesis that the observed data was not significantly different from the expected data for both the total swipes and the swipe ratio. It is important to note that this aligns with the objective of this chapter, which is to demonstrate and provide assurance that PCASP is behaving in a manner that is consistent to the real world.

9.1 Background: Validation Techniques

Chapters three to seven have addressed different parts of the same problem. That problem consists of how to formulate a system representing psychological influences.

Validating a hypothesis or model can often be carried out with empirical evidence. Particularly when human behaviour is concerned, the empirical results demonstrate reasoning for why people behave in a certain manner [91]. Validation does not have to be just empirical, Dash *et al.* used a data driven approach and achieved next place prediction for people in a physical space. By utilising known information such as the time of day, current location

and day of the week, they were able to accurately predict the next location of a person [40]. Applications for data driven models extend to security, Authentication graphs were derived from access logs to classify intruders on a network [72].

The tool PCASPallows for simulation of our rule based model. Any validation that we can provide gives us confidence that the tool is meeting its design objectives.

9.1.1 Validation Techniques

Validation provides confidence for a user that a tool is performing as it should. In a sense, it is a check for accuracy of a models representation against the real world system it is imitating.

The techniques used for validation are not a fully agreed concept. Sudeikat *et al.* highlights five avenues for validation of Multi-Agent Systems which are testing, run-time monitoring, static analysis, model checking and theorem proving where testing has the least amount of strength and theorem proving has the most strength [103]. On the other side, Sargent describes a more general approach to validation that is not specific to the program or model level, instead it focuses on operational validity which relates to the outputs the model provides [98]. In this case there are two modes of validation which are objective and subjective validation.

Program Validation

Testing Testing via user studies and assessing the accuracy of PCASP against real world data allows for confidence when considering specific use cases which relate to that real world data.

Run-Time Monitoring Depending on how PCASP is composed, confirming that properties that violate a *data-race* if not held is crucial. For instance, we do not want an agent to move to two different locations at the same time point.

Static Analysis An example of static analysis would be unit testing that we often see performed in various programming languages. For PCASP, such testing is important when verifying that a particular element such as a location behaves as intended.

Model Checking In model checking, we want to achieve correctness in the model and avoid deadlocks. One property of correctness for PCASP would be *An agent can always perform an action* and ensuring this deadlock does not occur would give confidence to the user that they designed a model that is deadlock-free.

Theorem Proving Theorem-proving ensures specific properties hold in a system. In PCASP, one such property might be that *removing a location has no negative impact on the compliance attitude of agents* and formally proving that maybe true or false dependent on the location that is removed.

Operational Validity

Objective An objective approach is performed by some form of trusted methodology. For example, statistical tests comparing the output of a simulation model when compared to real world data is one approach for objective validation. Often, the type of test and method for performing the test depends on if the model is underlying model is observable [98].

Subjective A subjective validity check would be performed by experts of the system who provide opinions on their observed usage of a simulation tool [98].

9.2 Preliminaries

In this chapter we use two different data sets for our comparisons, they are:

Expected Data This is the data collected from the smart card interventions study in Chapter 7. We make use of the total number of swipes and the swipe rate ratio.

Observed Data In this chapter we use PCASP to create a observed data set that takes the same format of the expected data and it generated on the basis of the three groups used to collect the expected data which are Control, Discrete+ and Continuous−.

Additionally we use three terminologies which are *people*, *agents* and *accuracies*:

People Any mention of people in this chapter refers to the people that participated in the user study from Chapter 7.

Agents Any mention of agents in this chapter refers to the agents in the simulation.

Accuracies The term accuracies or accuracy refers to how well fitted the observed data in the simulation is when comparing to the expected data from the empirical study in Chapter 6.

9.2.1 Analysis of Variance

An analysis of variance (ANOVA) is a set of statistical methods to analyse and approximate the differences between group means of some samples. In this chapter we used a one-way omnibus ANOVA which entails comparing the means of all the different experiment conditions against the simulation conditions and assessing if there is a difference.

9.2.2 Chi Squared Goodness of Fit

Chi Squared Goodness of Fit test The Chi Squared Test is used to determine how significantly different some observed data (for us, simulation data) is when compared to the expected data (for us, empirical data). For a Chi Squared Goodness of Fit test there are four assumption criteria that must be met:

Assumption 4 (Categorical Variable)**.** *One categorical variable (i.e., the variable can be dichotomous, nominal or ordinal). Examples of dichotomous variables include gender (2 groups: male or female), treatment type (2 groups: medication or no medication), educational level (2 groups: undergraduate or postgraduate) and religious (2 groups: yes or no).*

Assumption 5 (Independence of Observations)**.** *We should have independence of observations, which means that there is no relationship between any of the cases (e.g., participants).*

Assumption 6 (Mutually Exclusive)**.** *The groups of the categorical variable must be mutually exclusive.*

Assumption 7 (Expected Frequencies)**.** *There must be at least 5 expected frequencies in each group of your categorical variable.*

To clarify, in this chapter, all of the assumption criteria is met in order for us to proceed with a Chi-Squared Goodness of fit test.

9.3 Aims

Research Question 7 (Tool Accuracy)**.** *To what extent is* PCASP *accurate when comparing against real world expected data.*

Hypotheses on Total Number of Swipes The overall null hypothesis for this experiment is $H_{S1,0}$ *: There is no mean difference in the total number of swipes when comparing the expected data against the observed data.*

We have subordinate null hypotheses for each group:

$H_{S1,0,ObsC}$**:** *There is no mean difference in the total number of swipes when comparing the observed data against the expected data for the* Control *group.*

$H_{S1,0,ObsD+}$**:** *There is no mean difference in the total number of swipes when comparing the observed data against the expected data for the* Discrete+ *group.*

$H_{S1,0,ObsC-}$**:** *: There is no mean difference in the total number of swipes when comparing the observed data against the expected data for the* Continuous− *group.*

We have these as alternative hypotheses.

$H_{S1,1,ObsC}$**:** *The means of the observed data and expected data for total number of swipes are different for the* Control *group.*

$H_{S1,1,ObsD+}$**:** *The means of the observed data and expected data for total number of swipes are different for the* Discrete+ *group.*

$H_{S1,1,ObsC-}$**:** *The means of the observed data and expected data for total number of swipes are different for the* Continuous− *group.*

Hypotheses on Swipe Rate Ratio The overall null hypothesis for this experiment is $H_{S2,0}$ *: There is no mean difference in the swipe rate ratio when comparing the expected data against the observed data.*

We have subordinate null hypotheses for each group:

$H_{S2,0,ObsC}$**:** *There is no mean difference in the swipe rate ratio when comparing the observed data against the expected data for the* Control *group.*

$H_{S2,0,ObsD+}$**:** *There is no mean difference in the swipe rate ratio when comparing the observed data against the expected data for the* Discrete+ *group.*

$H_{S2,0,ObsC-}$**:** *There is no mean difference in the swipe rate ratio when comparing the observed data against the expected data for the* Continuous− *group.*

$H_{S2,1,ObsC}$**:** *The means of the observed data and expected data for the swipe rate ratio are different for the* Control *group.*

$H_{S2,1,ObsD+}$**:** *The means of the observed data and expected data for the swipe rate ratio are different for the* Discrete+ *group.*

$H_{S2,1,ObsC-}$**:** *The means of the observed data and expected data for the swipe rate ratio are different for the* Continuous− *group.*

Unlike the experiment in the Chapter 7, in this experiment we aim to accept the null hypothesis for both the total number of swipes and the swipe rate ratio. Accepting the null hypotheses would indicate that the expected and observed data is well fitted. Unfortunately, it will not tell us if PCASP is provided a level of accuracy, or if this set of data is by chance.

9.4 Method

9.4.1 PCASP Model

In order to generate the observed data, we first need a model capable of doing so. We define a swipe card model catering for the four events from the actual user study. We provide two different contexts for agents. They are compliant and nonCompliant. A model where no social influences are present is the following:

```
action enterSwipe:
        begin agent: //agentEnterSwiping
                loc:research;
                con:compliant;
                obs:swipe;
                newLoc:securityroom;
        end agent
end
action exitSwipe:
```

```
        begin agent: //agentExitSwiping
                loc:securityroom;
                con:compliant;
                obs:swipe;
                newLoc:research;
        end agent
end

action enterNoSwipe:
        begin agent: //agentEnterNoSwiping
                loc:research;
                con:nonCompliant;
                obs:noSwipe;
                newLoc:securityroom;
        end agent
end
action exitNoSwipe:
        begin agent: //agentExitNoSwiping
                loc:securityroom;
                con:nonCompliant;
                obs:noSwipe;
                newLoc:research;
        end agent
end
```

The actions are mirrored for the action of not swiping. The four actions defined capture the agents moving from the *offline* (Cyber Security Room) area to the *online* (Research) area and vice versa where they have the possibility to either swipe or not swipe. For the agents, the possible observations would then be the following:

```
property countSwipes:
        begin observeAgent:
                loc: securityroom;
                obs: swipe;
        end agent
```

```
end
property countNoSwipes:
        begin observeAgent:
                loc: securityroom;
                obs: noSwipe;
        end agent
end
```

For all of the agents, we can record each event as it happens and generate a data set which allows for some post processing where we can gather the total number of swipes per agent and the swipe rate ratio of each agent.

9.4.2 Model Parameters

There are three parameters to consider. They are the time intervals at which actions can occur, the contexts of agents, and the impact interventions have on agents behaviour.

Time When designing PCASP, we implemented it utilising as an engine the SysModels package in Julia [27]. The package deals with time by ensuring that agents behave concurrently according to a global clock. In order to specify the amount of time passing for actions we use a Poisson distribution. This distribution is established from the expected data to assess roughly how long agents would spend in the research and securityroom locations.

Context An agents context in the model presented in Section 9.4.1 can take the format of compliant or non − compliant. From the quantitative data of the user study we know that on average people swiped more often than not. For the purposes of this experiment we will use the swipe rate ratio of participants to initialise the contexts of agents. The same parameters for agents context used in the Control simulation are used for the Discrete+ and Continuous− simulations.

Interventions There are three different groups that we are comparing for the expected data, which are the Control, Discrete+ and Continuous−. We add to each model some criteria in order to replicate the behavioural interventions.

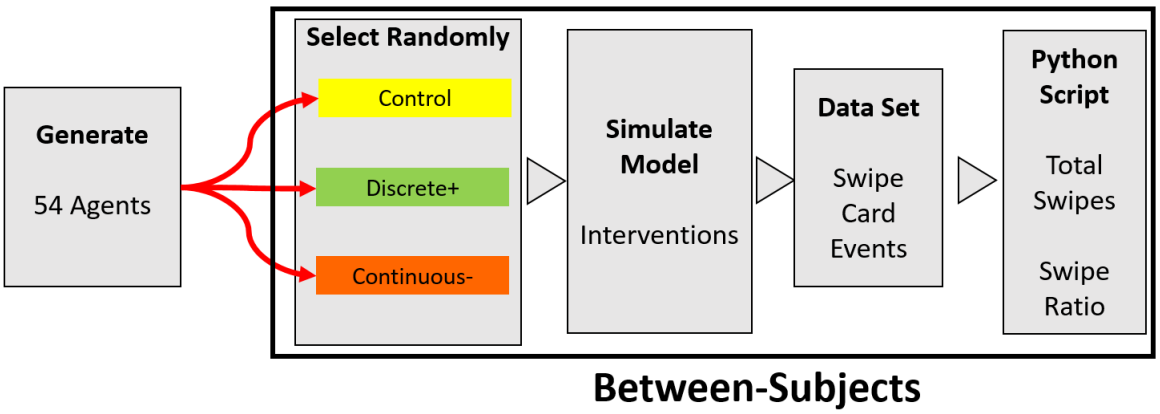


Figure 9.1 An overview of the simulation path.

