2013

Validated Agent-Based Model using Predictive Data Mining and Intervention Policy Testing Framework: A Case Study in Child Vehicle Safety

Ritwick Gupta
University of Windsor

Validated Agent-Based Model using Predictive Data Mining and Intervention Policy Testing Framework: A Case Study in Child Vehicle Safety

By

Ritwick Gupta

A Thesis
Submitted to the Faculty of Graduate Studies
through School of Computer Science
in Partial Fulfillment of the Requirements for
the Degree of Master of Science
at the University of Windsor

Windsor, Ontario, Canada

2013

Validated Agent-Based Model using Predictive Data Mining and Intervention Policy Testing Framework: A Case Study in Child Vehicle Safety

By

Ritwick Gupta

APPROVED BY:

_____

Dr. Christine Thrasher, External reader
School of Nursing

_____

Dr. Dan Wu, Internal reader
School of Computer Science

_____

Dr. Robert D. Kent, Co-Advisor
School of Computer Science

_____

Dr. Ziad Kobti, Co-Advisor
School of Computer Science

_____

Dr. Subir Bandyopadhyay, Chair of Defence
School of Computer Science

May 7th, 2013

DECLARATION OF PREVIOUS PUBLICATION

This thesis includes one original paper that has been previously accepted for publication in peer-reviewed conference proceeding, as follows:

Thesis Chapter	Publication title	Publisher	Conference	Publication status
Chapter 4,5	*Child Vehicle Safety Simulation using Regression analysis and Predictive Data Mining*	*IOS Press*	*KES-IDT 2013*	*In Press*

I certify that I haven't transferred the copyright of this work to the publication. I certify that the above material describes work completed during my registration as graduate student at the University of Windsor.

I declare that, to the best of my knowledge, my thesis does not infringe upon anyone's copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my thesis, published or otherwise, are fully acknowledged in accordance with the standard referencing practices.

I declare that this is a true copy of my thesis, including any final revisions, as approved by my thesis committee and the Graduate Studies office, and that this thesis has not been submitted for a higher degree to any other University or Institution.

ABSTRACT

Much work has been done on making and perfecting agent-based simulations on child safety measures in cars. These simulations, using algorithms based on social networks, cultural algorithms etc. try and predict what factors are responsible for the propagation of knowledge about child safety measures in a given society. One of the biggest factors being over-looked in these simulations is the validity of the model. In absence of validation against real data, these models may not be a true representation of a real world scenario, and the trends predicted though these simulations are questionable. This paper proposes a system design using regression analysis and predictive data mining on a survey done in the field of child safety. Using the result of this data mining process in the form of a decision tree, we can initialize our agent-based model with data from the survey and later validate the model comparing the results to the survey data. Consequently a framework is formed to test different agent profile based intervention techniques, so that a decision about selecting an intervention technique with a given cost can be demonstrated.

DEDICATION

Dedicated to

My Family and Friends

ACKNOWLEDGEMENTS

First and foremost, I am grateful to my advisors Dr. Robert Kent and Dr. Ziad Kobti, who gave me the opportunity to work in an exciting and challenging field of research. Their constant motivations, support, innovative ideas, own insight on research and enthusiasm have guided me toward successful completion of my thesis. My interactions with them have been of immense help in defining my research goals and in identifying ways to achieve them.

My sincere gratitude goes to Dr. Christine Thrasher for her valuable advice and helpful discussions during my thesis research. I would like to thank Dr. Dan Wu for his valuable comments and suggestions that further helped me in my research.

Finally I would like to thank my parents and my friends for their unconditional support and love.

TABLE OF CONTENTS

DECLARATION OF PREVIOUS PUBLICATION ..... **iii**

ABSTRACT ..... **iv**

DEDICATION ..... **v**

ACKNOWLEDGEMENTS ..... **vi**

LIST OF TABLES ..... **x**

LIST OF FIGURES ..... **xi**

CHAPTER 1 INTRODUCTION **1**

1.1 Current Research Motivation 2

1.2 Thesis Contribution 3

1.3 Thesis Outline 4

CHAPTER 2 LITERATURE REVIEW **6**

2.1 Validation in Agent-Based models 6

2.1.1 Empirical Validation 7

2.1.2 Predictive Validation 7

2.1.3 Structural Validation 7

2.2 Agent-based Models on Knowledge Flow Patterns and Prediction 9

2.2.1 Heterogeneity of Agents 10

2.2.2 Structure of Social Network 12

2.3 Child safety in Vehicles 16

2.4 Agent-based Models on Child Safety in Vehicles 19

2.4.1 Effects of Culture 19

2.4.2 Effects of Social Influence 22

2.4.3 Effects of Reputation 23

2.4.4 Predictive Data Mining 25

2.5 Domain of Thesis 26

CHAPTER 3 SOLUTION FRAMEWORK **28**

3.1 Initialization and Validation of Agent-Based Model 28
3.2 Intervention Policy Framework 29

CHAPTER 4 CHILD SAFETY SURVEY DATA ANALYSIS 31
4.1 Data Mining/Pre-processing 35
4.1.1 Data Cleaning 35
4.1.2 Data Pre-Processing 36
4.2 Regression Analysis 39
4.3 Decision Tree 44

