

Social-Influence-Based Control of Voluntary-Behaviour Adoption in Small Groups using Agent-Based Modelling

by

Keegan Aaron Fernandes

A thesis
presented to the University of Waterloo
in fulfilment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2024

© Keegan Aaron Fernandes 2024

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner:	Daniel Rivera Professor, School for Engineering of Matter, Transport, and Energy, Arizona State University
Supervisors:	Dan Davison Professor, Electrical and Computer Engineering, University of Waterloo David Wang Professor, Electrical and Computer Engineering, University of Waterloo
Internal Members:	George Freeman Professor, Electrical and Computer Engineering, University of Waterloo Rob Gorbet Professor, Knowledge Integration, University of Waterloo
Internal-External Member:	Carolyn MacGregor Professor, Systems Design Engineering, University of Waterloo

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

In this thesis, we take a systems control approach to explore various control influence techniques and investigate how effective they are at changing behaviour. We develop a psychology-based discrete-time behavioural change model of individuals in a group environment. The developed model incorporates work on individual behavioural change, namely the reasoned action approach, to represent how individual beliefs affect behaviour. Additionally, the model uses ideas from the study of social influence to account for social effects on individual beliefs such as conformity and internal biases. Feedback mechanisms are used to integrate the two models. Simulations of the developed model display characteristics of clustering and polarisation, consistent with empirical evidence. The impact of internal bias in driving this phenomenon is noted and developed into an external influence strategy. This strategy along with several others is evaluated using simulations of the model. The results of these simulations show that taking advantage of internal biases is an effective influence strategy. Additionally, results show that an influence strategy of spending time with a different individual in a group at each interaction, rotating to different individuals, can lead to more individuals performing a behaviour when compared to a “one to all” influence technique where an influencer tries to convince the whole group at once. The developed control strategies are deployed on a real-world test scenario and the behavioural change results are compared to those of simulation. Results show that the developed model produces behavioural outcomes whose trends are consistent with real-world results. These results provide evidence that the model can be used to predict the effects of different influence techniques and to investigate which influence strategies are effective at changing behaviour and which are not. The results also provide insight into how behaviours propagate through a group environment.

Acknowledgements

I would like to thank my supervisors, Professors Daniel Davison and David Wang, for their patience and incredible guidance. I would also like to thank my family and friends for their encouragement throughout my research journey.

Table of Contents

List of Figures	x
List of Tables	xii
1 Introduction	1
1.1 Research Goals	2
1.2 Thesis Outline	3
2 Background	5
2.1 Predicting Human Behaviour	5
2.1.1 Psychological Theories	5
2.1.2 Machine Learning Approaches	6
2.1.3 Computational Modelling	6
2.2 Computational Modelling Paradigms	7
2.2.1 Equation-Based Models	7
2.2.2 Agent-Based Models	7
2.3 Behavioural Change Theories	8
2.3.1 Social Cognitive Theory	9
2.3.2 Reasoned Action Approach	10
2.3.3 Summary	13
2.4 Social Influence and Internal Biases	14

2.4.1	Conformity	14
2.4.2	Internal Bias	15
3	Model Derivation	18
3.1	Reasoned Action Approach Model	18
3.1.1	Behaviour ($\mathbf{B}_i[\mathbf{k}]$) and Intention ($\mathbf{I}_i[\mathbf{k}]$)	19
3.1.2	Attitude Towards a Behaviour ($\mathbf{A}_i[\mathbf{k}]$)	20
3.1.3	Perceived Norm ($\mathbf{N}_i[\mathbf{k}]$)	21
3.1.4	Perceived Behavioural Control ($\mathbf{C}_i[\mathbf{k}]$)	22
3.1.5	Summary	23
3.2	Social Influence Model	23
3.2.1	Classes of Social Influence Models	23
3.2.2	Internal Bias	25
3.2.3	Social Influence Equations in the Model	27
3.2.4	Modeling Interventions as Social Influence	27
3.3	Behavioural Outcome and Feedback Dynamics	28
3.4	Model Overview and Limitations	30
3.4.1	Model Overview	30
3.4.2	Limitations	33
4	Model Verification and Validation	37
4.1	Single-Agent Model	37
4.1.1	Negative Behaviour Simulation	39
4.1.2	Positive Behaviour Simulation	41
4.1.3	Negative to Positive Behaviour Simulation	42
4.2	Two-Agent Model	43
4.2.1	Theoretical Results of a Simplified Multi-Agent System	44
4.2.2	Negative Belief Simulation	50

4.2.3	Positive Belief Simulation	51
4.2.4	Negative and Positive Belief Simulation	52
4.3	30-Agent Model	54
4.3.1	Mostly Negative Belief Simulation	54
4.3.2	Mixed Belief Simulation	55
4.4	Summary	56
5	Exploration of External Influence Control Techniques	57
5.1	Influence Techniques	58
5.1.1	One to All	58
5.1.2	Influencers	58
5.1.3	Rotating	59
5.1.4	Targeted	59
5.2	Theoretical Results	60
5.3	Simulations	63
5.3.1	“No-Intervention” Case	64
5.3.2	“One to All” Technique	65
5.3.3	“Influencer” Technique	68
5.3.4	“Rotating” Technique	68
5.3.5	“Targeted” Techniques	69
5.4	Summary	71
6	Experimental Findings	72
6.1	Simulation Results	72
6.1.1	“No Intervention” Case	74
6.1.2	“One to All” Influence	75
6.1.3	“Rotating” Influence	75
6.1.4	“Targeted” Influence	76

6.2	Experimental Findings	76
6.2.1	Procedure	76
6.2.2	Results	79
6.2.3	Statistical Analysis of Experimental Results	80
6.3	Summary	83
7	Conclusion	85
7.1	Summary	85
7.2	Contributions	86
7.3	Limitations	87
7.4	Preliminary Future Work	89
7.4.1	Result Feedback Variation	90
7.4.2	Social Structure Variation	92
7.4.3	Theoretical Analysis	94
7.5	Conclusions	94
	References	97
	APPENDICES	106
A	Simulation Plots	107

List of Figures

2.1	Schematic of the reasoned action approach model	11
3.1	Block diagram showing a simplified view of internal bias effects on behavioural belief interpretation	26
3.2	Block diagram for a two-person system	31
4.1	Block diagram for a one-person system	38
4.2	Simulation of the one-person system with negative behaviour	40
4.3	Simulation of one-person system with positive behaviour	42
4.4	Simulation of one-person system with behaviour transitioning from negative to positive	43
4.5	Simulation of two-person system with both individuals initially having negative beliefs	50
4.6	Simulation of two-person system with both individuals initially having positive beliefs	52
4.7	Simulation of two-person system with initially one individual having positive beliefs and one individual having negative beliefs	53
4.8	Polarisation of behaviours	54
4.9	Clustering of behaviours	56
5.1	Visual comparison of influence techniques	58
5.2	Violin plot comparing influence techniques	64
5.3	Sample behaviour for the “no-intervention” case	65

5.4	Sample behaviour for the “one to all” technique	66
5.5	Distribution comparison of varying bias levels	67
5.6	Sample behaviour for the “influencers” influence technique	68
5.7	Sample behaviour for the “rotating” technique	69
6.1	Violin plots comparing simulated influence techniques	74
6.2	Violin plots comparing experimental influence techniques	79
7.1	Simulation comparison of pulsed result feedback	91
7.2	Simulation comparison of negatively biased result feedback	92
7.3	Group divisions for 150 student classroom	93
7.4	Simulation of behaviour in a 150-student system	96
A.1	Simulation of one-person system with positive behaviour	108
A.2	Simulation of one-person system with behaviour transitioning from negative to positive	109
A.3	Simulation of two-person system with both individuals initially having negative beliefs	110
A.4	Simulation of two-person system with both individuals initially having positive beliefs	111
A.5	Simulation of two-person system with initially one individual having positive beliefs and one individual having negative beliefs	112
A.6	Simulation of 150-person system with negative initial beliefs on average . .	113
A.7	Simulation of 150-person system with mixed initial beliefs on average . . .	114
A.8	Simulation of 150-person system with a pulsed result feedback	115
A.9	Simulation of 150-person system with a biased result feedback	117

List of Tables

3.1	Important parameters and variables	36
4.1	Parameter values for the one-person model	39
6.1	Results for change in number of ideas from activity 1 to activity 2	82
6.2	Results for change in number of ideas from activity 1 to activity 3	83
7.1	Relative weights for student interactions	93

Chapter 1

Introduction

Predicting human behaviour is of great value to decision-makers at various levels. Policymakers and educators attempt to influence human behaviour in various ways, such as governments implementing public health campaigns to promote vaccination uptake [53], or schools adopting interventions to increase student engagement and academic performance [72]. Being able to predict how these policies affect human behaviour would allow decision-makers to effectively decide what policies are helpful and what policies could lead to unintended consequences. Some examples where investigating the effects of influence techniques could be helpful are:

Measles and Vaccine Hesitancy

Measles is a highly contagious disease that can lead to severe complications and death [24]. Measles had been all but eliminated in many Western countries, such as Canada and the U.S. thanks to the availability of an effective vaccine [21]. However, there has been a recent increase in the number of measles cases that are blamed on the decline of measles immunisation in children [69]. A leading cause of this reduction in immunisation is vaccine hesitancy due to the propagation of misinformation exacerbated by the effects of the COVID-19 pandemic [66].

Many conspiracies regarding the efficacy and safety of vaccines have been at the forefront of vaccine misinformation. Current theories for the cause of these conspiracies include a belief that science is untrustworthy, inconsistent feedback from authorities, and the fear of a lack of control, all of this amplified by echo chamber effects through social media [76]. Agencies have tried to dissuade these conspiracies by using information campaigns

to educate the public as well as by using controversial techniques such as trying to remove individuals that spread misinformation from social networks [46].

A tool to evaluate how to influence beliefs so that they can effectively reach large numbers of people about the benefits of vaccines would be valuable. Additionally, predicting whether actions, such as removing individuals from social networks, are beneficial or harmful is important. Some observers have suggested that such actions may cause those banned to unify and form a new more vocal misinformation group on alternate social media [76]. If the goal is to minimise the amount of misinformation being distributed, bans may be a poor solution and could exacerbate the problem.

Classroom Behaviour

Another example of decision-making is within a classroom. Instructors are constantly trying to find ways to improve student performance. One way they can do this is by encouraging good studying habits such as note-taking, scheduling, and self-regulated learning [54]. Instructors can use many strategies to try and encourage the use of these behaviours. They can encourage the class as a whole by providing information about the behaviour. Alternatively, they could spend time with individual students to try and encourage them to perform the behaviour [38]. Another option is that an instructor could take advantage of the class hierarchy and spend time encouraging the most influential students in the class with the hopes that this behaviour will propagate to the rest of the class. For example, an instructor could spend time with class representatives to teach them about useful behaviours and allow them to inform the rest of the class. It might be tempting to think that an instructor spending time with every individual in the class would yield better, more consistent results; however, an instructor has only a limited amount of time and resources, and may not be able to effectively help everyone.

Finding an effective strategy to increase the usage of a behaviour in a class can be difficult without a proper model to verify the effects of a decision. With a social behaviour model, simulations of different teaching scenarios could be carried out and the results used to guide teaching.

1.1 Research Goals

The research goal of this thesis is to investigate various intervention techniques and their effects on human behaviour change. We aim to explore scenarios involving small groups

of people, such as a classroom, where these people are being influenced to take up non-addictive and voluntary behaviours, such as spending more time brainstorming. With that in mind, a behavioural change computational model is developed using the well established reasoned action approach theory [33]. To account for social influence effects, present in the real world, theories of conformity and internal biases are also considered in the model. Most existing computational models based on the reasoned action approach [68, 77, 84] do not use interventions targeting an individual’s underlying beliefs or attitudes, instead focusing on increasing accessibility to resources. While greater access to resources can be beneficial, it is not always the case that people do not perform behaviour because of a lack of accessibility, but due to their underlying attitude towards a behaviour being poor. The model developed in this thesis uses the intervention ideas presented in the reasoned action approach of influencing an individual’s beliefs towards a behaviour.

Once the model has been developed and validated, based on existing psychology literature, it is used to simulate several different influence techniques. These simulations investigate how an external source of influence can use different techniques to more effectively convince a group of individuals to take up a target behaviour. Techniques include influencing everyone in the group (called “one-to-all”), influencing different individuals through the length of the simulation (called “rotating”), and even targeting specific individuals based on a given attribute (called “targeted”). The effectiveness of these attributes is evaluated and presented. Finally, to further validate the model, an empirical test is conducted. The model is used to simulate a replicable classroom scenario using different intervention techniques. These techniques are then applied in a real-world environment and the resulting change in behavioural uptake is compared to that in the simulation.

1.2 Thesis Outline

The chapters in this proposal are organised as follows:

- Chapter 2 provides background information on different modelling paradigms. Additionally, some behavioural change theories are reviewed, namely the reasoned action approach and social cognition theory. The social influence ideas of conformity and internal biases are also reviewed.
- Chapter 3 develops the behavioural change model. First, a mathematical model for the reasoned action approach is proposed. This is followed by integrating another model to account for social influence effects. These two models are then combined using feedback mechanisms to establish the completed model.

- In Chapter 4 the model is validated and verified by simulating basic zero input responses. Simulations are run for various scenarios involving 1, 2, and 30 individuals with the resulting simulations being compared to theoretical predictions.
- In Chapter 5 several different influence control techniques are explored and analysed for effectiveness. These techniques include a “one to all” case, an “influencer” case, a “rotating” case, and several different “targeted” cases.
- In Chapter 6 the model is further validated by comparing the results of simulated intervention techniques to those conducted in a real-world scenario.
- In Chapter 7 we provide a summary of the work done in this thesis as well as a list of contributions and limitations of this work. Avenues for future work and some preliminary results are also provided.

Chapter 2

Background

The prediction of human behaviour requires the development of a model. In this chapter, we review current methods of predicting human behaviour, and common modelling paradigms in the form of variable-based models and agent-based models. Additionally, we review important psychological phenomena needed as a basis for the behavioural change model. Namely, we explore behavioural change theories along with theories of social influence.

2.1 Predicting Human Behaviour

Predicting human behaviour is a complex problem and has been studied in various disciplines such as psychology, machine learning, and computational modelling. Research has resulted in several different approaches, each with unique attributes.

2.1.1 Psychological Theories

Psychological theories for human behavioural change attempt to explain why humans choose to perform certain behaviours and why those behaviours change. These models focus on different psychological factors that affect whether an individual chooses to perform a behaviour. Some focus on underlying human characteristics, such as the Big Five factors (openness, conscientiousness, extroversion, agreeableness, and neuroticism), to predict the likelihood that a person will perform certain behaviours [71]. Other theories focus on a person's intrinsic and extrinsic motivations to perform behaviours, such as with the health

action process approach [59]. Still others focus on the interaction of attitudes, beliefs and intentions to predict human behaviour [1]. These approaches provide a foundational basis for understanding human behaviour and are used to predict whether a person will perform a behaviour, as well as to identify how to try to convince a person to begin performing a behaviour [32, 58]. However, these models do not capture the dynamics of behaviour and how behaviour and its underlying determinants influence each other, whether through social means or the outcomes of their actions.

2.1.2 Machine Learning Approaches

Machine learning approaches use large swaths of data to predict whether a behaviour is likely or not. Machine learning methods try to identify features in a dataset that correlate to a specific outcome and then use these correlations to generalise predictions [22]. These methods can be useful in predicting behaviour [4] using existing individual data, and can sometimes be more useful than psychological methods where identifying individual characteristics can be difficult. However, to properly succeed, these methods require large amounts of appropriate data which do not always exist [65]. Additionally, machine learning methods can lack the nuance that psychological methods have in explaining why certain behaviours are performed and may not provide insight into potential interventions.

2.1.3 Computational Modelling

Computational modelling involves the application of mathematically defined systems to simulate and predict human behaviour. These models make simplifying assumptions about complex social systems to gain insight into how behaviours manifest and propagate [16]. These methods use psychological theories as their basis [68], generating an environment of different individuals in a system. These individuals each have different underlying variables that define them and their intention to perform a behaviour, usually based on psychology literature. These models then follow a set of rules where an individual’s behavioural outputs and underlying variables all influence each other. Additionally, these models take into account social interactions between individuals and the resulting influence between individuals [95]. Computational modelling is effective at analysing why individuals perform a behaviour since they are based on existing psychological work [80]. However, these simulations can be computationally intensive. This form of behaviour prediction is often used by policymakers to predict behaviours [51, 95]. We use computational models in this thesis due to their ease of implementation as well as their ability to capture the dynamic effects

of how behaviours evolve and analyse how this affects the system’s underlying variables, while also using psychology theory as its foundation.

2.2 Computational Modelling Paradigms

A model is an abstract representation of a real-world phenomenon that creates a simplified version of reality [80]. It is a tool that can be used to observe the consequences of assumptions about a system. There are two main modelling paradigms used for representing social systems. First, equation-based models involve a set of equations describing how a set of variables affects other variables. The second paradigm, agent-based models, involves individual agents and how the behaviour of these agents affects the system.

2.2.1 Equation-Based Models

Equation-based models (EBMs) use laws relating different parts of a system, defined by equations, that apply to the system as a whole [15]. EBMs typically use differential equations to mathematically describe how variables change over time. These models provide a high level of precision as well as an avenue to explore formal proofs of important properties [81]. However, if the modelled system increases in complexity analytical capability can be lost [80]. Additionally, a large drawback for EBMs is their inability to deal with heterogeneity. EBMs assume homogeneity in the system and tend to categorise the population of the system into a few states [73]. EBMs are often used in epidemiology to capture the macro-level dynamics of infectious disease outbreaks [44, 70, 89], often using a form of an SIR model in which the population is divided into susceptible individuals (S), infected individuals (I), and recovered individuals (R). The dynamics of these variables are dictated by differential equations. These models have been used to identify interventions to reduce disease spread as well as predict how diseases spread within a population.

2.2.2 Agent-Based Models

Agent-based models (ABMs) are a class of computational models where individual agents interact with one another and their environment. Agents’ individual variables can change based on rules given to them, the interactions of which can lead to unexpected and complex emergent behaviour [40]. ABMs allow for greater heterogeneity by allowing individuals in the model to have any combination of states desired and by allowing individual changes in

state, and their effects, to be observed. ABMs also allow for possible insight into complex phenomena by enabling direct observation of individual interactions [80]. However, ABM simulations entail computational costs as simulating a larger number of individuals leads to an exponential increase in the number of variables required. Additionally, ABMs have a reduced chance of establishing any formal proofs due to their emergent complexity [81]. Agent-based models are often used in social systems including the healthcare field. Agents usually model people, with a set number of traits, who make decisions based on their interactions with others as well as their environment. For example, some studies focus on how individuals weigh the monetary costs of driving and transit versus their desire to walk [51, 95]. Other studies look at how social pressure causes individuals to eat healthier [67] or consume less alcohol [34]. These models generally have a component describing how agents decide to perform a given behaviour and a means by which social influences change an agent's behaviour.

Due to the ability of ABMs to reveal insight into the underlying causes of resulting behaviour, and the ease at which it can generate heterogeneous agents, The ABM method is the method used for the developed model.

2.3 Behavioural Change Theories

To properly predict human behaviour, an understanding of what influences humans to perform certain behaviours is required. Behavioural Change Theories provide a psychological backing to predicting behaviour as well as a means of influencing change in behaviour. A core component of a large number of behaviour theories is the notion of self-efficacy, proposed by Albert Bandura [10]. Self-efficacy encapsulates the idea of one's own perceived ability to succeed in a situation. Bandura further developed his idea of self-efficacy into the social cognitive theory (SCT) [8]. SCT works on the idea that new behaviours are acquired by being exposed to external factors, especially through observing and imitating others. Another prominent behavioural theory is the reasoned action approach (RAA), developed by Martin Fishbein and Icek Ajzen [33]. The RAA is a framework that states that behavioural intention develops from a person's beliefs and perceived social norms. Most behavioural change models, including those such as the health action process approach or the Fogg behaviour model [36, 59], that rely on motivations or other characteristics to explain behaviour, can be described by using the methods described in SCT or RAA.

2.3.1 Social Cognitive Theory

Albert Bandura developed the idea of social learning theory, later renamed to the social cognitive theory (SCT), in 1969 [10]. In this theory, he proposed that behaviour is not determined only by internal forces, but by environmental factors as well. More specifically, SCT suggests that knowledge is acquired by observing others and the repercussions of their behaviours. The key ideas of this theory are that of “modelling”, outcome expectation, and self-efficacy; all of which interact with each other to produce a learned behaviour [56, 79].

Modelling

Modelling is the idea that people acquire knowledge by observing “models”, be that through interpersonal interaction or media sources. To demonstrate this idea Bandura conducted a series of experiments involving a Bobo doll [11, 12]. In these experiments, children were exposed to people exhibiting either aggressive or non-aggressive behaviour towards some toys. The children were then left in a room with the toys and a Bobo doll. The study found that children who were exposed to the aggressive “model” acted more aggressively towards the doll when compared to the other children. Bandura repeated the experiment; however, this time the children were exposed to videos of aggression and violence instead of a live person. This experiment produced similar results, with those that observed violent videos behaving more aggressively towards the Bobo doll. This further showed that “modelling” can occur with live, symbolic, or verbal models.

Outcome Expectations

Outcome expectations are the perceived consequences of one’s actions. The expectations are broken down into three main categories: physical, social, and self-evaluative [56]. Physical outcome expectation is the anticipation of what will be experienced after performing a behaviour, be it pleasant or unpleasant. This encapsulates both short and long-term effects. For example, a person who quits smoking will have the positive consequence of reduced coughing. However, they will also have the negative consequence of increased muscle tension. Social outcome expectation is the perceived social response given a performed behaviour. Smokers could expect a positive reaction from family who would support them quitting. They could also expect a negative reaction from friends who continue to smoke. Finally, self-evaluative outcome expectation is the anticipation of feelings, such as being ashamed or satisfied. This pertains to one’s internal beliefs about their self and the standards they set for themselves. Outcome expectations are heavily influenced by observed

consequences seen by other “models”. If a “model” receives positive or negative repercussions for a given behaviour, this can feed into a person’s expectations for performing that behaviour.

Self-Efficacy

Self-efficacy is one’s belief about their capability to successfully complete a task [9]. Bandura suggests that self-efficacy determines whether an individual intends to perform a behaviour, how much they are willing to try to achieve the behaviour, and how long they are willing to try [10]. People with low self-efficacy tend to be pessimistic about their accomplishments and what they can achieve. This in turn affects their motivation to perform a task. However, self-efficacy beliefs can be changed via three main means. First, accomplishments or mastery can greatly improve one’s perceived abilities. Succeeding at a task boosts one’s confidence in being able to complete similar future tasks and so improves self-efficacy. Secondly, vicarious experiences have been shown to improve self-efficacy. By observing a “model” person, who is similar to the individual, accomplish a task, the individual compares their abilities to that of the model and so improves their belief that they can accomplish the task. Lastly, self-efficacy can be boosted via persuasion. If a peer or superior reinforces the idea that the individual can accomplish the task, this improves their beliefs as well as their self-doubt.

SCT suggests that individuals set goals (intentions) for themselves which then manifest into behaviour. These goals are firstly influenced by outcome expectations. An individual decides whether a behaviour is worth doing based on the expected result of that behaviour. If the behaviour proves to be non-beneficial people would not set up the intention of doing it. Once an intention has been set a person is likely to then act on it and execute it. This is predicted by self-efficacy. Individuals with high self-efficacy are more likely to carry out an intention because of their higher level of confidence in succeeding at the behaviour.

2.3.2 Reasoned Action Approach

The reasoned action approach is a theory for predicting human behaviour developed by Martin Fishbein and Icek Ajzen [33]. It is an extension of their earlier work on the theory of planned behaviour [1] and the theory of reasoned action [31]. The reasoned action approach establishes that, given an identified behaviour, a person will choose to perform said behaviour following implicit consideration of underlying information and beliefs. These beliefs fall into three categories: behavioural beliefs, injunctive and descriptive normative

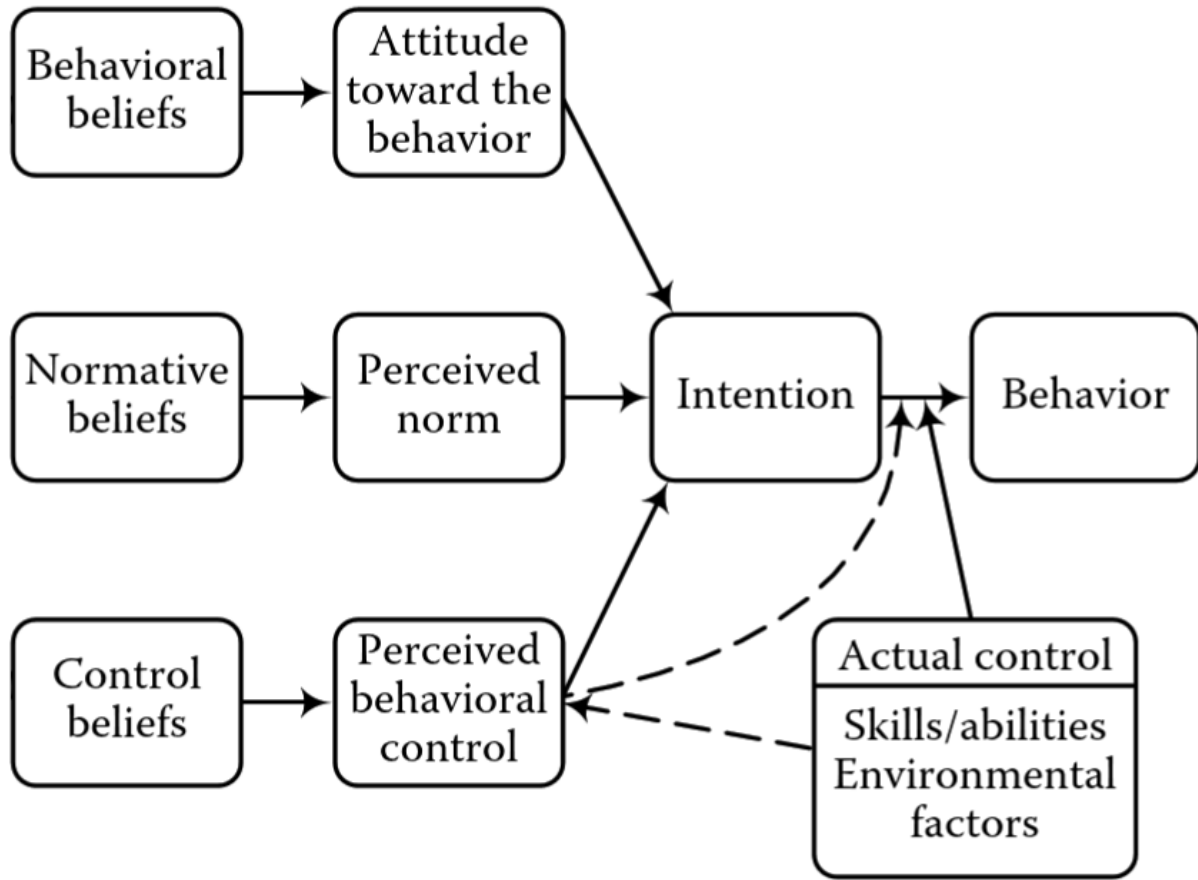


Figure 2.1: Schematic of the reasoned action approach model [33]

beliefs, and control beliefs. Each of these beliefs respectively determines a person’s attitude towards performing the behaviour, a perceived norm towards the behaviour, and a perceived behavioural control concerning the behaviour. The theory then states that a person’s attitudes, perceived norms, and perceived behavioural control guide their intention to perform the underlying behaviour. Whether a person performs a behaviour or not is directly related to their intent to do so as well as their actual control in being able to do so. RAA has been criticised for being too rational. The authors of the theory, however, argue that this is a misunderstanding of the theory and that the name only suggests that people’s behaviour is a rational result of their underlying beliefs, which are not necessarily rational [33]. A schematic of the reasoned action approach can be found in Figure 2.1, the details of which are explained below.

Behaviour and Intention

The first step in predicting human social behaviour is to define the behaviour of interest. This behaviour is an action that an individual may or may not perform, for example, whether or not a person chooses to exercise. Whether a person performs a behaviour is determined by a conjunction of an individual's intention and their actual control of the situation as seen in Figure 2.1. Intention represents an individual's willingness to perform a behaviour. The stronger the intention, the more likely it is that the behaviour will be carried out. However, it is possible that a lack of required skills and abilities, or environmental constraints, can prevent people from acting on their intentions. For example, a person, who is fully intent on exercising may be unable to do so due to a lack of resources, such as access to a gym or exercise equipment. These environmental constraints are called actual control in the RAA. The reasoned action approach states that behavioural intention is formed by a combination of an individual's attitudes, perceived norms, and perceived behavioural control towards the behaviour.

Attitude Towards a Behaviour

A person's attitude towards a behaviour is their evaluation as to whether performing said behaviour leads to positive or negative outcomes, i.e., if the behaviour is perceived to have a positive result, the attitude towards the behaviour will be favourable. This is similar to the idea of outcome expectation in the SCT. The reasoned action approach states that these attitudes can be further modelled by using an expectancy-value model [17, 25, 30, 52]. The expectancy-value model states that an individual's attitudes are determined by salient behavioural beliefs that are highly correlated with the behaviour and its results (Figure 2.1). An example of behavioural beliefs in the context of physical fitness is that exercise will lead to feeling healthy.

Perceived Norm

The next factor in determining an individual's intention to perform a behaviour is their perceived norm concerning the behaviour. Perceived norm is the perceived social pressure to engage or not to engage in the behaviour. Perceived norms are categorised as injunctive norms and descriptive norms [20]. Injunctive norms are the perceptions of what others think one should do concerning the behaviour. Descriptive norms, on the other hand, refer to the perception that others are performing the behaviour in question themselves. Like in the case of attitude, these norms can be further modelled using their respective injunctive

and normative beliefs, as seen in Figure 2.1. In the context of exercise, an example of an injunctive normative belief is a person's belief that their doctor or friends want them to exercise, while an example of a descriptive normative belief is the belief that their friends are themselves exercising.

Perceived Behavioural Control

The last factor determining behavioural intention is an individual's perceived behavioural control. Perceived behavioural control is the degree to which a person believes that they are capable of performing a behaviour as well as the degree to which they have control over its performance. Perceived behavioural control is very similar to Bandura's concept of self-efficacy [10] as described in his social cognitive theory. Perceived behavioural control can further be modelled by a person's control beliefs, as seen in Figure 2.1. These beliefs are similar to self-efficacy beliefs, and an example in the context of a project construction class is the belief that an individual is capable of brainstorming, or that they are fully in control as to whether the behaviour is performed.

The reasoned action approach states that behaviour is directly related to an individual's attitude towards the behaviour, perceived norms toward the behaviour, and perceived behavioural control concerning the behaviour. In general, a person is more likely to perform a behaviour if they have a favourable attitude towards it, they perceive a normative pressure to perform it, and they have a high level of perceived control over the behaviour.

2.3.3 Summary

SCT and the RAA both describe similar methods for explaining behaviour and their causes. They state that behaviour is made up of factors involving social norms, outcome expectations of a behaviour, and one's self-efficacy. SCT emphasises that behaviours are acquired by observing a model referent, while RAA suggests that changing behaviour can be accomplished by changing underlying beliefs. The RAA is the method employed in this thesis because the RAA's method for interventions is better described by the effects of social influence, which is the primary means being explored in this thesis. Additionally, the RAA has some established mathematical framework upon which a model can be built [33].

2.4 Social Influence and Internal Biases

Social influence is an aspect of human interaction whereby individuals modify their opinions, beliefs, attitudes, and behaviour towards those they are close to and interact with [74]. The reasons behind these influences vary from being persuaded [61] to a social pressure to conform with social norms [6, 29, 92]. The means of social influence is studied in the field of social psychology, where one of the key ideas is that of conformity [19]. Conformity is the idea that people want to have similar beliefs. This idea was expanded by Kelman into three main types [45]. Additionally, ideas such as an internal bias try to explain why some people tend to have polarised or different opinions [63].