- Control: There is no change to the model as there is no intervention in this group.
- Discrete+: There is an intervention set at a specific time point to influence agents to be more likely to swipe.
- Continuous−: There is a constant intervention in place to suppress agents willingness to swipe their smart card.

9.4.3 Experiment Setup

The experiment is setup using the expected data from Chapter 7 and compares like for like based on condition for a observed data set generated by the simulation tool PCASP. The expected data we already have, the observed still needs to be generated.

Faithful to empirical study In this chapter we set up the simulation models as faithful as possible to the empirical study. By faithful we mean that it attempts to mimic how the user study originally ran. We do this in five stages which are illustrated in Figure 9.1:

- **Generate:** We generate fifty four agents with unique ID's and base their compliance attitudes off the swipe rate for the Control group from the user study in Chapter 6.
- **Select Randomly:** We select randomly the agents and assign them to a simulated Control, Discrete+ or Continuous− group.

- **Simulate Model:** We simulate the each model once under it's particular condition with the appropriate intervention applied which we discuss in subsection 9.4.5.
- **Data Set:** From the simulation we then have an established list of swipe card events which are associated with each agents unique ID.
- **Python Script:** We run a python script over the data set to collect the total number of swipes and the swipe ratios for each agent.

9.4.4 Measurements

As per the user study in Chapter 7 we measure the same values in the observed data which is the total number of swipes and the swipe rate ratio (i.e. the number of total swipes against all events).

9.4.5 Implementing observed Interventions

Discrete+ In the user study, the Discrete+ intervention consisted of a reminder from the study leader at the twenty five minute mark. The purpose of the intervention is to increase the likelihood that a person would swipe their smart card.

To replicate the Discrete+ intervention in the simulation, we deploy a reminder to agents at the observed time of twenty five minutes in the simulation which increases the likelihood that an agent will swipe their smart card on entry and exit.

Continuous− In the user study, the Continuous− intervention consisted of a continuous messy room. The purpose of the intervention is to decrease the likelihood that a person would swipe their smart card.

To replicate the Continuous− intervention in the simulation, we suppress agents likelihood to swipe their smart card for the duration of the simulation.

9.4.6 Ethics

The expected data collected in the user study catered for the relevant ethics and in this experiment, there is no ethical consideration as we are including observed data.

9.5 Results

As a general rule, statistics were computed with a significance level of $\alpha = .05$. We used an omnibus ANOVA (Analysis of Variance) with planned contrasts. We assessed the difference in mean number of total swipes and mean swipe rate ratio between the expected data and observed data for each group type. Should an effect be apparent (a difference in means), we then calculate the correlation by means of a χ^2 (chi-squared) test for goodness of fit.

9.5.1 Participants

The number of participants of 54 from the user study is reflected in the observed data as 54 agents. Each group in the user study ran with in between 4-6 people. The same is true for the observed data where we run each study with the same number of agents to match the study.

9.5.2 Metric 1: Total Swipes

The measure total number of swipes is the number of times a participant swiped when entering and exiting the Cyber Security Room during each session.

Descriptive Statistics

Table 9.1 shows the mean and standard deviation of the three groups in the study for both the expected and observed data for the total swipes. In Tables 9.1 the values M_e and SD_e refer to the mean and standard deviation for the data collected in the user study and M_o and SD_o refer to the mean and standard deviation for the data collected in the simulations.

Table 9.1 Descriptive Statistics: Total Swipes

	Control	Discrete+	Continuous−
M_e	9.08	8.55	5.53
SD_e	4.48	3.75	2.29
M_o	9.81	10.25	7.01
SD_o	2.54	3.36	3.08

Table 9.2 ANOVA Results: Total Swipes (Control)

	Df	Sum Sq	Mean Sq	F-Value	p-Value
Control_e v. Control_o	1	17.29	17.289	6.351	.029
Residuals	11	29.94	2.72		

Table 9.3 ANOVA Results: Total Swipes (Discrete+)

	Df	Sum Sq	Mean Sq	F-Value	p-Value
Discrete+_e v. Discrete+_o	1	.29	.287	.023	.0.881
Residuals	18	225.46	12.526		

Table 9.4 ANOVA Results: Total Swipes (Continuous−)

	Df	Sum Sq	Mean Sq	F-Value	p-Value
$\text{Continuous}-_e$ v. $\text{Continuous}-_o$	1	22.8	22.798	2.451	.0.136
Residuals	17	158.2	9.303		

ANOVA - Total Swipes

We conducted an omnibus Analysis of Variance (ANOVA) to compare each condition against its observed data on the total number of swipes. We used with planned contrasts for each ANOVA:

- Control: Control_e v. Control_o
- Discrete+: Discrete+_e v. Discrete+_o
- Continuous−: $\text{Continuous}-_e$ v. $\text{Continuous}-_o$

Control The ANOVA for the Control data shows a statistically significant difference between means of the expected and observed data, $F(1,11) = 6.35$, $p = .029$. We thereby reject the null hypothesis $H_{\text{S1,0,ObsC}}$ that there is no mean difference in the total number of swipes when comparing the observed data against the expected data for the Control group.

Discrete+ The ANOVA for the Discrete+ data shows no significant difference between means of the expected and observed data, $F(1,18) = .023$, $p = .0881$. We thereby accept the null hypothesis $H_{\text{S1,0,ObsD}}$ that there is no mean difference in the total number of swipes when comparing the observed data against the expected data for the Discrete+ group.

Table 9.5 Descriptive Statistics: Swipe Rate Ratio

	Control	Discrete+	Continuous−
M_e	.92	.90	.81
SD_e	.15	.17	.24
M_o	.81	.83	.59
SD_o	.21	.27	.25

Continuous− The ANOVA for the Continuous− data shows no significant difference between means of the expected and observed data, $F(1,17) = 2.45$, $p = .136$. We thereby accept the null hypothesis $H_{\mathsf{S1,0,ObsC-}}$ that there is no mean difference in the total number of swipes when comparing the observed data against the expected data for the Continuous− group.

χ^2 Test - Total Swipes

A chi-square goodness of fit was calculated comparing the occurrence of the expected data for total swipes against the occurrence of the simulated data for total swipes. No significant deviation from the simulated values was and the result of the test is: $\chi^2(2) = .716, p = .699$. We thereby fail to reject the overall null hypothesis $H_{\mathsf{S1,0}}$ that there is no mean difference in the total number of swipes when comparing the expected data against the observed data.

9.5.3 Metric 2: Swipe Rate Ratio

The *Swipe Rate Ratio* is the rate at which participants successfully swiped their smart card on entering and exiting the Cyber Security Room. A swipe rate ratio of 1 indicates that the participant swiped every time they entered and exited.

Descriptive Statistics

Table 9.5 shows the mean and standard deviation of the three groups in the study for both the expected and observed data for the swipe rate ratio.

Table 9.6 ANOVA Results: Swipe Ratio (Control)

	Df	Sum Sq	Mean Sq	F-Value	p-Value
Control$_e$ v. Control$_o$	1	.102	.102	5.696	.0036
Residuals	11	0.2	0.02		

Table 9.7 ANOVA Results: Swipe Ratio (Discrete+)

	Df	Sum Sq	Mean Sq	F-Value	p-Value
Discrete+$_e$ v. Discrete+$_o$	1	.07	.07	.095	.343
Residuals	18	1.31	0.07		

Table 9.8 ANOVA Results: Swipe Ratio (Continuous−)

	Df	Sum Sq	Mean Sq	F-Value	p-Value
Continuous−$_e$ v. Continuous−$_o$	1	.14	.14	2.34	.144
Residuals	17	1.05	0.06		

ANOVA - Swipe Ratio

We conducted the same ANOVA style tests for the swipe ratio as we did for the total number of swipes:

Control The ANOVA for the Control data shows a statistically significant difference between means of the expected and observed data, $F(1,11) = 5.696$, $p = .0036$. We thereby reject the null hypothesis $H_{\mathsf{S2,0,ObsC}}$ that there is no mean difference in the swipe ratio when comparing the observed data against the expected data for the Control group.

Discrete+ The ANOVA for the Discrete+ data shows no significant difference between means of the expected and observed data, $F(1,18) = .095$, $p = .343$. We thereby accept the null hypothesis $H_{\mathsf{S2,0,ObsD}}$ that there is no mean difference in the total number of swipes when comparing the observed data against the expected data for the Discrete+ group.

Continuous− The ANOVA for the Continuous− data shows no significant difference between means of the expected and observed data, $F(1,17) = 2.34$, $p = .144$. We thereby accept the null hypothesis $H_{\mathsf{S2,0,ObsC-}}$ that there is no mean difference in the total number of

swipes when comparing the observed data against the expected data for the Continuous− group.

χ^2 Test - Swipe Ratio

A chi-square goodness of fit was calculated comparing the occurrence of the expected data for swipe ratio against the occurrence of the simulated data for swipe ratio. No significant deviation from the simulated values was and the result of the test is: $\chi^2(2) = .465, p = .792$. We thereby fail to reject the overall null hypothesis $H_{S2,0}$ that there is no mean difference in the swipe ratio when comparing the expected data against the observed data.

9.6 Discussion

9.6.1 Limitations

The parametrisation of the model relied heavily on the quantitative information from the expected data.

In this experiment we were fortunate to have the data to inform the model about peoples behaviour and define a model accordingly to represent them as simulated agents. Unfortunately, this would not be the case for the majority of the time when considering the evolution of compliance attitudes. Nevertheless, we feel that the first version of PCASP provides a starting point for assessing how validation concepts can be utilised to provide a ground truth.

The model used did not exploit fully the use of observations

The model we used for simulation was a naive reflection of reality. From watching the video footage of participants in the empirical study there is an indication that proximity of participants has an influence on another participants chance to swipe. These observations are something we did not capture in the model and may well improve the result we get from the chi squared goodness of fit test. Nevertheless, we did make use of the observations with the simulated Discrete+ model demonstrating that some agents will have noticed and changed their behaviour due to the intervention part way through the simulation.

9.6.2 Ecological Validity

The empirical data that we have collected in Chapter 6 is already limited due to the restricted population that we used via convenience sampling. Because of this method it further limits the validation in this experiment and makes the work less generalisable than if the sample had been larger and from a wider pool of potential participants.

9.7 Chapter Conclusion

In this chapter we addressed research question 6 with an experiment to validate the accuracy of PCASP when comparing to empirical data. Our results showed that there is no significant difference in the observed and expected data for both metrics of the total swipes and the swipe ratio in two of the conditions.

Failing to reject both null hypotheses in this chapter is considered a success for addressing the research question. Whilst we did not measure to what extent PCASP is accurate, we managed to show that in four out of six ANOVAs carried out that there was no significant difference in the mean values.

This chapter demonstrates that validation techniques can be applied to PCASP and is another contribution towards the holistic body of research presented in this thesis towards compliance attitudes for security policies.

In this chapter we did not model the control case where the observed data fit with the expected data. This then leads us to question whether or not we actually achieved the aim for this chapter. With one of the conditions failing to fit, we cannot reliably say that the model simulation is in any way reflective of the real world. However, it does give us a base to go on as we can now assess what was wrong with the control case. Perhpas the parameterisation was not optimised to suit for a model of this type, which leads us to wonder what would the ideal/optimal parameters be as a base for each of the cases.

A key insight for this chapter is that PCASP can be compared against real world data. This step towards validation is crucial for the tool. Unfortunately we can't say that it is accurate, we can only claim that the data sets themselves are significantly different which allows us to have faith that the tool is heading towards a good fit for what we want.

An important part of this work was creating the synthetic data set which required us to run simulations. The simulation was flexible enough to allow us to produce the format of the data we needed ensuring we could perform automated analysis trivially.

A limitation of this work is that different parameters for the simulations would produce different results. It would usually not be a problem, however, without any knowledge of what those parameters were, such as compliance attitude of agents, it means our estimates of an even distribution have no grounding in reality.

Any future work in this area needs to ensure that any parameters used for the simulation is faithful to the real world. In essence, this is the correctness and completeness problem which is consistent throughout this thesis. As the majority of this work is mainly a proof of concept, it creates that foundation for future work to build on. By assessing and observing the collection of real world data, we could establish the correct parameters to use in the simulation models.

Chapter 10

Conclusion

The work in thesis addressed the problem of security policies in organisations where employees have a choice to comply. The area we concerned ourselves with is that of *compliance attitudes*, which are the internal mechanisms guiding a persons compliance behaviour. We stated and provided evidence that a person's compliance attitude can be subject to change from social influences from other people. The importance of assessing social influences towards compliance attitudes allows an organisation to make more informed decisions about the current risk environment that they are in. Furthermore, if an organisation could predict the impact of social influences, then it would inform behavioural interventions should they be needed. We derived the following aim:

> ***Aim**: To provide tools and methodologies for an organisation to predict and analyse the impact social influences have towards compliance attitudes for security policies.*

We chose to address the aim by identifying and focusing on six research questions that indicated unknown areas when understanding social influences towards compliance attitudes. By no means does answering these six research questions provide a finished view of social influences over compliance attitudes, however, we feel that our holistic approach achieves the aim.

The contributions of this thesis are the chapters addressing the research questions. Whilst we do not define a *main* contribution, the simulation tool PCASP, which at the moment is a proof on concept language that is manually translated to Julia for simulations is certainly

one output of the thesis that paves the way for assessing compliance attitudes. As a first version tool it offers a user the ability to assess compliance attitudes towards security policies. Chapters 4, 6, 7 and 8 all contribute in some shape towards PCASP whether that is through design, implementation or a user study for validation. The semantics of the rule based model demonstrates agents that have the capability to observe the actions of other agents. The implementation as the model is ported to the first version of the tool PCASP. Finally, the collection and validation of the model in a trial study for compliance attitudes towards swiping smart cards.

The modelling to represent of social influences demonstrates the impact that different social influences have on compliance attitudes. In particular the propagation of an influence and the uncertainty of compliance attitudes that an adversary faces when exploiting social influences. We demonstrated propagation in Chapter 3 and briefly touched on the subject in Chapter 4 when discussing system states.

We demonstrated by machine learning the applications of decision trees to try and identify compliance attitudes when given a set of agent traces. We are not able to claim that decision trees were accurate for identifying compliance attitudes. As the complexity of the traces evolved, the accuracy of the decision trees tended to decrease.

Overall, the methodologies we use in this thesis demonstrate that there does not exist one approach to solving the problem surrounding social influences towards compliance attitudes. Our holistic approach with its range of contributions demonstrates requirements for a wide assessment when addressing social influences towards compliance attitudes.

10.1 Future Work

Formal Verification If PCASP is to undergo further development, then we would build it upon a formally verified model to ensure integrity is maintained from design of the model to implementation as a simulation tool. The work we have carried out in this thesis is mainly a descriptive overview of how agent observations can be modelled and provides no material towards formal verification.

Deep Machine Learning In Chapter 5 we assessed the accuracy of decision trees to identify hidden compliance attitudes. We used a limited amount of data in order to replicate what one could typically gather if they were to build up the traces through observations in an organisation. The application of deep Machine Learning may provide further insights into what sort of data is needed in order to identify hidden compliance attitudes.

Validation: Whilst we have provided a level of validation, it was for one specific user study where the tests returned values that only indicated that the output of the observed data matched the expected empirical data from the user study. Further steps to enhance validation techniques and identify the key requirements is a research area that would need a lot of attention.

Commercial Tool The tool PCASP is a first version tool that is currently not applicable for anyone outside of the academic community. Research to both establish the tool as usable for an organisation and identifying the fundamentals that the tool would need in order to attract users would provide an impact for PCASP. Furthermore, commercialising the tool in order to provide organisations with confidence that they can easily predict how compliance attitudes evolve amongst their employees would be an end goal of this body of research.

Bibliography

[1] (2017). Identification badge policy and procedure for employees. http://www.hullccg.nhs.uk/wp-content/uploads/2017/06/identification-badge-policy-and-procedure.pdf. [Online; accessed 08-June-2018].

[2] (2018). Behavioural insights team. https://www.behaviouralinsights.co.uk/. [Online; accessed 08-June-2018].

[3] (2018). Cyber security breaches survey 2018. https://www.gov.uk/government/statistics/cyber-security-breaches-survey-2018. [Online; accessed 08-Sept-2018].

[4] (2018). Laerd statistics. https://statistics.laerd.com/. [Online; accessed 26-November-2018].

[5] (2018). Research institute in science of cyber security. https://www.riscs.org.uk/.

[6] (2018). The TREsPASS Project. https://www.trespass-project.eu/. Accessed: 2018-11-27.

[7] Abrams, D., Ando, K., and Hinkle, S. (1998). Psychological attachment to the group: Cross-cultural differences in organizational identification and subjective norms as predictors of workers' turnover intentions. *Personality and Social psychology bulletin*, 24(10):1027–1039. Sage Publications Sage CA: Thousand Oaks, CA.

[8] Ajzen, I. (2009). The theory of planned behaviour: reactions and reflections. *Psychology & Health*, 26(9):1113–1127. Routledge.

[9] Asch, S. E. (1955). Opinions and social pressure. *Scientific American*, 193(5):31–35. JSTOR.

[10] Association of Certified Fraud Examiners (2018). Coca-Cola data breach highlights importance of laptop security. https://www.acfe.com/fraud-examiner.aspx?id=4294986501. Accessed: 2018-11-27.

[11] Bada, M., Sasse, A. M., and Nurse, J. R. (2019). Cyber security awareness campaigns: Why do they fail to change behaviour? *arXiv preprint arXiv:1901.02672*.

[12] Bandura, A. (1989). Human agency in social cognitive theory. *American psychologist*, 44(9):1175. American Psychological Association.

[13] Barrick, M. R. and Mount, M. K. (1991). The big five personality dimensions and job performance: a meta-analysis. *Personnel psychology*, 44(1):1–26. Wiley Online Library.

[14] Bartsch, S. and Sasse, M. A. (2013). How users bypass access control - and why: The impact of authorization problems on individuals and the organization. In *ECIS*.

[15] Beautement, A., Sasse, M. A., and Wonham, M. (2009). The compliance budget: managing security behaviour in organisations. In *Proceedings of the 2008 New Security Paradigms Workshop*, pages 47–58. ACM.

[16] Blythe, J., Koppel, R., and Smith, S. W. (2013). Circumvention of security: Good users do bad things. *IEEE Security & Privacy*, 11(5):80–83. IEEE.

[17] Boer, H. and Seydel, E. (1996). Protection motivation theory. In *Predicting Health Behaviour: Research and Practice with Social Cognition Models*, pages 95–120. Open University Press.

[18] Brantingham, P. J., Brantingham, P. L., et al. (1981). *Environmental criminology*. Sage Publications Beverly Hills, CA.

[19] Bratman, M. (1987). Intention, plans, and practical reason.

[20] Breiman, L. (2017). *Classification and regression trees*. Routledge.

[21] Bulgurcu, B., Cavusoglu, H., and Benbasat, I. (2010). Information security policy compliance: an empirical study of rationality-based beliefs and information security awareness. *MIS quarterly*, 34(3):523–548. Society for Information Management and The Management Information Systems Research Center.

[22] Bullée, J.-W. H., Montoya, L., Pieters, W., Junger, M., and Hartel, P. H. (2015). The persuasion and security awareness experiment: reducing the success of social engineering attacks. *Journal of experimental criminology*, 11(1):97–115. Springer.

[23] Carmichael, P. and Morisset, C. (2017). Learning decision trees from synthetic data models for human security behaviour. In *International Conference on Software Engineering and Formal Methods*, pages 56–71. Springer.

[24] Carmichael, P., Morisset, C., and Groß, T. (2016). Influence tokens: analysing adversarial behaviour change in coloured petri nets. In *Proceedings of the 6th Workshop on Socio-Technical Aspects in Security and Trust*, pages 29–40. ACM.

[25] Carmichael, P., Morisset, C., and Groß, T. (2018a). Interventions over smart card swiping behaviour. In *STAST (Socio-Technical Aspects of Security and Trust)*. In publication.

[26] Carmichael, P., Morisset, C., and Groß, T. (2018b). Simulating influencing human behaviour in security. In *STAST (Socio-Technical Aspects of Security and Trust)*. In publication.

[27] Caufield, T. (2017). Sysmodels package. https://github.com/tristanc/SysModels. [Online; accessed 08-June-2017].

[28] Caufield, T. and Parkin, S. (2016). Case study: Predicting the impact of a physical access control intervention. In *STAST (Socio-Technical Aspects of Security and Trust)*. In publication.

[29] Caulfield, T., Baddeley, M., and Pym, D. (2016). Social learning in systems security modelling. *constructions*, 14(15):3. Self-Published.

[30] Caulfield, T. and Pym, D. (2015). Improving security policy decisions with models. *IEEE Security & Privacy*, 13(5):34–41. IEEE.

[31] Cheh, C., Chen, B., Temple, W. G., and Sanders, W. H. (2017). Data-driven model-based detection of malicious insiders via physical access logs. In *International Conference on Quantitative Evaluation of Systems*, pages 275–291. Springer.

[32] Cialdini, R. (2016). *Pre-Suasion: A revolutionary way to influence and persuade*. Simon and Schuster.

[33] Cialdini, R. B. (2007). *Influence: The psychology of persuasion*. Harper Business.

[34] Cialdini, R. B., T. M. R. (1998). Social influence: Social norms, conformity and compliance. In *The handbook of social psychology*, pages 151–192. In D. T. Gilbert, S. T. Fiske, G. Lindzey (Eds.).

[35] Cinnirella, M. and Green, B. (2007). Does 'cyber-conformity'vary cross-culturally? exploring the effect of culture and communication medium on social conformity. *Computers in Human Behavior*, 23(4):2011–2025. Elsevier.

[36] Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Routledge.

[37] Crampton, J. and Morisset, C. (2012). Ptacl: A language for attribute-based access control in open systems. In *International Conference on Principles of Security and Trust*, pages 390–409. Springer.

[38] Cremers, C., Horvat, M., Hoyland, J., Scott, S., and van der Merwe, T. (2017). A comprehensive symbolic analysis of tls 1.3. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 1773–1788. ACM.

[39] Das, S., Kim, T. H.-J., Dabbish, L. A., and Hong, J. I. (2014). The effect of social influence on security sensitivity. In *Proc. SOUPS*, volume 14.

[40] Dash, M., Koo, K. K., Gomes, J. B., Krishnaswamy, S. P., Rugeles, D., and Shi-Nash, A. (2015). Next place prediction by understanding mobility patterns. In *Pervasive Computing and Communication Workshops (PerCom Workshops), 2015 IEEE International Conference on*, pages 469–474. IEEE.

[41] Dhamija, R., Tygar, J. D., and Hearst, M. (2006). Why phishing works. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 581–590. ACM.

[42] Diekmann, O. and Heesterbeek, J. A. P. (2000). *Mathematical epidemiology of infectious diseases: model building, analysis and interpretation*, volume 5. John Wiley & Sons.

[43] Dolan, P., Hallsworth, M., Halpern, D., King, D., Metcalfe, R., and Vlaev, I. (2012). Influencing behaviour: The mindspace way. *Journal of Economic Psychology*, 33(1):264–277. Elsevier.

[44] Dolan, P., Hallsworth, M., Halpern, D., King, D., and Vlaev, I. (2010). Mindspace: Influencing behaviour for public policy. Institute of Government.

[45] Doyen, S., Klein, O., Pichon, C.-L., and Cleeremans, A. (2012). Behavioral priming: it's all in the mind, but whose mind? *PloS one*, 7(1):e29081. Public Library of Science.