2.4.1 Conformity

Conformity is the phenomenon of people changing their behaviour or beliefs to match those around them [19]. It is a social effect that arises from a norm being established by a group of individuals. People tend to conform to this norm rather than their own desires because it is easier to do so and are sometimes persuaded to do so via social pressures [19]. A key experiment validating the idea was conducted by Asch in 1951 [6, 7]. In this study, a group of people were shown a series of lines with varying lengths and were asked to find the line which matched the length of a reference line shown. In each group all participants, but one, were actors. These actors intentionally provided incorrect responses to gauge how this would affect the participants' responses. The results of the experiment suggested that individuals were highly likely to conform to the group and provide an incorrect response, even when the correct response was obvious. The effectiveness of this conformity was determined by how obvious the answer was as well as how many others in the group provided an incorrect response. Conformity was further expanded into three main types by Kelman [45]. These types are compliance, identification, and internalisation. Of these, compliance and internalisation have received greater attention in social psychology research. Identification, which is an individual's desire to conform to a person that they respect, leading them to want to be like that individual, has been studied less thoroughly.

Compliance

Compliance, also called normative influence, occurs when a person conforms to be liked or accepted by a group [64]. People change their beliefs to fit in due to a desire for positive relationships and to avoid the negative repercussions of deviating from the group [78]. The

beliefs and behaviours that a group exhibits are called social norms. A person undergoing compliance conformity will begin to display these beliefs or behaviours, though they may not accept them internally. The degree to which an individual is influenced to accept a belief can be determined by social impact theory [60]. Psychologist Latané suggests that the amount of influence that a group has on an individual is determined by the group's strength, immediacy, and size [50]. A group's strength is the set of factors that make an individual from the group influential, such as charisma, wealth, etc. Immediacy is the recency with which the group has demonstrated a belief or behaviour. Finally, the impact of a group is determined by the number of sources of influence in the group.

Internalisation

Internalisation, also called informational influence, is the acceptance of a belief or behaviour from a credible source [88]. This occurs when an individual looks for information from a member of their group, that they accept as correct. This leads to people both publicly and privately accepting the belief. When individuals are confronted with a situation with which they are unfamiliar, they tend to look to people with experience, or who they believe will help them best. This leads to experts being looked at for information [39]. In the context of exercising, individuals are likely to look to their doctors for information.

2.4.2 Internal Bias

Social influence work also incorporates the idea that not all individuals conform to each other. There are situations in which individuals have differing opinions and do not necessarily come to a consensus. In an extreme case, individuals get polarised and have diametrically opposing views. Internal biases are a possible explanation for this phenomenon. Internal bias can be the tendency for people to find evidence that validates their existing beliefs [63]. This can either be in the form of selecting only supportive information, or by ignoring contradictory information. Common types of internal bias are a biased search for information, a biased interpretation, and a biased memory.

Biased Search for Information

Biased search for information is the behaviour displayed by individuals who search for evidence that supports their cause while also avoiding evidence that refutes it [47]. People will phrase questions in ways that are more likely to lead to a desired answer. An example

of this, in a health context, would be an individual who has a negative view towards exercising asking “*is exercising bad for you?*,” “*does exercising lead to injuries?*,” etc. On the other hand, someone who views this behaviour positively would phrase these questions accordingly, i.e., “*is exercising good for you?*,” “*does exercising reduce the likelihood of injuries?*.” The phrasing of these questions invariably leads to biased responses [48].

Biased Interpretation

This bias leads individuals, who are presented with the same information, to interpret the information differently based on their existing beliefs, in a way that favours their view [87]. Evidence for this bias was presented by an experiment conducted at Stanford University [55]. In this study, a group of participants were divided into two halves. The first half was made up of people who were in favour of capital punishment and the others were against it. The two groups were then given an identical series of fictional studies regarding the topic. Each group stated that the studies strengthened their existing beliefs and rated supportive studies as better than contradictory ones.

Biased Memory

Finally, biased memory recall is the effect of remembering evidence that supports an individual’s expectations. Some believe that this is due to the effect of being able to more easily store information that matches expectations. Others suggest that surprising information is more memorable. Both theories have been shown to be valid [86]. An example of this bias is presented in a study where participants were asked to read a profile of a woman [82]. This profile contained information on both introverted and extroverted behaviours. The participants were later asked to recall evidence of introversion and extroversion. One group was told that the woman was applying to be a librarian, and this group found more evidence of introversion. The other group was told the woman was applying for a real estate position, and they found more evidence of extroversion.

Internal bias leads individuals, who are presented with the same information, to interpret the information differently based on their existing beliefs, favouring their view [87]. For example, there could be two individuals, Alice and Bob, each who believe that brainstorming is strongly effective and strongly ineffective, respectively. If Alice and Bob both have large biases they may completely misinterpret each other’s opinions about brainstorming. Alice will interpret Bob’s negativity towards brainstorming as more positive and could believe Bob thinks brainstorming is only somewhat ineffective. Similarly, Bob will

interpret Alice's attitude as more negative and could believe Alice thinks brainstorming is only somewhat effective. These internal biases can lead to significantly differing opinions with perceived consensus.

Chapter 3

Model Derivation

In Section 3.1 we develop a mathematical model of individual behaviours based on the reasoned action approach (RAA). We also model the effects of social interaction and how it pertains to the RAA. In Section 3.2 we develop a mathematical model of social influence. The important role of feedback is discussed in Section 3.3. The complete model is presented in Section 3.4.

3.1 Reasoned Action Approach Model

When modelling the effects of behavioural change, Section 2.3 shows that both the SCT and the RAA provide very similar approaches. Both take into account social interactions and their effect on behaviour, both take into account the effects of a behaviour's expected result and its effects on beliefs, and both consider a self-efficacy-like component as key to predicting behaviour. However, RAA has a predefined framework that more easily lends itself to a mathematical model. The main ideas of the theory, shown in Figure 2.1, revolve around the concept that behaviour is influenced by a person's intention to perform a behaviour, which is itself informed by their attitudes, perceived norms, and perceived behavioural control. These psychological constructs are in turn influenced by a set of respective beliefs and their associated value to the individual. This section seeks to put these ideas into a mathematical framework using foundational beliefs to predict an individual's behaviour. A method for measuring these states is also included following the approach of Fishbein and Ajzen [33]. The model developed is a discrete-time dynamical system where each step k is a point by which all members in a group have had an opportunity to

interact. For example, in a classroom setting, k would represent each day when students and teachers interact in a class.

3.1.1 Behaviour ($B_i[k]$) and Intention ($I_i[k]$)

In Section 2.3.2, the RAA states that we must first define the behaviour of interest. An example of this behaviour could be “*exercising*”. Whether an individual, i , performs a behaviour, $B_i[k] \in [-1, 1]$, is stated to be influenced by the interaction between a person’s intention to perform the behaviour, $I_i[k] \in [-1, 1]$, and their actual control of the situation. A person may not perform a behaviour, even if they do intend on doing so, because they may not have actual control of their environment, i.e., being unable to exercise because of inadequate gym availability. For this thesis, we consider that all behaviours being evaluated are actionable, and so assume that all persons have full actual control of their behaviour. This assumption implies we can ignore the effects of the actual control block shown in Figure 2.1. This further allows us to state that one’s behavioural intent is equivalent to one’s behaviour, simply stated: $B_i[k] = I_i[k]$. Here, when $B_i[k] = 1$, individual i intends to perform the behaviour and will perform the behaviour, whereas if $B_i[k] = -1$ the behaviour is not performed, $B_i[k] = 0$ implies that the individual only occasionally performs the behaviour. An alternative explanation for $B_i[k]$ is the probability that an individual performs the behaviour. In this paradigm, when $B_i[k] = 1$, individual i performs the behaviour with 100% certainty, whereas if $B_i[k] = -1$ the behaviour is not performed, $B_i[k] = 0$ implies that the individual performs the behaviour 50% of the time. All values in between represent the degree to which, or the frequency with which, the behaviour is carried out. Which paradigm is chosen depends on the way the behaviour is measured. An individual’s behaviour can be measured using a series of statements, that are ranked for level of accuracy, known as a Likert scale. Examples of statements that can be used to directly assess behavioural intention include “*I will spend time exercising 3 times a week (never-occasionally-always)*”, “*I plan on exercising 3 times a week (never-occasionally-always)*.”, and “*how likely are you to exercise 3 times a week (not likely-somewhat likely-definitely likely)*.”

The RAA states that behavioural intention is formed by how an individual’s attitude towards a behaviour, perceived norms towards a behaviour, and perceived behavioural control towards a behaviour lead to a behaviour being performed. The behaviour is calculated using a simple linear regression analysis between a person’s measured attitude, perceived norms, perceived behavioural control, and their measured behaviour leading to a function

of the following form [33],

$$\begin{aligned} B_i[k] &= \lambda_i A_i[k] + \beta_i N_i[k] + \gamma_i C_i[k], \\ \lambda_i + \beta_i + \gamma_i &= 1 \text{ and } \lambda_i, \beta_i, \gamma_i > 0, \end{aligned} \tag{3.1}$$

where $A_i[k]$ represents the attitude, $N_i[k]$ represents the perceived norms, and $C_i[k]$ represents the perceived behavioural control of agent i , which are discussed further in the following sections. The gains $\lambda_i, \beta_i, \gamma_i$ weigh how important these components are to determining the agent's behaviour. The more positive an agent's attitude, the more they believe that their peers approve of the behaviour, and the more the agent believes they are capable of performing the behaviour, the more likely they are to perform the behaviour. A limitation of the RAA is that it may be unable to accurately predict and influence addictive and compulsive behaviours. Therefore we assume that all behaviours being studied are voluntary and under an individual's full control.

3.1.2 Attitude Towards a Behaviour ($A_i[k]$)

The first component of intention that is considered is a person's attitude towards a behaviour. Section 2.3.2 states that those individuals who expect a positive result from a behaviour will have a positive attitude towards that behaviour, and vice versa. Attitude towards a behaviour can be directly measured using a Likert scale with survey questions such as "*My exercising is (good-bad)*" or "*My exercising is (pleasant-unpleasant)*." The RAA states that these attitudes can be further modelled by using an expectancy-value model [17, 25, 30, 52] where behavioural belief influences attitude.

The expectancy-value model states that an individual's attitudes are determined by the strength of a list of beliefs $b_{i,l}[k] \in [-1, 1]$, and their associated evaluation as to how important ($e_{i,l} \in [0, 1]$) the belief is to determining the end attitude, where l refers to a specific belief about the result of the behaviour at hand. When $b_{i,l}[k] = 1$ the individual, i , thinks that there is a 100% chance that the behaviour results in the outcome defined by the belief, when $b_{i,l}[k] = -1$ the individual thinks there is a 0% chance that the behaviour results in the outcome defined by the belief, and when $b_{i,l}[k] = 0$ the individual thinks there is a 50% chance that the behaviour results in the outcome defined by the belief. The value of $e_{i,l}$ determines how much of a contribution the associated behavioural belief, l , has to the individual's attitude towards the behaviour. Additionally, the relative importance of each belief, $e_{i,l}$, determines how much each behavioural belief affects attitude. When $e_{i,l} = 1$, the individual associates the outcome of the belief $b_{i,l}$ as being important for determining their attitude towards the behaviour and when $e_{i,l} = 0$, the belief $b_{i,l}$ has no impact towards

the individual's attitude towards the behaviour. We assume that $e_{i,l}$ remains constant for all individuals for simplicity.

The specific beliefs that are highly associated with attitude are called salient beliefs and can be identified by asking individuals to list the advantages and disadvantages of performing the behaviour in question [33]. Examples of salient behavioural beliefs, with regards to health and fitness, are that this exercise leads to better health, as well as that it does not lead to injury. Strength in a belief and its associated evaluation can be gauged using a Likert survey, with questions like “*exercising will help me feel healthier (unlikely-likely)*,” “*feeling healthy is ... (good-bad)*,” or “*exercising will not lead to injury (unlikely-likely)*,” “*getting injured is... (good-bad)*.” Finally, we can combine these behavioural beliefs and evaluations to generate an attitude using the sum of their products [25]:

$$A_i[k] = \sum_{l=1}^{L^b} b_{i,l}[k]e_{i,l}, \quad \sum_{l=1}^{L^b} e_{i,l} = 1 \quad (3.2)$$

where L^b is the number of salient behavioural beliefs.

3.1.3 Perceived Norm ($N_i[k]$)

Intention is also influenced by a person's perceived norms concerning the behaviour. Perceived norm is the perceived social pressure to engage or to not engage in the behaviour. There are two kinds of perceived norms, an injunctive norm and a descriptive norm [20] as stated in Section 2.3.2. Injunctive norms are the perceptions of what others think one should do concerning the behaviour and are measured using Likert style questions like “*Most people who are important to me think I should exercise (true-false)*.” Descriptive norms on the other hand refer to the perception that others are performing the behaviour in question themselves and are measured using questions such as “*Most people like me would exercise (agree-disagree)*.” Like behavioural attitude, perceived norm, $N_i[k]$, can be further modelled as follows [1],

$$N_i[k] = \frac{1}{2} \sum_{l=1}^{L^I} n_{i,l}^I[k]m_{i,l}^I + \frac{1}{2} \sum_{l=1}^{L^D} n_{i,l}^D[k]m_{i,l}^D, \quad (3.3)$$

$$\sum_{l=1}^{L^I} m_{i,l}^I = 1, \text{ and } \sum_{l=1}^{L^D} m_{i,l}^D = 1.$$

The $n_{i,l}^I$ and $n_{i,l}^D$ are the injunctive and descriptive normative beliefs respectfully, and $m_{i,l}^I$ and $m_{i,l}^D$ are the injunctive and descriptive motivations to comply, l is the individual's

specific belief referents (be it their parents, friends etc.) and L^I and L^D represent the total number of injunctive and descriptive normative beliefs respectively. For simplicity, we assume that $m_{i,l}^I$, and $m_{i,l}^D$ remain constant for all individuals. Again, like in the case of attitude, salient normative belief referents ($n_{i,l} \in [-1, 1]$) can be identified using lines of questioning such as “Please list individuals or groups that would approve/disapprove of you exercising” or “Please list all of the individuals who would/would not exercise.” Each of these normative belief referents is then measured using a Likert scale, i.e. “My family/friends think I should exercise (probable-improbable)” or “My friends would exercise (probable-improbable).” These normative beliefs are then scaled by an individual’s motivation to comply, $m_{i,l} \in [0, 1]$, which can be measured with questions such as “When it comes to health, I want to do what my family thinks I should do (agree-disagree)” or “When it comes to health, how much do you want to be like your friends (very much-not at all).”

3.1.4 Perceived Behavioural Control ($C_i[k]$)

Perceived behavioural control is the final component that contributes to behavioural intention. Perceived behavioural control is the degree to which a person believes that they are capable of performing a behaviour as well as the degree to which they have control over its performance. However, using the assumption that an individual has full control of performing the behaviour, we can say that perceived behavioural control, $C_i[k] \in [-1, 1]$, is only a function of one’s assumed capabilities. Perceived behavioural control can be directly measured using a Likert scale, i.e. “If I really wanted to, I could exercise (likely-unlikely),” or “I am confident that I can exercise (true-false),” or “Me exercising is completely up to me (agree-disagree).” Furthermore, the value of one’s perceived behavioural control can be modelled similarly as attitude and perceived norms [1] as seen in Figure 2.1,

$$C_i[k] = \sum_{l=1}^{L^c} c_{i,l}[k] p_{i,l}, \text{ with } \sum_{l=1}^{L^c} p_{i,l} = 1. \quad (3.4)$$

Here $c_{i,l} \in [-1, 1]$ is the belief that a control factor, l , will be present and $p_{i,l} \in [0, 1]$ is the power the factor has to facilitate the performance of the behaviour. The variable L^c represents the total number of control beliefs for the individual. Again, We assume that $p_{i,l}$ remains constant for all individuals for simplicity. Similar to identifying salient behavioural and normative beliefs, salient control beliefs can be identified by asking individuals to list the factors that would enable/impede them from performing a behaviour. These responses would enable us to identify factors whose control beliefs and power would need to be measured. Control belief measures would be measured via Likert style questions such as

“I will get enough sleep (likely-unlikely).” Likewise, the power of the control factor can be measured using questions such as *“Getting enough sleep would enable me to exercise (agree-disagree).”*

3.1.5 Summary

In summary, the RAA states that behaviour is directly related to an individual’s attitude towards the behaviour, perceived norms toward the behaviour, and perceived behavioural control concerning the behaviour. Each of these variables can be constructed using a series of beliefs and their associated weighting factors. In general, the more favourable a person’s attitude towards a behaviour, the more they perceive a normative pressure to perform a behaviour, and the higher their perceived control of the behaviour the more likely they are to actually perform the behaviour. Each of these attribute contributions is weighted to result in the final behavioural frequency and these weights vary among individuals and environments. Therefore, these weights (i.e. $e_{i,l}$, $m_{i,l}$, $p_{i,l}$, λ_i , β_i , and γ_i) will be treated as parameters that need to be identified.

3.2 Social Influence Model

Other aspects of the model that need to be considered are the dynamics of social influence and its effects on the core beliefs stated in the RAA.

3.2.1 Classes of Social Influence Models

Social influence is an aspect of human interaction whereby individuals modify their opinions, beliefs, attitudes, and behaviour towards those they are close to and interact with [74]. The reasons behind these influences are discussed in Section 2.4 and vary from a unifying force, described by conformity[6, 29, 92], to a repulsive force that can be explained by confirmation bias[61], to a social pressure to conform with social norms. There exist many means of modelling social influence in the literature; however, Flache et al. [35] state that all these social influence theories follow three main classes of models: models of assimilative social influence, models with similarity-based influence, and models with repulsive influence.

Assimilative models assume that all individuals that are connected in some way have an influence on each other and can be represented by [35],

$$o_i[k+1] = o_i[k] + \mu \sum_j \omega_{ij}(o_j[k] - o_i[k]), \quad \sum_j \omega_{ij} = 1, \quad 0 < \mu < 1 \quad (3.5)$$

where $o_i[k] \in [-1, 1]$ is the opinion of individual i at interaction k , μ is the rate of opinion convergence, and ω_{ij} represent the impact that individual j has on i 's opinion. According to Flache, this form of social influence leads to the eventual consensus of opinion in all individuals in the network and is similar to the ideas of conformity discussed in Section 2.4.1, where μ represents the rate of convergence.

Similarity-based models follow a related form; however, the only individuals capable of influencing opinion, in this case, are those with sufficiently similar views [35]. This model is represented as follows,

$$o_i[k+1] = o_i[k] + \mu \sum_j f(o_i[k], o_j[k])(o_j[k] - o_i[k]), \quad 0 < \mu < 1$$

$$f(o_i[k], o_j[k]) = \begin{cases} \omega_{ij}, & |o_i[k] - o_j[k]| \leq \epsilon_i \\ 0, & \text{otherwise} \end{cases}, \quad \sum_j \omega_{ij} = 1 \quad (3.6)$$

where ϵ_i is the threshold at which the i th individual is willing to adapt to another individual's opinion. This model states that only if another individual has an opinion that is sufficiently close to the main person will the person be willing to consider their opinion. This idea is related to internal bias, discussed in Section 2.4.2. This model results in opinions converging to several clusters depending on the size of the threshold. The assimilative model is a special case of the similarity model with $\epsilon_i = 1$.

Finally, repulsive influence models incorporate the idea that if individuals are too dissimilar they can influence each other towards opposing views. This model takes on a modification of the similarity model [35],

$$o_i[k+1] = o_i[k] + \mu \sum_j f(o_i[k], o_j[k])(o_j[k] - o_i[k]), \quad 0 < \mu < 1$$

$$f(o_i[k], o_j[k]) = \omega_{i,j}(1 - 2|o_j[k] - o_i[k]|), \quad \sum_j \omega_{ij} = 1 \quad (3.7)$$

Here, the effects of someone with an opinion sufficiently different enough, ± 0.5 in this case, will lead to individuals diverging in opinion eventually resulting in fully bi-polarised views among the individuals, while individuals within the threshold will converge [35].

In general, a hybrid model approach, where all three of these models are combined, leads to a better approximation of human influence. A hybrid approach would allow for

several clusters to form, from the similarity-based model, along with a concentration of viewpoints at the centre, from the assimilative model, as well as a grouping of opinions at either end of the spectrum, from the repulsive model. This is shown in [35], where a combined model best described the political landscape of a large range of countries, with a large proportion having a centrist opinion with clusters of opinions forming on both the left and right end of the political spectrum. However, in the case of this thesis, only the similarity model is used, for simplicity, to explain conformity-like pressures; with the clustering and polarising effects captured using the internal bias discussed in Section 2.4.2.

The impact, ω_{ij} , of an agent j on agent i is an important factor in the social influence models. In the real world, one can expect the relative impact a person has on someone else evolves based on factors such as trust and importance. However, for simplicity, for this model, we assume that the weights remain static removing some dynamic effects that might arise from social identity theory or social impact theory. This assumption can be valid in short time-scale environments where a person’s relative trust has been established and opinions about others do not necessarily change, such as in a fitness class with a group of friends.

3.2.2 Internal Bias

In Section 2.4.2 we discussed how individuals do not solely conform their opinions with their surrounding group. This effect can be achieved by the assimilative and repulsive-based influence models described in Section 3.2.1; however, we use a different approach in the form of internal bias. Internal biases cause inconsistencies in perceived information compared to reality. Most research focuses on using a biased search for information, or a confirmation bias, where more weight is placed on information or opinions that agree with the individual and where opposing information is ignored [3, 18, 83, 93], akin to the assimilative and repulsive influence models described previously. As an alternative, we propose using the biased interpretation of information described in 2.4.2 to study its effects. Biased interpretation results in individuals skewing how they view information presented to them, making the information seem more aligned with their current beliefs. This internal bias is included as a non-linear component in the model by adding a signed offset when an agent interprets any information presented to them. This offset is added in the direction that supports the agent’s underlying belief; that is, it is positive if the agent’s belief is positive and vice versa. This leads to agents interpreting information as more aligned with their beliefs. This effect is considered whenever an agent tries to interpret

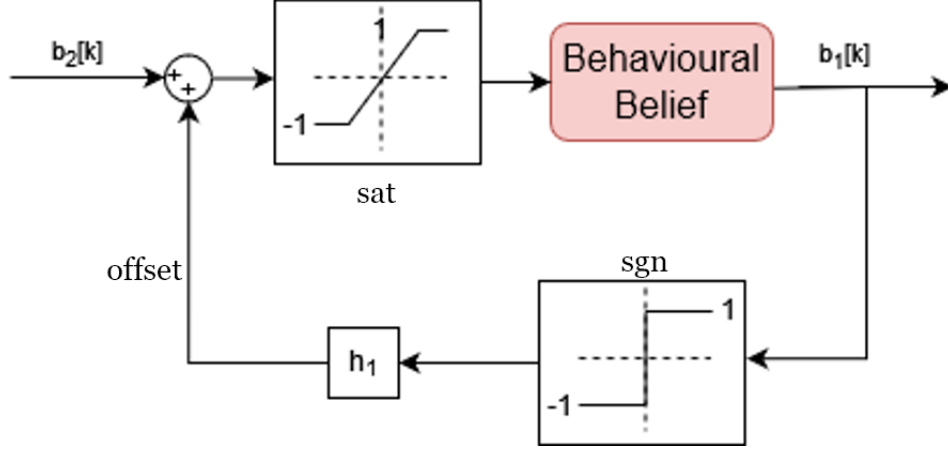


Figure 3.1: Block diagram showing a simplified view of internal bias effects on behavioural belief interpretation

another agent's beliefs. Equations pertaining to this bias are as follows:

$$\text{sat}(x) = \min(\max(x, -1), 1) \quad (3.8)$$

$$R_{i,l}[k]_{\text{bias}} = \text{sat}(R_{i,l}[k] + \text{sgn}(b_{i,l}[k])h_i), \quad (3.9)$$

$$B_i^N[k]_{\text{bias}} = \text{sat}(B_i[k] + \text{sgn}(n_{i,l}^D[k])h_i), \quad (3.10)$$

$$A_i[k]_{\text{bias}} = \text{sat}(A_{i,l}[k] + \text{sgn}(n_{i,l}^I[k])h_i), \quad (3.11)$$

$$B_i^C[k]_{\text{bias}} = \text{sat}(B_i[k] + \text{sgn}(c_{i,l}[k])h_i), \quad (3.12)$$

$$\hat{b}_{j,l}[k]_{\text{bias}} = \text{sat}(\hat{b}_{j,l}[k] + \text{sgn}(b_{i,l}[k])h_i), \quad (3.13)$$

$$\hat{c}_{j,l}[k]_{\text{bias}} = \text{sat}(\hat{c}_{j,l}[k] + \text{sgn}(\hat{c}_{i,l}[k])h_i). \quad (3.14)$$

Here a saturation function, $\text{sat}(x)$, ensures that, when the bias is added, the signals do not exceed the range $[-1, 1]$. Parameter h_i represents the strength of the internal bias for a given agent. $R_{i,l}$ is the outcome of performing the desired behaviour and how it affects behavioural belief l . The resulting variables are the biased interpretations agent i has on the information presented to them. These variables are used for the social influence effects described in Section 3.2.3 as well as in outcome feedback updates shown in Section 3.3. A simplified example block diagram showing how bias is implemented in (3.13) can be seen in Figure 3.1.

3.2.3 Social Influence Equations in the Model

With regard to the RAA, the effects of social influence are already partially included through an individual's perceived social norm. However, by taking advantage of the broader use of social influence we can directly account for changes in the core beliefs that are influenced by social pressures. One aspect of the RAA that can be influenced is an individual's behavioural beliefs. These beliefs are influenced by others in one's social circle. Additionally, an individual's control beliefs can be influenced by others' control beliefs. This idea comes from Bandura's self-efficacy theory [10] in which he states that self-efficacy is influenced by a person's performance outcome, other people's behaviour, as well as verbal persuasion from others. The effects of others on an individual's beliefs can be modelled using assimilative social influence. In our model, behavioural, normative and control beliefs are influenced by conformity pressures and are represented as

$$b_{i,l}[k] = \hat{b}_{i,l}[k] + \mu_i \sum_j \omega_{i,j} (\hat{b}_{j,l}[k]_{bias} - \hat{b}_{i,l}[k]), \quad (3.15)$$

$$n_{i,l}^D[k+1] = n_{i,l}^D[k] + \mu_i \sum_j \omega_{i,j} (\alpha B_j[k]_{bias} - n_{i,l}^D[k]), \quad (3.16)$$

$$n_{i,l}^I[k+1] = n_{i,l}^I[k] + \mu_i \sum_j \omega_{i,j} (\alpha A_j[k]_{bias} - n_{i,l}^I[k]), \quad (3.17)$$

$$c_{i,l}[k] = \hat{c}_{i,l}[k] + \mu_i \sum_j \omega_{i,j} (\hat{c}_{j,l}[k]_{bias} - \hat{c}_{i,l}[k]), \quad (3.18)$$

where μ_i is the suggestibility of agent i and dictates how readily and agents changes their opinion to match their peers, $\omega_{i,j}$ is the relative influence agent j has on agent i with $\sum_j \omega_{i,j} = 1$ and tells us how influential a given agent is. Parameter $\alpha \in [0, 1]$ is a gain representing how behaviours and attitudes affect normative and control beliefs. $\hat{b}, \hat{c}, \hat{b}_{bias}, \hat{c}_{bias}, \hat{B}_{bias}$, and $\hat{A}_{bias}$ represent the effects of internal bias, described in Section 3.2.2, and outcome feedback which is discussed further in Section 3.3. An overview of these components can be found in Section 3.4

3.2.4 Modeling Interventions as Social Influence

In the RAA, the means of changing people's behaviour are usually interventions designed to inform an individual about their respective beliefs [33]. In the model designed in this chapter, such interventions would appear in the form of additional sources of influence in the belief components. Therefore, our control inputs would appear as additional individuals in

our multi-agent system, also called control agents. This implies that the set of individuals that interact with a person would now include an additional source of influence that can sway the person’s beliefs. For example, in a health and fitness example, a doctor would be able to influence the patient with their own beliefs concerning exercising. By informing the patient of the benefits of exercising they may be able to sway the patient’s beliefs about the behaviour. A result of this means of influence is that it focuses on small-scale personal interactions with a control agent. Larger scale dynamics such as through advertising and social networks may require different modelled intervention approaches, and so, we assume that systems being studied involve smaller groups with personal connections to the control agent.

3.3 Behavioural Outcome and Feedback Dynamics

Another important component of the model is the inclusion of an outcome feedback mechanism. The RAA mentions that certain attributes such as behavioural beliefs, normative beliefs, and control beliefs are susceptible to feedback based on the performed behaviour [33]. We first look at behavioural beliefs. These beliefs are based on an expected outcome of a behaviour. If performing a behaviour leads to the outcome in question it should affect the belief that the outcome will occur. Therefore, we can state that the belief that an outcome can occur should converge to whether or not the outcome in question does occur. For example, if an individual expects to feel healthy due to exercise, and their health increases as a result of performing this behaviour, in that case, their belief that this outcome will occur should remain high. However, if exercising results in feeling less healthy, this should result in a reduction in their belief that exercise leads to feeling healthier. Additionally, if a behaviour is not being performed effectively ($B_i[k] \leq 0$), it is reasonable to expect that the individual will not associate any result to the behaviour, ignoring its effect. For example, if a person does not exercise, and they feel unhealthy, their belief that exercise leads to better health should not be affected. To model the behavioural belief outcome update step we first need to establish how behaviour translates into outcomes.

The outcome component of the model establishes the extent to which performing a behaviour truly leads to the preconceived outcomes, established by behavioural beliefs. Performing a behaviour may not fully result in the outcome stated by the individual’s belief $b_{i,l}$. The result function establishes the degree to which the belief outcome is achieved. This is represented by the following equation,

$$R_{i,l}[k] = f_l(B_i[k]) \quad (3.19)$$

where $R_{i,l}[k]$ is the outcome component for agent i associated with belief l . The function f_l may be determined by performing a regression analysis between a person's measured behaviour along with their measured results. For this thesis the following simplification is made to relate behaviours to outcomes:

$$f_l(B_i[k]) = \begin{cases} +B_i[k] & \text{if } l \text{ is a positive behavioural belief} \\ -B_i[k] & \text{if } l \text{ is a negative behavioural belief} \end{cases} \quad (3.20)$$

If an outcome positively correlates with a belief, for example, the belief that exercising *does* lead to improved health is true, then a positive linear relationship is assumed. On the other hand, if an outcome negatively correlates with a belief, for example exercising *does not* lead to soreness, which is incorrect, then a negative linear relationship is assumed. Therefore, in this example, if an individual chooses to exercise ($B_i[k] > 0$) then they will feel healthier and the more they do the behaviour the better their health will be. However, the more they exercise the more sore they will feel, contradicting their belief that exercise does not lead to soreness.

Now that the outcome component has been established we can continue looking at the behavioural belief outcome update step. The equation representing this is:

$$\hat{b}_{i,l}[k+1] = b_{i,l}[k] + \mathbb{1}_{[0,1]}(B_i[k])\mu_i(\alpha R_{i,l}[k]_{bias} - b_{i,l}[k]), \quad (3.21)$$

$$\mathbb{1}_{[0,1]}(B_i[k]) = \begin{cases} 1 & \text{if } B_i[k] > 0 \\ 0 & \text{otherwise } l \text{ is a negative behavioural belief} \end{cases} \quad (3.22)$$

Here, the agent's belief is updated to track the outcome of the behaviour they have performed. To account for the fact that outcome feedback is only accounted for when the agent performs the behaviour sufficiently an indicator function, $\mathbb{1}_{[0,1]}(B_i)$, is used. The indicator function ensures that an agent updates their behavioural beliefs only if they have sufficiently performed the behaviour. If an agent has not performed the behaviour they will not make any claims that the behaviour has resulted in better or worse performance.

Additionally, a person's perceived behavioural control is susceptible to a whole host of effects as described by Bandura [10] and stated previously. For instance, other people's behaviour, as well as a person's own behaviour, affect their self-efficacy. If other similar individuals have performed a behaviour it will lead to the main person improving their self-efficacy. This effect impacts an individual's perceived behavioural control, as this is an analogue of Bandura's self-efficacy. Therefore, we get the following,

$$\hat{c}_{i,l}[k+1] = c_{i,l}[k] + \mu_i(\alpha B_i^C[k]_{bias} - c_{i,l}[k]) \quad (3.23)$$

This equation implies that, if person i performs a behaviour, and if individuals j perform the behaviour, then person i is more likely to believe that they can perform the behaviour in future.

3.4 Model Overview and Limitations

With all the model components described in this chapter, we now present an overview of the model. A simplified, two-agent, model block diagram can be found in Figure 3.2. This figure shows the RAA components in blue and the social influence components in red. The signals from outside agent 2 are fed into the system for agent 1 and vice versa. This model can be extended to an arbitrary number of agents. A summary of important parameters and variables is shown in Table 3.1.

3.4.1 Model Overview

We will now describe, in brief, the steps that the model iterates through in simulation. First, a set of initial conditions for the system are established, for most simulation setups these initial conditions are randomly selected, uniformly, from the variable domains stated in Table 3.1.

Next, the agents' behavioural beliefs are updated using a two-step process, an outcome feedback step and a social influence step, both of which are included in the behavioural belief components in Figure 3.2. Agents make a biased evaluation of the outcomes of any previous behaviour and update their beliefs to match them as described in Section 3.3,

$$R_{i,l}[k]_{bias} = \text{sat}(R_{i,l}[k] + \text{sgn}(b_{i,l}[k])h_i),$$

$$\hat{b}_{i,l}[k+1] = b_{i,l}[k] + \mathbb{1}_{[0,1]}(B_i)\mu_i(\alpha R_{i,l}[k]_{bias} - b_{i,l}[k]),$$

The agent then updates their belief by conforming to their biased interpretation of others in their network,

$$\hat{b}_{i,j,l}[k]_{bias} = \text{sat}(\hat{b}_{j,l}[k] + \text{sgn}(b_{i,l}[k])h_i),$$

$$b_{i,l}[k] = \hat{b}_{i,l}[k] + \mu_i \sum_j \omega_{i,j}(\hat{b}_{i,j,l}[k]_{bias} - \hat{b}_{i,l}[k]).$$

This behavioural belief update step is followed by updates in the normative beliefs components seen in Figure 3.2. The descriptive norm beliefs are updated based on the

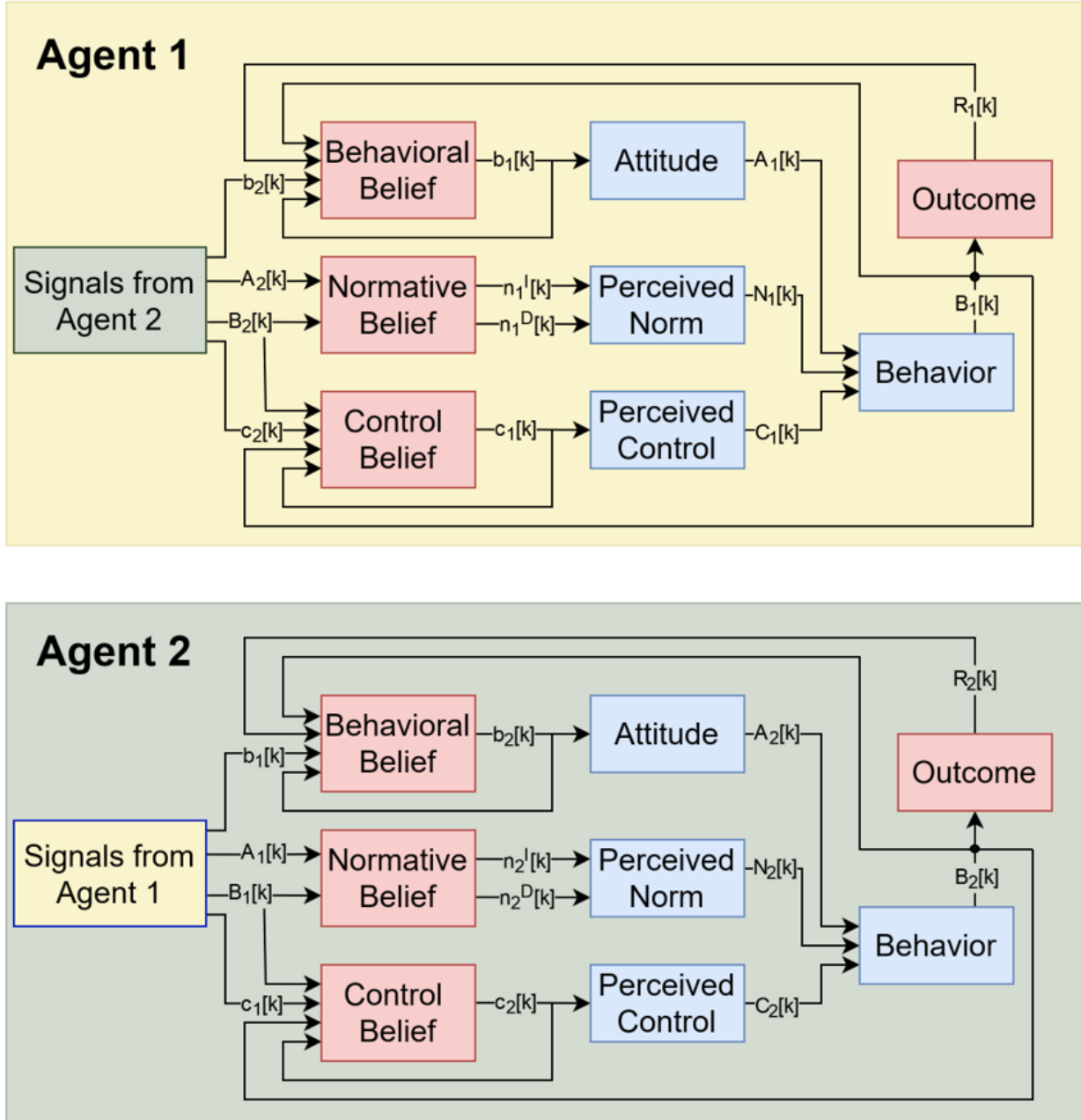


Figure 3.2: Block diagram for a two-person system

perceived behaviour of an agent's peers, i.e. if an agent's peers are performing the behaviour then their beliefs update to reflect this.

$$B_{i,j,l}^N[k]_{bias} = \text{sat}(B_j[k] + \text{sgn}(n_{i,l}^D[k])h_i),$$

$$n_{i,l}^D[k+1] = n_{i,l}^D[k] + \mu_i \sum_j \omega_{i,j}(\alpha B_{i,j,l}^N[k]_{bias} - n_{i,l}^D[k]),$$

The agent's injunctive beliefs are then updated based on the attitude of others in the system, i.e. if an agent's peers think that the behaviour is helpful then their beliefs are updated to track that.

$$A_{i,j,l}[k]_{bias} = \text{sat}(A_j[k] + \text{sgn}(n_{i,l}^I[k])h_i),$$

$$n_{i,l}^I[k+1] = n_{i,l}^I[k] + \mu_i \sum_j \omega_{i,j}(\alpha A_{i,j,l}[k]_{bias} - n_{i,l}^I[k])$$

Next, the control belief component updates the agent's control beliefs. Similar to behavioural beliefs, the agent updates their belief to track whether they have performed the behaviour or not, this is to account for the fact that if the agent performed the behaviour this reinforces their self-efficacy about being able to complete the behaviour in the future.

$$B_{i,l}^C[k]_{bias} = \text{sat}(B_i[k] + \text{sgn}(c_{i,l}[k])h_i),$$

$$\hat{c}_{i,l}[k+1] = c_{i,l}[k] + \mu_i(\alpha B_{i,l}^C[k]_{bias} - c_{i,l}[k]).$$

Additionally, the agent changes their belief based on social influence effects in (3.4.1), if others that are like them think they can perform the behaviour, they too can perform the behaviour.

$$\hat{c}_{i,j,l}[k]_{bias} = \text{sat}(\hat{c}_{j,l}[k] + \text{sgn}(\hat{c}_{i,l}[k])h_i),$$

$$c_{i,l}[k] = \hat{c}_{i,l}[k] + \mu_i \sum_j \omega_{i,j}(\hat{c}_{i,j,l}[k]_{bias} - \hat{c}_{i,l}[k]).$$

Next, the agent's attitudes, perceived norms, and perceived behavioural control are established using an expectancy-value model as described in Section 3.1.

$$A_i[k] = \sum_{l=1}^{L^b} b_{i,l}[k]e_{i,l}, \text{ with } \sum_{l=1}^{L^b} e_{i,l} = 1$$

$$\begin{aligned}
N_i[k] &= \frac{1}{2} \sum_{l=1}^{L^I} n_{i,l}^I[k] m_{i,l}^I + \frac{1}{2} \sum_{l=1}^{L^D} n_{i,l}^D[k] m_{i,l}^D, \\
\sum_{l=1}^{L^I} m_{i,l}^I &= 1, \text{ and } \sum_{l=1}^{L^D} m_{i,l}^D = 1 \\
C_i[k] &= \sum_{l=1}^{L^c} c_{i,l}[k] p_{i,l}, \text{ with } \sum_{l=1}^{L^c} p_{i,l} = 1.
\end{aligned}$$

Using these calculated values the agent's behaviour is then established in the behaviour components shown in Figure 3.2 using a convex combination of attitude, perceived norms, and perceived behavioural control,

$$\begin{aligned}
B_i[k] &= \lambda_i A_i[k] + \beta_i N_i[k] + \gamma_i C_i[k], \\
\lambda_i + \beta_i + \gamma_i &= 1 \text{ and } \lambda_i, \beta_i, \gamma_i > 0,
\end{aligned}$$

Finally, the outcome of performing the behaviour is determined in the outcome components shown in Figure 3.2 using the methods described in Section 3.3,

$$\begin{aligned}
R_{i,l}[k] &= f_l(B_i[k]) \\
f_l(B_i[k]) &= \begin{cases} +B_i[k] & \text{if } l \text{ is a positive behavioural belief} \\ -B_i[k] & \text{if } l \text{ is a negative behavioural belief} \end{cases}
\end{aligned}$$

This process is then repeated as necessary for each step k .

3.4.2 Limitations

The model developed in this chapter, like any model, has limitations, including the following:

- We assume that the RAA can accurately describe how beliefs affect human behaviour. While the RAA has been used to explain various behavioural conditions, it may not be suitable to explain behaviours that arise from addictive, not fully voluntary compulsions and the influence techniques suggested in this thesis may not be suitable to tackle these problems. The RAA also states that behaviour is limited by an individual's actual control of a situation; if a person intends to perform a behaviour but cannot access the means to perform the behaviour, they will not perform the behaviour. These ideas are captured in the following set of assumptions:

Assumption 3.1. *All behaviours being studied are voluntary and of an individual's own volition.*

Assumption 3.2. *All persons have full actual control of their behaviour.*

Assumption 3.3. *The RAA is a sufficient explanation of fully voluntary human behaviour.*

- The RAA relies on several underlying parameters that determine the importance of different salient beliefs to a person's end behaviour. For simplicity, in the model, we assume that these parameters remain constant for every individual in the system. The process of measuring an individual's underlying parameters and beliefs can also be very difficult to conduct and requires several surveys that can be noisy and unreliable measures of a person's true beliefs. The following assumption is made:

Assumption 3.4. *An individual's belief evaluations $e_{i,l}$, motivations to comply $m_{i,l}^I$, $m_{i,l}^D$, power of control $p_{i,l}$, and behaviour weights $\lambda_i, \beta_i, \gamma_i$ remain constant.*

- The methods of capturing social influence in the model are a similarity-based model and an internal bias. We assume that these components can accurately explain the social dynamics of individuals in a system and how they propagate. For simplicity, we assume that the social weights in this model remain constant. Other factors such as social identity and status dynamics are ignored to simplify the model and allow for easier analysis. The means of influence in this model, where an external source of influence influences belief change in individuals, mainly focuses on relationships that are personal. The influence dynamics on larger scale systems, such as through social networks and advertising, have not been investigated and may not apply.

Assumption 3.5. *The similarity model and internal biases are sufficient explanations of social influence dynamics.*

Assumption 3.6. *The relative impact of ω_{ij} , of an agent j on agent i remains constant.*

Assumption 3.7. *The systems being studied involve small groups with personal connections to the control agent in the system.*

- We assume that the outcome of a behaviour follows the linear response stated in Equation (3.20). This linear relationship is unlikely for many scenarios where factors such as diminishing returns or other nonlinear effects like time delays may be a

factor. However, to simplify analysis, this assumption is made. The outcome feedback component can be substituted for further analysis. Additionally, we assume that an individual takes into account the outcomes of a behaviour only if they sufficiently perform the behaviour (i.e. $B_i[k] > 0$).

Assumption 3.8. *The outcome of a behaviour follows Equation (3.20).*

Assumption 3.9. *Individuals take into account the outcome of a behaviour only if they sufficiently perform the behaviour (i.e. $B_i[k] > 0$).*

Table 3.1: Important parameters and variables

Symbol	Description
$B_i[k] \in [-1, 1]$	Behaviour of agent i
$A_i[k] \in [-1, 1]$	Attitude of agent i
$N_i[k] \in [-1, 1]$	Perceived social norm of agent i
$C_i[k] \in [-1, 1]$	Perceived behavioural control of agent i
$R_i[k] \in [-1, 1]$	Behavioural outcome of agent i , $R_i[k] = [R_1 R_2 \dots R_{L^b}]^T$ for L^b behavioural beliefs
$b_i[k] \in [-1, 1]$	Behavioural beliefs of agent i , $b[k] = [b_1 b_2 \dots b_{L^b}]^T$ for L^b behavioural beliefs
$n_i^I[k] \in [-1, 1]$	Injunctive normative beliefs of agent i , $n^I[k] = [n_1^I n_2^I \dots n_{L^I}^I]^T$ for L^I injunctive normative beliefs
$n_i^D[k] \in [-1, 1]$	Descriptive normative beliefs of agent i , $b[k] = [n_1^D n_2^D \dots n_{L^D}^D]^T$ for L^D descriptive normative beliefs
$c_i[k] \in [-1, 1]$	control beliefs of agent i , $c[k] = [c_1 c_2 \dots c_{L^c}]^T$ for L^c control beliefs
$e_{i,l} \in [0, 1]$	Value of a behavioural belief l held by agent i
$m_{i,l} \in [0, 1]$	Motivation to comply of a normative belief l held by agent i
$p_{i,l} \in [0, 1]$	Power of a control belief l held by agent i
$\mu_i \in (0, 1)$	Suggestibility (Rate of convergence) for agent i
$\omega_{i,j} \in [0, 1]$	Relative influence of an agent j on another agent i , $\sum_j \omega_{i,j} = 1$
$h_i \in [0, 1)$	Internal bias offset for agent i

Chapter 4

Model Verification and Validation

In this chapter, we verify whether the model created in Chapter 3 is correctly implemented and validate that its results are consistent with observations from the literature. We begin by simulating a zero-input response for an exercise scenario. For this section, the behaviour being enforced in simulation is exercising three times per week. The simulation is run on three different scenarios involving 1, 2, and 30 individuals respectively. The results of these simulations are compared with expected behaviour to check for consistency.

4.1 Single-Agent Model

The single agent model ignores all social aspects of the original combined model proposed in Section 3.4 and shown in Figure 3.2. This approach is used to reduce the model complexity as much as possible, allowing us to analyse the simplest response. A block diagram of the model with only one agent can be found in Figure 4.1. From this figure, we can see that we are left with what is essentially the RAA model from Section 3.1 combined with feedback components from Section 3.3 resulting in the following simplified model:

$$b_{i,l}[k+1] = \hat{b}_{i,l}[k+1] \quad (4.1)$$

$$n_{l,i}^D[k+1] = n_{l,i}^D[k] \quad (4.2)$$

$$n_{i,l}^I[k+1] = n_{i,l}^I[k] \quad (4.3)$$

$$c_{i,l}[k+1] = c_{i,l}[k] + \mu_i^{CB}(\alpha^{CB} B_{i,l}^C[k]_{bias} - c_{i,l}[k]) \quad (4.4)$$

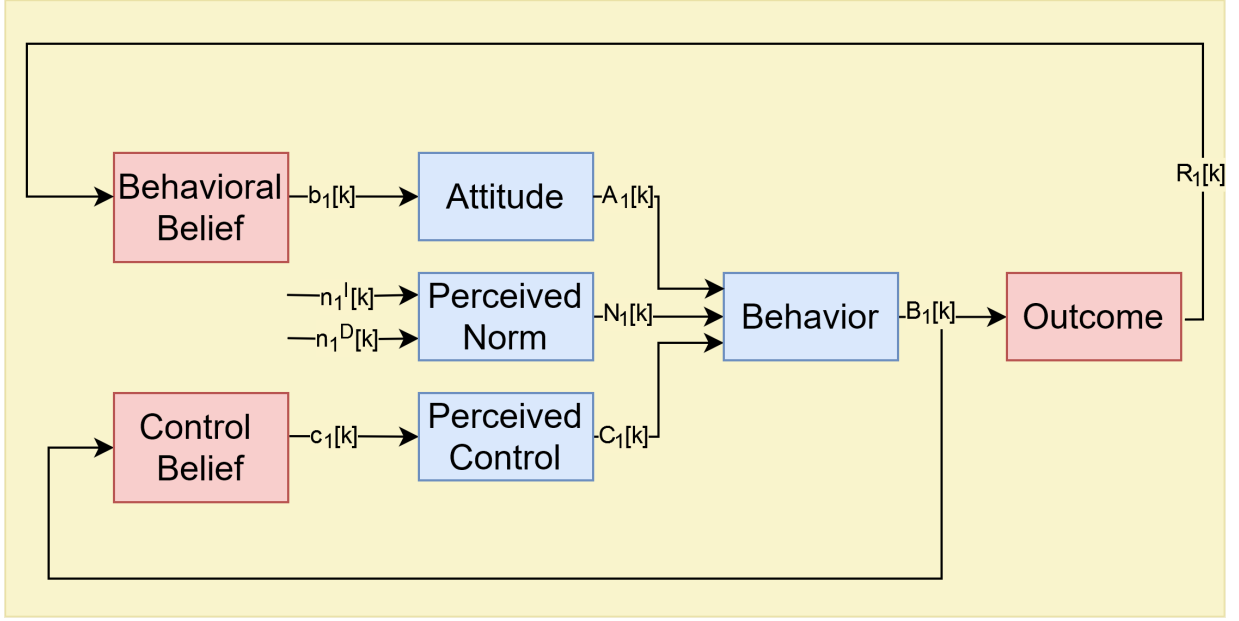


Figure 4.1: Block diagram for a one-person system

For the system involving exercise as a behaviour, we consider two salient behavioural beliefs, one salient normative referent, and one control factor. The first behavioural belief is the idea that exercise will lead to feeling healthy. This belief has a positive evaluation with regards to the behaviour, i.e., feeling healthy is considered a good thing. Additionally, there is a second belief that exercise will lead to more soreness. In this example, when $B_i[k] = -1$ we assume the agent is not exercising, if $B_i[k] = 0$ the agent exercises three times a week every other week, and if $B_i[k] = 1$ the agent exercises three times a week every week.

In the one-person model, behavioural beliefs can be affected only by the results of performing a particular behaviour. For simplicity, in this simulation, we assume that the individual's behaviour has a linear relationship, with a gain of 1, to the observed result for each of their behavioural beliefs. A positive relationship is held between exercise and feeling healthier and a negative relationship is established with feeling sore. I.e. people feel healthier after exercising and more sore. This leads to the following equation:

$$\begin{bmatrix} R_{i,1}[k] \\ R_{i,2}[k] \end{bmatrix} = \begin{bmatrix} B_i[k] \\ -B_i[k] \end{bmatrix} \quad (4.5)$$

Since the one-person model has no dynamic social aspect, the only influences possible

are those that are external. In this case, we assume the normative referent to be the individual’s parents. The descriptive norm, in this case, is associated with their parents doing the behaviour, while the injunctive norm is defined by whether their parents approve of the behaviour. For simplicity, we assume that the dynamics of the external influences are unchanging.

Finally, the salient control factor associated with the behaviour is whether or not the individual will have time to perform the behaviour. Again, since this model involves only one individual, this belief is not affected by social influences. Therefore, the only component that changes this belief is whether the individual has performed the behaviour or not.

The values for the parameters in the one-person model were chosen arbitrarily (shown in Table 4.1), but within the domains listed in Table 3.1 to demonstrate typical observations. Three simulations are analysed; one where the desired behaviour is not being performed, one where the behaviour is being performed, and finally one where the behaviour transitions from one state to the other.

Table 4.1: Parameter values for the one-person model

Parameter	Value
λ	$2/3$
β	$1/3$
γ	$1/3$
$\begin{bmatrix} e_{1,1} \\ e_{1,2} \end{bmatrix}$	$\begin{bmatrix} 1 \\ 0.5 \end{bmatrix}$
$m_{1,1}^I$	0.8
$m_{1,1}^D$	0.8
$p_{1,1}$	1
h^R	0.1
h^{BC}	0.1
μ	0.2

4.1.1 Negative Behaviour Simulation

The first simulation is one under which the individual initially has very poor belief in the behaviour having a positive outcome. Additionally, the individual’s normative referent (i.e., their parents) thinks poorly of the behaviour as well. Finally, the individual does not

believe they will have time to accomplish the tasks required for the behaviour. The results are shown in Figure 4.2.

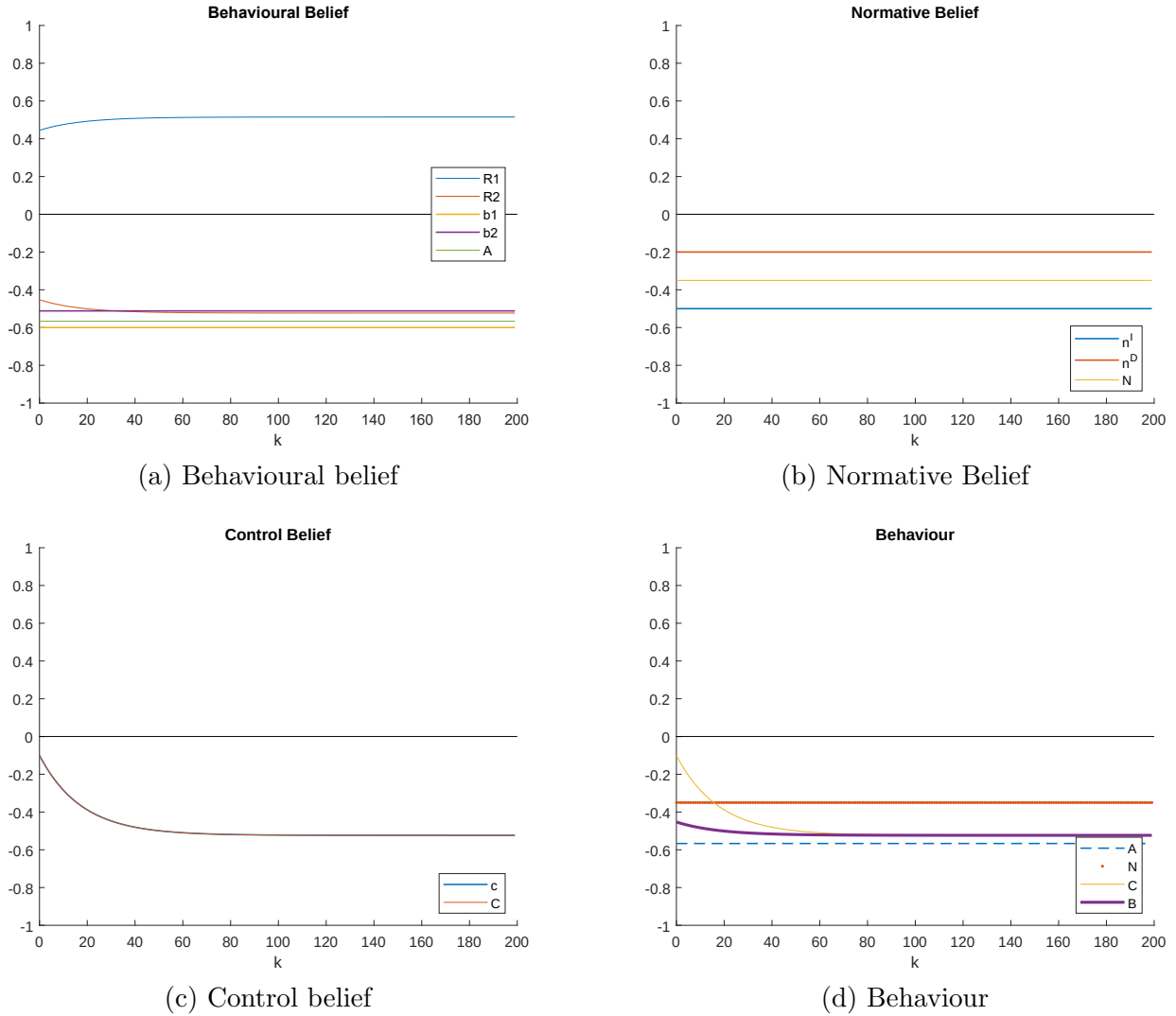


Figure 4.2: Simulation of the one-person system with negative behaviour

The first thing to notice from the simulation is that very little changes. It is obvious from the setup of the one-person model that normative beliefs (n^I, n^D) are unchanging due to the lack of individuals to interact with and the assumption that the individual's unmodeled referents have a constant normative effect. On the other hand, the reason for the

behavioural belief signals ($b_{1,1}, b_{1,2}$) remaining constant is due to the nonlinear effects of the result feedback mechanism. Behavioural beliefs are changed by result feedback based on the behaviour but not by social influences, which are non-existent in this scenario. Additionally, the nature of accepting the result of the behaviour only if it is being reasonably performed (i.e., $B[k] > 0$) means that, as long as $B[k] < 0$, results will have no impact on the belief of the individual. Finally, control beliefs (c) are only affected by whether the individual is performing the behaviour or not, ignoring social effects. Figure 4.2d shows how this belief tracks with behaviour.

The results in Figure 4.2 align with what would be expected in an exercise scenario with only one individual. Their normative beliefs would be unchanged due to the lack of changing social influences. An individual with certain behavioural beliefs concerning exercising, regardless of them being positive or negative, has no reason to change those beliefs if they do not reasonably try to perform the behaviour and view how it affects the expected results. Lastly, since control beliefs are related to self-efficacy, as explained in Section 2.3.2, not performing the behaviour brings into question whether the individual is capable of doing the behaviour and so lowers their perceived behavioural control (in this case the belief that they have enough time to perform the behaviour).

4.1.2 Positive Behaviour Simulation

We now simulate a situation where the individual believes that exercise will lead to feeling healthy and that it is unlikely to cause soreness. The individual's normative referents also approve of the behaviour, and the individual also believes that they are in control of being able to perform the behaviour. Results of the individual's behavioural beliefs and behaviour signals are shown in Figure 4.3. To see all of the signals refer to Figure A.1.

This simulation has a few more interesting aspects at play. Again, as expected, the normative belief signal (N) does not vary, and the control belief signal (C) tracks with the individual's behaviour. However, the individual's behavioural beliefs ($b_{1,1}, b_{1,2}$) do change. They seem to track the results of performing the behaviour ($R_{1,1}, R_{1,2}$) with an offset; this outcome is expected. As the individual performs the behaviour, it results in them feeling healthier; however, it also results in an increase in soreness. This result is also biased by the individual's belief in what the result should be, leading to an internal bias. The combination of these factors leads the individual to update their initial behavioural beliefs about exercise, resulting in the outputs shown in Figure A.1.

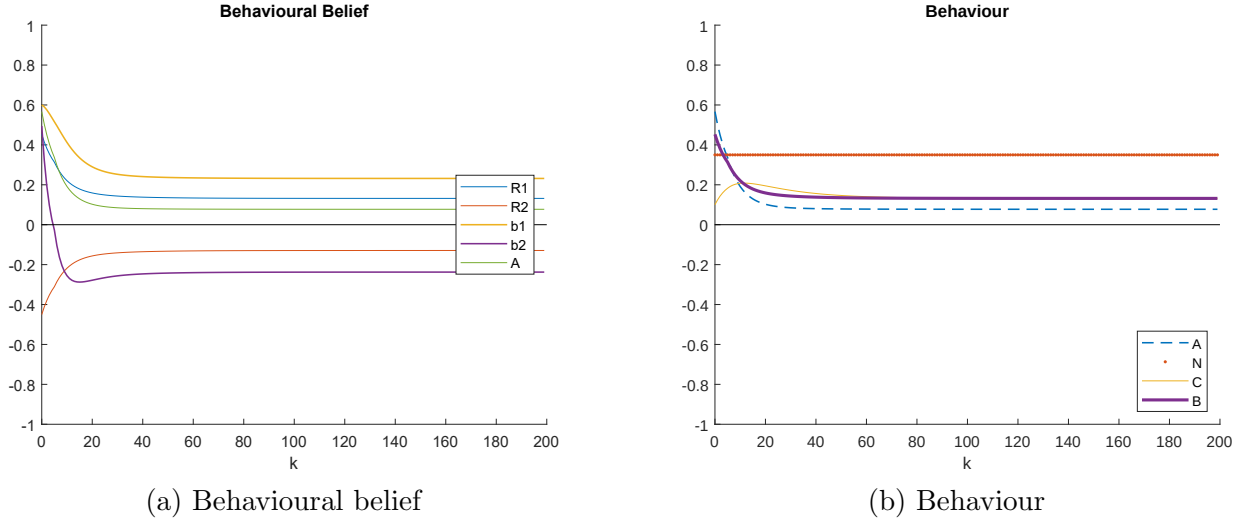


Figure 4.3: Simulation of one-person system with positive behaviour

4.1.3 Negative to Positive Behaviour Simulation

We now look at a situation where the individual initially has a low belief that exercise will lead to feeling healthy but a high belief that it will cause soreness. Their normative belief is extremely high, indicating that their parents fully support exercising; however, their control belief is extremely low, indicating that the individual doesn't believe they will have time to do the behaviour. With this simulation, we wish to observe the transition behaviour of the system when an individual goes from performing the behaviour infrequently to performing it more frequently. The main results are shown in Figure 4.4. To see all of the signals refer to Figure A.2.

As before, the normative beliefs (N) are unchanged. Likewise, control beliefs (C) track the behaviour output. The more interesting dynamics occur in the behavioural belief signals ($b_{1,1}, b_{1,2}$) of the individual (Figure 4.4a). As time goes by the individual's control belief signal corrects itself to reflect how frequently the individual actually performs the behaviour, and so increases. This causes the individual to perform the behaviour more frequently. When the individual begins to do the behaviour more than every other week ($B[k] > 0$) then they begin to consider how exercising relates to their behavioural beliefs. In this situation, they realise that exercising makes them more sore than they initially expected. This results in the individual updating their beliefs to reflect the result outcome of performing the behaviour (in this case with a positive offset to reflect the information bias of their belief). Their belief that their health is unlikely to improve because of the behaviour

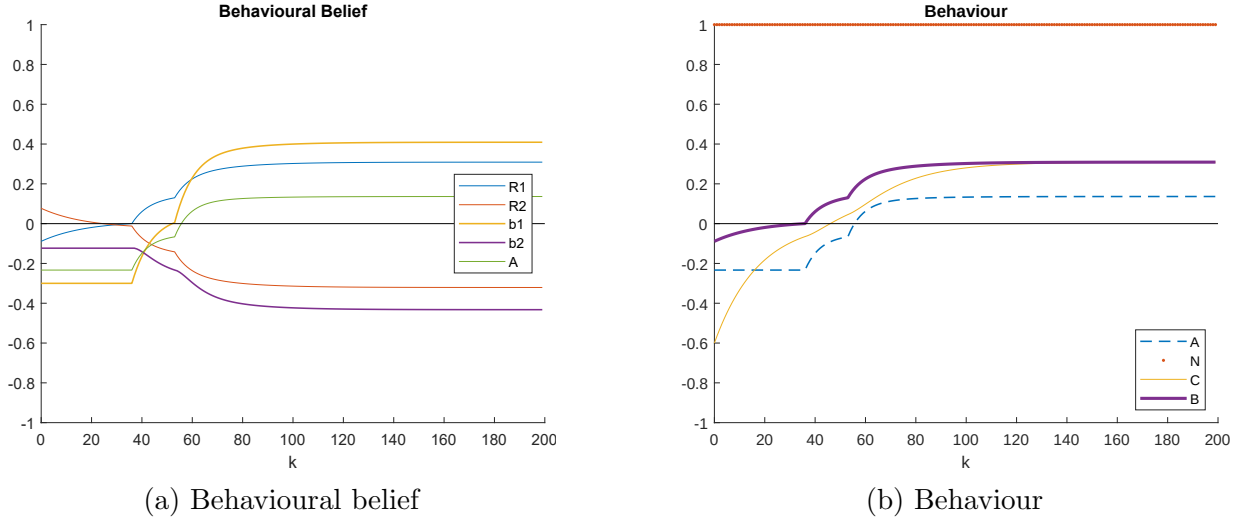


Figure 4.4: Simulation of one-person system with behaviour transitioning from negative to positive

is also updated to reflect the results observed by performing the behaviour. Initially, the result concerning health is biased negatively to reflect the individual's negative belief at the time; however, at around $k = 60$ (i.e. the 60th interaction), the individual realises that exercise is more likely to lead to feeling healthy and so has a positive information bias when getting results. This causes a boost in attitude towards the belief, leading to a higher frequency of exercising.