[46] Eloff, J. H. P., Labuschagne, L., and Badenhorst, K. P. (1993). A comparative framework for risk analysis methods. *Computers & Security*, 12(6):597–603.

[47] Field, A., Miles, J., and Field, Z. (2012). *Discovering statistics using R*. Sage publications.

[48] Frank, R. H. (1987). If homo economicus could choose his own utility function, would he want one with a conscience? *The American Economic Review*, pages 593–604. JSTOR.

[49] French, J. (2011). Why nudging is not enough. *Journal of Social Marketing*, 1(2):154–162. Emerald Group Publishing Limited.

[50] Friedkin, N. E. (2006). *A structural theory of social influence*, volume 13. Cambridge University Press.

[51] Gellert, A. and Vintan, L. (2006). Person movement prediction using hidden markov models. *Studies in Informatics and control*, 15(1):17. INFORMATICS AND CONTROL PUBLICATIONS.

[52] Georgeff, M., Pell, B., Pollack, M., Tambe, M., and Wooldridge, M. (1998). The belief-desire-intention model of agency. In *International Workshop on Agent Theories, Architectures, and Languages*, pages 1–10. Springer.

[53] Guo, K. H., Yuan, Y., Archer, N. P., and Connelly, C. E. (2011). Understanding nonmalicious security violations in the workplace: A composite behavior model. *Journal of management information systems*, 28(2):203–236. Routledge.

[54] Han, X. and Tan, Q. (2010). Dynamical behavior of computer virus on internet. *Applied Mathematics and Computation*, 217(6):2520–2526. Elsevier.

[55] Hansson, H. and Jonsson, B. (1994). A logic for reasoning about time and reliability. *Formal aspects of computing*, 6(5):512–535. Springer.

[56] Hazari, S., Hargrave, W., and Clenney, B. (2008). An empirical investigation of factors influencing information security behavior. *Journal of Information Privacy and Security*, 4(4):3–20. Taylor & Francis.

[57] Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, 6(2):107–128. Sage Publications Sage CA: Thousand Oaks, CA.

[58] Herath, T. and Rao, H. R. (2009). Protection motivation and deterrence: a framework for security policy compliance in organisations. *European Journal of Information Systems*, 18(2):106–125. Taylor & Francis.

[59] Hinkley, D. V. and Cox, D. (1979). *Theoretical statistics*. Chapman and Hall/CRC.

[60] Hoaglin, D. C., Iglewicz, B., and Tukey, J. W. (1986). Performance of some resistant rules for outlier labeling. *Journal of the American Statistical Association*, 81(396):991–999. Taylor & Francis Group.

[61] Hosmer, D. W. and Lemesbow, S. (1980). Goodness of fit tests for the multiple logistic regression model. *Communications in statistics-Theory and Methods*, 9(10):1043–1069. Taylor & Francis.

[62] Hu, Q., Dinev, T., Hart, P., and Cooke, D. (2012). Managing employee compliance with information security policies: The critical role of top management and organizational culture. *Decision Sciences*, 43(4):615–660. Wiley Online Library.

[63] Ifinedo, P. (2012). Understanding information systems security policy compliance: An integration of the theory of planned behavior and the protection motivation theory. *Computers & Security*, 31(1):83–95. Elsevier.

[64] Ifinedo, P. (2014). Information systems security policy compliance: An empirical study of the effects of socialisation, influence, and cognition. *Information & Management*, 51(1):69–79. Elsevier.

[65] James, W. (2013). *The principles of psychology*. Read Books Ltd.

[66] Jensen, K. (2013). *Coloured Petri nets: basic concepts, analysis methods and practical use*, volume 1. Springer Science & Business Media.

[67] Jensen, K., Kristensen, L. M., and Wells, L. (2007). Coloured Petri Nets and CPN tools for modelling and validation of concurrent systems. *International Journal on Software Tools for Technology Transfer*, 9(3-4):213–254.

[68] Johnston, A. C. and Warkentin, M. (2010). The influence of perceived source credibility on end user attitudes and intentions to comply with recommended it actions. *Journal of Organizational and End User Computing (JOEUC)*, 22(3):1–21. IGI Global.

[69] Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

[70] Kahneman, D., Knetsch, J. L., and Thaler, R. H. (1991). Anomalies: The endowment effect, loss aversion, and status quo bias. *Journal of Economic perspectives*, 5(1):193–206.

[71] Kammüller, F. and Probst, C. W. (2017). Modeling and verification of insider threats using logical analysis. *IEEE systems journal*, 11(2):534–545. IEEE.

[72] Kent, A. D., Liebrock, L. M., and Neil, J. C. (2015). Authentication graphs: Analyzing user behavior within an enterprise network. *Computers & Security*, 48:150–166. Elsevier.

[73] Kothari, V., Blythe, J., Smith, S., and Koppel, R. (2014). Agent-based modeling of user circumvention of security. In *Proceedings of the 1st International Workshop on Agents and CyberSecurity*, page 5. ACM.

[74] Kubrin, C. E. and Weitzer, R. (2003). New directions in social disorganization theory. *Journal of research in crime and delinquency*, 40(4):374–402. Sage Publications.

[75] Kwiatkowska, M., Norman, G., and Parker, D. (2011). Prism 4.0: Verification of probabilistic real-time systems. In *International Conference on Computer Aided Verification*, pages 585–591. Springer Berlin Heidelberg.

[76] Lee, S. M., Lee, S.-G., and Yoo, S. (2004). An integrative model of computer abuse based on social control and general deterrence theories. *Information & management*, 41(6):707–718. Elsevier.

[77] Lenzini, G., Mauw, S., and Ouchani, S. (2015). Security analysis of socio-technical physical systems. *Computers & electrical engineering*, 47:258–274. Elsevier.

[78] Lenzini, G., Mauw, S., and Ouchani, S. (2016). Analysing the efficacy of security policies in cyber-physical socio-technical systems. In *International Workshop on Security and Trust Management*, pages 170–178. Springer.

[79] Leonard, L. N., Cronan, T. P., and Kreie, J. (2004). What influences it ethical behavior intentions—planned behavior, reasoned action, perceived importance, or individual characteristics? *Information & Management*, 42(1):143–158. Elsevier.

[80] Lior, R. et al. (2014). *Data mining with decision trees: theory and applications*, volume 81. World Scientific.

[81] Michie, S., Atkins, L., and West, R. (2014a). The behaviour change wheel. *A guide to designing interventions. 1st ed. Great Britain: Silverback Publishing*.

[82] Michie, S., van Stralen, M. M., and West, R. (2011). The behaviour change wheel: a new method for characterising and designing behaviour change interventions. *Implementation science*, 6(1):42. BioMed Central.

[83] Michie, S., West, R., Campbell, R., Brown, J., and Gainforth, H. (2014b). *ABC of behaviour change theories*. Silverback Publishing.

[84] Milgram, S. (1963). Behavioral study of obedience. *The Journal of abnormal and social psychology*, 67(4):371. American Psychological Association.

[85] Miller, G. R. and Hewgill, M. A. (1964). The effect of variations in nonfluency on audience ratings of source credibility. *Quarterly Journal of Speech*, 50(1):36–44. Taylor & Francis.

[86] Milner, R. (1982). *A Calculus of Communicating Systems*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

[87] Mitnick, K. D. and Simon, W. L. (2011). *The art of deception: Controlling the human element of security*. John Wiley & Sons.

[88] Murata, T. (1989). Petri nets: Properties, analysis and applications. *Proceedings of the IEEE*, 77(4):541–580. IEEE.

[89] Ng, B.-Y., Kankanhalli, A., and Xu, Y. C. (2009). Studying users' computer security behavior: A health belief perspective. *Decision Support Systems*, 46(4):815–825. Elsevier.

[90] Nurse, J. R., Buckley, O., Legg, P. A., Goldsmith, M., Creese, S., Wright, G. R., and Whitty, M. (2014). Understanding insider threat: A framework for characterising attacks. In *Security and Privacy Workshops (SPW), 2014 IEEE*, pages 214–228. IEEE.

[91] Pahnila, S., Siponen, M., and Mahmood, A. (2007). Employees' behavior towards is security policy compliance. In *System sciences, 2007. HICSS 2007. 40th annual Hawaii international conference on System Sciences*, pages 156b–156b. IEEE.

[92] Pedregosa, F. et al. (2019). Cross-validation. https://scikit-learn.org/stable/modules/cross_validation.html. [Online; accessed 18-August-2019].

[93] Probst, C. W., Kammüller, F., and Hansen, R. R. (2016). Formal modelling and analysis of socio-technical systems. In *Semantics, Logics, and Calculi*, pages 54–73. Springer.

[94] Ramos, J. and Torgler, B. (2012). Are academics messy? testing the broken windows theory with a field experiment in the work environment. *Review of Law & Economics*, 8(3):563–577. De Gruyter.

[95] Rao, H., Greve, H. R., and Davis, G. F. (2001). Fool's gold: Social proof in the initiation and abandonment of coverage by wall street analysts. *Administrative Science Quarterly*, 46(3):502–526. SAGE Publications.

[96] Reisig, W. (2012). *Petri nets: an introduction*, volume 4. Springer Science & Business Media.

[97] Rogers, R. W. (1975). A protection motivation theory of fear appeals and attitude change1. *The journal of psychology*, 91(1):93–114. Taylor & Francis.

[98] Sargent, R. G. (2013). Verification and validation of simulation models. *Journal of simulation*, 7(1):12–24. Taylor & Francis.

[99] Serazzi, G. and Zanero, S. (2004). Computer virus propagation models. In *Performance Tools and Applications to Networked Systems*, pages 26–50. Springer.

[100] Shaw, R. S., Chen, C. C., Harris, A. L., and Huang, H.-J. (2009). The impact of information richness on information security awareness training effectiveness. *Computers & Education*, 52(1):92–100. Elsevier.

[101] Sommestad, T., Hallberg, J., Lundholm, K., and Bengtsson, J. (2014). Variables influencing information security policy compliance: a systematic review of quantitative studies. *Information Management & Computer Security*, 22(1):42–75. Emerald Group Publishing Limited.

[102] Stern, P. C. (1999). Information, incentives, and proenvironmental consumer behavior. *Journal of consumer Policy*, 22(4):461–478. Springer.

[103] Sudeikat, J., Braubach, L., Pokahr, A., Lamersdorf, W., and Renz, W. (2006). Validation of bdi agents. In *International Workshop on Programming Multi-Agent Systems*, pages 185–200. Springer.

[104] Team, C. O. B. I. (2015). Applying behavioural insights to organ donation: preliminary results from a randomised controlled trial. Technical report, Cabinet Office.

[105] Thaler, R. H. (2000). From homo economicus to homo sapiens. *The Journal of Economic Perspectives*, 14(1):133–141. JSTOR.

[106] Thaler, R. H. and Sunstein, C. R. (2009). *Nudge: Improving decisions about health, wealth, and happiness*. Penguin.

[107] Turland, J., Coventry, L., Jeske, D., Briggs, P., and van Moorsel, A. (2015). Nudging towards security: Developing an application for wireless network selection for android phones. In *Proceedings of the 2015 British HCI conference*, pages 193–201. ACM.

[108] Übelacker, S. (2013). *Security-aware organisational cultures as a starting point for mitigating socio-technical risks*.

[109] Uebelacker, S. and Quiel, S. (2014). The social engineering personality framework. In *Socio-Technical Aspects in Security and Trust (STAST), 2014 Workshop on*, pages 24–30. IEEE.

[110] Van der Hoek, W. and Wooldridge, M. (2008). Multi-agent systems. *Foundations of Artificial Intelligence*, 3:887–928. Elsevier.

[111] Vroom, C. and Von Solms, R. (2004). Towards information security behavioural compliance. *Computers & Security*, 23(3):191–198. Elsevier.

[112] Warkentin, M. and Willison, R. (2009). Behavioral and policy issues in information systems security: the insider threat. *European Journal of Information Systems*, 18(2):101–105. Springer.

[113] White, A. P. and Liu, W. Z. (1994). Bias in information-based measures in decision tree induction. *Machine Learning*, 15(3):321–329. Springer.

[114] Wilson, J. Q. and Kelling, G. L. (1982). Broken windows. *Atlantic monthly*, 249(3):29–38.

[115] Wooldridge, M. (2009). *An introduction to multiagent systems*. John Wiley & Sons.

[116] Yevseyeva, I., Morisset, C., and Van Moorsel, A. (2016). Modeling and analysis of influence power for information security decisions. *Performance Evaluation*, 98:36–51. Elsevier.

[117] Zhu, F., Carpenter, S., Kulkarni, A., and Kolimi, S. (2011). Reciprocity attacks. In *Proceedings of the Seventh Symposium on Usable Privacy and Security*, pages 9–23. ACM.

Appendix A

Experiment Pre-Registration

A.1 Structured Abstract

Background. A social influence of a messenger or the broken-window effect may impact the compliance level for security policies [44, 81, 114].

Aim. We investigate the effect of socio-environmental interventions on smart card swiping behaviour.

Method. We will conduct a study with three conditions in a between-subjects design on a sample of university students owning a university-issued access-control smart card.

In all three groups, participants will be asked to complete a set of Capture the Flag challenges in our Cyber Security Room where they had no Internet access. Outside of the room, the participants could access the Internet. Thereby, participants will be compelled to leave and enter the room, without disclosing the experiment purpose.

We will ask participants to swipe their smart card on entering and exiting the Cyber Security Room. The Control group had no intervention. The Discrete+ experiment group will be exposed to a *Messenger* influence [44]. The Continuous− experiment group will be exposed to a an *untidy* Cyber Security Room, which was inspired by the broken-window Theory [114]. We measured swiping behavior on entry and exit, computing total swipes and swipe rate ratio as key metrics subjected to an Analysis of Variance.

Anticipated Results. We expect to see a shift in the security decision making of the individuals. We will request a debrief questionnaire from subjects to understand their

Table A.1 Operationalisation of Study: Interventions on Smart Card Swiping Behaviour

	Groups	Intervention	Instrument
IV: Condition	Control Discrete+ Continuous−	None *Messenger* [44] broken-window [94]	None Reminder Messy Room
DV: Swipe	All	Visual Log	Video Camera Adafruit PN532 Reader

self awareness towards security decision making. In particular, we expect that under the negative continuous intervention, the total number of swipes decreases. Whereas, the discrete intervention will increase the number of swipes in a given time interval.

Anticipated Conclusions. We expect that an intervention can change the security decision making of a group.

A.2 State of Data Collection

Has any data been collected for this study yet?

(a) ☑ **NO** data has been collected.

(b) □ Some data has been collected, but not analyzed.

(c) □ Some data has been collected and analyzed.

A.3 Aims

Research Question 8 (Messenger Effect)**.** *To what extent does a Messenger effect have on the swipe rate of participants?*

Table A.1 provides an overview of the operationalisation for this research question. As an independent variable (IV), we have selected the use of the *Messenger* effect [44].

Research Question 9 (broken-window Effect)**.** *To what extend does the broken-window effect have an effect on the swipe rate of participants?*

Table A.1 also provides the info for the second research question. As the IV for this case, we consider a messy ('untidy') room as the intervention.

For both research questions, we intend to compare them against a Control group which has no experimental treatment. For all three groups (Control, Discrete+ and Continuous−), we will measure the dependent variable (DV) by use of a video camera and smart card logging system. Using both of these metrics, we will generate an event log for all participants.

Hypotheses on Total Number of Swipes The overall null hypothesis for this experiment is $H_{\mathsf{T},0}$ *: There is no mean difference in the total number of swipes for participants exposed to an intervention.*

We have subordinate null hypotheses for each condition:

$H_{\mathsf{T},0,\mathsf{D}+}$**:** *The* Discrete+ *intervention does not impact the mean total number of swipes.*

$H_{\mathsf{T},0,\mathsf{C}-}$**:** *: The* Continuous− *intervention does not impact the mean total number of swipes.*

We have these as alternative hypotheses.

$H_{\mathsf{T},1,\mathsf{D}+}$**:** *The* Discrete+ *intervention (Messenger effect) impacts the total number of swipes.*

$H_{T,1,\mathsf{C}-}$**:** *A* Continuous− *messy ('untidy') room (broken-window effect) impacts the total number of swipes.*

Hypotheses on Swipe Rate Ratio The overall null hypothesis for this experiment is $H_{\mathsf{R},0}$ *: There is no mean difference in the swipe rate ratio for participants exposed to an intervention.*

The subordinate null hypotheses for each condition are:

$H_{\mathsf{R},0,\mathsf{D}+}$**:** *The* Discrete+ *intervention does not impact the mean swipe rate ratio.*

$H_{\mathsf{R},0,\mathsf{C}-}$**:** *The* Continuous− *intervention does not impact the mean swipe rate ratio.*

$H_{\mathsf{R},1,\mathsf{D}+}$**:** *The* Discrete+ *intervention (Messenger effect) impacts the swipe rate ratio.*

$H_{R,1,\mathsf{C}-}$**:** *A* Continuous− *messy ('untidy') room (broken-window effect) impacts the swipe rate ratio.*

A.4 Methods

This section lists the methods of the experiment construction and provides a clear guide on what we implemented.

As part of the experiment, subjects will complete a variety of tasks over the course of an hour. Each task has some general theme of security and will be discussed in this section.

A.4.1 Study Groups

Participants are randomly assigned to three groups:

- Control: Our first group is our baseline where participants perform a set of tasks over fifty five minutes. No intervention is in place here. The room is set in a defined 'tidy' state.
- Discrete+: The first experiment group is exposed to a discrete intervention. By discrete we mean at one time point the group is exposed to some intervention. The purpose of this intervention is to positively influence group behaviour around swipe compliance. The room is kept in a defined 'tidy' state. We anticipate that this will increase participants swipe levels.
- Continuous−: The second experiment group is exposed to a continuous intervention present in the *offline* Cyber Security Room. The intervention is specifically to set the room in a defined 'untidy' state. The purpose of this intervention is to negatively influence group behaviour around swipe compliance. We anticipate that this will decrease participants swipe levels.

In each group we will ensure a short break is in place at the 25 minute mark. At this point we gathered all of the participants in the clean room and ask them how they are getting on. The instructor will swipe their smart card as they enter and say *"How is everyone finding the tasks?"*. The instructor will then swipe their smart card then leave the room. Each groups session time will be fifty five minutes ($\pm$20s).

A.4.2 Experimental Environment

The experiment will be conducted in a purpose-built environment for cyber-security capture-the-flag tasks.

In the experiment, we designate *offline* and *online* areas. The *offline* area had no Internet connectivity, which would be considered a high-security clean-room. The *online* area had

Raspberry Pis provided to allow for Internet connectivity. Subjects must swipe their student smart card as they enter and exit the *offline* area.

Offline Area: The *offline* area includes a Cyber Security Room and contains sensitive experimental equipment. During the buildings construction, the room was not fitted with a smart card reader. We make the following changes to this room:

- Using a Raspberry Pi we design and program an Adafruit PN532 RFID Card Reader to work with Smart Cards.
- We fit a Video Camera to record the experiment. The camera is inside the offline room, points at the door, and observes the smart card reader.

Online Area: Subjects will use this area to collect supporting information to assist the completion of the Capture-The-Flag challenges in the cyber security room. This area has Raspberry Pis with a mouse, keyboard, monitor and wireless network connection.

A.4.3 Study Groups

For the purposes of this study we are considering different influences towards compliant behaviour. We split the study into three groups which are the following:

- ***Control Group:*** Our first group is our baseline where subjects perform a set of tasks over an hour. No intervention is in place here. To ensure consistency the room setup will be maintained for each Control Group.
- ***Experiment Group One:*** The first experiment group is a discrete intervention. By discrete we mean at one time point the group is exposed to some intervention. The purpose of this intervention is to negatively influence group behaviour around swipe compliance.
- ***Experiment Group Two:*** The second experiment group is a continuous intervention. This is an intervention which is always present in our *clean* area. The purpose of this intervention is to negatively influence group behaviour around swipe compliance.

In each group we will ensure a short break is in place at the 25 minute mark. At this point we will gather all of the subjects in the clean room and ask them how they are getting on.

As the instructor I will swipe my smart card as I enter and say *"How is everyone finding the tasks?"*

A.4.4 Experiment Group One - Discrete Intervention Positive

The goal of the discrete intervention is to create a positive outlook on swiping a smart card for subjects. The re-enforcement from us as the instructor provides that authoritative message that this is the behaviour we expect.

At the short break, once all subjects are in the clean room we will provide our discrete intervention. As the instructor, I will swipe my card as I enter and say to all participants *"Just a reminder to make sure you all are swiping as you enter and exit. How is everyone finding the tasks?"*

A.4.5 Experiment Group Two - Continuous Intervention Negative

The goal of the continuous intervention is to reduce the swiping of individuals by creating a untidy environment.

The continuous intervention will consist of making the clean room *untidy* by placing the following in the room:

- Out of order signs on unused machine
- Collection of monitors and machines on the floor.
- A wastebin full of plastic items.

The break at the 25 minute mark will still take place. As the instructor I will swipe my smart card as I enter and say *"How is everyone finding the tasks?"*.

A.4.6 Ecological Validity

In order to misdirect subjects away from the true purpose of the experiment, we must ensure the ecological validity of the task relates to the security policy in place.

The subjects will be told the following:

- *The experiment is seeking to understand how subjects manage secure information. In specifics, when you have no Internet connectivity in one location and do in another.*
- *We are monitoring movement between locations and in this setting a security requirement is for everyone to swipe their smart card on entry and exit.*

A.5 Independent Variables (IVs)

Describe the conditions (for an experimental study) or predictor variables (for a correlational study).

Our manipulation is the introduction of an intervention in the experiment groups. This intervention will be either discrete/continuous.

We thereby have one IV (condition) with three levels:

Control Group: Statement of the security policy only.

Experiment Group One: Discrete intervention, positive: Experimenter entering the room stating a reminder of the policy.

Experiment Group Two: Continuous intervention, negative: Room put in a specified untidy state.

A.6 Dependent Variables (DVs)

Dependent variables: Describe the key dependent variable(s) specifying how they will be measured.

We measure the consistency of swipes for entry and exit in a secure room against visual logs which give us the ground truth. The swipes are collected by the smart card reader and the visual logs from the IP camera.

A.7 Mediator Variables

Describe any variables you expect to mediate the relationship between your IV's and DV. Specify how they will be measured.

N/A

A.8 Moderator Variables

Describe any variables you expect to moderate the relationship between your IV's and DV. Specify how they will be measured.

A.9 Control of Confounding Variables (CVs)

We employ the following methods to ensure the control of confounding variables:

- Random assignment of participants for groups. Participants will be randomly assigned to either a control or experiment group, from there they will be offered randomly selected slots which they choose to attend.
- Random assignment of tasks for participants. Due to the nature of the tasks, it is not clear in advance which tasks will cause more people to leave the room, which could skew our results. As such, we randomly assign tasks to ensure that subjects are not all completing the same tasks and we have an even distribution for how people move over the course of a group.

A.10 Data Preparation

Describe what measures will be taken to check assumptions and label outliers.

We will measure who and when subjects swipe their smart cards.

With these two overarching measurements we capture:

Table A.2 Example Frequency table for study results.

Freq.	Swipes	Missed
Control	$swipes^{+}(g_1)$	$swipes^{-}(g_1)$
Experiment	$swipes^{+}(e_1)$	$swipes^{-}(e_1)$
Experiment	$swipes^{+}(e_2)$	$swipes^{-}(e_2)$

- The relationship between non-compliance and main location in the room.
- When is data correct?
- Do we miss any data?
- Do we check for outliers?