4.2 Two-Agent Model

We now move on to a model involving only two individuals interacting with each other. The goal of these simulations is to observe the social influence aspects of the model in the simplest possible situation. This will make it easier to view the dynamics at play when individuals interact with each other. The block diagram for this situation is shown in Figure 3.2. For ease of interpretation, states that are passed from one individual to the other are outputted from independent state output blocks. These blocks (Signals from Agent 1/2) take the output states from Agent 1 and feed them into Agent 2's system where necessary, and vice versa. Since we are now including social influence effects, the equations for the system revert to those listed in Section 3.4. Additionally, we keep the parameters listed in Table 4.1. Lastly, we again use the result formula from Equation (4.5). For the

case with two individuals, we run three simulations, one where both individuals have a negative belief about the behaviour, one where both individuals have a positive belief, and finally, a simulation where one individual has a positive belief and the other has a negative belief.

Before proceeding to the simulations of the multiagent system that includes bias and outcome feedback, we can first examine what would happen to the system if some of the non-linear components were removed. In this case, the system is only driven by conformity effects from the assimilative social influence mode. Given this simplified system, a few theoretical results can be found.

4.2.1 Theoretical Results of a Simplified Multi-Agent System

We first examine a system with both internal bias and outcome feedback disabled.

Lemma 1. *Consider the system of the form (3.5) with the condition that if an agent i is influenced by agent j then agent j is also influenced by agent i . This represents the belief dynamics of the system when only conformity is in effect. Then the beliefs of the agents in the system will converge to a set of $\mathbf{M}$ clusters, where $1 \leq \mathbf{M} \leq \mathbf{N}$ and $\mathbf{N}$ is the number of agents in the system.*

Proof. The belief dynamics of the system when only conformity is in effect, and not including internal biases or outcome feedback, is described as follows,

$$\begin{aligned} o_i[k+1] &= o_i[k] + \mu_i \sum_{j=1}^{\mathbf{N}} \omega_{ij} (o_j[k] - o_i[k]), \quad \sum_j \omega_{ij} = 1, \quad 0 < \mu < 1 \\ &= (1 - \mu_i) o_i[k] + \mu_i \sum_{j=1}^{\mathbf{N}} \omega_{ij} o_j[k] \end{aligned} \tag{4.6}$$

where o_i represents a belief of agent i in the RAA model, w_{ij} represents the relative influence agent j has on agent i , and μ_i is the suggestibility of agent i .

We can rewrite this system in matrix form, as follows,

$$o[k+1] = W o[k], \tag{4.7}$$

$$W = \begin{bmatrix} 1 - \mu_1 & \mu_1 \omega_{12} & \cdots & \mu_1 \omega_{1\mathbf{N}} \\ \mu_2 \omega_{21} & 1 - \mu_2 & \cdots & \mu_2 \omega_{2\mathbf{N}} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{\mathbf{N}} \omega_{\mathbf{N}1} & \mu_{\mathbf{N}} \omega_{\mathbf{N}2} & \cdots & 1 - \mu_{\mathbf{N}} \end{bmatrix}. \tag{4.8}$$

In this form, we can see that matrix W is a stochastic matrix whose rows sum to 1. Further, because of the assumption that every agent i that influences agent j is also influenced by agent j , we can rearrange the system, grouping agents into irreducible groups that influence each other:

$$\bar{o}[k+1] = \bar{W}\bar{o}[k], \quad (4.9)$$

$$\bar{W} = \begin{bmatrix} \bar{W}_1 & 0 & \cdots & 0 \\ 0 & \bar{W}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \bar{W}_M \end{bmatrix} \quad (4.10)$$

where $1 \leq M \leq N$ is the number of irreducible groups, and so the number of clusters that form. Each $\bar{W}_i$ is an irreducible stochastic matrix which includes all the agents that form a chain of influence. The only way to be not included in a chain with agent i is by being in a separate chain that does not influence any agent in i 's chain.

Now that we have established M irreducible chains, $\bar{W}_i$, we can show that each chain is aperiodic by showing that the maximum belief, o_{max} , in an irreducible chain decreases over time unless all beliefs are the same.

Let $o_{max}[k] = \max_i(o_i[k])$. First, if all beliefs are the same then we have $o_i[k] = o_{max}[k]$ for all i . Then, using (4.6),

$$o_i[k+1] = (1 - \mu_i)o_i[k] + \mu_i \sum_{j=1}^N \omega_{ij}o_j[k] \quad (4.11)$$

$$= (1 - \mu_i)o_{max}[k] + \mu_i \sum_{j=1}^N \omega_{ij}o_{max}[k] \quad (4.12)$$

$$= (1 - \mu_i)o_{max}[k] + \mu_i o_{max}[k] \quad (4.13)$$

$$= o_{max}[k]. \quad (4.14)$$

Therefore, if all beliefs in the chain are equal then the system is aperiodic and the system stays at the constant state.

Now, if instead all beliefs are not the same, there exists an agent j with belief $o_j[k] < o_{max}[k]$ that is interacting with an agent r with belief $o_i[k] = o_{max}[k]$. We can first show

that any agent with a belief less than the maximum will never reach the maximum:

$$o_j[k+1] = (1 - \mu_j)o_j[k] + \mu_j \sum_{i=1}^N \omega_{ji} o_i[k]. \quad (4.15)$$

If we assume all agent interacting with agent j have the maximum belief ($o_{max}[k]$), as a best case scenario, then,

$$o_j[k+1] \leq (1 - \mu_j)o_j[k] + \mu_j \sum_{i=1}^N \omega_{ji} o_{max}[k] \quad (4.16)$$

$$\leq (1 - \mu_j)o_j[k] + \mu_j o_{max}[k]. \quad (4.17)$$

Subtracting with $o_{max}[k]$ we get,

$$o_{max}[k] - o_j[k+1] \geq o_{max}[k] - (1 - \mu_j)o_j[k] - \mu_j o_{max}[k] \quad (4.18)$$

$$\geq o_{max}[k] - o_j[k] + \mu_j o_j[k] - \mu_j o_{max}[k] \quad (4.19)$$

$$\geq (1 - \mu_j)(o_{max}[k] - o_j[k]) > 0. \quad (4.20)$$

Therefore,

$$o_{max}[k] > o_j[k+1] \quad (4.21)$$

We can now show that the belief of agent r reduces by using (4.6),

$$o_r[k+1] = o_r[k] + \mu_r \sum_{i=1}^N \omega_{ri} (o_i[k] - o_r[k]) \quad (4.22)$$

$$= o_{max}[k] + \mu_r \sum_{i=1}^N \omega_{ri} (o_i[k] - o_{max}[k]). \quad (4.23)$$

as a best case scenario agent r interacts with all but one agent j whose belief is not the maximum, expanding (4.23) we get,

$$o_r[k+1] \leq o_{max}[k] + \mu_r [\omega_{r1}(0) + \dots + \omega_{rj}(o_j[k] - o_{max}[k]) + \dots + \omega_{rN}(0)] \quad (4.24)$$

$$\leq o_{max}[k] - \mu_r \omega_{rj} (o_{max}[k] - o_j[k]) \quad (4.25)$$

$$< o_{max}[k] \quad (4.26)$$

Therefore, any agent with a belief equal to the maximum that interacts with an agent with a lower belief will have their belief reduced and will never return to having the maximum

belief they started with. Thus, since the initial state of the system can never repeat the system is aperiodic.

Since each $\bar{W}_i$ is an irreducible and aperiodic stochastic matrix, it has an eigenvalue of 1 with multiplicity 1 and all other eigenvalues are less than 1. Further, the eigenvector of the matrix is $\vec{1}$ and so all the states associated with the sub-matrix converge. This can be applied to each sub-matrix, $\bar{W}_i$, leading to $\mathbf{M}$ clusters of beliefs. $\square$

Using Lemma 1 we can see that the beliefs of agents that are accessible to one another (i.e. they are connected to each other) will converge into at most $\mathbf{M}$ clusters, and so their beliefs will converge as well. For example, if a group of friends decides to take part in some exercising behaviour, then if the group can communicate fully with one another; then, the group will converge onto an opinion about exercising. However, if there are smaller groups of friends in a fitness class, then groups that can communicate with each other will converge in opinion, while some groups, that are not in contact with the others, may converge to a different opinion.

Next, we can examine a system where internal bias effects are accounted for.

Lemma 2. *Given a system with only conformity and bias effects, shown in (3.4.1), if all agents have positive(negative) beliefs, then the beliefs of all agents will polarise to +1(−1).*

Proof. Under the assumptions of the lemma the equations of the system (3.4.1) have the following equivalent form:

$$o_j[k]_{st} = \text{sat}(o_j[k] + \text{sgn}(o_i[k])h_i), \quad o_j[k]_{st} \in [-1, 1] \quad (4.27)$$

$$o_i[k+1] = o_i[k] + \mu_i \sum_j \omega_{i,j}(o_j[k]_{st} - o_i[k]), \quad (4.28)$$

with

$$h_i \in [0, 1], \quad \mu_i \in (0, 1), \quad o_i \in [-1, 1], \quad \sum_j \omega_{i,j} = 1, \quad (4.29)$$

where o_i represents a belief of agent i in the RAA model. Conditions in (4.29) match those of the system, shown in Table 3.1.

We will prove the lemma for only the positive belief case as proving the negative belief case is a trivial extension.

Let us identify agent r such that $o_r[k] = o_{\min}[k] = \min_i(o_i[k])$. Since beliefs are positive for all agents we can see that $o_j[k]_{st} \geq o_j[k]$. We will first prove the trivial case where all

beliefs are equal to 1 and so $o_r[k] = o_{min}[k] = 1$. From (4.27) we see that, for all j , $o_j[k]_{st} = 1$. Further, using (4.28),

$$o_r[k+1] = o_r[k] + \mu_i \sum_j \omega_{i,j} (o_j[k]_{st} - o_r[k]) \quad (4.30)$$

$$= o_r[k] + \mu_i \sum_j \omega_{i,j} (1 - 1) \quad (4.31)$$

$$= o_r[k] = 1 \quad (4.32)$$

Therefore the system stays at the polarised maximum.

Now we consider the case where $o_r[k] \neq 1$. Let us define $h_{min} = \min_i(h_i)$ and $\mu_{min} = \min_i(\mu_i)$. Using (4.27) we see that, for all j , $o_j[k]_{st} > o_{min}[k]$. If we assume $1 > o_{min}[k] + h_{min}$, from (4.28) we get,

$$o_i[k+1] = o_i[k] + \mu_i \sum_j \omega_{i,j} (o_j[k]_{st} - o_i[k]) \quad (4.33)$$

$$\geq o_i[k] + \mu_i \sum_j \omega_{i,j} (o_{min}[k] + h_{min} - o_i[k]) \quad (4.34)$$

$$\geq (1 - \mu_i) o_i[k] + \mu_i (o_{min}[k] + h_{min}) \quad (4.35)$$

$$\geq (1 - \mu_i) o_{min}[k] + \mu_i (o_{min}[k] + h_{min}) \quad (4.36)$$

$$\geq o_{min}[k] + \mu_i h_{min} \quad (4.37)$$

$$\geq o_{min}[k] + \mu_{min} h_{min} \quad (4.38)$$

Which implies,

$$o_{min}[k+1] \geq o_{min}[k] + \mu_{min} h_{min}, \quad (4.39)$$

and that,

$$o_{min}[k+a] \geq o_{min}[k] + a \mu_{min} h_{min}. \quad (4.40)$$

We can define $K = \lceil \frac{1-h_{min}-o_{min}[k]}{\mu_{min} h_{min}} \rceil > 0$, then

$$o_{min}[k+K] \geq o_{min}[k] + K \mu_{min} h_{min} \quad (4.41)$$

$$\geq 1 - h_{min} \quad (4.42)$$

We also know that $o_i[k]$ is upper bounded by 1 for all k ,

$$o_i[k+1] = o_i[k] + \mu_i \sum_j \omega_{i,j} (o_j[k]_{st} - o_i[k]) \quad (4.43)$$

$$\leq o_i[k] + \mu_i \sum_j \omega_{i,j} (1 - o_i[k]) \quad (4.44)$$

$$\leq (1 - \mu_i) o_i[k] + \mu_i \quad (4.45)$$

$$\leq 1 \quad (4.46)$$

$$(4.47)$$

Therefore, there exists a $\bar{K} < K$ such that $1 \geq o_{min}[k + \bar{K}] \geq 1 - h_{min}$. Since the minimum value is now at least greater than $1 - h_{min}$ we have,

$$o_i[k + \bar{K} + 1] = o_i[k + \bar{K}] + \mu_i \sum_j \omega_{i,j} (o_j[k + \bar{K}]_{st} - o_i[k + \bar{K}]) \quad (4.48)$$

$$= o_i[k + \bar{K}] + \mu_i \sum_j \omega_{i,j} (1 - o_i[k + \bar{K}]) \quad (4.49)$$

$$= (1 - \mu_i) o_i[k + \bar{K}] + \mu_i \quad (4.50)$$

The solution to this dynamic equation is,

$$o_i[k + \bar{K} + a] = (1 - \mu_i)^a o_i[k + \bar{K}] + \mu_i \sum_{l=0}^{a-1} (1 - \mu_i)^{k-l-1} \quad (4.51)$$

$$(4.52)$$

taking the limit we get,

$$\lim_{a \rightarrow \infty} o_i[k + \bar{K} + a] = \lim_{a \rightarrow \infty} (1 - \mu_i)^a o_i[k + \bar{K}] + \mu_i \lim_{a \rightarrow \infty} \sum_{l=0}^{a-1} (1 - \mu_i)^{k-l-1} \quad (4.53)$$

$$= 0 + \mu_i \frac{1}{1 - (1 - \mu_i)} = 1 \quad (4.54)$$

Therefore, the minimum value of the system is strictly increasing and tends towards 1, unless it is already equal to 1, and so becomes polarised.

This proof can be extended to the negative case by simply taking the maximum value and showing that the upper bound decreases to -1 . $\square$

Therefore, we can see, that in a situation where all agents in the system have a general agreement on their beliefs about a behaviour, their beliefs will polarise to an extreme thanks to the addition of a signed internal bias.

With these theoretical results, we can get a deeper understanding of the dynamics at play in the system. Now, we analyse some simulation results with all non-linearities enabled.

4.2.2 Negative Belief Simulation

We first simulate a situation where both individuals in the classroom initially have negative beliefs concerning exercise. Both individuals do not believe that exercise is helpful, they do not believe that their parents support the behaviour, and they do not believe that they have time to perform the behaviour. The main results of this simulation are shown in Figure 4.5. To see all of the signals refer to Figure A.3.

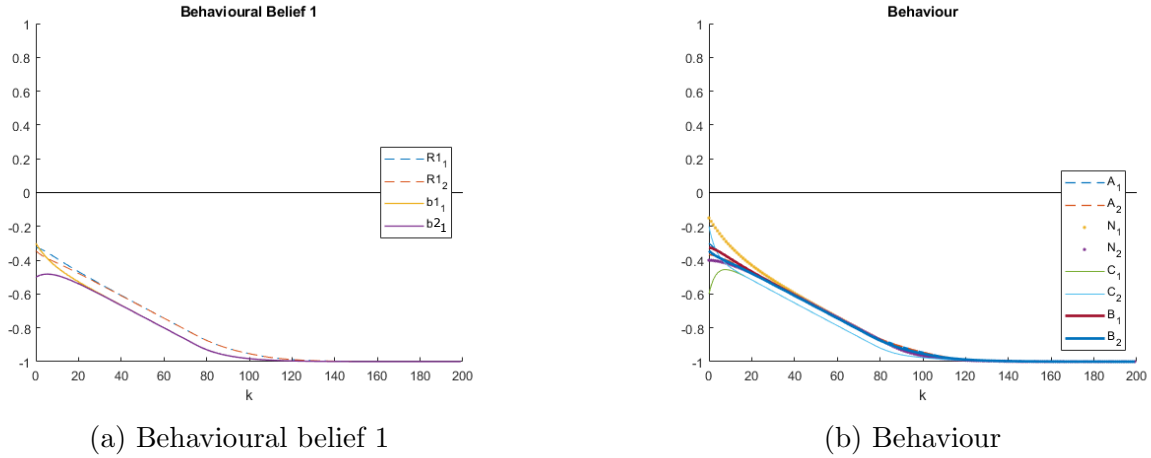


Figure 4.5: Simulation of two-person system with both individuals initially having negative beliefs

As expected, the inclusion of social influence in the two-agent model results in significantly different behaviour from the one-agent model. Since neither individual is performing the behaviour with regular consistency ($B[k] < 0$) their behavioural beliefs ($b_{i,1}, b_{i,2}$) are not influenced by any results feedback ($R_{i,1}, R_{i,2}$) associated with performing the behaviour. Consequently, their behavioural beliefs converge to one another. However, since the individuals have a negative information bias, they interpret that the other individual has a

stronger negative belief than is true. This in turn makes the individual think exercising is worse than it is with regards to feeling healthy or in causing soreness. Combined, this causes the individuals to have a feedback loop leading to beliefs that spiral down.

As a result of their strong negative beliefs, the individuals' attitudes (A_1, A_2) fall, causing their behaviour (B_1, B_2) to fall in tandem. Additionally, since normative beliefs (N) are affected by the other individual's behaviours and attitudes, these also begin to drop, leading to each individual believing that the other does not perform the behaviour readily, nor do they approve of the behaviour. Since control beliefs (C) are affected by individuals performing the behaviour, this also falls, aided by the social influence of the other individual's falling control belief.

This spiralling behaviour leading to strongly polarised beliefs is reminiscent of group polarisation [62]. Group polarization is the tendency for a group of people to have more extreme behaviours and beliefs than that of the individuals in the group, leading to the individuals eventually having a stronger belief than what they stated. In this case, both individuals believe that exercise is not helpful and so they feed into each other's beliefs resulting in a strong belief that the behaviour will not be useful. Group polarisation is sometimes explained by informational influences [91], where the sharing of new information helps an individual become more convinced of their initial beliefs. The emergence of this effect is understandable due to the information bias involved in the model and lends credence to its validity.

4.2.3 Positive Belief Simulation

Our next simulation is similar to the first; however, in this case, both individuals initially have a positive belief concerning exercise. They both believe that exercise will lead to feeling healthy and will not lead to much soreness. They also believe that the other individual approves of the behaviour and are themselves doing it. Finally, both individuals believe that they will have enough time to perform the behaviour. The main results are shown in Figure 4.6. To see all of the signals refer to Figure A.4.

The main difference in this situation is that, because both individuals have positive beliefs, they are performing the behaviour more readily ($B_1, B_2 > 0$) and so take into account the results (R_1, R_2) of them performing the behaviour. This means that each individual's behavioural beliefs ($b_{i,1}, b_{i,2}$) are not only influenced by social effects but are also affected by the results of the behaviour ($R_{i,1}, R_{i,2}$). The outcomes act like a mediator holding the social influences in check with reality and so limit the amount of polarisation that occurs, which is an observation of note. The boundedness of behavioural beliefs

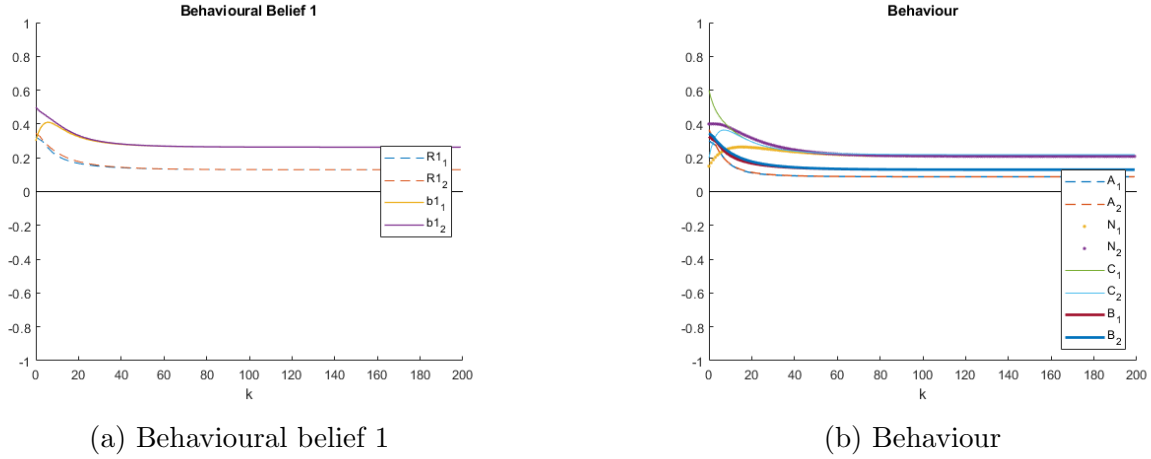


Figure 4.6: Simulation of two-person system with both individuals initially having positive beliefs

influences the other states to remain bounded as well. This situation reflects what you would expect in reality, where even though you might hold a strong belief about something, the reality of the situation would normally correct your beliefs on the matter.

4.2.4 Negative and Positive Belief Simulation

Finally, we simulate a situation where the first individual initially believes that exercise is not beneficial and that others are not supportive of doing the behaviour, and that they do not have time to do the behaviour. On the other hand, the second individual initially believes the opposite. The main results are shown in Figure 4.7. To see all of the signals refer to Figure A.5.

In these results, we see that individual 1 initially has a negative behaviour signal ($b_{1,1}, b_{1,2}$) but eventually switches to being positive. When the individual has negative behaviour they are only being affected by social influences. This causes their belief that exercise is helpful in feeling healthy to increase towards that of Individual 2. On the other hand, Individual 2's health belief drops both due to the social influence induced by Individual 1 and because the results of performing the behaviour show a lower health benefit result than expected, even with the bias. Eventually, at the 38th interaction, individual 1 goes from being uncertain of the helpfulness of the behaviour to believing that the behaviour is useful in increasing health. This causes the individual to search for information that supports their argument and so leads to a positive internal bias when interpreting the

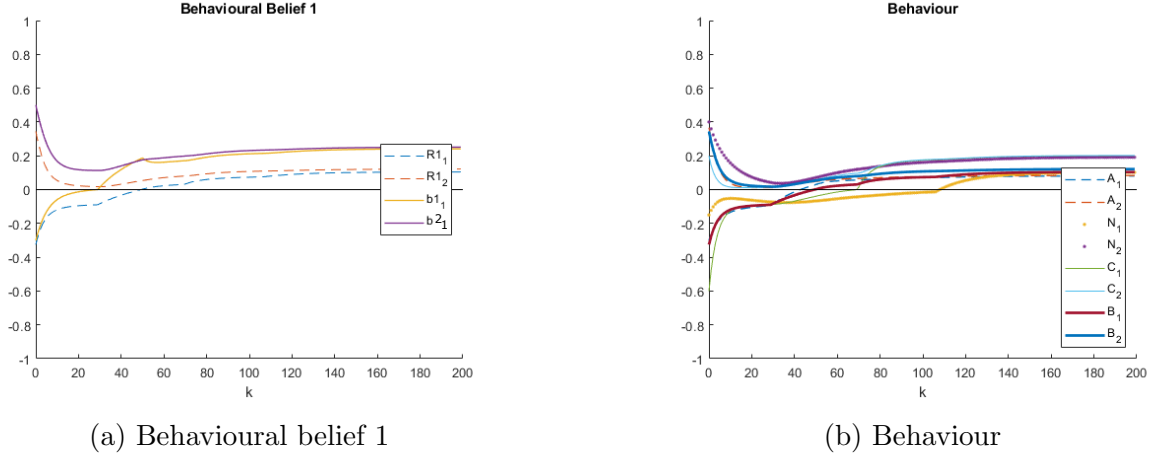


Figure 4.7: Simulation of two-person system with initially one individual having positive beliefs and one individual having negative beliefs

other individual's beliefs. At the 50th interaction mark the individual starts performing the behaviour more readily ($B_1 > 0$) and so takes into account the results of their actions ($R_{1,1}, R_{1,2}$), leading to a slight reduction in their belief.

Similarly, individual 2 initially has a belief that exercising does not lead to soreness. However, this is quickly corrected by the fact that their results show otherwise. They are also influenced by individual 1's beliefs. The individuals' beliefs don't initially converge because individual 1 is only affected by individual 2's beliefs, which they have a bias in interpreting, and so track this biased value. However, at the 50th interaction mark individual 1 begins to take into account the results of their behaviour which mediates their initial beliefs. Eventually, the individuals converge to a point that is slightly higher than their biased result-based influence, due to their biased social interpretation.

The individual's normative beliefs and control beliefs follow along as expected from a real-life situation. One individual doesn't believe a behaviour is useful and so doesn't perform it. However, the other individual does think it is useful. This leads to the first individual gaining insight into the behaviour through observing the second individual, slowly convincing them of its usefulness. Meanwhile, individual 2 slowly loses faith in the behaviour because of what they observe from individual 1's attitudes and behaviour; however, because the second individual has sufficient evidence of the behaviour working, through their results, they are not swayed. Eventually, the first individual tries the behaviour for themselves and observes that it is indeed helpful and so continues performing the behaviour and improves their beliefs about it.

4.3 30-Agent Model

We now move on to a larger-scale simulation with 30 individuals interacting with each other. This would represent a person with a large group of friends who are all trying to be healthier. The block diagram of this simulation is an extension of the block diagram shown in Figure 3.2 with 30 agents. The equations for this system are the same as that of the two-person model, except generalised for 30 individuals. We will look at two main simulations with this model, one with most individuals initially having a negative belief about the behaviour and one where individuals initially have mixed opinions regarding it.

4.3.1 Mostly Negative Belief Simulation

We first look at a simulation where most individuals initially have a negative opinion about exercise. Because there is a large number of individuals involved in this simulation, each individual's initial conditions are generated using a random uniform distribution ranging from -1 to 1 . Two simulations where most of the initial conditions were set to be less than 0 are shown in Figure 4.8. To see a simulation with 150 agents see Figure A.6. The

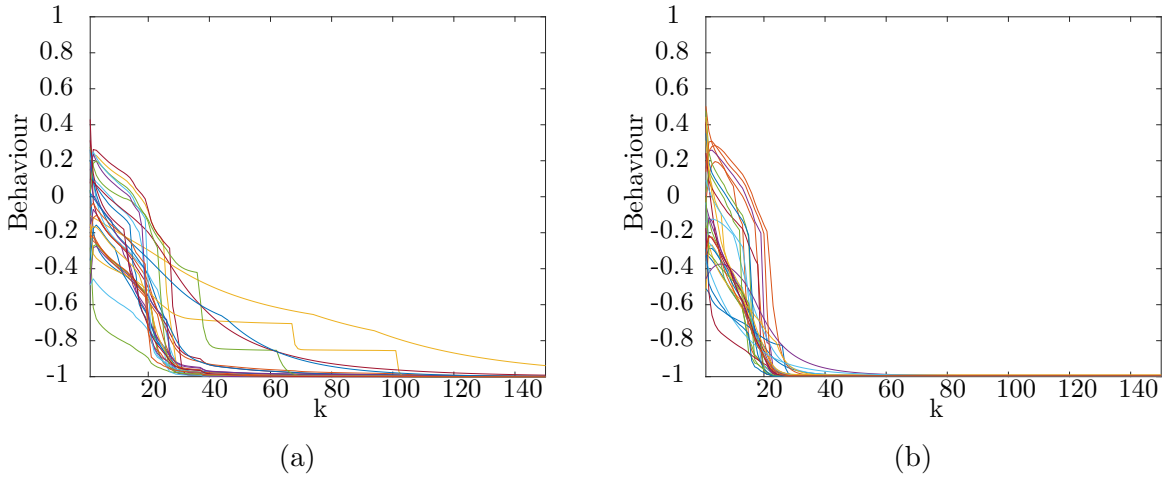


Figure 4.8: Polarisation of behaviours

figure clearly shows that with these initial conditions, the whole system saturates and everyone in the group ends up not performing the behaviour. We can observe that the individuals begin with a spread of initial beliefs and behaviours. However, in this case, the parameters of these individuals were such that the individuals with negative beliefs

eventually overwhelmed those with positive beliefs and led to a negative outcome. We see that, initially, the individuals begin to converge in beliefs, moving towards the average of the group. However, as individuals begin to change beliefs from being optimistic about the behaviour to being pessimistic, the effects of internal biases cause significant changes in beliefs as the individuals begin to interpret information in a way that suits their newfound direction. This confirmation bias leads to individuals interpreting that their peers' beliefs are more negative than they actually are, leading to a vicious feedback loop that causes beliefs to spiral further and further towards the negative.

This spiralling polarisation behaviour is similar to what was observed in Figure 4.5 of the two-person system; however, the effects are even more dramatic. The increased number of agents in the system means that there are more negative agents in the system who are likely to convince you that the behaviour is bad. Therefore, there is a stronger negative pressure pulling any given agent downward. The 150-person simulation in Figure A.6 provides further evidence of this.

4.3.2 Mixed Belief Simulation

Finally, we look at a mixed belief simulation. In this case, we again randomly generate initial conditions using a uniform distribution; however, the initial conditions were chosen so half the individuals positively believe in the behaviour and the other half do not. Some simulations of the resultant behaviour are shown in Figure 4.9. To see a 150-person version of the model refer to Figure A.7.

We see some interesting dynamics at play in this mixed simulation. Immediately, the results look quite different from the previous simulation, with large groups of individuals having differing opinions from other individuals. We can see that with a mixture of initial conditions, the individuals eventually get to a point where they have clusters of varying beliefs forming. The main driving force of this is the confirmation bias in each individual. While there is a convergence effect that causes individuals to try and reach a consensus, some individuals have a large confirmation bias causing them to misinterpret their peers' beliefs. This leads to individuals having a perceived consensus of beliefs where in reality they have different beliefs.

This clustering of beliefs is reminiscent of what is seen in a real group of individuals forming varying sets of beliefs about a subject and is reflective of what would be expected from similarity-based social models [35], further validating the model.

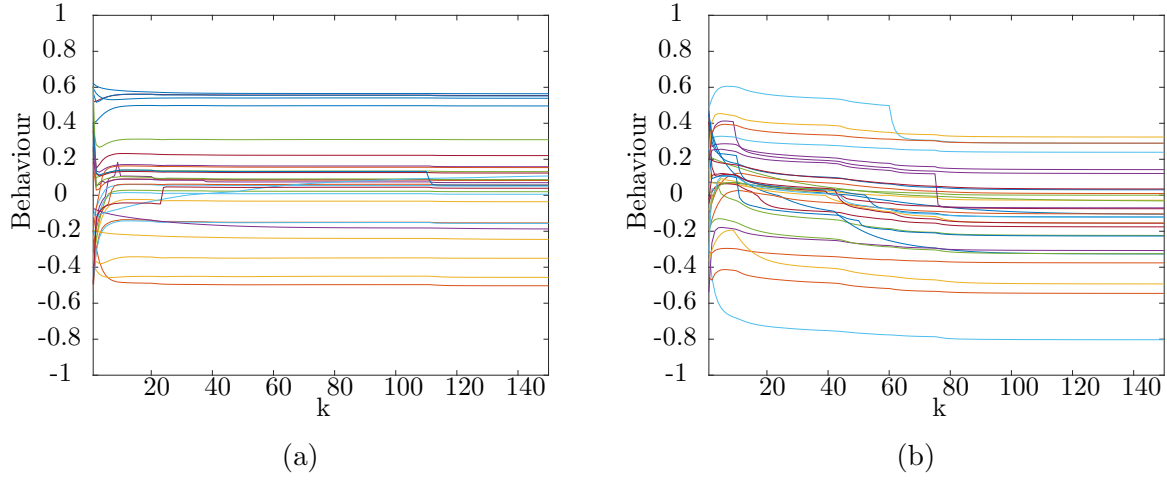


Figure 4.9: Clustering of behaviours

4.4 Summary

From the results in this chapter, we can conclude that using the mathematical models of the reasoned action approach and social influence, established in Sections 3.1 and 3.2, the derived model produces results that are consistent with the underlying psychological theories. Additionally, attributes that were not directly accounted for, such as group polarisation and clustering, appear in the results, further confirming the model's validity.

Chapter 5

Exploration of External Influence Control Techniques

Now that we have established that the model developed in Section 3 is a reasonable method for simulating the behaviour of individuals in a group environment, we can explore a variety of simulation scenarios with this model. A goal of this research is to evaluate and discover what techniques are effective at influencing groups to take up beneficial behaviours. In Section 3.2 it was established that an external source of influence, also called a control agent, such as an instructor or a manager, is a means by which individuals can be presented information to convince them to participate in a behaviour. With that in mind, we evaluate the effects of four types of external influence techniques: influencing all individuals in a group at the same time (“one to all”), focusing on influencing the most influential individuals in a group (“influencer”), focusing on one different person at a time (“rotating”), and a variety of targeted techniques that focuses on individuals that meet certain criteria (“targeted”). An illustration of these techniques can be seen in Figure 5.1. To measure the effectiveness of each influence technique we will look at the distribution of the number of agents that are performing a behaviour sufficiently ($B > 0$). Techniques that result in more agents with $B > 0$ will be considered more effective. For the purposes of this thesis, we make the following assumption,

Assumption 5.1. *The control agent is able to continuously exert a constant influence on the other agents in the system.*

In all simulations, the control agent will exert the maximal influence on a belief (i.e. the agent’s belief will be 1 for all beliefs)

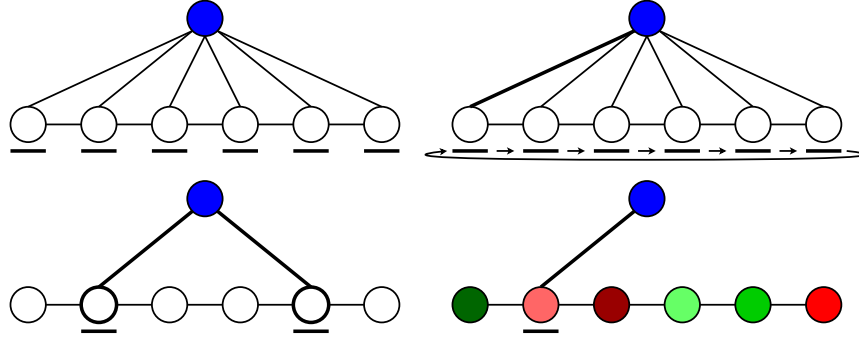


Figure 5.1: Top Left: control agent weakly influencing many individuals (one to all). Top Right: control agent interacting with individuals one at a time in rotation, identified by moving underlines (rotating). Bottom Left: control agent strongly influencing few individuals, who in turn influence more individuals (influencer). Bottom Right: control agent strongly influencing individuals with a key attribute (targeted)

5.1 Influence Techniques

Here we describe each of the four techniques.

5.1.1 One to All

For the first influence technique, we look at a technique where a control agent is trying to convince their entire group, simultaneously, of the benefits of a behaviour. In the “one to all” technique, the control agent has less ability to convince any one individual in the group; that is, the control agent has less influence on each individual due to a division of their social impact [49]. To account for this effect, we set the amount of influence the control agent to be inversely proportional to the number of individuals they interact with.

5.1.2 Influencers

The second technique examined takes advantage of the existing influential members in the group to allow the control agent to more easily spread beliefs. The control agent will focus on a subset of individuals that are the most influential in the group, measured by finding agents with the maximum $\omega_{i,j}$ to all agents in the system. More influential individuals can sway the beliefs of their peers more easily than less influential individuals. In this

technique, the control agent focuses their attention on only these m individuals, with the hope that, by more effectively convincing the influential members that the behaviour is beneficial, these members can then convince their peers of its benefit. By approaching only these select individuals, the control agent’s relative influence is kept high, since their attention is not being spread thin by attending to all of the people in the group.

5.1.3 Rotating

The third technique is one where the control agent spends time with one individual at each step k , repeatedly cycling through all of the individuals. This allows the control agent to convince each individual to the best of their ability for the short time they spend with them. However, this technique poses the risk of the individual reverting to their original state in the time between visits.

5.1.4 Targeted

The final technique being examined is that of targeting individuals with key state attributes. If an individual meets the state requirements then the control agent focuses on that agent until the agent is convinced of the behaviour’s benefits. We consider four versions of the targeted technique.

The first version is a belief-based technique, as introduced in [26], inspired by the impact of internal bias shown in Section 4.3 to try and maximise the number of individuals that flip from having negative behavioural beliefs to positive behavioural beliefs. To achieve these flips, this technique identifies individuals that have negative *beliefs* but are close to flipping (i.e., the least negative beliefs). The control agent should, in theory, be able to flip these individuals’ beliefs with little effort, since they are close to changing their minds already. Additionally, this technique relies on the idea that, once an individual has positive beliefs, we can count on their internal biases to reinforce their beliefs and sustain their now positive attitude towards the behaviour.

The second targeted technique extends on the belief technique to target an agent’s behaviour. This technique identifies and influences individuals with *behaviour* that is close to flipping from negative to positive. The appeal of this technique is that, unlike the first targeted technique, the control agent needs to measure only the behaviour of agents which is more practical to implement in a real-life situation.

Finally, the third and fourth targeted techniques are both *bias* based approaches. As we demonstrated in Chapter 4, internal biases have a strong impact on the ability to change an

individual's behaviour. With this in mind, we employ a technique that targets individuals who are not performing the behaviour and meet certain bias criteria. In one technique we target individuals with the lowest bias. The idea behind this technique is that it should be easier for a control agent to convince low-bias individuals to change behaviours; and, once they have changed beliefs, the control agent now has additional sources of social pressure to convince individuals with higher bias to change their behaviours as well. In the other bias targeting technique we identify individuals with high bias. The idea behind this technique is that high-bias individuals who can be influenced to change their behaviour will more strongly remain positive about a behaviour because of their high bias. They now also provide additional pressure on those still not performing the behaviour.

5.2 Theoretical Results

Having established that the control objective of this section is to drive the behaviour of as many agents as possible to something positive (i.e. $B > 0$), an obvious question arises: is this ever possible? The following lemma establishes sufficient conditions to guarantee that a control agent will be able to persuade a target agent to switch from $B < 0$ to $B > 0$ and remain there with consistent persuasion.

Lemma 3. *Consider the system in (3.4.1) with parameter conditions shown in Table 3.1 and a target agent i that has some negative belief, be it a behavioural belief, normative belief, or control belief at time step $\tilde{k}$, denoted by o_i (i.e. $o_i[\tilde{k}] < 0$). We include an additional control agent p as an external source of influence. This agent is like other agents in the model; however, they are not influenced by other agents in the system and are in full control of their states. We assume that the control agent p exerts the maximum positive belief to influence the target agents (i.e. $\hat{b}_{p,l} = B_p = A_p = \hat{c}_{p,l} = 1$). Then, if the suggestibility of the target agent (μ_i) is large enough, their internal bias (h_i) is low enough, and the relative weight of influence of the control agent ($\omega_{i,p}$) is large enough and meets the condition $1 \geq \omega_{i,p} > \frac{2-\mu_i}{2-h_i}$, there exists some K such that $o_i[k] > 0$ for all $k \geq K$.*

Proof. Under the assumptions of the lemma the equations of the system (3.4.1) have the

following equivalent form:

$$M_i[k]_{st} = \text{sat}(M_i[k] + \text{sgn}(o_i[k])h_i), \quad M_i[k]_{st} \in [-1, 1] \quad (5.1)$$

$$\hat{o}_j[k]_{st} = \text{sat}(\hat{o}_j[k] + \text{sgn}(o_i[k])h_i), \quad \hat{o}_j[k]_{st} \in [-1, 1] \quad (5.2)$$

$$\hat{o}_p[k]_{st} = \text{sat}(\hat{o}_p[k] + \text{sgn}(o_i[k])h_i), \quad \hat{o}_p[k]_{st} \in [-1, 1] \quad (5.3)$$

$$\hat{o}_i[k+1] = o_i[k] + \mu_i(M_i[k]_{st} - o_i[k]), \quad (5.4)$$

$$o_i[k+1] = \hat{o}_i[k+1] + \mu_i \sum_j \omega_{i,j}(\hat{o}_j[k+1]_{st} - \hat{o}_i[k+1]) + \mu_i \omega_{i,p}(\hat{o}_p[k]_{st} - \hat{o}_i[k+1]), \quad (5.5)$$

with

$$h_i \in [0, 1], \quad \mu_i \in (0, 1), \quad o_i, M_i \in [-1, 1], \quad \sum_j \omega_{i,j} + \omega_{i,p} = 1, \quad (5.6)$$

where M represents any feedback terms in the system, and o_i represents a belief of agent i in the RAA model. Conditions in (5.6) match those of the system, shown in Table 3.1.

We can rewrite equations (5.4) and (5.5), using the fact that $\sum_j \omega_{i,j} + \omega_{i,p} = 1$, as

$$\begin{aligned} \hat{o}_i[k+1] &= (1 - \mu_i)o_i[k] + \mu_i M_i[k]_{st}, \\ o_i[k+1] &= (1 - \mu_i)\hat{o}_i[k+1] + \mu_i \sum_j \omega_{i,j}\hat{o}_j[k+1]_{st} + \mu_i \omega_{i,p}\hat{o}_p[k]_{st}. \end{aligned}$$

Combining these equations, we get

$$o_i[k+1] = (1 - \mu_i)^2 o_i[k] + (\mu_i - \mu_i^2) M_i[k]_{st} + \mu_i \sum_j \omega_{i,j} \hat{o}_j[k+1]_{st} + \mu_i \omega_{i,p} \hat{o}_p[k]_{st}. \quad (5.7)$$

We now look at the difference between $o_i[k+1]$ and $o_i[k]$,

$$\begin{aligned} o_i[k+1] - o_i[k] &= (-2\mu_i + \mu_i^2)o_i[k] + (\mu_i - \mu_i^2)M_i[k]_{st} \\ &\quad + \mu_i \sum_j \omega_{i,j}\hat{o}_j[k+1]_{st} + \mu_i \omega_{i,p}\hat{o}_p[k]_{st}. \end{aligned} \quad (5.8)$$

Since we have full control over agent p we set $\hat{o}_p[k] = 1$ for all k . Minimizing equations (5.1)-(5.3) we get

$$\hat{o}_p[k]_{st} = 1 - h_i, \quad (5.9)$$

$$M_i[k]_{st} = -1, \quad (5.10)$$

$$\hat{o}_j[k]_{st} = -1. \quad (5.11)$$

Lower bounding (5.8) using these saturated values, and the fact that $\sum_j \omega_{i,j} = 1 - \omega_{i,p}$, we get

$$\begin{aligned} o_i[k+1] - o_i[k] &\geq (-2\mu_i + \mu_i^2)o_i[k] + (\mu_i - \mu_i^2)(-1) + \mu_i(1 - \omega_{i,p})(-1) + \mu_i\omega_{i,p}(1 - h_i) \\ &\geq (-2\mu_i + \mu_i^2)o_i[k] + (\mu_i - 2)\mu_i + (2 - h_i)\mu_i\omega_{i,p}. \end{aligned}$$

If $o_i[k] < 0$ and $\omega_{i,p} > \frac{2-\mu_i}{2-h_i}$ then

$$\begin{aligned} o_i[k+1] - o_i[k] &> (-2\mu_i + \mu_i^2)o_i[k] + (\mu_i - 2)\mu_i + (2 - h_i)\mu_i \frac{2-\mu_i}{2-h_i} \\ o_i[k+1] - o_i[k] &> (-2\mu_i + \mu_i^2)o_i[k] > 0 \end{aligned}$$

Therefore, if $o_i[k] < 0$, $\hat{o}_p[k] = 1$, and $\omega_{i,p} > \frac{2-\mu_i}{2-h_i}$, then $o_i[k+1] > o_i[k]$. Assuming these conditions hold, we can define $\tilde{A}$ as

$$\tilde{A} = (\mu_i - 2)\mu_i + (2 - h_i)\mu_i\omega_{i,p}.$$

We can now define K such that $K > 1$ and $o_i[\tilde{k} + a] < 0$ for all $a \in [0, K)$. From this, we can see that

$$o_i[\tilde{k} + a] > o_i[\tilde{k}] + a\tilde{A}.$$

If we define $\bar{K} = \lceil \frac{-o_i[\tilde{k}]}{\tilde{A}} \rceil > 0$, then

$$o_i[\tilde{k} + \bar{K}] > o_i[\tilde{k}] - \frac{o_i[\tilde{k}]}{\tilde{A}}\tilde{A} = 0$$

Therefore, we know that $1 < K \leq \bar{K}$, i.e. K exists and is less than $\bar{K}$.

Now that we have shown that with $\hat{o}_p[k] = 1$, and $\omega_{i,p} > \frac{2-\mu_i}{2-h_i}$ a control agent can change a target agent's belief from negative to positive given enough time we need to show that this positive belief will remain.

Taking equation (5.7) and applying the lower bound from taking conditions (5.9)-(5.11) we get

$$o_i[k+1] \geq (1 - \mu_i)^2 o_i[k] + (\mu_i - 2)\mu_i + (2 - h_i)\mu_i\omega_{i,p}$$

if we assume that $o_i[k] > 0$ and that $\omega_{i,p} > \frac{2-\mu_i}{2-h_i}$ holds, then

$$o_i[k+1] > (1 - \mu_i)^2 o_i[k] + (\mu_i - 2)\mu_i + (2 - h_i)\mu_i \frac{2-\mu_i}{2-h_i} > 0.$$

Therefore, $o_i[k] > 0$ for all k . Thus a control agent will be able to drive a target agent i 's beliefs to $o_i[k] > 0$ for all $k \geq K$ if the control agent has $\hat{o}_p[k] = 1$ and $\omega_{i,p} > \frac{2-\mu_i}{2-h_i}$. $\square$

By Lemma 3, the control agent will be able to drive any target agent i 's beliefs to be positive at some time step K if they meet the sufficient criteria. Furthermore, since behaviour is a weighted average of beliefs, the control agent will also be able to convince any target agent to have $B_i[k] > 0$ for all $k \geq K$. For the “one to all” technique, since the control agent is applying pressure to every agent in the system, as long as the control agent's relative influence for every agent in the system meets the criteria of Lemma 3, they will be able to convince all agents to perform the behaviour regularly (i.e. $B_i > 0$ for all i) given enough time. The other influence techniques involve targeting a single agent and so, as long as the above criteria are met, the control agent will be able to convince any given target over time.

5.3 Simulations

For this section, we use the model to simulate a classroom environment, as an example. In this environment there is a class of 30 students that are fully connected, i.e. all students can communicate with all other students in the class. An instructor acts as an external source of influence that is trying to convince their students that brainstorming is beneficial. Brainstorming has been chosen as a target behaviour because it has been shown to lead to a greater breadth of quality ideas and is seen as a useful skill for students [2]. There are two behavioural beliefs selected (“brainstorming will lead to higher quality ideas” and “brainstorming does not take time away from completing a project”), two normative beliefs (“my peers are performing brainstorming” and “my peers believe that brainstorming is beneficial”), and one control belief (“I can access resources to achieve the behaviour”). Each simulation was run for $k = 0, \dots, 200$ to allow for the effects of the influence techniques to take hold, where k is a class day where all students have had the chance to communicate with each other. Further, each influence technique was simulated 100,000 times with random initial conditions and parameters. These values were selected from a uniform distribution spanning the parameter domains, indicated in Table 3.1. It is important to note that, while this simulation uses a classroom for simply explaining the simulation results, the results can be adapted to other small-group scenarios in which a group of individuals communicate with each other and have an external source of influence, such as a workplace or a coaching environment.

To measure the effectiveness of each influence technique we look at the distribution of the number of students in each simulation that were performing the behaviour regularly ($B > 0$) averaged over $k = 190$ to $k = 200$. Figure 5.2 shows a violin plot of the results for each set of influence techniques. The violin plot shows a histogram of the results, oriented

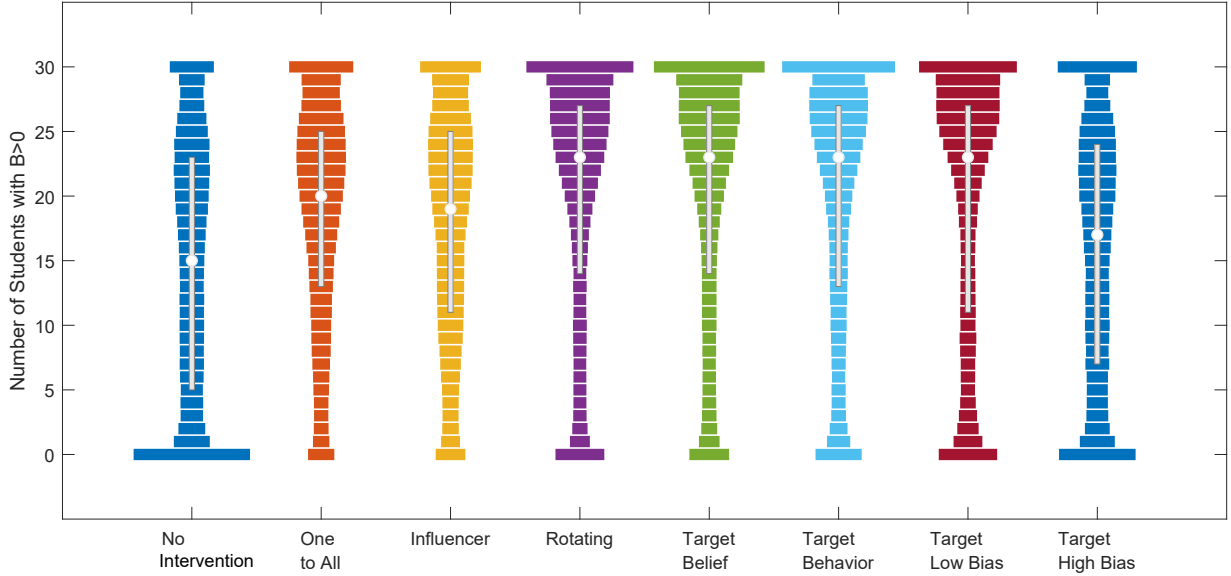


Figure 5.2: Violin plot comparing the results of different influence techniques in a class of 30 students. White dots display the median of each technique.

vertically, overlaid with a box plot of the data. The more effective an influence technique is the more students will be convinced to perform the behaviour regularly; therefore, better influence techniques will have top heavy distributions.

5.3.1 “No-Intervention” Case

We first simulate a technique with no control being applied to get a baseline value. Figure 5.2 shows that, for this technique, the median number of students performing the target behaviour regularly is 15, with the first and third quartile at 5 and 23 students respectively. This result is reasonable since the initial conditions for the system are chosen randomly from a uniform distribution, and we can expect, on average, approximately half of the 30 students with $B > 0$ in most simulations. Additionally, we see that a large number of simulations resulted in no students with $B > 0$. This can be explained by taking into account internal biases leading to polarisation effects combined with the impact of outcome feedback, which normally mediates any errant perceptions, not being considered in runs where behaviour is negative. Similarly, there is a noticeable number of runs where all 30 students are performing the behaviour regularly. This is explained by the fact that the greater the number of students performing the behaviour the more social pressure there is

to perform the behaviour. Figure 5.3 provides an example of behaviour results from one of the simulation runs.

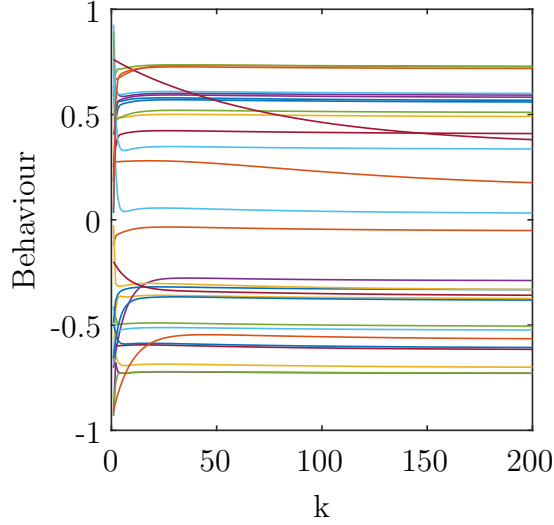


Figure 5.3: Sample behaviour for the “no-intervention” case

5.3.2 “One to All” Technique

Next, we look at the “one to all” influence technique. From Figure 5.2 we see that there is a noticeable improvement in the behaviour distribution compared to the “no-intervention” baseline. The median, first, and third quartile measure at 20, 13, and 25 students respectively. Not only does the median increase by 5 students but so do the lower and upper quartiles. This shows that trying to convince a class of 30 students of the benefits of a behaviour by lecturing the class as a whole does have some positive benefits. Even though the instructor’s influence over any given student is reduced they may still have a non-zero impact; however, as the number of students in a class increases, this technique will have a diminishing impact as the instructor’s influence is spread thinly over the class. Figure 5.4 provides an example of behaviour results from one of the simulation runs. From this figure, we can see that the technique was able to improve the belief of the students slightly when compared to the no-intervention case. The largest contribution to this improvement comes from a student who flips from having a negative belief to having a positive belief

at $k = 5$. When this happens, the now positive internal bias of the student kicks in and causes the student to quickly increase in their belief. This is a key observation.

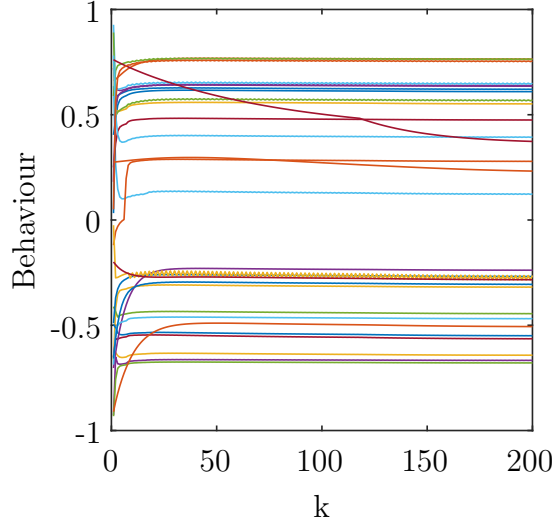


Figure 5.4: Sample behaviour for the “one to all” technique

To further illustrate the impact of internal bias on influence techniques we simulate two additional scenarios: “low-bias” and “high-bias”. In the low-bias simulations, student biases are randomly chosen with a uniform distribution spanning $0 - 0.3$, while in the high-bias simulations student biases are randomly chosen with a uniform distribution spanning $0.7 - 1$. The mixed bias simulations represent the standard “one to all” technique with biases sampled uniformly from $0 - 1$. Figure 5.5 shows a violin plot of the results for the mixed, low, and high-bias simulation scenarios. From this figure, we can see that, in the low-bias scenarios, the instructor was able to convince most students that brainstorming is beneficial to them. The median value of 30 with a large first quartile value of 25 means that 75% of the time at least 25 students ended up brainstorming. The high-bias scenarios show that it is indeed much more difficult to convince a persistent class to change their behaviour. On average only 15 students were convinced to brainstorm and there were almost always a few students who never started brainstorming. The mixed bias scenarios present an interesting middle ground between the two scenarios. It has a median value between the other two simulation scenarios, as well as the highest spread of all scenarios.

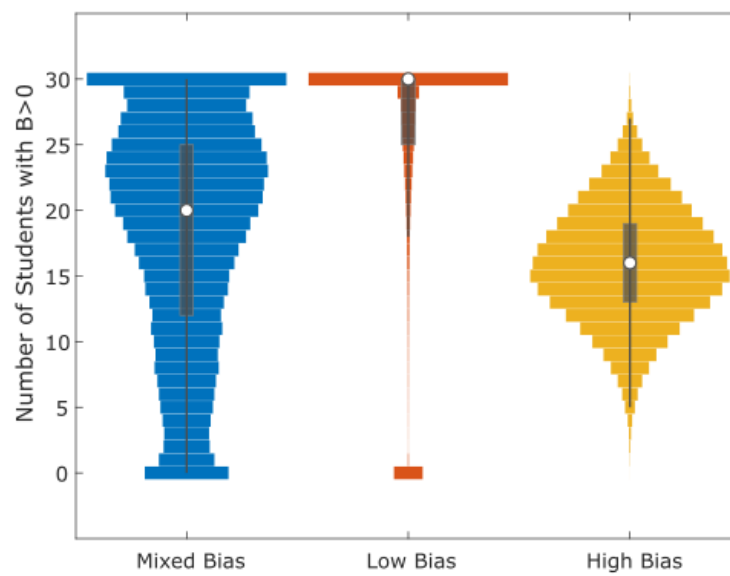


Figure 5.5: Distribution of behaviours in a 30-student classroom, while under instructor influence, with mixed, low, and high bias values. Measured using the number of students, out of 30, who have positive behaviour.

5.3.3 “Influencer” Technique

The “influencer” technique was simulated with the instructor targeting the four most influential students in the class. Figure 5.2 shows that the “influencer” technique performs slightly worse compared to the “one to all” technique, with median, first, and third quartile values of 19, 11, and 25 students. Relying on the influence of a few prominent members of a class to propagate positive beliefs does not appear to be as effective as the generalised “one to all” influence technique. From the sample run shown in Figure 5.6, we see the “influencer” influence technique performs slightly worse than the “one to all” technique. Similar to the “one to all” technique, the students improve in their beliefs, with the largest contribution happening from a student flipping from negative to positive belief.

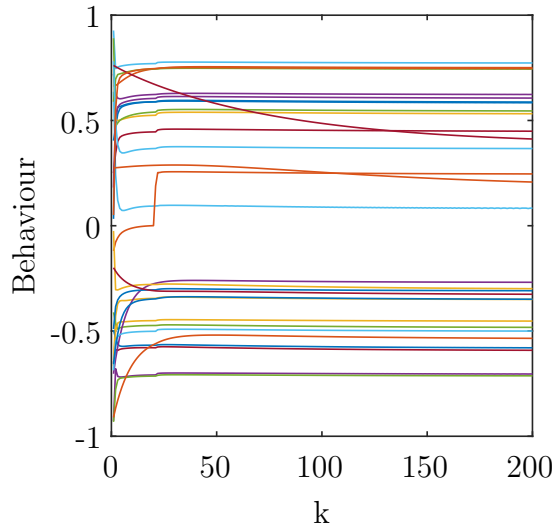


Figure 5.6: Sample behaviour for the “influencers” influence technique

5.3.4 “Rotating” Technique

The “rotating” technique shows a large improvement over the previously mentioned influence techniques. Figure 5.2 shows that this technique has a median value of 23 students performing the behaviour regularly, with first and third-quartile values of 14 and 27 students. The “rotating” technique results in more students with $B > 0$ more often, with an upper tail heavy distribution. This implies that spending time with a different student

each day can be more effective at behavioural change compared to the traditional “one to all”, stand in front of the class and lecture, technique. The “rotating” influence technique may suffer when the class is made up of more students as the instructor will spend less time per student throughout the simulation, allowing the student’s beliefs to regress in the meantime. Figure 5.7 demonstrates how the “rotating” technique operates. At first glance, this graph is very noisy, but upon closer inspection, it is clear that there are large, periodic increases in a student’s belief whenever the instructor spends time with them. The instructor can impart a large amount of influence on a student, causing their belief to spike. As the instructor moves over to a different student this spike diminishes over time, until the instructor comes by once again. These large spikes in belief can occasionally cause students to flip from a negative belief to a positive one leading to the bias effects previously demonstrated. With the “rotating” technique, the instructor’s large influence is, in general, able to pull more students over from having a negative belief to a positive belief.

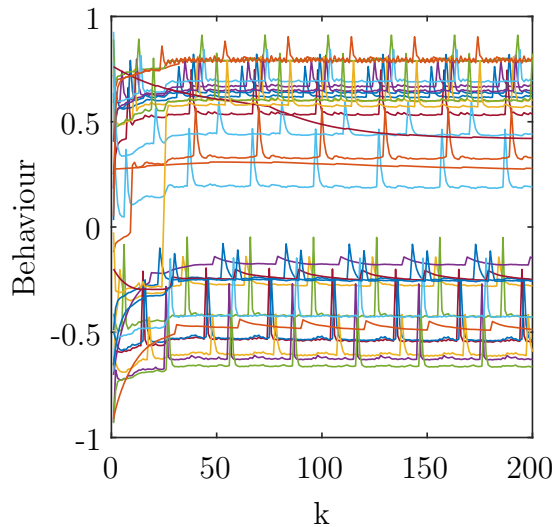


Figure 5.7: Sample behaviour for the “rotating” technique

5.3.5 “Targeted” Techniques

The “target belief” technique results in large improvements that are similar to those of the “rotating” technique, with identical median, first quartile, and third quartile values

of 23, 14, and 27 students respectively, as seen in Figure 5.2. These improvements show the effectiveness of exploiting students’ internal biases. As a student begins to believe that brainstorming is good for them, they interpret information as supporting their belief, further reinforcing the benefits of brainstorming to the student. This technique would not suffer as much when more students are added to a class since it targets key students; however, as this technique is more difficult to implement (identifying students that are on the verge of switching beliefs can be difficult) the “rotating” technique, which is straightforward to implement, might be preferred.

We would expect the “target behaviour” technique to perform similarly to the “target belief” technique, and the median and third quartile values in Figure 5.2 show that it does with values of 23 and 27 students, respectively. However, we see that the “target behaviour” technique has a slightly lower first quartile at 13 students. This can be explained by the fact that behaviour is not directly affected by internal biases. Internal biases affect the underlying beliefs that cause behaviour and so if a student begins to perform a behaviour regularly, it is not guaranteed that their internal biases will reinforce the behaviour. For example, a student could have a strong positive belief that brainstorming leads to better ideas, but at the same time, the student might have a strong belief that brainstorming can take time away from completing a task. These two beliefs could balance each other leading to a behaviour only slightly below zero. Now, with instructor influence, the student’s beliefs both slightly improve; they think that brainstorming is still good for ideas and is not as bad at taking time away from an activity, though it still takes some time. Since the student’s beliefs have not flipped sides their biases remain and so the negative bias about the time aspect of brainstorming takes hold and pulls the student back into not performing the behaviour. In short, this technique has a harder time maintaining a student’s positivity towards a behaviour when compared to the “target belief” technique. However, when the behaviour technique does lead to students maintaining positive attitudes about the behaviour it tends to pull more students to do so as well; this can be attributed to the fact that this technique ensures that a student is performing the behaviour before switching to a different student.

Finally, we look at the two “target bias” techniques. From Figure 5.2 it is evident that the influence technique that targets low-bias students is significantly more effective than the influence technique that targets high-bias students. The low-bias technique has median, first quartile, and third quartile values of 23, 11, and 27 students and performs almost as well as the “target belief” and “target behaviour” techniques, though with a higher spread. The high-bias technique on the other hand has median, first quartile, and third quartile values of 17, 7, and 24 and performs only a little better than the “no-intervention” case. With the low-bias technique, the instructor can more easily get students to flip behaviours,

as their low bias means the instructor is spending less time fighting biases. However, once those students flipped they had a higher chance of switching back because their low bias values would not be as effective compared to other techniques, leading to a larger spread in the distribution. The high-bias technique, on the other hand, proves difficult for the instructor to even begin to switch student’s behaviours, leading to a low median value. This again shows that targeting low-bias individuals does have some meaningful benefit as a concept and could theoretically be used in conjunction with other techniques for added gain.

5.4 Summary

In this chapter, we used the model developed in Chapter 3 to analyse the impact of several intervention influence techniques that can be employed by control agents trying to convince a group to take up desired behaviours. We were able to derive a sufficient requirement to ensure that an external influence would be able to convince an agent to change behaviour; if the control agent has relative weight $w_{i,p} > \frac{2-\mu_i}{2-h_i}$ then they will be able to change agent behaviour from $B < 0$ to $B \geq 0$. With this result we examined the effectiveness of seven different influence techniques: “one to all”, “influencer”, “rotating”, “targeted belief”, “targeted behaviour”, “targeted low-bias”, and “targeted high-bias”. We observed that the “rotating”, “targeted belief”, and “targeted behaviour” techniques had a marked improvement in behavioural change when compared to a “one to all” and even an “influencer” technique. Simulations for the “bias targeted” techniques showed that identifying individuals with low biases is an important technique for convincing more people about the benefits of a behaviour and it is advisable to target these individuals over high-bias individuals. The conclusions about the relative strengths of each intervention influence technique, in general, depend on the number of agents in the system n as well as the distribution of initial conditions and parameters. An advantage of having a model is that simulations can be easily rerun for each particular situation.