We first consider the function $\beta : S \to N$, which corresponds to the number of times the smartcard corresponding to a given subject has been swiped. This information is directly collected by the smartcard reader. (**Any correction for errors? inconsistencies? what about multiple swipes?**).

Given a group of subjects $G \subseteq S$, we define the number of recorded swipes as:

$$swipes^{+}(G) = \sum_{s \in G} \beta(s)$$

We then consider a function $\alpha : S \to N$, which corresponds to the number of visual swipes done by a particular subject. This data is recorded through CCTV/webcam monitoring (**manual inspection? automatic recording? Any room for errors?**). We can the define the metrics:

$$swipes^{-}(G) = \sum_{s \in G} \alpha(s) - \sum_{s \in G} \beta(s)$$

The frequency data for the study is shown in Table A.2.

A.11 Data Analysis

The data we collect from running each study group consists of smart card logs, video footage, pre-study consent form with a personality test for each subject and a post study questionnaire.

A.11.1 Consistency Check

Of the four data sets we collect for each study, we will use the smart card logs and the video footage to generate the ground truth, which creates our fifth data set named visual logs. This empirical evidence will provide us with a high level of certainty for the events that have occurred.

To generate the visual logs we define a process to ensure that all studies are treated with the same level of consistency and rigor:

1. **Synchronise** the times from the video footage and smart card logs.
2. **Correlate** the visual initial swipe of a subject with the hashed smart card log for that swipe.
3. **Generate** the list of events for the timed part of the study, i.e. where subjects are completing challenges.

Sychronise: The clock on the video footage begins at 00:00 and the smart card logs time as per Greenwich Mean Time. For each experiment, the study leader will swipe their smart card after the camera is switched to record mode. From this we can gather the exact timings of swipes with regards to the smart card logs. This will also assist when we generate the logs for the visual logs data set, as we can provide an absolute time.

Correlate: As all participants smart cards are hashed, we have no association between the names of subjects and their smart card number. To ensure that the visual logs can be generated we correlate the initial swipe of a subject with visual features. We use the following metrics to classify a subject and address conflict resolution:

- Are they wearing glasses?
- What colour top/t-shirt/coat are they wearing?
- What colour hair do they have?
- If these three features clash with another participant in that group, do they have any unique features that are easily identifiable?

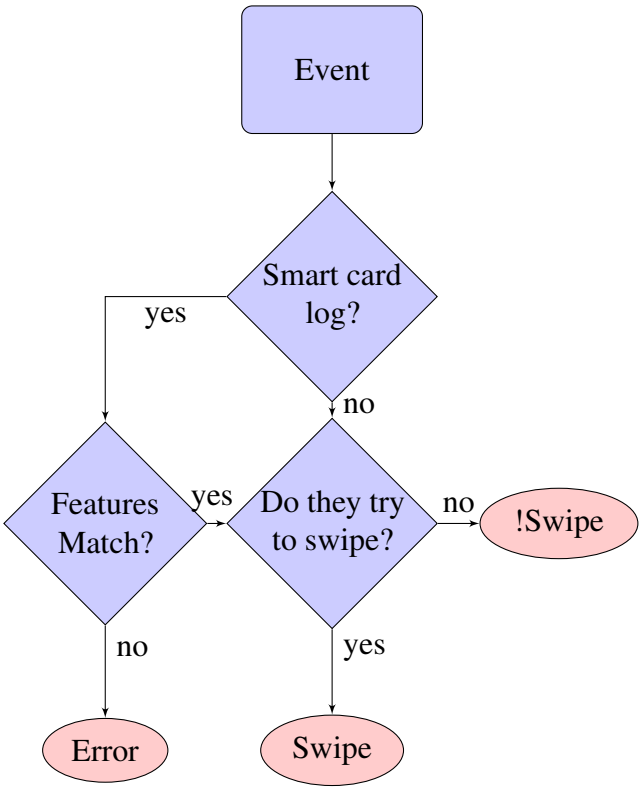


Figure A.1 Process for recording an event, where Swipe refers to the expected behaviour of someone swiping, !Swipe is someone entering or exiting with swiping. Finally, Malicious refers to someone swiping with a different card.

- What is the hash of their smart card?

We should know the hash of each user's smart card, as they will enter in a specific ordering, which is associated with a number for where they are sat. This initial swipe will be before the timed section of the study begins and we will use this to correlate their features to their smart card hash value.

Generate: In order to generate visual logs we make use the video footage to gather information for each event that occurs. Where an event is a person entering or exiting the room.

Figure A.1 describes the process for recording an event so that it can be logged in the visual logs data set. An outcome will create a log with the timestamp and user number/smart card hash for the event. After all events have been processed, the visual logs files will provide the ground truth for what occurred. We make the following assumptions for the data:

Assumption 1 (Multiple Swipes)**.** *If the smart card log contains identical user swipes within 1 seconds of each other, these are removed and treated as the card reader being too sensitive or the subject tried to swipe for too long of a period.*

Assumption 2 (Accurate Smart Card Logs)**.** *We assume that the smart card reader is infallible and does not produce an incorrect log when a smart card is swiped.*

Assumption 3 (Accountable Smart Card Logs)**.** *We assume that the smart card reader only produces a smart card log when a RFID card is presented. It will not create phantom logs when no RFID is present.*

Any entry generated will take one of four formats of:

$$time, group, hash, entrySwipe \tag{A.1}$$

$$time, group, hash, exitSwipe \tag{A.2}$$

$$time, group, hash, entryNoSwipe \tag{A.3}$$

$$time, group, hash, exitNoSwipe \tag{A.4}$$

Where $time$ is the time of the swipe as adjusted due to the clock synchronisation, $group$ refers to either Control, Discrete or Continuous, $hash$ is the hashed smart card value providing a unique identifier and the final entry refers to the type of event that occurred.

A.12 Main Analyses

Describe what analyses (e.g., t-test, repeated-measures ANOVA) you will use to test your main hypotheses.

We will make use of the ANOVA test then use the planned comparisons to calculate effect sizes.

Our format will be as follows:

1. Run the ANOVA:
    - Assess the F-value and F-test results.
2. Planned Comparisons (Table A.3)

- Control vs Discrete
- Control vs Continuous

3. Calculate Effect Sizes

Table A.3 Planned Comparisons against Control condition.

Contrast	**CTRL**	**CONT-**	**DISC+**
1	1	-1	0
2	1	0	-1

A.12.1 ANOVA Testing

We will use an analysis of variance testing to ensure that we gather the F-value and F-test results for the comparisons of mean between Control, Continuous- and Discrete+.

A.12.2 Inclusion of Outliers

As a final data preparation we perform multiple tests which vary the size of an outliers labelling rule where we use the interquartile range as part of resistant rules to generate inner fences for an inclusion zone:

$$IF_L = LQ - 1.5(IQ) \tag{A.5}$$

$$IF_L = LQ - 3(IQ) \tag{A.6}$$

$$IF_U = UQ + 1.5(IQ) \tag{A.7}$$

$$IF_U = UQ + 3(IQ) \tag{A.8}$$

Where IF is the inner fence, LQ and UQ are the lower and upper quartile and IQ is the inter–quartile range [60].

Once outliers have been detected we will then cap them at the 5th and 95th percentile of data to ensure that results and effect sizes are not heavily influenced by one data point.

A.12.3 Calculating Effect Sizes

We will calculate the effect sizes for our planned comparisons using two methods. The first will be Cohen's d which we use to describe the standardized mean difference of an effect.

The second is the Hedges-g which is often referred to as the corrected effect size.

A.13 Intermediary Check

We will perform an Intermediary check at the halfway point, roughly 50-60 participants to assess whether or not it is worth investing continuing resources into the study.

A.14 Pilot Study

We ran a Pilot Study to understand the type of data we receive and to ensure the experiment would run as expected. It consisted of four PhD students all from Cyber Security. The participants will be aware of the experiment and its aim before participating. Therefore, the pilot study does not give us any indication towards the expected data, just ability to validate the experiment runs. Due to two camera failures we only have data for twenty minutes.

A.14.1 Pilot Study: Results

Table A.4 Pilot Study: Results

User	**Swipes**		**Visual**		**Ratio**	**IV**
	In	*Out*	*In*	*Out*		
1	4	3	4	3	1	NA
2	3	2	3	2	1	NA
3	3	2	3	2	1	NA
4	1	0	1	0	1	NA

Our results show us that all of the participants swiped as they entered and exited the *Clean* room.

After speaking with the participants, we found that they admitted to forgetting to swipe, however, this was after the twenty minute period where we had no visual log due to multiple camera failures.

Table A.4 shows the results from the pilot study. Appendix 3 (last one) shows the type of data we are getting back from the swipe card system.

A.15 Secondary Analyses

Describe what secondary analyses you plan to conduct (e.g., order or gender effects).

As a secondary analysis, we plan a binomial logistic regression with the experiment condition as main predictor and the pre-experiment personality traits as co-variate predictors. We model as response variable the outcomes of the swiping events. Primarily, we are interested in a binomial logistic regression, that is, collapsing the possibly multinomial cases of swiped/not-swiped—with-own-card/with-another's-card to a single nominal DV of swiped/not-swiped, encoded as 1 and 0. We expect the vast majority of observed cases to be cleanly projectable onto a binomial response variable.

In this analysis, we consider the effect of the Big Five Personality Traits [13], where we would expect, for instance, agreeableness to amplify the compliance to the experimenter's request under the positive/discrete experiment condition.

Furthermore, we are measuring a physical security policy, where the actions of individuals are public. Attributional biases within individuals may be influenced by those who do or don't comply.

A.16 Validation

Describe what diagnostics or validation methods you plan to employ to check the soundness of the analyses.

For the the χ^2 tests of independence, we will check the expected cell counts after having established the contingency tables.

For the binomial logistic regressions, we will use the list of assumptions accounted for by Laerd Statistics [4] as guidance, along with Andy Field's account of checks for logistic regression assumptions [47].

A major validation analysis will be on the residual after the model has been obtained. We will consider standardized and studentized residuals, leverage, and DFBeta.

To consider multicoliniarity, we will further consider the variance inflation factor (VIF), per predictor as well as the average VIF. We will test the linearity of the logit following Field [47, Section 8.8.2].

Overall, we will use the R package car for the regression diagnostics.

In terms of variance explained of the logistic regression model, we will compute pseudo-R^2 metrics: Hosmer and Lemeshow, Cox and Snell, as well as Nagelkerke.

We will compute a validation of the logistic regression model in an 80:20 random resampling. We will use a Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) as metrics for accuracy.

A.17 Sample

Where and from whom will data be collected? How will you decide when to stop collecting data (e.g., target sample size based on power analysis or accuracy in parameter estimation, set amount of time)? If you plan to look at the data using sequential analysis, describe that here.

The total sample size of the groups are expected to be around one hundred. We split the subjects randomly across the experiment and control group. Each group will then be allocated a number of sessions where subjects will perform the outlined tasks.

A.18 Exclusion Criteria

Who will be excluded (e.g., outliers, participant who fail manipulation check, demographic exclusions)? Will they be replaced by other participants?

We will exclude event observations that are inconclusive, that is, cases in which the observations from the RFID and CCTV1 sensors are contradictory and cannot be resolved with our decision algorithm.

We will also exclude participants who do not exit the room during the time of the study. That is, they never swipe.

A.19 Exception Handling

Should exceptions from the planned study occur (e.g., unexpected effects observed), how will they be handled?

Exceptions will be documented explicitly. Unexpected effects and further analyses will be considered exploratory and documented as such.

A.20 Sign-Off

Pre-registration written by (initials): P.C. / T.G.

Pre-registration reviewed by (initials): T.G.

Appendix B

User Study Material

Documentation for user study starts on next page.

RESEARCH PARTICIPANT INFORMATION SHEET

Dear prospective participant: Please read this information sheet carefully and ask as many questions as you like before you decide whether you want to participate in this research study. You are free to ask questions at any time before, during, or after your participation in this research.

Project Title:	Cyber Security Challenges
Principal Investigator:	Thomas Gross
Location:	Newcastle University, School of Computing

PURPOSE OF THIS RESEARCH STUDY
You are being asked to participate in a research study designed to find out how users manage solving security challenges when faced with no internet connectivity.

PROCEDURES
You will be asked to fill in of personality traits (5 minutes).
You will be asked to fill in a survey on your general attitude towards the study (5 minutes).
You will be asked to solve a set of security challenges where one of the rooms you work in is partially recorded. (We provide a supplemental information sheet about camera data) (60 minutes)
The total time of the experiment should not take any longer than 75 minutes.

CONFIDENTIALITY AND DATA PROTECTION
Only your age and gender will be recorded along with the experimental data. This data will be recorded anonymously, i.e. your name will not be associated with them. Your experimental data as well as the video will be stored securely on an encrypted device. Your smart card is stored using encryption.

COMPENSATION
We will compensate you for the time spent in the experiment according to the conventions of Newcastle Psychology Schools, a *£10* Amazon Voucher.

TERMINATION OF RESEARCH STUDY
You are free to choose whether or not to participate in this study. You can choose to cease participation at any time.

AVAILABLE SOURCES OF INFORMATION
Any further questions you have about this study will be answered by the PhD student:
Name: *Peter Carmichael* (*p.j.carmichael@ncl.ac.uk*)

You can also approach the Principal Investigator, Dr. Thomas Gross (thomas.gross@ncl.ac.uk).

User Number ……..

How I am in general

Here are a number of characteristics that may or may not apply to you. For example, do you agree that you are someone who *likes to spend time with others*? Please write a number next to each statement to indicate the extent to which **you agree or disagree with that statement.**

1 Disagree Strongly	2 Disagree a little	3 Neither agree nor disagree	4 Agree a little	5 Agree strongly

I am someone who…

1. _____ Is talkative
2. _____ Tends to find fault with others
3. _____ Does a thorough job
4. _____ Is depressed, blue
5. _____ Is original, comes up with new ideas
6. _____ Is reserved
7. _____ Is helpful and unselfish with others
8. _____ Can be somewhat careless
9. _____ Is relaxed, handles stress well.
10. _____ Is curious about many different things
11. _____ Is full of energy
12. _____ Starts quarrels with others
13. _____ Is a reliable worker
14. _____ Can be tense
15. _____ Is ingenious, a deep thinker
16. _____ Generates a lot of enthusiasm
17. _____ Has a forgiving nature
18. _____ Tends to be disorganized
19. _____ Worries a lot
20. _____ Has an active imagination
21. _____ Tends to be quiet
22. _____ Is generally trusting
23. _____ Tends to be lazy
24. _____ Is emotionally stable, not easily upset
25. _____ Is inventive
26. _____ Has an assertive personality
27. _____ Can be cold and aloof
28. _____ Perseveres until the task is finished
29. _____ Can be moody
30. _____ Values artistic, aesthetic experiences
31. _____ Is sometimes shy, inhibited
32. _____ Is considerate and kind to almost everyone
33. _____ Does things efficiently
34. _____ Remains calm in tense situations
35. _____ Prefers work that is routine
36. _____ Is outgoing, sociable
37. _____ Is sometimes rude to others
38. _____ Makes plans and follows through with them
39. _____ Gets nervous easily
40. _____ Likes to reflect, play with ideas
41. _____ Has few artistic interests
42. _____ Likes to cooperate with others
43. _____ Is easily distracted
44. _____ Is sophisticated in art, music, or literature

RESEARCH PARTICIPANT INFORMATION SHEET

Dear prospective participant: Please read this information sheet carefully and ask as many questions as you like before you decide whether you want to participate in this research study. You are free to ask questions at any time before, during, or after your participation in this research.

Project Title:	Cyber Security Challenges
Principal Investigator:	Thomas Gross
Location:	Newcastle University, School of Computing

THE USE OF VIDEO CAMERAS IN EXPERIMENTS
During this experiment, a high-resolution Webcam will video people moving in and out of one location.

PURPOSE OF CAPTURED VIDEOS
We record your movement to identify how you moved around during the security challenges.

CONFIDENTIALITY AND DATA PROTECTION
We consider the videos of your face as sensitive information and take special care to protect it.
We only use the videos for the analysis of information scores and for no other purpose.
While we are processing the videos they will be stored anonymously, that is, without association to your name, and stored securely on an encrypted device.

Once the experiment is complete and we have processed the data, we will delete the captured videos.

Under no circumstances will the video itself ever be used in a report, a publication or any other research output.

OPT-OUT OF VIDEO CAPTURE
If you wish to opt out of the video capture then you will opt of the study entirely, you are free to do this.

AVAILABLE SOURCES OF INFORMATION
Any further questions you have about this study will be answered by the PhD Student:
Name: *Peter Carmichael* (*p.j.carmichael@ncl.ac.uk*)

You can also approach the Principal Investigator, Dr. Thomas Gross (thomas.gross@ncl.ac.uk).

Informed Consent Form

Cyber Security Study
I, the undersigned, confirm that (please tick box as appropriate):

1.	I have read and understood the information about the project, as provided in the Information Sheet dated _____.	☐
2.	I have been given the opportunity to ask questions about the project and my participation.	☐
3.	I voluntarily agree to participate in the project.	☐
4.	I understand I can withdraw at any time without giving reasons and that I will not be penalised for withdrawing nor will I be questioned on why I have withdrawn.	☐
5.	The procedures regarding confidentiality and data protection have been clearly explained to me (e.g. use of names/pseudonyms, anonymisation of data, etc.).	☐
6.	If applicable, separate terms of consent for interviews, audio, video or other forms of data collection have been explained and provided to me.	☐
7.	The use of the data in research, publications, sharing and archiving has been explained to me.	☐
8.	I understand that other researchers will have access to this data only if they agree to preserve the confidentiality of the data and if they agree to the terms I have specified in this form.	☐
9.	Select only **one** of the following: • I would like that what I have said or written as part of this study will be used in reports, publications and other research outputs. • I do not want what I have said or written as part of this study used in reports, publications or other research outputs.	☐ ☐
10.	I, along with the Researcher, agree to sign and date this informed consent form.	☐

Participant:

_____ _____ _____
Name of Participant Signature Date

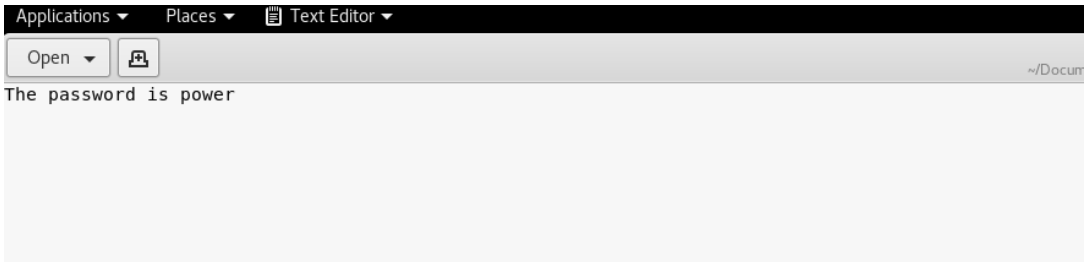
Researcher:

_____ _____ _____
Name of Researcher Signature Date

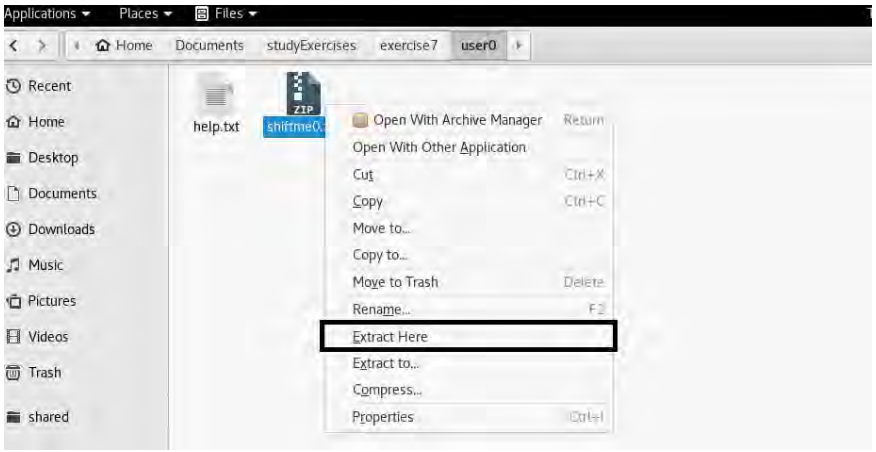
Project Title:
Main Contact: Peter Carmichael (p.j.carmichael@ncl.ac.uk)

How to complete a task

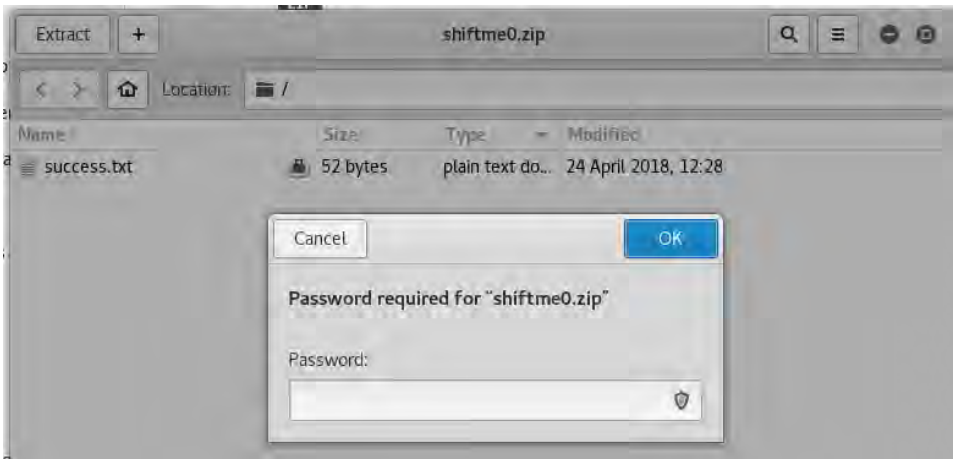
1. Get the password of the zipped file.



2. Click Extract Here



3. Click Extract twice
4. Enter Password



Questionnaire – Post Study

User Number: ……..