Chapter 6

Experimental Findings

The influence techniques explored in Chapter 5 produce interesting results. Namely, they display that “targeted” and “rotating” influence techniques appear to be more effective at convincing students to take up brainstorming behaviours than a more traditional “one to all” classroom setup of presenting at the front of the class. To further validate the efficacy of the model we test these influence techniques in a real-world scenario. This scenario is first simulated using some of the various influence techniques discussed previously. The test scenario is then implemented in the real world and the trends are then compared to the simulation results [28].

The scenario chosen is the brainstorming classroom scenario previously discussed, adapted to a practical application in a real classroom. In this scenario, classes are divided into groups and are asked to complete a series of team activities. Class sections are each advised on the benefits of brainstorming using different influence techniques before they begin the team activity. The results of their brainstorming activity are collected at the end of the activity and the experiment is repeated with a different activity at a later date.

6.1 Simulation Results

We first simulate the effects of three influence techniques an instructor might use to influence students in a classroom and discuss the relative performance of these results when compared to a control scenario where no intervention technique is applied. The influence techniques analysed are the “one to all” technique to replicate a “normal” approach to teaching, the “rotating” technique due to its effectiveness shown in Chapter 5, and finally,

the “targeted belief” technique, again because of the effectiveness displayed previously. The equations in (3.4.1) describe the model used for the simulations. The simulation is set up to mimic what can be realistically achieved in a real-world classroom scenario.

The simulation is set to represent a classroom environment with 140 students. These students are divided into 35 groups of four. We assume students are more influenced by others in their group and have only a few influences from their peers outside their group, simulated by setting a 10% chance a connection is made outside of the group. The behaviour being investigated is brainstorming and consists of two assumed salient behavioural beliefs (“brainstorming will lead to higher quality ideas” and “brainstorming does not take time away from completing a project”), two normative beliefs (“my peers are brainstorming” and “my peers believe brainstorming is good”), and one control belief (“I can access resources to brainstorming”). An instructor is included as an additional agent in the system whose beliefs are unaffected by the students in the classroom and who can fully control their state. In these simulations, a simulation step k represents a moment where all students in the system have a unique set of interactions with their group. For example, during the course of a school term, a student will have interactions with their class every day they have a class, and so, an interaction will represent each of these class days.

The simulation is devised to mimic the experimental setup described in Section 6.2, to follow. The simulation is set to emulate students going through three separate team activities. At the beginning of these team activities, the instructor influences the students via one of the three previously stated influence techniques. The students then go on to perform the activity and any amount of brainstorming they desire. Finally, the students receive the outcomes of performing their behaviour at the end of the activity. This is then repeated for the next team activity.

To get an overview of how the stated influence techniques perform, each technique is simulated 100,000 times with randomised initial conditions and parameters spanning the parameter domains, indicated in Table 3.1. Every student’s initial beliefs (b, c, n), belief weights (e, m, p), internal bias (h), suggestibility (μ), and the relative influence of students on one another (ω) are selected from a uniform distribution. By observing the distribution of all of these results we can get a view of how these techniques tend to perform. For our analysis, we look at the change in behaviour from a baseline, taken after Activity 1, to measure the relative improvement in behaviour caused by the implemented influence technique. The results of these simulations can be found in Figure 6.1 as violin plots of the data which displays a vertically oriented distribution of the change in brainstorming behaviour from Activity 1 to Activity 2 and from Activity 1 to Activity 3 for all simulated techniques; the distributions are overlaid with a box-plot of the data. The left axis of Figure

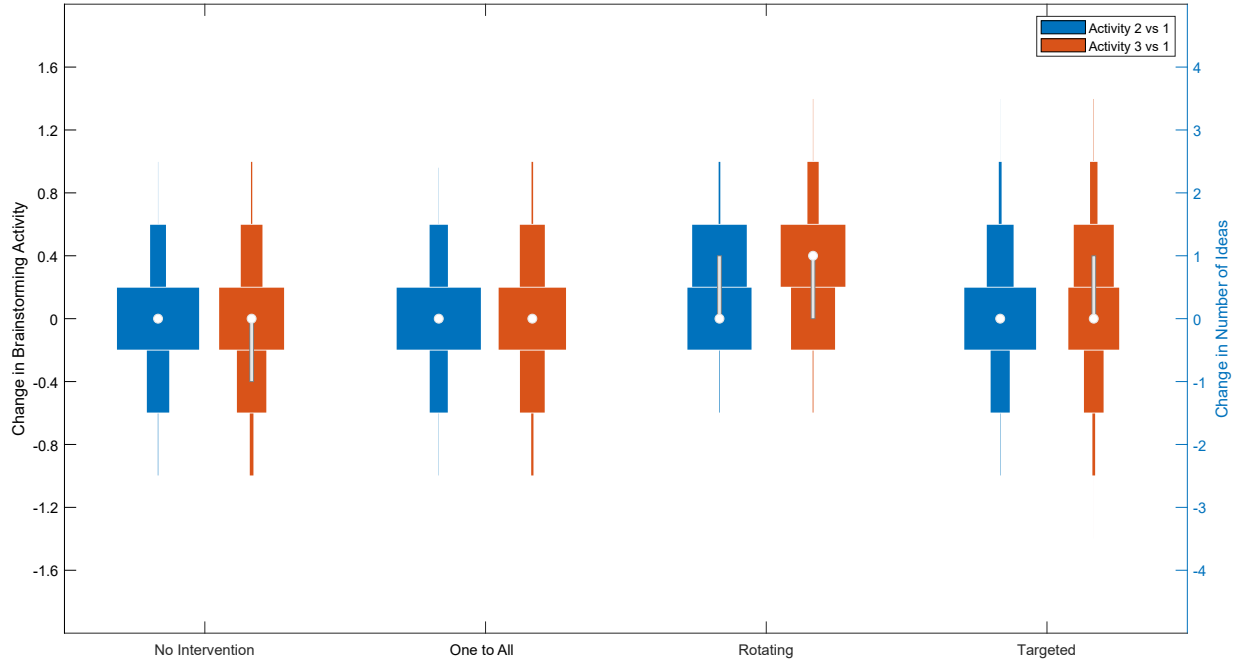


Figure 6.1: Violin plot comparing the simulated change in number of ideas between Activity 1 to Activity 2 and from Activity 1 to Activity 3.

6.1 shows the value of the change in behaviour B for each team activity, as outputted from the model, while the right axis shows an equivalent conversion to the change in number of ideas when normalised to match results collected from experimental testing in Section 6.2.

6.1.1 “No Intervention” Case

To begin, a set of simulations is performed with no intervention applied. These results provide a baseline for any trends. The students are allowed to evolve their opinions on brainstorming with no external influences present. From Figure 6.1 we see that for Activity 2 the students mostly do not change their behaviour compared to Activity 1, with a median of 0. However, on the third activity, the students begin to reduce the amount of brainstorming on average. There are some scenarios where students increase the amount of brainstorming they perform; however, we can see that the lower quartile of the distribution grows longer, indicating more students reduce their performance of brainstorming. This reduction is explained by students constantly spending time brainstorming but only observing any benefits of brainstorming when results are presented at the end of the activity.

So, in general, a student is more exposed to the disadvantages of brainstorming than to the advantages.

6.1.2 “One to All” Influence

Next, the “one to all” influence technique is simulated. In this simulation scenario, the instructor tries to convince the whole class simultaneously of the benefits of brainstorming. However, because the instructor is dividing their effort among a large number of people, they have less ability to convince any one student in the class due to a division of their social impact [49]. To account for this effect, we set the amount of influence the instructor has on each student to be inversely proportional to the number of students they interact with. Activity 1 was simulated with no intervention applied to get baseline data for the class, to represent what will be done in the real-life experiment described in Section 6.2. Following this, Activities 2 and 3 had interventions applied at the start of the activity, after which behaviour is measured and outcome feedback is presented. Figure 6.1 shows the resulting distribution in the change in behaviour for the simulations. From this figure, we see that the “one to all” technique performs very similarly to the “no intervention” case through all team activities. After Activity 2 there is no change in behaviour in most students, similar to that seen in the “no intervention” case. However, after Activity 3 there is not as large a drop in behaviour compared to the “no intervention” case, leading to a marginal improvement in performance. Overall, the “one to all” influence technique is ineffective at changing the students’ behaviour.

6.1.3 “Rotating” Influence

The “rotating” influence technique is one where the instructor influences groups of four students at a time. The instructor cycles through all 35 groups in the class before they begin a team activity. Because the instructor is influencing fewer students than they do with the “one to all” technique they can exert more influence on students in each group. However, there is a potential drawback of students reverting to their pre-influenced states once the instructor is no longer present. As with the “one to all” technique, Activity 1 is used as a baseline and no intervention is applied. Figure 6.1 shows that the “rotating” influence technique is much more effective at improving the amount of time students spend brainstorming compared to the “no intervention” and “one to all” techniques. After Activity 2 the upper quartile is higher when compared to the other influence techniques and this performance is further increased after Activity 3, with the median behaviour improving to 0.5.

6.1.4 “Targeted” Influence

The final intervention being simulated is the “belief targeted” influence technique described in Section 5.1. In the “targeted” technique the instructor identifies the student with negative beliefs that are closest to flipping opinions (i.e., the least negative beliefs). The instructor should be able to flip this student’s beliefs with ease since the student is close to changing their mind already. Additionally, once the student has flipped opinions, their internal biases will reinforce their newfound beliefs and sustain their attitude. This technique would allow instructors to focus on one individual student at a time but at the cost of implementing an appropriate measurement instrument to identify this student. Figure 6.1 shows the resulting simulation distribution for this technique. From this figure, we can see that the “targeted” technique performs well. After Activity 2 there is not much of an increase in behaviour, though the upper portion of the distribution is larger than the lower portion. After Activity 3, however, there is a marked increase in brainstorming by students with the upper quartile increasing. This shows the importance of influencing students who are ambivalent about performing a behaviour. A caveat of this “targeted” technique is the difficulty of implementing the strategy in the real world. To measure every student’s beliefs would require significant surveying and is very time-intensive. Therefore, although the “targeted” technique shows good potential in simulation, its difficulty in implementation combined with the limited number of test groups available led it to be excluded in experimental testing.

6.2 Experimental Findings

With the simulation knowledge at hand, we now implement the influence techniques in a real-world classroom to verify if the model’s predictions are accurate. We were able to apply these influence techniques to students in a freshman Electrical and Computer Engineering class on engineering profession and practice, at the University of Waterloo. The data collected in this study was provided as a secondary use of information and received ethics clearance from the Office of Research Ethics at the University of Waterloo.

6.2.1 Procedure

The class consisted of 401 students with three sections of 126, 150, and 125 students respectively. The students in each section were then organised randomly into teams of

three or four, based on availability, for their team activities with Section 1 having 37 groups, Section 2 having 38 groups, and Section 3 having 33 groups.

Prior to any team activities, students were familiarised with the concept of brainstorming and its phases, such as the synthesis and analysis phases, were described by their instructor, via a thirty minute online lecture on engineering design as described in [5]. Each section was then asked to participate in three team activities over one four-month term. The team activities involved a design problem that needed to be solved by constructing an object with limited provided resources; the first activity involved building a rocket using a balloon and a paper cup, the second a tower of paper that needed to support a marshmallow, and the third a paper bridge that can support a rolling ball. Each team activity took place over 1 hour, 20 minutes of which were used by students to develop and construct ideas. At the end of the activity, students were ranked based on the performance of their construction; no marks were awarded for participation to remove the effects of external rewards.

Each group was asked to write down their group number and all ideas they had during the activity on a poster board that was collected at the end of the team activity. The number of ideas recorded on the boards was used as an approximate proxy to quantify the amount of brainstorming each group performed. This has limitations in that not all students would write down their ideas and some would discuss them instead. Due to the practical constraints of not being able to extensively measure beliefs, only two of the three simulated influence techniques were applied: the “one to all” technique, applied to Section 1, and the “rotating” technique, applied to Section 2. Section 3 was used as a control (no intervention) case. All three sections were led by the same instructor.

During the first team activity, no interventions were applied to all three sections to get a baseline value for the number of ideas each group would write down. Following this, during the second and third team activities, Section 1 continued to be undisturbed as a control group while Sections 2 and 3 had their respective influence techniques applied. The effectiveness of these methods is evaluated by comparing the change in the number of ideas each group wrote down on their poster boards. The talking points used by the instructor for influenced sections, before the second and third activities, to convince students of the benefits of brainstorming were as follows:

- **Team Activity 2:** “Try to brainstorm before putting things together. Brainstorming is the process of coming up with multiple ideas, before choosing one to go with. This is known to create better designs. Do this before you start building your final prototype.”

- **Team Activity 3:** “As you saw in the Engineering Design lecture, you should always brainstorm before building your prototype. This is a very important part of design. Make sure you come up with several new ideas before building the final prototype.”

Each team activity began with the instructor asking students to sit with their groups in pre assigned locations. The instructor then went on to describe the goals of the activity for the day. Students were instructed to not use any external devices, such as cell phones or laptops, and only use the materials provided by the instructor to complete their task. They were then asked to write their group numbers on their poster boards and write down all the ideas they had regarding the team activity onto the board. Students were then told they could begin the team activity. Once the students started any desired influence techniques were then deployed using the script mentioned above. The number of ideas written on a bristle board was qualitatively measured by identifying unique concepts and diagrams. Specific procedures for each section are described in the subsections bellow.

“No Intervention” case

For the “no intervention” case, students were brought into the classroom, the activity for the day was described, and the students were all asked to record any ideas they may have regarding their task on a provided poster board. No information was given regarding brainstorming. Once the team activity was completed the poster boards were collected and the number of ideas listed was measured and recorded.

“One to All” Influence

For the “one to all” influence technique, once the activity had been described and the students began the activity, the instructor took a traditional, front-of-the-class approach to encouraging students. At the beginning of Activity 2 and 3, the instructor spent one minute reminding the class of the benefits of brainstorming to generate more creative ideas, using the script described above. The students went on to participate in the activity and the number of ideas was again recorded.

“Rotating” Influence

In the third, “rotating” technique, the instructor was more involved with encouraging brainstorming. Once the activity for the day was described, and students started their

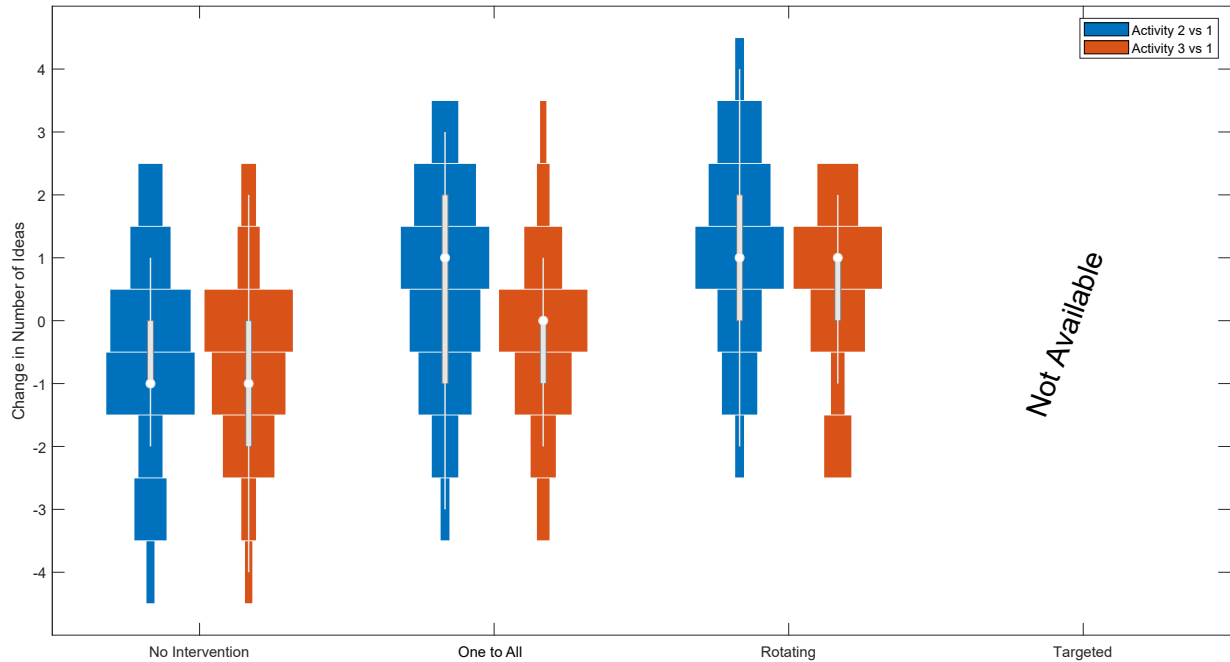


Figure 6.2: Violin plots comparing the experimental change in the number of ideas between Activity 1 to Activity 2 and from Activity 1 to Activity 3.

activity, the instructor began Activities 2 and 3 by visiting each group individually, for approximately 15 seconds, to discuss the benefits of brainstorming using the script mentioned above. The students continued to participate in the activity and the number of ideas was recorded. As a result of the way this intervention was deployed some groups of students had less time post-intervention, and more time pre-intervention, to complete the activity, which is a limitation of this approach.

6.2.2 Results

The distribution of the change in the number of ideas for all the experiments is shown in Figure 6.2 as a violin plot.

“No Intervention” case

For the “no intervention” case, Figure 6.2 shows that, during Activities 2 and 3, the groups recorded fewer ideas compared to Activity 1. A possible explanation for this is that the

performance improvement as a result of brainstorming was not notable enough compared to the time spent on idea formulation, especially given the narrow time window for the activities, leading students to reduce the amount of time spent on brainstorming. The general trend of brainstorming less in these results is consistent with the trend noted in the “no intervention” simulations with brainstorming reducing over time.

“One to All” Influence

For the “one to all” influence technique, the distribution of results shown in Figure 6.2 shows that during Activity 2 groups began to list a lot more ideas compared to Activity 1, which is a large improvement over the no-intervention case. However, for Activity 3, the number of ideas falls, with most students reducing the number of written ideas, though there is still a slight improvement compared to the control case. This phenomenon is partially observed in the simulated data where the “one to all” technique resulted in a slightly positive mean change in number of ideas after Activity 2 of 0.0013, 95% CI [0.0002,0.0024], which then drops to a slightly negative value after Activity 3 of -0.0090 , 95% CI $[-0.0104, -0.0076]$ as seen in Figure 6.1. This drop may be explained by groups being initially encouraged by the instructor to brainstorm; however, seeing that the results of brainstorming were not significant when compared to the time spent, the groups began to abandon this pursuit.

“Rotating” Influence

In the third, “rotating” technique, Figure 6.2 shows that for Activity 2 groups increased the number of written ideas the most when compared to the other techniques. For Activity 3 the number of ideas fell slightly but was significantly higher than what was observed for the other influence techniques. It appears that by spending more individual time with groups the instructor was able to more effectively convince students that brainstorming is beneficial. However, this comes at the cost of additional time and energy being spent by the instructor visiting each individual group of students.

6.2.3 Statistical Analysis of Experimental Results

We wish to see if the trends of the experimental results match those of the simulations in Section 6.1. The expected trend from the simulated results is that the “rotating” technique provides the largest improvement in the number of ideas written down per group. To test

this expectation we establish a set of hypotheses. Let $\Delta x_1, \Delta x_2$, and Δx_3 be the change in the number of ideas written down by each group, compared to Activity 1, for the “no intervention” section, the “one to all” section, and the “rotating” section respectively. Let $\tilde{m}()$ be the median operator. Then,

Hypothesis 0 (H0): *The median change in the number of ideas does not differ for the “no intervention” section ($\tilde{m}(\Delta x_1)$), the “one to all” section ($\tilde{m}(\Delta x_2)$), and the “rotating” section ($\tilde{m}(\Delta x_3)$). That is, $\tilde{m}(\Delta x_1) = \tilde{m}(\Delta x_2) = \tilde{m}(\Delta x_3)$.*

Hypothesis 1 (H1): *The median change in the number of ideas of the “no intervention” section ($\tilde{m}(\Delta x_1)$), the “one to all” section ($\tilde{m}(\Delta x_2)$), and the “rotating” section ($\tilde{m}(\Delta x_3)$) are not all equal.*

If Hypothesis 1 is true then:

Hypothesis 2 (H2): *The “rotating” section has the largest increase in the median number of ideas. That is, $\tilde{m}(\Delta x_3) > \tilde{m}(\Delta x_1)$ and $\tilde{m}(\Delta x_3) > \tilde{m}(\Delta x_2)$.*

To evaluate Hypotheses H0-H2, a Kruskal-Wallis test [23] was performed on the recorded data. This test was chosen since it is a non-parametric test that does not make assumptions about the distribution of the underlying data. The change in the number of ideas each group wrote down, relative to Activity 1, was used as the response variable, with the influence technique set as the treatment group. We first look at the change in the number of ideas from Activity 1 to Activity 2; Table 6.1 provides the mean ranks for each of the influence techniques used for the Kruskal-Wallis test as well as the p-values for the pairwise comparisons for each section. When comparing the results of Activity 2, the Kruskal-Wallis test found that at least one section had a significant difference in the median change in the number of ideas with a p-value of 0.00006, consistent with Hypothesis 1. We further analysed this result by running a post-hoc Wilcoxon test with Bonferroni correction. The results show that the “one to all” and “rotating” techniques have a higher median change in ideas when compared to the “no intervention” control group, with p-values of 0.0144 and 0.00004 respectively. However, there is no statistical difference between the “one to all” and “rotating” techniques. This means that the “one to all” and “rotating” techniques led to an improvement in the median number of ideas written down by students, with mean ranks 57.72 and 69.65 respectively, when compared to the “no intervention” case, with a mean rank of 37.68. This is somewhat in line with simulated data where the “one to all” and “rotating” techniques performed better than the “no intervention” case.

We now look at the change in the number of ideas from Activity 1 to Activity 3. We again use the Kruskal-Wallis test which indicates at least one section has a significant difference in the median change in the number of ideas listed by a group, p-value of 0.0006,

Table 6.1: Results for change in number of ideas from activity 1 to activity 2

(a) Kruskal-Wallis Ranks and p-Value

Influence	NumGroups	Mean Rank (95% CI)	Std. Error
No Intervention	37	37.68 (29.11,46.24)	5.06
One to All	38	57.72 (49.27,66.18)	4.99
Rotating	33	69.65 (60.57,78.73)	5.38
p-value	0.00006		

(b) p-Values of pairwise comparisons

Influence	No Intervention	All	Rotating
No Intervention			
One to All	0.0144		
Rotating	0.00004	0.3103	

consistent with Hypothesis 1. Further running the post-hoc Wilcoxon test, the results of which are shown in Table 6.2, we find that the number of ideas when using the “rotating” technique is different from both the “no intervention” control case and the “one to all” technique with p-values of 0.00043 and 0.0395 respectively. Additionally, the mean rank of the “rotating” technique (70.35) is significantly higher than the “no intervention” (42.58) and “one to all” (52.34) techniques, which is consistent with Hypothesis 2 that the “rotating” technique has a higher change in the median number of recorded ideas when compared to the “no intervention” and “one to all” techniques.

The results of the statistical analysis, in conjunction with the distribution of results shown in Figure 6.2, show that the “rotating” influence technique is indeed more impacting at influencing students to increase brainstorming activity when compared to the “one to all” and “no intervention” control techniques. Additionally, the results are consistent with the trends found in our simulations with “rotating” leading to an increase in brainstorming, “one to all” leading to an insignificant change, and “no intervention” leading to a decline, lending credence to the usefulness and partial validity of our model.

Table 6.2: Results for change in number of ideas from activity 1 to activity 3

(a) Kruskal-Wallis Ranks and p-Value

Influence	NumGroups	Mean Rank (95% CI)	Std. Error
No Intervention	37	42.58 (34.09,51.07)	5.02
One to All	38	52.34 (43.96,60.73)	4.95
Rotating	33	70.35 (61.35,79.35)	5.31
p-value	0.0006		

(b) p-Values of pairwise comparisons

Influence	No Intervention	All	Rotating
No Intervention			
One to All	0.4984		
Rotating	0.00043	0.0395	

6.3 Summary

In this chapter, we explored the application of several influence techniques to a real-world classroom scenario where an instructor attempts to increase the use of brainstorming in their class. We evaluated the benefits of a “one to all” influence technique, where an instructor speaks to the entire class on the benefits of a behaviour, a “rotating” technique, where an instructor visits individual groups to more strongly convince them of a behaviour’s merits at the cost of more time spent, and a “targeted” influence technique, which involved the instructor identifying students on the verge of changing opinions about a behaviour and encouraging them to perform the behaviour.

Simulation results show that these influence techniques do have an impact on how many students will perform brainstorming, with most benefits achieved in the “rotating” and “targeted” techniques. The “targeted” technique takes advantage of a student’s internal bias to help maintain positive sentiment about a behaviour; however, implementing this technique is difficult since it requires being able to identify target students. With the knowledge of the simulation results, the “one to all” and “rotating” techniques were implemented, along with a “no intervention” control group, in a real-world classroom sit-

uation. Three class sections were established, each receiving one of the proposed influence techniques, and were asked to participate in three team activities. The students were asked to record their ideas for these team activities on a poster board; the number of these ideas was used to quantify brainstorming performance. This experiment resulted in students subjected to the “rotating” technique having a significantly higher increase in the number of recorded ideas when compared to students subject to no intervention and the “one to all” technique. The “one to all” technique was shown to have only a temporary benefit over the control group, after which the two techniques matched. These results were consistent with the simulation results. Further, the experimental results align with the trends observed in the simulation results presented, providing additional validation of the model, beyond the results from Chapter 4, and displaying the usefulness of the model as a tool to analyse the effectiveness of different influence techniques. The results presented show that traditional teaching, where instructors lecture from the front of a classroom, may be less effective in influencing students than other techniques that give instructors a better chance at connecting with individual students and more strongly convincing students of the benefits of advantageous behaviours. A rotating styled flipped classroom, where students are asked to watch lectures before class and class time is used to solve problems in a collaborative environment [13], may provide instructors with a higher degree of influence to convince students to participate in beneficial behaviours, assuming instructors rotate through their students, spending time with all of them.

There are some limitations to this study. The simulation results presented in Section 6.1 were generated by analysing the model’s response while randomly sampling parameters using a random uniform distribution over each parameter’s domain. This method helps explore all possible outcomes of a given influence technique; however, this may produce results that are not reflective of typical human behaviour. For example, in simulation, it is possible to have agents in the system whose beliefs instantly switch from very positive to very negative, and vice versa, because they have a large suggestibility value, μ , which is likely to be very rare in the real world. A set of simulations that more precisely mimic human behaviour could be attained by identifying the distribution of parameters that are more likely to occur in reality. An additional limitation was the length of time that the experiment was conducted. The experiment was run over three, hour-long sessions, of team activities, and only 20 minutes of each hour was used by students to brainstorm and construct their ideas. This time pressure could have incentivized students to spend less time brainstorming than they would have otherwise liked to. This, in addition to having more team activities, would provide more robust measurements of students’ behaviour over the long term.

Chapter 7

Conclusion

7.1 Summary

The goal of this thesis was to explore various influence techniques that can effectively convince a group of individuals to take up a desired target behaviour. To achieve this we developed an agent-based model of behavioural change based on foundational psychology theory, partially validated the model using expected psychological outcomes, simulated several influence techniques using the model, and then tested our results using a real-world empirical study.

Chapter 1 provided an introduction to the problem at hand and motivated the need for tools to predict human behaviour and to test different intervention techniques before they are implemented.

Chapter 2 explored the literature on existing modelling paradigms, equation-based and agent-based. Further, we examined psychology literature on behavioural change theories such as the social cognitive theory and the reasoned action approach. Social influence theories of conformity and internal bias were also discussed.

In Chapter 3, we developed the agent-based behavioural change model utilised in this thesis. The reasoned action approach was selected as the foundational behaviour model while an assimilative social influence model was used to capture conformity effects. Internal bias in the form of a signed offset was added as a means to counter conformity. Lastly, an outcome feedback mechanism was implemented to replicate the effects behavioural outcomes have on people's beliefs.

Chapter 4 demonstrated the validity of the developed model by comparing its simulated results to expected theoretical results. The system was simulated, with zero input, in 1-person, 2-person, and 30-person scenarios with a wide set of initial conditions. The results of these simulations demonstrated clustering and polarisation effects that are strongly driven by internal bias factors. Additionally, some theoretical results were provided for a simplified system.

In Chapter 5, we used the now partially validated model to test various influence techniques. A theoretical result regarding sufficient conditions for an external influence source to be able to change opinions was provided. Then “one to all”, “influencer”, “rotating”, and “targeted” influence techniques were simulated using a classroom scenario where an instructor is trying to convince their students to spend time brainstorming. The results showed that the “rotating” and “targeted” techniques were very effective at changing behaviours when compared to the “one to all” and “influencer” techniques.

In Chapter 6, to further validate the model, we extended the influence techniques described to a real-world test scenario. We simulated an environment that resembled a real-world classroom where students were asked to complete team activities and received varying styles of encouragement to spend time brainstorming. Once simulations were completed the experiment was replicated in a real-world classroom over several team activities. The results of the experiments showed that the “rotating” influence technique performed better than the “one to all” influence technique and better than the “no intervention” case at convincing students to brainstorm, matching the trends of the simulated results.

7.2 Contributions

This thesis aimed to study the effects of social influence techniques on human behaviour. The following contributions were made in pursuit of that goal:

- In Chapter 3 we furthered the study of agent-based models in a social psychology context by introducing and analysing a model involving a specific combination of the reasoned action approach for behavioural prediction and to explain behavioural change, social influence effects in the form of conformity and described by a similarity model, and internal biases in the form of a signed offset to counteract some conformal pressures. This unique combination of methodologies provided insight into how and why behaviours propagate.
- We found that, in our simple model, internal bias was capable of producing polarisation and clustering in Chapter 4. This shows that a simple model can be used

to produce complex emergent results, and also tells us that internal biases are an important factor in driving human social behaviour.

- We established that internal biases can be taken advantage of to drive human behaviour in Section 5.3. By relying on internal bias to sustain an individual’s beliefs we were able to demonstrate that spending time to get individuals to believe in a behaviour, by using a “rotating” or “targeted” influence technique, was more effective than trying to convince a whole group simultaneously to believe in the behaviour, by using a “one to all” technique.
- We provided some preliminary theoretical analysis of the model described in this thesis. In Lemma 1 we found that, in a system without internal bias or outcome feedback, the system would converge to a set of clusters of opinion. Additionally, we showed in Lemma 2, that a system without outcome feedback and including internal bias will lead to positively (or negatively) polarised beliefs if all agents in the system have positive(or negative) beliefs about the behaviour. Finally, in Lemma 3, we showed that as long as a control agent is large enough, the suggestibility of an individual is large enough, and/or the internal bias of the individual is low enough, i.e. $\omega_{i,p} > \frac{2-\mu_i}{2-h_i}$, the control agent will be able to convince an agent to switch behaviours from negative to positive.
- In Chapter 6 we provided experimental findings to validate the developed model. We compared simulated results of the model to empirical findings and found that the trends in improvements in behaviour matched, with a “rotating” influence technique resulting in larger improvements than a “one to all” technique.

7.3 Limitations

In previous chapters, we discussed some limitations and assumptions of the work done in this thesis. Several simplifying assumptions have been made to be able to compute the model given finite resources; it would be infeasible to model every single aspect of human behaviour for example. Here we summarise and reiterate the limitations:

- While the RAA has been used to explain several behavioural conditions, it may not be suitable to explain behaviours that arise from addictive, not fully voluntary compulsions and the influence techniques suggested in this thesis may not be suitable to tackle these problems; leading to Assumption 3.1 that all behaviours are voluntary

and of an individual’s own volition. The RAA also states that behaviour is limited by an individual’s actual control of a situation, if a person intends to perform a behaviour but cannot access the means to perform the behaviour, they will not perform the behaviour, resulting in Assumption 3.2 that all persons have full control of their behaviour. Finally, we assume that the RAA can accurately describe how beliefs affect human behaviour under the mentioned constraints with Assumption 3.3.

- The RAA relies on several underlying parameters that determine the importance of different salient beliefs to a person’s end behaviour. For simplicity, in the model, we assume that these parameters remain constant for every individual in the system in Assumption 3.4. This limits the accuracy of our predictions as, for most people, the beliefs they care about change over time. Additionally, this limits forms of intervention that target changing a person’s salient beliefs. Expanding the model to have dynamic parameters could be an avenue for future work
- The methods of capturing social influence in the model are a similarity-based model and an internal bias. We assume that these components can accurately explain the social dynamics of individuals in a system and how they propagate in Assumption 3.5. Additionally, For simplicity, we assume that the social weights in this model remain constant in Assumption 3.6. This assumption ignores phenomena such as social identity dynamics or effects of status change which is not fully accurate. For example, if you found that your instructor was constantly lying to you by saying that exercising you will never get sore, you will eventually catch on and lose trust in their opinion. The extension of this research to include dynamic weights is also of interest in future work.
- The means of influence in this model, where an external source of influence influences belief change in individuals, mainly focus on relationships that are personal. The influence dynamics on larger scale systems, such as through social networks and advertising, have not been investigated and may not apply, resulting in Assumption 3.7 that the system being studied involves small groups with personal connections to the control agent.
- In Assumption 3.8 we assume that the outcome of a behaviour follows the linear response stated in Equation (3.20). This linear relationship is unlikely for many scenarios where factors such as diminishing returns or other nonlinear effects like time delays may be a factor. However, to simplify analysis, this assumption is made. The outcome feedback component can be substituted for further analysis. Additionally,

in Assumption 3.9 we assume that an individual takes into account the outcomes of a behaviour only if they sufficiently perform the behaviour (i.e. $B_i[k] > 0$).

- This model doesn't account for any noise or uncertainty in measurement that different agents may have. When an agent interprets an external variable, it is likely that they do not accurately interpret this signal. Internal bias accounts for some misinterpretation, however, adding in random noise could account for other uncertainties.
- A limitation of the influence techniques discussed in Chapter 5, is the use of a high-impact weight for the instructor in the system. From Lemma 3 we know that high weight is important in being able to influence agents. While this high weight might be true of an instructor in a classroom, it is not necessarily the case. For example, a door-to-door salesman trying to convince you to buy some goods would have a relatively low impact, in theory.
- Another limiting factor regarding the control techniques discussed is Assumption 5.1 that control agents have an unlimited pool of resources to influence others with. It is assumed that the control agent can influence others indefinitely, however, in reality, a human source of influence, such as an instructor, will get tired eventually. Being able to identify optimal control influence strategies associated with a cost function associated with the control agent is an interesting avenue for future work.
- The model simulations use random parameters selected from a uniform distribution. While this method captures the general idea of how a system will behave, it can be over-general and select parameters that may be outside what a typical human might present. For example, having a very high suggestibility can lead to people changing opinions rapidly whenever they encounter new information, which is fairly unrealistic. Narrowing down the range of parameter values that more closely match human characteristics would provide better insights into typical human behaviour. Additionally, the model has not been fully validated. To validate the model fully would require significant empirical testing with large amounts of survey data, conducted on several different test groups. This is an immense task which can be the topic of future work.

7.4 Preliminary Future Work

In this section, we discuss some preliminary work done that could spark future research investigations. There are possibilities for analysing different result feedback techniques,

social structures, or variations in internal bias.

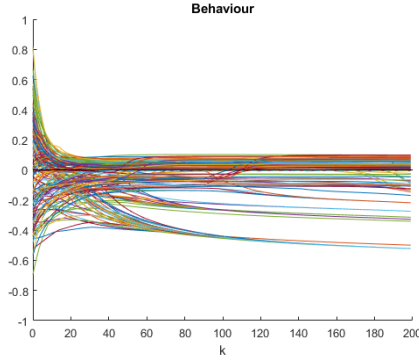
7.4.1 Result Feedback Variation

Using the classroom scenario, we first look at how changing the way students receive feedback can affect how they decide to pursue the targeted behaviour. The variations used are those commonly employed in a classroom. In the first scenario, we limit result feedback related to the belief that brainstorming will lead to better ideas to only a few times a term. In the second scenario, we add a negative bias to some of these results and see how that affects students' behaviours compared to a continuous feedback case.

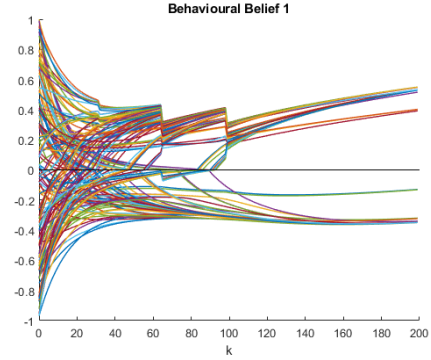
Pulsed Result Feedback

In most classrooms, students are only able to gauge the effectiveness of their habits when they receive the results from an assessment. In our scenario, this would mean that the belief that brainstorming leads to better ideas will be validated only when they get the results of an assessment, assuming that better ideas will lead to better grades. To simulate this, we will assume that results regarding the first belief (b_1) will come into play at only three different times: at $k = 33$, at $k = 66$, and finally at $k = 99$. We will treat the first result as a small team activity and so will have a weaker effect on the behaviour convergence (i.e., $\mu_i = 0.2$) compared to the other two feedback pulses, which will be considered a larger group projects respectively (i.e., $\mu_i = 0.9$). Figure 7.1 shows an sample of the behaviour for the 150-person system. For a longer simulation, with all signals displayed, see Figures A.7 and A.8.

An interesting observation can be made comparing the results to that of the continuous feedback response shown in Figure 7.1. Firstly, we can see that the behaviour tends to the positive in the pulsed feedback scenario while they remain mixed or close to zero in the continuous feedback scenario. Looking closer at the pulsed feedback results, the student's behaviour starts to group, mainly concentrated in the positive half of the scale. However, at $k = 33, 66$, and 99 , when their results come in, the student's beliefs drop; this is because the behaviour results in outcomes that are lower than what the students believed would happen. However, once the result is no longer in play and the dynamics are only affected by social influences, the student's beliefs start to increase again, eventually tending to the positive. This is intriguing and implies that depending on the beliefs of the group, the timing of the result feedback may have a detrimental or beneficial effect. For example, if a group of students has an overall popular opinion regarding the behaviour,



(a) Behaviour with continuous feed-back



(b) Behaviour with pulsed feedback

Figure 7.1: Simulation comparison of pulsed result feedback

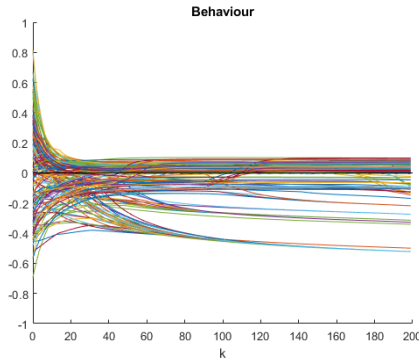
and there is nothing to hold that belief in check (i.e., results) then the student's beliefs will propagate due to conformity pressures and cause polarisation due to internal biases. Therefore, introducing a result that shows them otherwise could have a negative impact on behaviour performance. However, this is also true in the reverse scenario, where students who have negative beliefs could cause everyone to stop performing the behaviour unless the results encourage them otherwise.

Biased Result Feedback

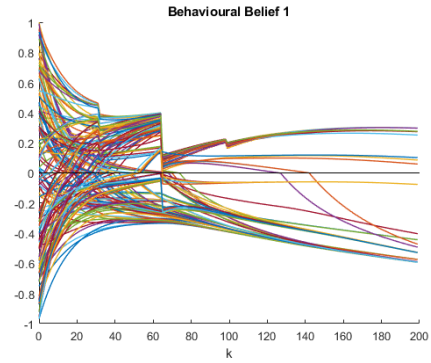
Some instructors have the belief that students can be motivated to study harder if they are faced with difficulties at the beginning of their term. This mentality can manifest itself by making the student's assessments in the first half of the term far more difficult than what would be expected and easing up on the students in the last half of the course. To simulate this we will use the parameters from the previous section with the caveat that results from the first two assessments have a bias of -0.2 added on to Equation (3.4.1) resulting in the following:

$$R_{i,l}[k]_{bias} = \text{sat}(R_{i,l}[k] + \text{sgn}(b_{i,l}[k])h_i) - 0.2, \quad (7.1)$$

Figure 7.2 shows the results of the behaviour. For a longer simulation, with all signals displayed, see Figure A.9. These simulations reveal the drastic effects a bias in the feedback can have. Although the biased behaviour in Figure 7.2 seems to initially appear the same as their pulsed feedback counterparts in Figure 7.1, the negative influence they have in the



(a) Behaviour with continuous feedback



(b) Behaviour with negatively biased feedback

Figure 7.2: Simulation comparison of negatively biased result feedback

first two assessments is enough that, even though the final exam has no bias, the students cannot recover from the mentality that their behaviour does not help enough with their idea generation. As Figure A.9 shows, the students eventually stop believing in this behaviour's usefulness and stop doing it. Therefore, misleading students about their results can have negative ramifications when it comes to reinforcing good behaviours.

These results show that varying the method of feedback can have a dramatic effect on the end effect on agent behaviours and warrants further study.

7.4.2 Social Structure Variation

Another aspect of a classroom setting that could have an impact on the way behaviours propagate is the network structure through which they interact. To explore this we set up a 150-person classroom with randomly selected parameters. These simulations take into account two main levels of social interaction. The class is first divided into 30 groups of five, called teams. Additionally, there are five leaders in the classroom, called class representatives. Each of these representatives is in charge of six unique teams; we will call this batch of teams the class rep's group. The strength of the interaction of the people in these groups, which represents the non-normalised w_{ij} portion in the equations, is listed in Table 7.1.

The weights in Table 7.1 show that class representatives have a large amount of influence on people in their groups while not being strongly influenced by others, except for other

Table 7.1: Relative weights for student interactions

Interaction	Non-normalised Weight
Class rep to class rep. influence	0.8
Class rep influence on their six groups	0.8
Six groups' influence on class rep	0.08
Teams members' influence on one another	0.2

class representatives. Team members also have a stronger relationship amongst themselves, when compared to others in the class. Finally, to ease analysis, we assume that students outside of their teams and class representatives have zero interactions with each other and so do not influence each other. For visualisation purposes see Figure 7.3. In this figure, each

1	16	31	46	61	76	91	106	121	136
2	17	32	47	62	77	92	107	122	137
3	18	33	48	63	78	93	108	123	138
4	19	34	49	64	79	94	109	124	139
5	20	35	50	65	80	95	110	125	140
6	21	36	51	66	81	96	111	126	141
7	22	37	52	67	82	97	112	127	142
8	23	38	53	68	83	98	113	128	143
9	24	39	54	69	84	99	114	129	144
10	25	40	55	70	85	100	115	130	145
11	26	41	56	71	86	101	116	131	146
12	27	42	57	72	87	102	117	132	147
13	28	43	58	73	88	103	118	133	148
14	29	44	59	74	89	104	119	134	149
15	30	45	60	75	90	105	120	135	150

Figure 7.3: Group divisions for 150 student classroom

number represents a student. Students in a team are encircled by a thin border. Students in a class rep's group are encircled by a thick border. Finally, the class representatives themselves are highlighted in yellow.

Figure 7.4 shows the first belief state of all 150 students at interactions $k = 1, 20, 40, 60, 80$, and 100 . Teams are divided by black borders and groups are divided by purple lines. The class representative for each group is the top left-most student in each group. From these results, we can see that students tend to align their beliefs within their teams first and foremost. They then influence their class representatives who influence the rest of their

group and the other class representatives. These teams and groups reach an equilibrium where all other social influences and results attract them equally.

These results show that social structure is indeed an important aspect of the established model and the results show a lot of insight into how beliefs propagate through a group. Further study could reveal strategies involving efficient group setup to best propagate desired behaviours.

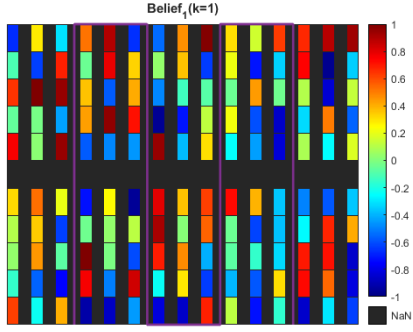
7.4.3 Theoretical Analysis

We have established some theoretical work through Lemmas 1, 2, 3; however, there is much more that can be analysed. Several aspects of the model are difficult to theoretically analyse, due to non-linear components of the model, and are open for future work. Some characteristics of interest are the conditions required to maintain behaviours without an external influence; when the control agent is no longer present in the system, such as at the end of a class term when the instructor is no longer present, what conditions are necessary to maintain an agent's beliefs about a behaviour. Additionally, investigating the conditions required for the full non-linear model to polarise or to form equilibrium clusters when no control agent is present is of interest. With the effects of outcome feedback, the dynamics of the system complicate and it becomes difficult to know when an agent might change opinions because performing the behaviour may result in the agent realising that the behaviour is not beneficial and cause them to stop, however, then social pressures might convince them the behaviour is beneficial. If there are conditions that can inform us of when polarisation might occur or not occur, that would be insightful. Analysis on the full non-linear model, in general, is difficult without a control agent driving the system in one direction or another; any theoretical analysis involving this model would be useful. An additional analysis of interest is the investigation of optimal control techniques. Currently, the influence models suggested do not account for any costs on the control agent or make any claims on optimality. Investigating influence techniques which account for constraints such as a limited pool of energy, or a limited rate of influence on the control agent would be valuable for further determining the effectiveness of different influence techniques.

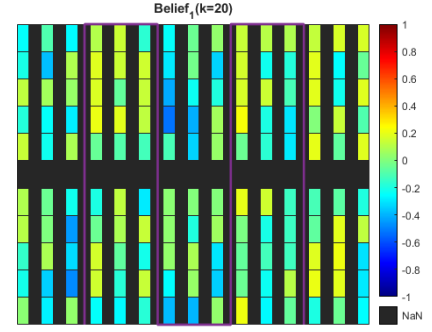
7.5 Conclusions

Many decisions are being made, involving human behaviour, that would benefit from evaluative testing before they are implemented. Two examples include government health-care policy to encourage exercise, and educator teaching strategies to encourage students'

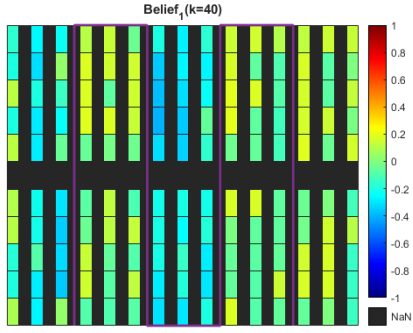
studying behaviours to achieve better grades. Decision-makers will make better choices if they can assess and compare candidate policies, and their potential outcomes, before implementing them. This thesis demonstrates that a behavioural change model can be used as a tool to provide insights into influence techniques. The model developed demonstrates the importance of internal bias in causing polarisation or clustering of opinions. Taking advantage of internal bias we are able to develop an influence technique that is effective at changing behaviour. Additionally, we observe that individual-focused influence techniques, which take advantage of internal biases to sustain behaviours, can be more effective at influencing agents to perform a behaviour compared to a front-of-the-class and lecture technique, suggesting that alternative teaching methods, such as a flipped classroom, may be beneficial for students. The insights provided by this model show that being able to explore the effects of influence techniques is a key step in identifying and developing unique influence techniques.



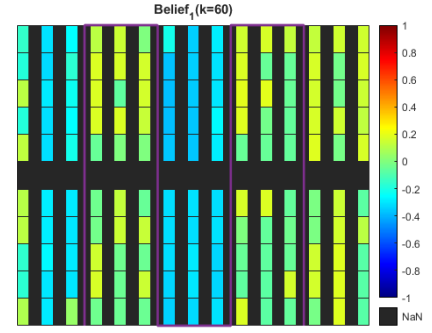
(a) $k = 1$



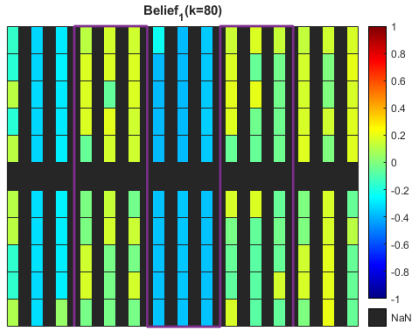
(b) $k = 20$



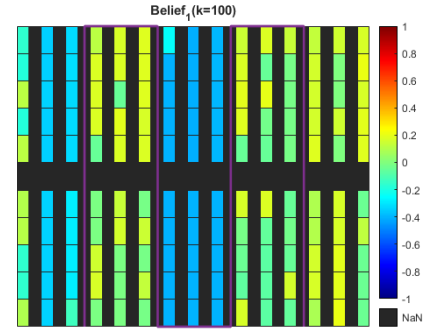
(c) $k = 40$



(d) $k = 60$



(e) $k = 80$



(f) $k = 100$

Figure 7.4: Simulation of behaviour in a 150-student system

References

- [1] Icek Ajzen. The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, 50(2):179 – 211, 1991. Theories of Cognitive Self-Regulation.
- [2] Hosam Al-Samarraie and Shuhaila Hurmuzan. A review of brainstorming techniques in higher education. *Thinking Skills and creativity*, 27:78–91, 2018.
- [3] Mário S Alvim, Artur Gaspar da Silva, Sophia Knight, and Frank Valencia. A multi-agent model for opinion evolution under cognitive biases. *arXiv preprint arXiv:2402.17615*, 2024.
- [4] Mayank Anand, Arun Velu, and Pawan Whig. Prediction of loan behaviour with machine learning models for secure banking. *Journal of Computer Science and Engineering (JCSE)*, 3(1):1–13, 2022.
- [5] Gordon C Andrews, J Dwight Aplevich, Roydon Andrew Fraser, Carolyn G MacGregor, and Herbert C Ratz. *Introduction to professional engineering in Canada*. Pearson, 2018.
- [6] Solomon E Asch. Studies of independence and conformity: I. a minority of one against a unanimous majority. *Psychological monographs: General and applied*, 70(9):1, 1956.
- [7] Solomon E Asch. Effects of group pressure upon the modification and distortion of judgments. In *Documents of gestalt psychology*, pages 222–236. University of California Press, 1961.
- [8] Albert Bandura. Social cognitive theory of mass communication. *Media psychology*, 3(3):265–299, 2001.
- [9] Albert Bandura et al. Guide for constructing self-efficacy scales. *Self-efficacy beliefs of adolescents*, 5(1):307–337, 2006.

- [10] Albert Bandura, WH Freeman, and Richard Lightsey. *Self-efficacy: The exercise of control*. Springer, 1999.
- [11] Albert Bandura, Dorothea Ross, and Sheila A Ross. Transmission of aggression through imitation of aggressive models. *The Journal of Abnormal and Social Psychology*, 63(3):575, 1961.
- [12] Albert Bandura, Dorothea Ross, and Sheila A Ross. Imitation of film-mediated aggressive models. *The Journal of Abnormal and Social Psychology*, 66(1):3, 1963.
- [13] J Edward Bates, Hasan Almekdash, and Maggie J Gilchrest-Dunnam. The flipped classroom: A brief, brief history. *The flipped college classroom: Conceptualized and re-conceptualized*, pages 3–10, 2017.
- [14] B Douglas Bernheim. A theory of conformity. *Journal of political Economy*, 102(5):841–877, 1994.
- [15] Georgiy V. Bobashev, D. Michael Goedecke, Feng Yu, and Joshua M. Epstein. A hybrid epidemic model: Combining the advantages of agent-based and equation-based approaches. In *2007 Winter Simulation Conference*, pages 1532–1537, 2007.
- [16] Joanna J Bryson, Yasushi Ando, and Hagen Lehmann. Agent-based modelling as scientific method: a case study analysing primate social behaviour. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1485):1685–1699, 2007.
- [17] Earl R Carlson. Attitude change through modification of attitude structure. *The Journal of Abnormal and Social Psychology*, 52(2):256, 1956.
- [18] Xi Chen, Panayiotis Tsaparas, Jefrey Lijffijt, and Tijn De Bie. Opinion dynamics with backfire effect and biased assimilation. *PLOS ONE*, 16:1–17, 09 2021.
- [19] Robert B Cialdini and Noah J Goldstein. Social influence: Compliance and conformity. *Annu. Rev. Psychol.*, 55:591–621, 2004.
- [20] Robert B Cialdini, Raymond R Reno, and Carl A Kallgren. A focus theory of normative conduct: recycling the concept of norms to reduce littering in public places. *Journal of personality and social psychology*, 58(6):1015, 1990.
- [21] FT Cutts, A-M Henao-Restrepo, and JM Olive. Measles elimination: progress and challenges. *Vaccine*, 17:S47–S52, 1999.

- [22] Yonghui Dai and Tao Wang. Prediction of customer engagement behaviour response to marketing posts based on machine learning. *Connection Science*, 33(4):891–910, 2021.
- [23] Yadolah Dodge. *Kruskal-Wallis Test*, pages 288–290. Springer New York, New York, NY, 2008.
- [24] JJ Dunn, F Baldanti, E Puchhammer, M Panning, O Perez, H Harvala, et al. Measles is back—considerations for laboratory diagnosis. *Journal of Clinical Virology*, 128:104430, 2020.
- [25] Norman T Feather. *Expectations and actions: Expectancy-value models in psychology*. Routledge, 1982.
- [26] Keegan Fernandes, Daniel Davison, and David Wang. An agent-based model to explore belief and behavioural change in a classroom. In *Proceedings of the 2022 Conference of The Computational Social Science Society of the Americas*. Springer International Publishing, 2023.
- [27] Keegan Fernandes, Daniel Davison, and David Wang. Impact of confirmation bias on group dynamics. In *Proceedings of the 2023 Conference of The Computational Social Science Society of the Americas*. Springer International Publishing, in press.
- [28] Keegan Fernandes, Daniel Davison, and David Wang. An agent-based model of group behavioural dynamics: Strategies to influence behaviours and experimental findings. *IEEE Transactions on Computational Social Systems*, under review.
- [29] Leon Festinger, Stanley Schachter, and Kurt Back. *Social pressures in informal groups; a study of human factors in housing*. Harper, 1950.
- [30] Martin Fishbein. A consideration of beliefs and their role in attitude measurement. *Readings in attitude theory and measurement*, pages 257–266, 1967.
- [31] Martin Fishbein. A theory of reasoned action: some applications and implications. *Nebraska Symposium on Motivation*, 1979.
- [32] Martin Fishbein. A reasoned action approach to health promotion. *Medical Decision Making*, 28(6):834–844, 2008.
- [33] Martin Fishbein and Icek Ajzen. *Predicting and changing behavior: The reasoned action approach*. Taylor & Francis, 2011.

- [34] Ben G Fitzpatrick, Jason Martinez, Elizabeth Polidan, and Ekaterini Angelis. On the effectiveness of social norms intervention in college drinking: The roles of identity verification and peer influence. *Alcoholism: clinical and experimental research*, 40(1):141–151, 2016.
- [35] Andreas Flache, Michael Mäs, Thomas Feliciani, Edmund Chattoe-Brown, Guillaume Deffuant, Sylvie Huet, and Jan Lorenz. Models of social influence: Towards the next frontiers. *Journal of Artificial Societies and Social Simulation*, 20(4), 2017.
- [36] Brian J Fogg. A behavior model for persuasive design. In *Proceedings of the 4th international Conference on Persuasive Technology*, pages 1–7, 2009.
- [37] Michael C Friedman. Notes on note-taking: Review of research and insights for students and instructors. *Harvard Initiative for Learning and Teaching*, pages 1–34, 2014.
- [38] Mary Beth Gilboy, Scott Heinerichs, and Gina Pazzaglia. Enhancing student engagement using the flipped classroom. *Journal of nutrition education and behavior*, 47(1):109–114, 2015.
- [39] Serge Guimond. Attitude change during college: Normative or informational social influence? *Social Psychology of Education*, 2(3):237–261, 1997.
- [40] Elizabeth Hunter, Brian Mac Namee, and John D Kelleher. A comparison of agent-based models and equation based models for infectious disease epidemiology. In *26th AIAI Irish Conference on Artificial Intelligence and Cognitive Science*. Technological University Dublin, 2018.
- [41] Roland Imhoff and Pia Lamberty. A bioweapon or a hoax? the link between distinct conspiracy beliefs about the Coronavirus disease (COVID-19) outbreak and pandemic behavior. *Social Psychological and Personality Science*, 11(8):1110–1118, 2020.
- [42] Kathleen Hall Jamieson and Dolores Albarracín. The relation between media consumption and misinformation at the outset of the SARS-CoV-2 pandemic in the US. *The Harvard Kennedy School Misinformation Review*, 2020.
- [43] Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I*, pages 99–127. World Scientific, 2013.
- [44] Rowland R Kao. The role of mathematical modelling in the control of the 2001 FMD epidemic in the UK. *TRENDS in Microbiology*, 10(6):279–286, 2002.

- [45] Herbert C Kelman. Compliance, identification, and internalization three processes of attitude change. *Journal of conflict resolution*, 2(1):51–60, 1958.
- [46] Chandan Khanna. Twitter to permanently ban users who spread COVID misinformation. <https://www.aljazeera.com/news/2021/3/2/twitter-to-permanently-ban-users-who-spread-covid-misinformation>, 2024. Accessed: 2021-02-07.
- [47] Asher Koriat, Sarah Lichtenstein, and Baruch Fischhoff. Reasons for confidence. *Journal of Experimental Psychology: Human learning and memory*, 6(2):107, 1980.
- [48] Ziva Kunda, Geoffrey T Fong, Rasyid Sanitioso, and Emily Reber. Directional questions direct self-conceptions. *Journal of Experimental Social Psychology*, 29(1):63–86, 1993.
- [49] Bibb Latané. The psychology of social impact. *American psychologist*, 36(4):343, 1981.
- [50] Bibb Latané and Sharon Wolf. The social impact of majorities and minorities. *Psychological Review*, 88(5):438, 1981.
- [51] Pablo D Lemoine, Juan Manuel Cordovez, Juan Manuel Zambrano, Olga L Sarmiento, Jose D Meisel, Juan Alejandro Valdivia, and Roberto Zarama. Using agent based modeling to assess the effect of increased bus rapid transit system infrastructure on walking for transportation. *Preventive medicine*, 88:39–45, 2016.
- [52] Rensis Likert. A technique for the measurement of attitudes. *Archives of psychology*, 1932.
- [53] Nathan C Lo and Peter J Hotez. Public health and economic consequences of vaccine hesitancy for measles in the United States. *JAMA pediatrics*, 171(9):887–892, 2017.
- [54] Marty Lobdell. *Study Less, Study Smart: A Guide to Effective Study Techniques and Enhanced Learning*. Createspace Independent Publishing Platform, 2015.
- [55] Charles G Lord, Lee Ross, and Mark R Lepper. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of personality and social psychology*, 37(11):2098, 1979.
- [56] Aleksandra Luszczynska and Ralf Schwarzer. Social cognitive theory. *Predicting health behaviour*, 2:127–169, 2005.

- [57] Donald L McCabe and Linda Klebe Trevino. Academic dishonesty: honor codes and other contextual influences. In *Academic ethics*, pages 545–561. Routledge, 2017.
- [58] Rosemary McEachan, Natalie Taylor, Reema Harrison, Rebecca Lawton, Peter Gardner, and Mark Conner. Meta-analysis of the reasoned action approach (RAA) to understanding health behaviors. *Annals of Behavioral Medicine*, 50(4):592–612, 2016.
- [59] Maren M Michaelsen and Tobias Esch. Motivation and reward mechanisms in health behavior change processes. *Brain research*, 1757:147309, 2021.
- [60] Brian Mullen. Effects of strength and immediacy in group contexts: Reply to jackson. *Journal of Personality and Social Psychology*, 50:514–516, 1986.
- [61] David G Myers. Polarizing effects of social interaction. *Group decision making*, pages 125–161, 1982.
- [62] David G Myers and Helmut Lamm. The group polarization phenomenon. *Psychological bulletin*, 83(4):602, 1976.
- [63] Raymond S Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175–220, 1998.
- [64] Jessica M Nolan, P Wesley Schultz, Robert B Cialdini, Noah J Goldstein, and Vladas Griskevicius. Normative social influence is underdetected. *Personality and social psychology bulletin*, 34(7):913–923, 2008.
- [65] Noratikah Nordin, Zurinahni Zainol, Mohd Halim Mohd Noor, and Lai Fong Chan. Suicidal behaviour prediction models using machine learning techniques: A systematic review. *Artificial intelligence in medicine*, 132:102395, 2022.
- [66] Douglas J Opel, Noel T Brewer, Alison M Buttenheim, Timothy Callaghan, Richard M Carpiano, Chelsea Clinton, Jad A Elharake, Lisa C Flowers, Alison P Galvani, Peter J Hotez, et al. The legacy of the COVID-19 pandemic for childhood vaccination in the USA. *The Lancet*, 401(10370):75–78, 2023.
- [67] Mark G Orr, Sandro Galea, Matt Riddle, and George A Kaplan. Reducing racial disparities in obesity: simulating the effects of improved education and social network influence on diet behavior. *Annals of epidemiology*, 24(8):563–569, 2014.
- [68] Mohammad H Pakravan and Nordica MacCarty. An agent-based model for adoption of clean technology using the theory of planned behavior. *Journal of Mechanical Design*, 143(2):021402, 2021.

- [69] Abhishek Pandey and Alison P Galvani. Exacerbation of measles mortality by vaccine hesitancy worldwide. *The Lancet Global Health*, 11(4):e478–e479, 2023.
- [70] Liuyong Pang, Shigui Ruan, Sanhong Liu, Zhong Zhao, and Xinan Zhang. Transmission dynamics and optimal control of measles epidemics. *Applied Mathematics and Computation*, 256:131–147, 2015.
- [71] Sampo V Paunonen. Big Five factors of personality and replicated predictions of behavior. *Journal of personality and social psychology*, 84(2):411, 2003.
- [72] Harm Peters, Marko Zdravkovic, Manuel João Costa, Antonio Celenza, Kulsoom Ghias, Debra Klamen, Liz Mossop, Michael Rieder, Vishna Devi Nadarajah, Danai Wangsaturaka, et al. Twelve tips for enhancing student engagement. *Medical teacher*, 41(6):632–637, 2019.
- [73] Hazhir Rahmandad and John Sterman. Heterogeneity and network structure in the dynamics of diffusion: Comparing agent-based and differential equation models. *Management science*, 54(5):998–1014, 2008.
- [74] Lisa Rashotte. Social influence. *The Blackwell encyclopedia of sociology*, 2007.
- [75] Daniel Romer and Kathleen Hall Jamieson. Conspiracy theories as barriers to controlling the spread of COVID-19 in the us. *Social Science & Medicine*, 263:113356, 2020.
- [76] Lisa Rosenbaum. Escaping catch-22—overcoming covid vaccine hesitancy, 2021.
- [77] Andrea Scalco, Andrea Ceschi, Itad Shiboub, Riccardo Sartori, Jean-Marc Frayret, and Stephan Dickert. The implementation of the theory of planned behavior in an agent-based model for waste recycling: A review and a proposal. *Agent-based modeling of sustainable behaviors*, pages 77–97, 2017.
- [78] Stanley Schachter. Deviation, rejection, and communication. *The Journal of Abnormal and Social Psychology*, 46(2):190, 1951.
- [79] Dale H Schunk. Social cognitive theory. In *APA Educational Psychology Handbook: Vol. 1. Theories, Constructs, and Critical Issues*. American Psychological Association, 2012.
- [80] Paul Smaldino. *Modeling social behavior: Mathematical and agent-based models of social dynamics and cultural evolution*. Princeton University Press, 2023.

- [81] Eliot R. Smith and Frederica R. Conrey. Agent-based modeling: A new approach for theory building in social psychology. *Personality and Social Psychology Review*, 11(1):87–104, 2007. PMID: 18453457.
- [82] Mark Snyder and Nancy Cantor. Testing hypotheses about other people: The use of historical knowledge. *Journal of Experimental Social Psychology*, 15(4):330–342, 1979.
- [83] Pawel Sobkowicz. Opinion dynamics model based on cognitive biases of complex agents. *Journal of Artificial Societies and Social Simulation*, 21(4), 2018.
- [84] Shravan Sogani, Rukmini Muduganti, Henry Hexmoor, and Fred Davis. Introducing agent based implementation of the theory of reasoned action: a case study in user acceptance of computer technology. In *International Conference on Integration of Knowledge Intensive Multi-Agent Systems, 2005.*, pages 507–511. IEEE, 2005.
- [85] Kevin Spieser. Stabilizing the psychological dynamics of people in a crowd. Master’s thesis, University of Waterloo, 2008.
- [86] Charles Stangor and David McMillan. Memory for expectancy-congruent and expectancy-incongruent information: A review of the social and social developmental literatures. *Psychological Bulletin*, 111(1):42, 1992.
- [87] Keith E Stanovich, Richard F West, and Maggie E Toplak. Myside bias, rational thinking, and intelligence. *Current Directions in Psychological Science*, 22(4):259–264, 2013.
- [88] Stephanie Watts Sussman and Wendy Schneier Siegal. Informational influence in organizations: An integrated approach to knowledge adoption. *Information systems research*, 14(1):47–65, 2003.
- [89] NK Vaidya, M Morgan, T Jones, Lapin Miller, S Lapin, and EJ Schwartz. Modelling the epidemic spread of an H1N1 influenza outbreak in a rural university town. *Epidemiology & Infection*, 143(8):1610–1620, 2015.
- [90] Elizabeth Ruth-Anne Vanderwater. A dynamic control approach to modeling and analysing the effects of rewards on behaviour and attitude change. Master’s thesis, University of Waterloo, 2010.
- [91] Amiram Vinokur and Eugene Burstein. Effects of partially shared persuasive arguments on group-induced shifts: A group-problem-solving approach. *Journal of Personality and Social Psychology*, 29(3):305, 1974.