1. How often did the study leader enter the room and remind you to swipe your smart card?

   1-5 (1 – Very Often; 2 – Somewhat Often; 3 – Did not notice; 4 – Not that often; 5 – Not at all)

   ……

2. How often did you swipe your smartcard during the completion of the cyber security challenges (i.e. the past hour)?

   1-5 (1 – Very Often; 2 – Somewhat Often; 3 – Did not notice; 4 – Not that often; 5 – Not at all)

   ……

3. How diligently have you swiped your smartcard? On a scale of 1-5 where 1 is not diligently to 5 which is very diligently

   1-5 (1 – Not diligently; 2 – Somewhat diligently; 3 – Neither diligently or not diligently; 4 – Diligently; 5-Very diligently)

   ……

4. *Do you agree that all participants in the group swiped their smart card every time then entered or exited the room?*

   1-5 (1 – Strongly Disagree; 2 – Somewhat Disagree; 3 – Neither Disagree or Agree; 4 – Somewhat Agree; 5 – Strongly Agree)

   ……

5. *How often did you swipe with someone else's card?*

   1-5 (1 – Very Often; 2 – Somewhat Often; 3 – Did not notice; 4 – Not that often; 5 – Not at all)

   ……

6. *Indicate to what extent you felt during the past hour of this study? (Do this for all emotions)*

   1 – Very Slightly or Not at All; 2 – A little; 3- Moderately; 4 – Quite a Bit; 5 Extremely

_____ 1. Interested	_____ 2. Distressed	_____ 3. Excited	_____ 4. Upset
_____ 5. Strong	_____ 6. Guilty	_____ 7. Scared	_____ 8. Hostile
_____ 9. Enthusiastic	_____ 10. Proud	_____ 11. Irritable	_____ 12. Alert
_____ 13. Ashamed	_____ 14. Inspired	_____ 15. Nervous	_____ 16. Determined
_____ 17. Attentive	_____ 18. Jittery	_____ 19. Active	_____ 20. Afraid

7. *How often did you collaborate with another participant during the study?*

   1-5 (1 – Very Often; 2 – Often; 3 – Did not notice; 4 – Not that often; 5 – Not at all)

   ……

8. *How was your behaviour influenced by other participants being in the room?*

   1-5 (1 – Very Often; 2 – Often; 3 – Did not notice; 4 – Not that often; 5 – Not at all)

   ……

9. *How did you perceive it when the study leader entered the room to remind you of smartcard swiping?*

   ……

10. *What were your motivations to swipe or not swipe your smartcard?*

    ……

11. *How important do you perceive it is to swipe the smartcard for security?*

    1-5 (1 – Very Important; 2 – Somewhat Important; 3 – Neither Important or Unimportant; 4 – Somewhat Unimportant 5 – Very Unimportant)

    ……

Secure Data Management

Project - User Behaviour on Security Challenges

Peter Carmichael, Thomas Gross & Charles Morisset

Overview

This project runs many studies where in each study, participants swipe their smart card and are recorded as they move between two different rooms on the sixth floor of the Urban Sciences Building.

Data Management

We recognise that both the smart card data and video footage must be effectively managed. The data we collect in this study does not directly link to identity; however, it is still sensitive.

Smart Card Data

The smart card data is one-way hashed and we do this on the UID (Unique Identifier). We do not record the student number, picture or name from the card.

The purpose of this collection is to manage how regularly users move between locations. We do the following with the smart card data:

1. The data is collected on a Raspberry Pi with a micro SD card during each study.
2. At the end of each study the data is transferred from the SD card to a Newcastle University Laptop, where it is stored locally. It is then removed from the SD card.
3. Once collected at the end of each study, we use a programming script to match a unique pseudonym to the hashed UID, which completely removes the opportunity for anyone to recover the original UID.
4. The data is then safe for processing and using in our research.

Video Recording

The study records the area where participants swipe their smart card. We use a HD handheld camera to record and the video is stored directly onto a micro SD card.

The purpose of this collection is to manage how regularly users move between locations. We do the following with the video recordings:

1. The video recordings are collected on a HD handheld camera.
2. At the end of each study the recordings are transferred to a Newcastle University Laptop. It is possible we may have to use an external hard drive. In which case, the hard drive is encrypted and stored in a locked cabinet. The recording is then wiped from the SD card.
3. Within reasonable time, the video is processed and used to log the events in the study.
4. After processing, the video is deleted as it is no longer required.

For any more questions please contact Peter Carmichael by email (p.j.carmichael@ncl.ac.uk).

University Ethics Form Version 2.1.1

Date submitted
19/04/2018 17:48:43

Applicant Details

Is this approval for a:
Staff Project [A1]

Name of Principal Researcher:
Thomas Gross

Please enter your email address
p.j.carmichael@ncl.ac.uk

Please select your school / academic unit
School of Computing [A1]

Project Details

Project Title
Security Challenge Human Behaviour

Project Synopsis
Recent research investigated the effect of an influence on a private security decision, however, we do not understand how public security decisions influence a group culture towards security policy compliance. We investigate the impact of an influencing intervention on smart card swiping behaviour.

Project start date
25/04/2018

Project end date
31/07/2018

Is the project externally funded?
No [A3]

Does your project involve collaborators outside of the University?
No [N]

Existing Ethics, Sponsorship & Responsibility

Has ethical approval to cover this proposal already been obtained?
No [N]

Will anyone be acting as sponsor under the NHS Research Governance Framework for Health and Social Care?
No [N]

Do you have a Newcastle upon Tyne Hospitals (NUTH) reference?
No [N]

Will someone other than you (the principal investigator) or your supervisor (for student projects) be responsible for the conduct, management and design of the research?
No [N]

The Animals (Scientific Procedures) Act defines protected animals as: 'any living vertebrate other than man…in its foetal, larval or embryonic form……from the stage of its development when— (a)in the case of a mammal, bird or reptile, half the gestation or incubation period for the relevant species has elapsed; and (b)in any other case, it becomes capable of independent feeding'.

In practice 'Protected' animals are all living vertebrates (other than man), including some immature forms, and cephalopods (e.g. octopus, squid, cuttlefish).

Using this definition, does your research involve the observation, capture or manipulation of animals or their tissues?

No [N]

Will the study involve participants recruited by virtue of being NHS patients or service users, their dependents, their carers or human tissues or the use of NHS & Health/Social Care Facilities or otherwise require REC approval?

No [N]

Does the research involve human participants e.g. use of questionnaires, focus groups, observation, surveys or lab-based studies involving human participants?

Yes [Y]

Does the study involve any of the following? [a. The study involves children or other vulnerable groups; as defined in Section 59 of the Safeguarding Vulnerable Adults Act 2006 as those who are relatively or absolutely incapable of protecting their own interests, or those in unequal relationships e.g. participants who are subordinate to the researcher(s) in a context outside the research?]

Does the study involve any of the following? [b. The study requires the co-operation of a gatekeeper (defined as someone who can exert undue influence) for initial access to the groups or individuals to be recruited e.g. students at school, members of a self-help group, or residents of a nursing home? NB. The IoN & School of Psychology volunteer pools are not considered gatekeepers in this case.]

Does the study involve any of the following? [c. It is necessary for participants to take part in the study without their knowledge and consent e.g. covert observation of people in non-public places?.]

Does the study involve any of the following? [d. Deliberately misleading participants in any way?]

Does the study involve any of the following? [e. Discussion of sensitive topics e.g. sexual activity or drug use?]

Does the study involve any of the following? [f. The administration of drugs, placebos or other substances (e.g. food substances, vitamins) to the study participants.]

Does the study involve any of the following? [g. Invasive, intrusive or potentially harmful procedures of any kind?*]

Does the study involve any of the following? [h. Obtaining blood or tissue samples?*]

Does the study involve any of the following? [i. Pain or more than mild discomfort?]

Does the study involve any of the following? [j. Psychological stress, anxiety, harm or negative consequences beyond that encountered in normal life?]

Does the study involve any of the following? [k. Prolonged or repetitive testing i.e. more than 4 hours commitment or attendance on more than two occasions?]

Does the study involve any of the following? [l. Financial inducements (other than reasonable expenses and compensation for time)?]

Does the research involve the viewing, usage or transfer of Sensitive Personal Data as defined by the Data Protection Act 1998 or data governed by statute such as the Official Secrets Act 1989 / Terrorism Act 2006, commercial contract or by convention e.g. client confidentiality? (If you are unsure please tick YES and complete the sub-questions).

No [N]

Will the study cause direct or indirect damage to the environment or emissions outside permissible levels or be conducted in an Area of Special Scientific Interest or which is of cultural significance?

No [N]

Will the research be conducted outside of the European Economic Area (EEA) or will it involve international collaborators outside the EEA?
No [N]

Next Steps

Based on your responses your project has been categorised as (ethically) low risk and no further review is required before you start work. You will receive a formal approval email on submission of this form. Should your project change you may need to apply for new ethical approval.

Supporting Documentation

Please upload any documents (not uploaded elsewhere in the application) which you think are relevant to the consideration of your application.

filecount - Please upload any documents (not uploaded elsewhere in the application) which you think are relevant to the consideration of your application.
0

Thank you for completing the University's Ethical Review Form. Based on your answers the University is satisfied that your project has met its ethical expectations and grants its ethical approval. Please be aware that if you make any significant changes to your project then you should complete this form again as further review may be required. Confirmation of this decision will be emailed to you. Please complete the declaration to submit your application. Declaration I certify that: [the information contained within this application is accurate.]
Yes [Y]

Thank you for completing the University's Ethical Review Form. Based on your answers the University is satisfied that your project has met its ethical expectations and grants its ethical approval. Please be aware that if you make any significant changes to your project then you should complete this form again as further review may be required. Confirmation of this decision will be emailed to you. Please complete the declaration to submit your application. Declaration I certify that: [the research will be undertaken in line with all appropriate, University, legal and local standards and regulations.]
Yes [Y]

Thank you for completing the University's Ethical Review Form. Based on your answers the University is satisfied that your project has met its ethical expectations and grants its ethical approval. Please be aware that if you make any significant changes to your project then you should complete this form again as further review may be required. Confirmation of this decision will be emailed to you. Please complete the declaration to submit your application. Declaration I certify that: [I have attempted to identify the risks that may arise in conducting this research and acknowledge my obligation to (and rights of) any participants.]
Yes [Y]

Thank you for completing the University's Ethical Review Form. Based on your answers the University is satisfied that your project has met its ethical expectations and grants its ethical approval. Please be aware that if you make any significant changes to your project then you should complete this form again as further review may be required. Confirmation of this decision will be emailed to you. Please complete the declaration to submit your application.

Declaration

I certify that:

[no work will begin until all appropriate permissions are in place.]

Yes [Y]

Appendix C

Simulation Models

C.1 Challenge Model

$$C : \text{normsNonCompliant}, \text{compliant}, \text{nonCompliant}, \text{fearCompliant}$$

$$S : \text{wearingID}, \text{carryingID}, \text{noID}$$

$$Link : (\text{hall}, \text{off}_1)(\text{off}_1, \text{hall})(\text{hall}, \text{off}_2)(\text{off}_2, \text{hall})(\text{hall}, \text{sec})(\text{sec}, \text{hall})$$

```
agent: // We repeat this for many agents
        id:a_1;
        con:subjectiveNonCompliant;
        status:wearingID;
        observations:empty;
        loc:off_1;
end
action wearID:
        begin agent: //WearID
                con:fearCompliant;
                status:carryingID;
                newStatus:wearingID;
        end agent
end
action chall:
```

```
        begin agent: //challenger
                con:compliant;
                status:wearingID;
                loc:hall;
                obs:obsChall;
        end agent
        begin agent: //challengee
                status:carryingID;
                loc:hall;
                newStatus:wearingID;
        end agent
end
action noChall:
        begin agent: //!challenging.
                con:nonCompliant;
                id:id1|id2
                loc:hall;
                obs:obsNoChall;
                newLoc:off1
        end agent
        begin agent: //agent without ID.
                loc:hall;
                status:!wearingID
        end agent
end
observation obsChall:
        loc:hall;
        newObs:(a,chall);
end
change auth: //Auth
        con:nonCompliant;
        observe:(a1,chall) in a.obs;
        newCon:compliant;
```

```
end


property strongCompliance:
        began evalAgent:
                con:compliant|fearCompliant;
                status:wearingID
        end
end
property weakCompliance:
        began evalAgent:
                con:compliant|fearCompliant|
                normsNonCompliant;
        end
end
```

C.2 Challenge: Observation Intervention Model

We only provide the change in the model, not the complete model:

```
observation obsChall:
        loc:hall|off_1;
        newObs:(a,chall);
end
```

C.3 Challenge: Forced Challenge Model

We only provide the change in the model, not the complete model:

```
action forcedChall:
        begin agent: //challenger
                id: a1|a18|a21
                status:wearingID|carryingID;
                loc:hall;
                obs:obsChall;
                newStatus:wearingID
        end agent
        begin agent: //challengee
                status:carryingID;
                loc:hall;
                newStatus:wearingID;
        end agent
end
```