- [92] Wendy Wood. Attitude change: Persuasion and social influence. *Annual review of psychology*, 51, 2000.
- [93] Bo Xu, Renjing Liu, and Weijiao Liu. Individual bias and organizational objectivity: An agent-based simulation. *Journal of Artificial Societies and Social Simulation*, 17(2):2, 2014.
- [94] Yong Yang. A narrative review of the use of agent-based modeling in health behavior and behavior intervention. *Translational behavioral medicine*, 9(6):1065–1075, 2019.
- [95] Yong Yang, Amy H Auchincloss, Daniel A Rodriguez, Daniel G Brown, Rick Riolo, and Ana V Diez-Roux. Modeling spatial segregation and travel cost influences on utilitarian walking: Towards policy intervention. *Computers, environment and urban systems*, 51:59–69, 2015.

APPENDICES

Appendix A

Simulation Plots

In this appendix we include simulation results from the main document with extended time periods as well as with signals that were not as important for analysis purposes.

In Figure [A.1](#) shows the simulation signals plotted for the one-person with positive beliefs with regards to the behaviour seen in Section [4.1.2](#).

Figure [A.2](#) shows the simulations result of a one-person system as their beliefs transition from negative to positive values, as seen in Section [4.1.3](#).

Figure [A.3](#) shows the result of a two-person system where both people have negative beliefs with regard to the behaviour, as described in Section [4.2](#).

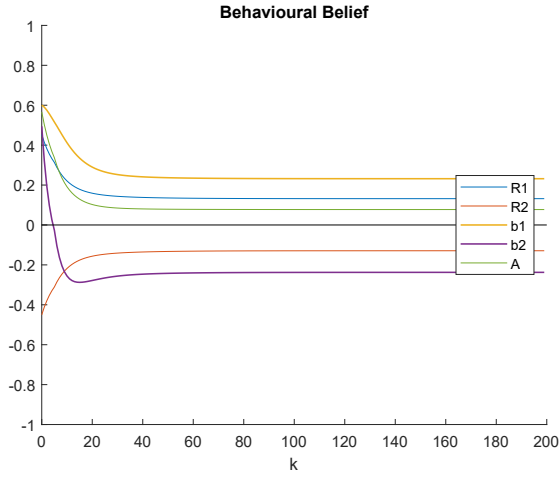
Figure [A.4](#) shows the result of a two-person system where both people have positive beliefs with regard to the behaviour, as described in Section [4.2](#).

Figure [A.5](#) shows the result of a two-person system where both people initially have negative beliefs with regard to the behaviour and then transition to positive beliefs, as described in Section [4.2](#).

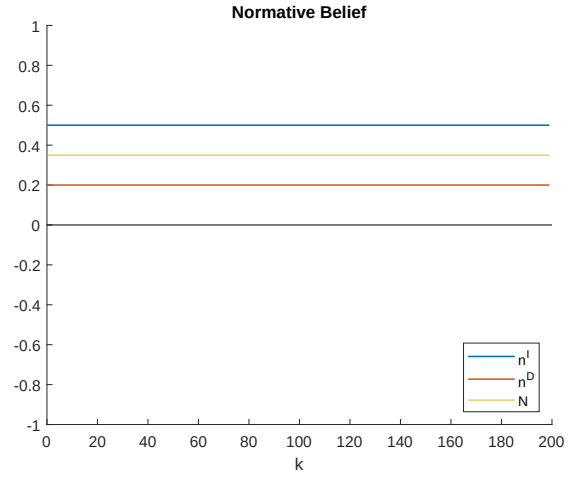
Figure [A.6](#) shows the result of a 150-person system where most people have negative initial beliefs with regard to the behaviour, as described in Section [4.3](#).

Figure [A.7](#) shows the result of a 150-person system where there is an even distribution initial beliefs with regard to the behaviour, as described in Section [4.3](#).

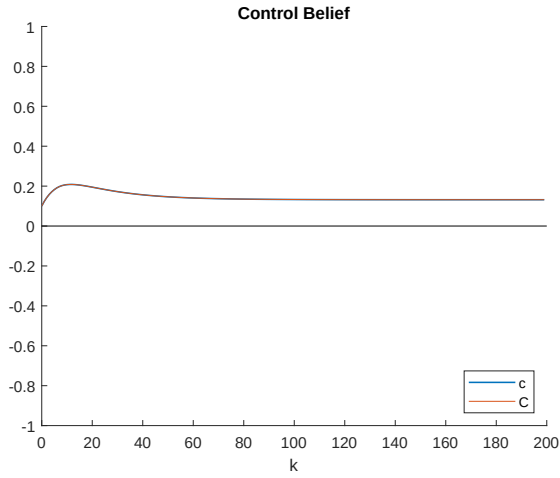
Figure [A.8](#) shows the result of a 150-person system with mixed initial beliefs with regard to the behaviour and a pulsed result feedback, as described in Section [5.3](#).



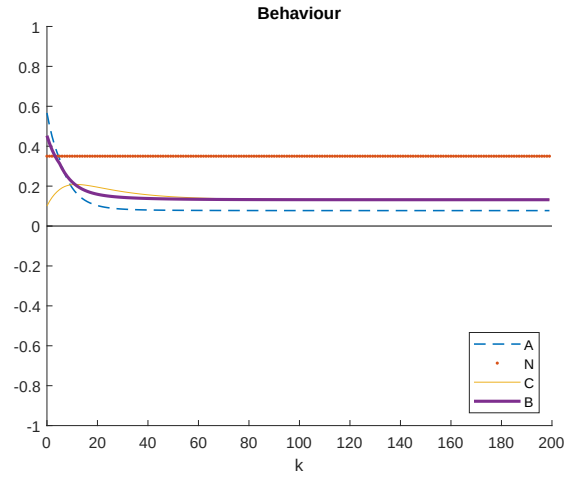
(a) Behavioural belief



(b) Normative Belief

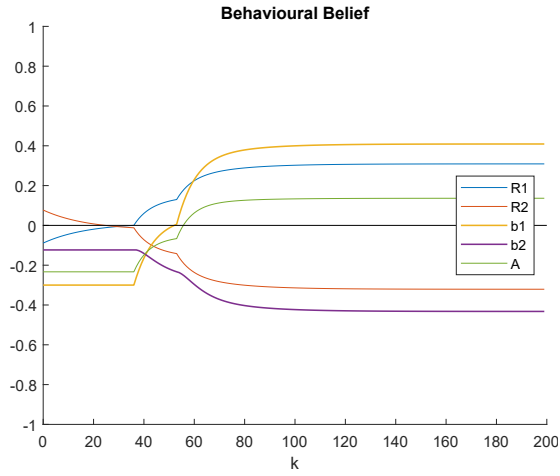


(c) Control belief

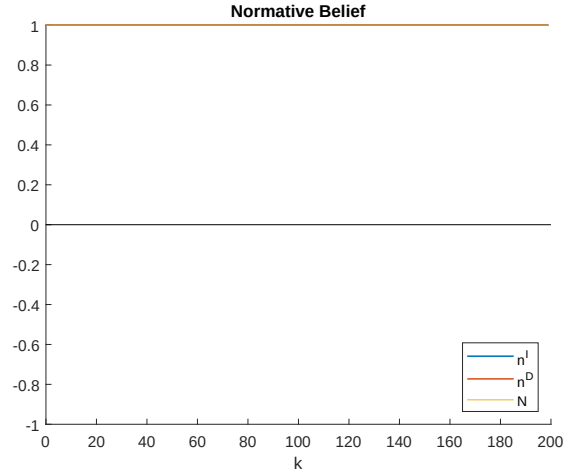


(d) Behaviour

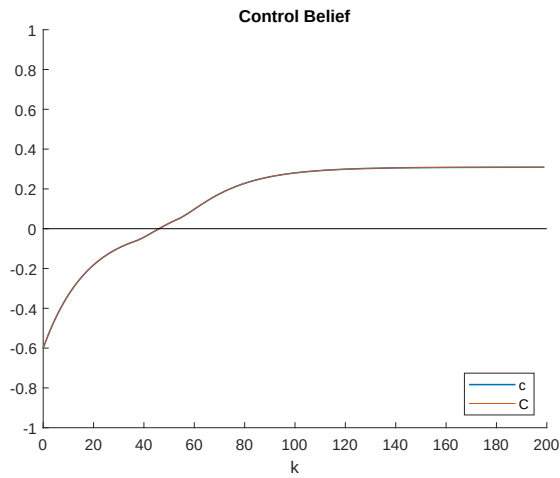
Figure A.1: Simulation of one-person system with positive behaviour



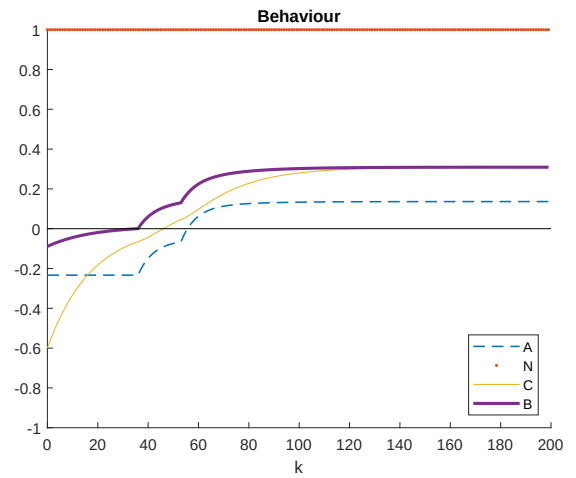
(a) Behavioural belief



(b) Normative Belief

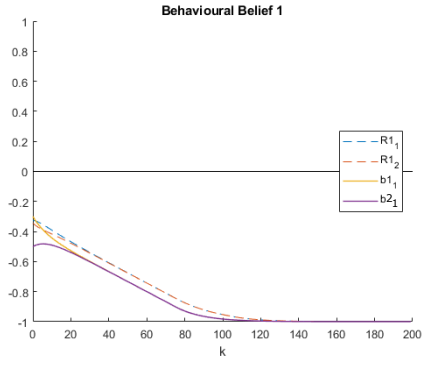


(c) Control belief

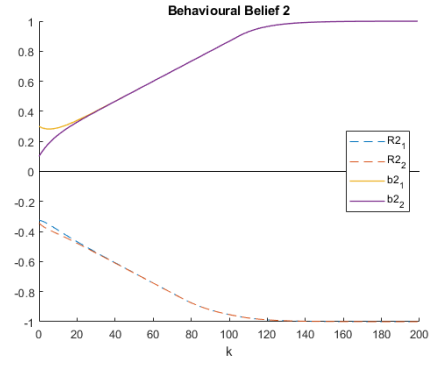


(d) Behaviour

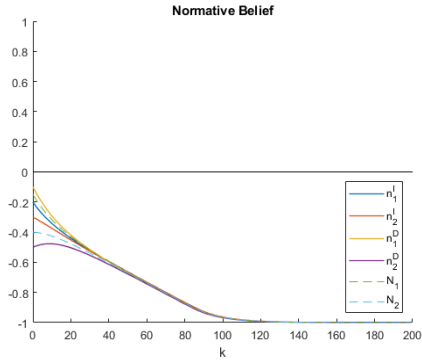
Figure A.2: Simulation of one-person system with behaviour transitioning from negative to positive



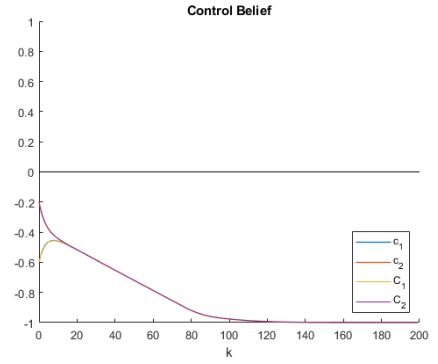
(a) Behavioural belief 1



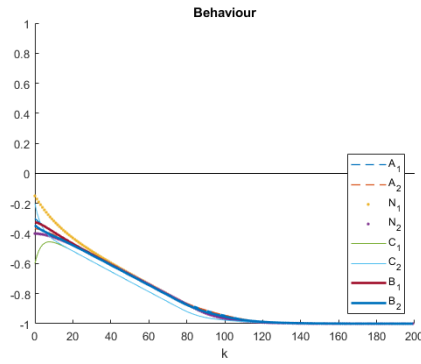
(b) Behavioural belief 2



(c) Normative belief

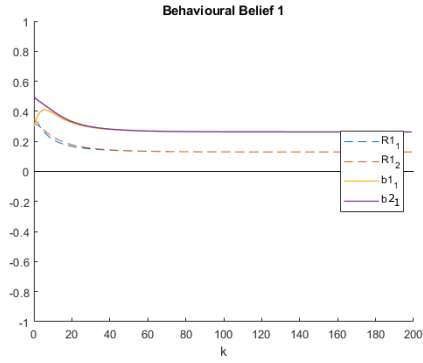


(d) Control belief

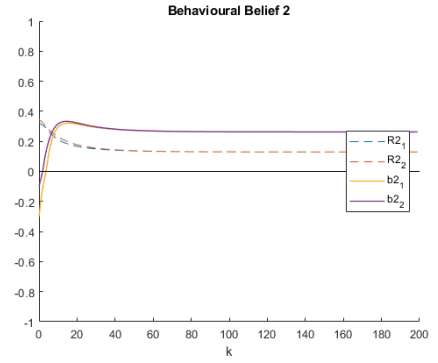


(e) Behaviour

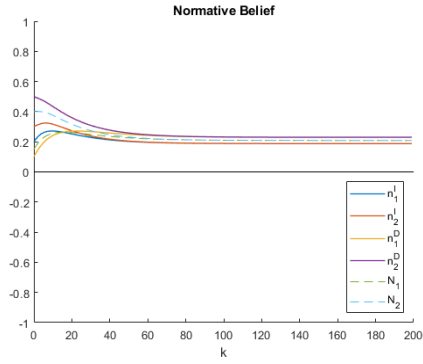
Figure A.3: Simulation of two-person system with both individuals initially having negative beliefs



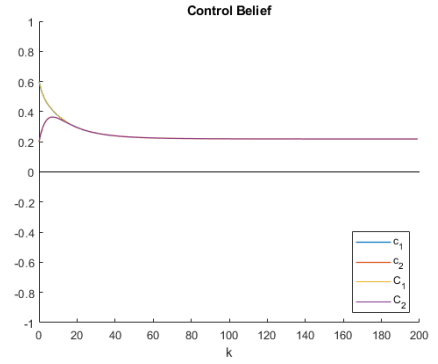
(a) Behavioural belief 1



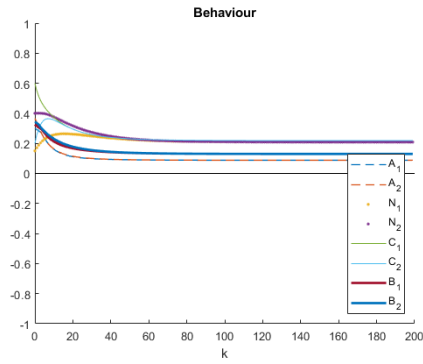
(b) Behavioural belief 2



(c) Normative belief

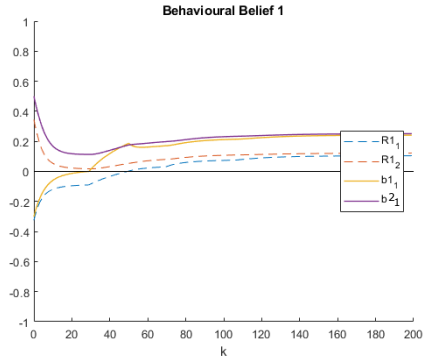


(d) Control belief

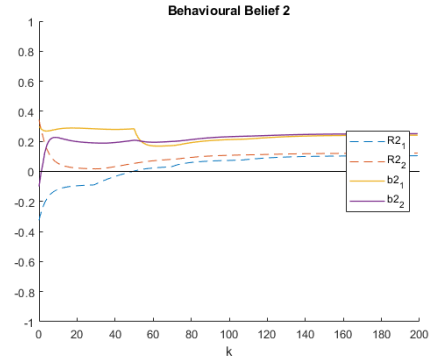


(e) Behaviour

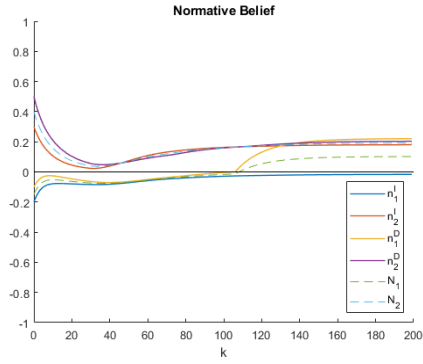
Figure A.4: Simulation of two-person system with both individuals initially having positive beliefs



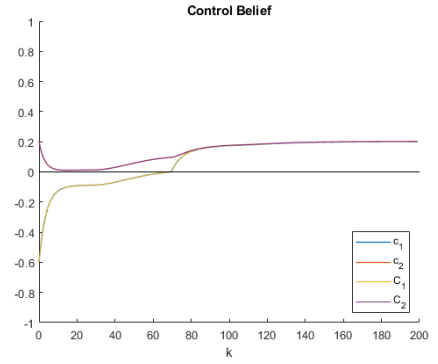
(a) Behavioural belief 1



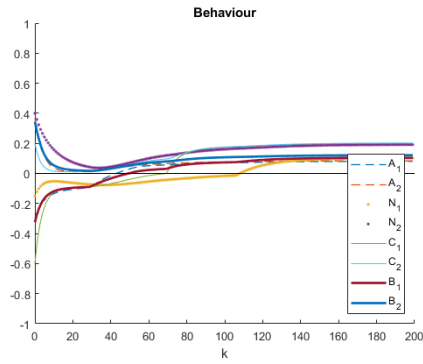
(b) Behavioural belief 2



(c) Normative belief

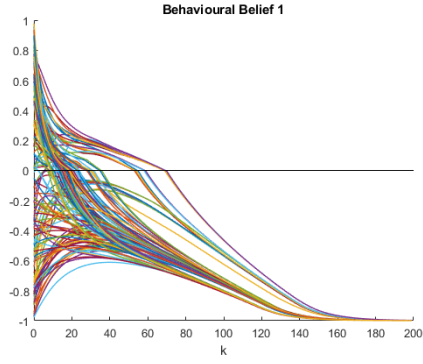


(d) Control belief

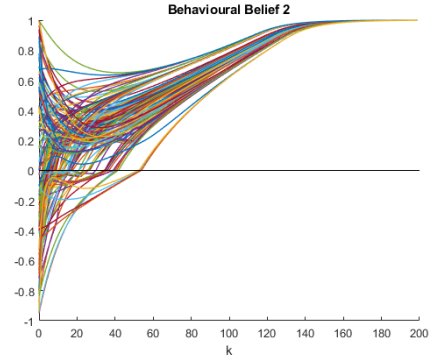


(e) Behaviour

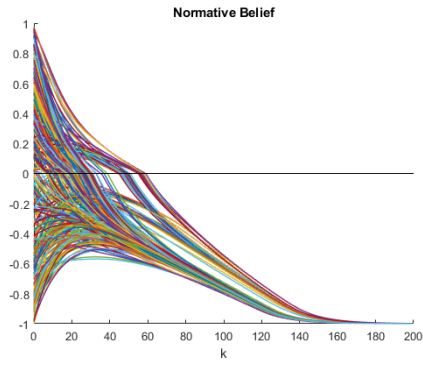
Figure A.5: Simulation of two-person system with initially one individual having positive beliefs and one individual having negative beliefs



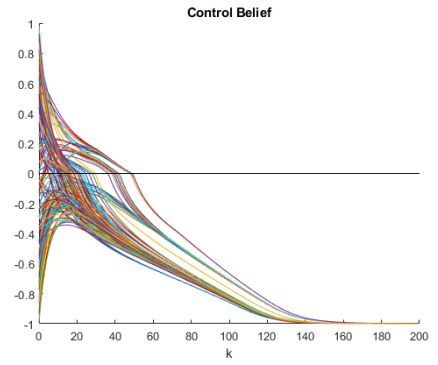
(a) Behavioural belief 1



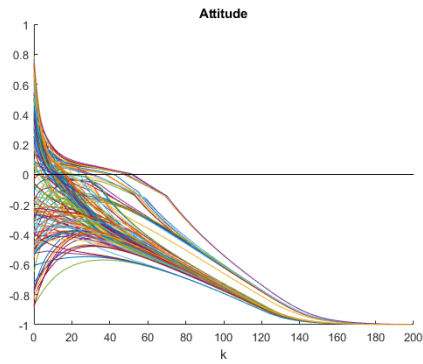
(b) Behavioural belief 2



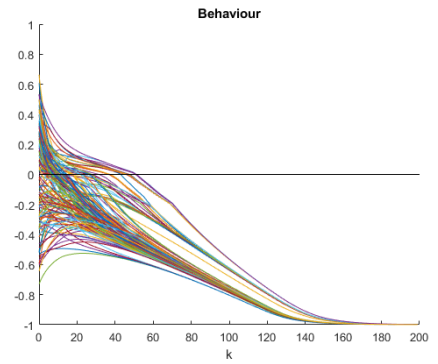
(c) Normative belief



(d) Control belief

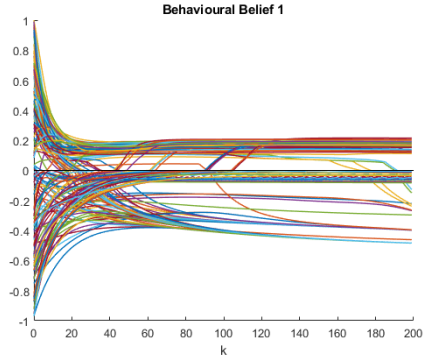


(e) Attitude

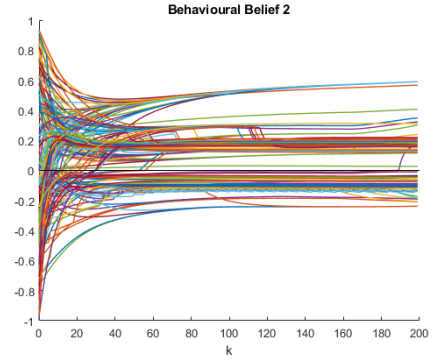


(f) Behaviour

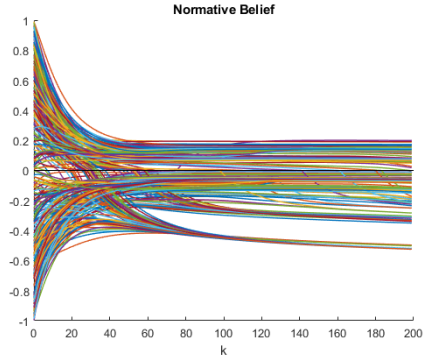
Figure A.6: Simulation of 150-person system with negative initial beliefs on average



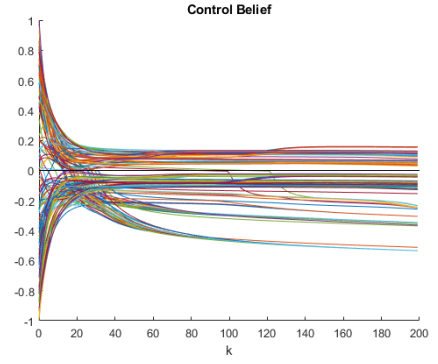
(a) Behavioural belief 1



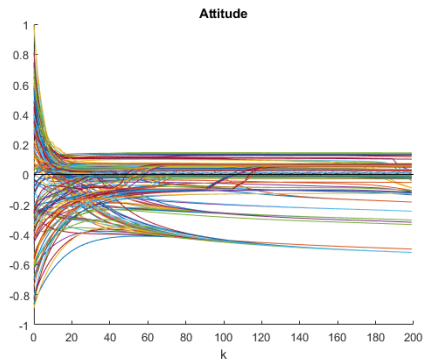
(b) Behavioural belief 2



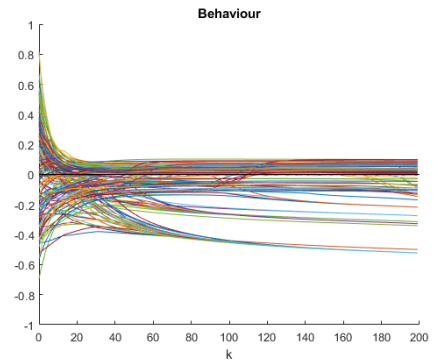
(c) Normative belief



(d) Control belief

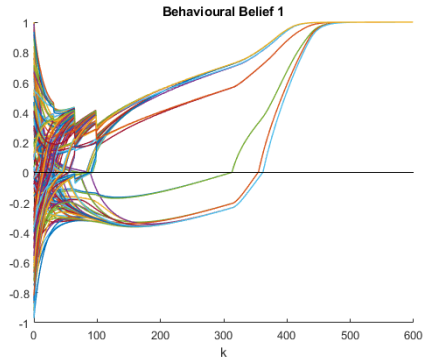


(e) Attitude

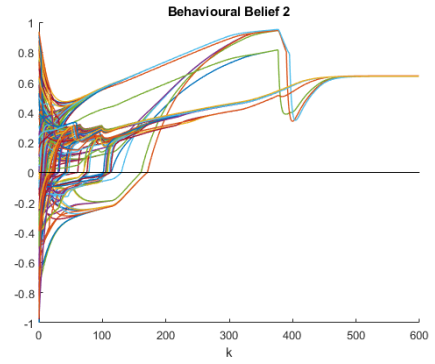


(f) Behaviour

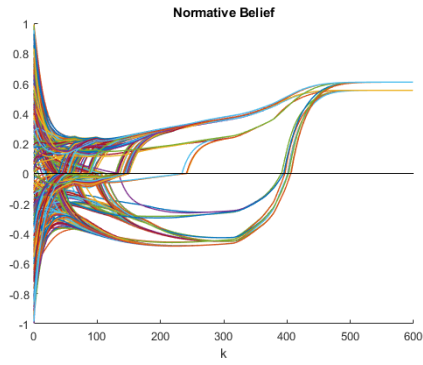
Figure A.7: Simulation of 150-person system with mixed initial beliefs on average



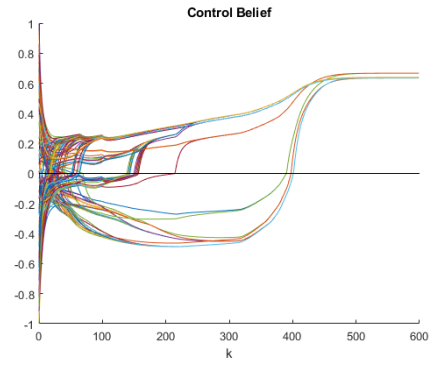
(a) Behavioural belief 1



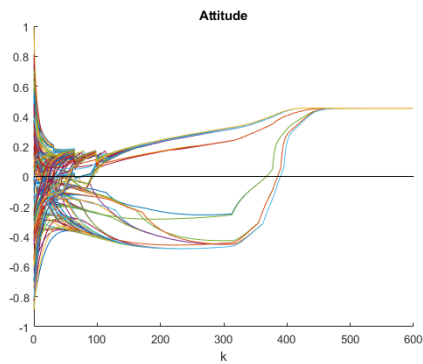
(b) Behavioural belief 2



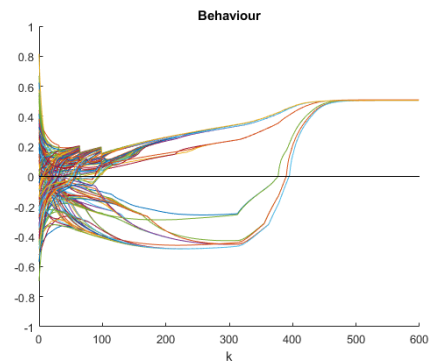
(c) Normative belief



(d) Control belief



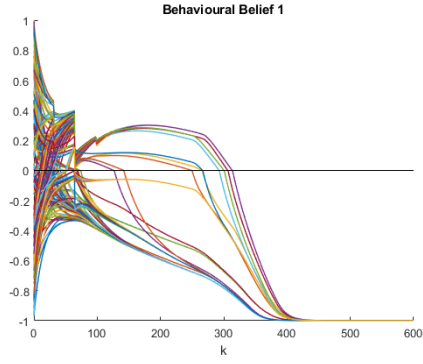
(e) Attitude



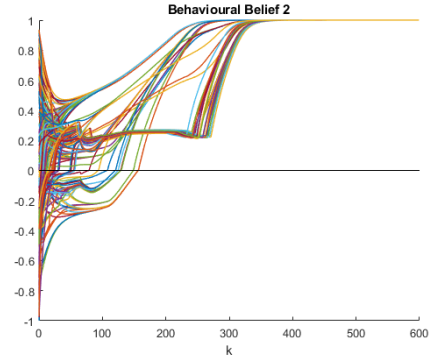
(f) Behaviour

Figure A.8: Simulation of 150-person system with a pulsed result feedback

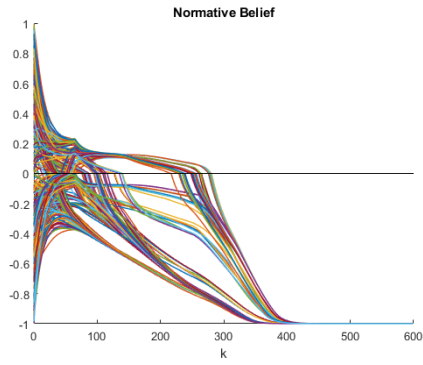
Figure [A.9](#) shows the result of a 150-person system with mixed initial beliefs with regard to the behaviour and a pulsed result feedback with a negative bias, as described in Section [5.3](#).



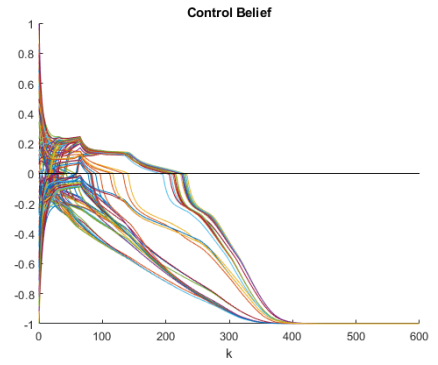
(a) Behavioural belief 1



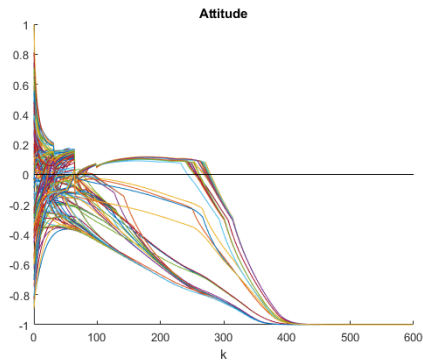
(b) Behavioural belief 2



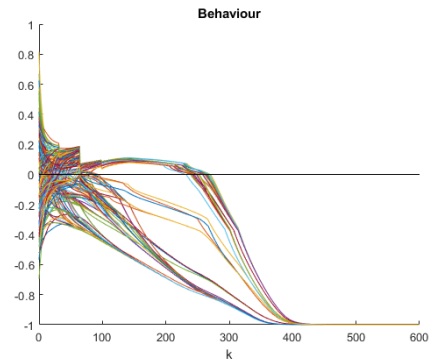
(c) Normative belief



(d) Control belief



(e) Attitude



(f) Behaviour

Figure A.9: Simulation of 150-person system with a biased result feedback