CHAPTER 5 AGENT-BASED MODEL 47
5.1 Repast 47
5.1.1 General Repast Setup 48
5.1.2 Repast Setup for AutoSimModel 49
5.2 Initialization of Agent parameters/attributes 50
5.2.1 Learning Rate 50
5.2.2 Knowledge Deterioration Rate 51
5.2.3 Driving probability 52
5.2.4 Accident Rate 52
5.2.5 Reputation 52
5.3 Algorithms in the Simulation 53
5.3.1 Basic Intervention Framework 53
5.3.2 Cultural Algorithm 53
5.3.3 Social Network 54
5.4 Flow of Simulation 55
5.4.1 Algorithm 56
5.4.2 Flowchart of Agent-Based Model 58

CHAPTER 6 INTERVENTION POLICY FRAMEWORK 59
6.1 Intervention Policy 59
6.2 Brute force Method 64
6.3 Genetic Algorithm 65

CHAPTER 7 EXPERIMENTS AND RESULTS 67

7.1 Agent-Based Model 67
7.2 Intervention Policy: Brute Force Method 69
7.3 Intervention Policy: Genetic Algorithm 74
7.4 Explanation of methodology and results on an abstract level 76

CHAPTER 8 CONCLUSION AND FUTURE WORK **77**

REFERENCES/BIBLIOGRAPHY **80**

APPENDICES **88**

Appendix A: Best Intervention Policies (Brute Force Method) 88
Appendix B: Best Intervention Policies (Genetic Algorithm) 93

VITA AUCTORIS **97**

LIST OF TABLES

Table 4.1 Regression Analysis……43

Table 4.2 Decision Tree for Learning rate of agents……...46

Table 7.1 Sensitization Table for Genetic Algorithm……75

LIST OF FIGURES

Figure 2.1 The S curves of diffusion varying with degree of heterogeneity [13]……...10

Figure 2.2 Different types of networks and threshold [6]……...13

Figure 2.3 Speed of diffusion varying randomness and personal threshold (*h*) [13]……..15

Figure 2.4 The Cultural Algorithm [18]……..20

Figure 2.5 Average health loss in children in presence of social network …….21

(bottom) and its absence (top) [18]

Figure 2.6 Average health loss in children in presence of belief space (top)……..23

and both belief space and intervention (bottom) [19]

Figure 2.7 Predictive Data Mining: The Architecture [5]……...26

Figure 2.8 Domain of Thesis……...27

Figure 3.1 Framework for Validation of Agent-based Model……....29

Figure 3.2 Intervention Policy Framework…….30

Figure 5.1 Cultural Algorithm and Social Network……55

Figure 5.2 Flowchart of Agent-based model……...58

Figure 6.1 Process of Crossover……..66

Figure 7.1 Average knowledge of Child Safety over 180 days……...68

Figure 7.2 Comparison of Average Final Knowledge in different Age groups……..70

Figure 7.3 Comparison of Average Final Knowledge in different Gender groups…….…71

Figure 7.4 Comparison of Average Final Knowledge in different City Population……...71

Figure 7.5 Comparison of Average Final Knowledge in different Income Levels…….…72

Figure 7.6 Comparison of Average Final Knowledge in different Countries……..…72

of Primary driver training

Figure 7.7 Comparison of Average Final Knowledge in different Education Level……..73

CHAPTER 1
INTRODUCTION

Road accidents are one of the leading causes of deaths in children around the world [39]. Car seats are used to prevent children from serious injuries. These car seats reduce the risk of injuries in children by a significant amount, yet misuse of these seats is high, even in developed countries like Canada [27]. There has been a lot of research done by many government and non-government agencies to investigate the reasons for this misuse and to reduce it in an effort to increase child safety in cars. Many agent-based models have been developed for the same purpose, which predict what factors play a major role when it comes to improper use of child safety measures in cars. These models and simulations make use of concepts such as cultural learning, social networks, reputation of agents etc. Most of these models aim to predict the extent of spread of knowledge about child safety measures in cars over a period of time. These simulations present the user with a set of parameters in order to define and control various characteristics and behaviors. These parameters are used to drive the algorithms being used in the simulation. Some examples of these parameters are the learning rate, accident rate etc.

With the rise in better computing power, researchers and computer scientists have developed many simulations to dig deep into knowledge propagation [6-16,21-28], and hence use of agent-based modeling has increased for the same, as it has ability to model complex emergent phenomena, that more traditional modeling approaches cannot capture easily. In agent-based model, the individual or agent is the atomic model element, rather than the social system as a whole. Modeling of heterogeneous agents, their decision-

making processes and social interactions are very explicit in agent-based models. The macro-level dynamics of the social system emerge dynamically from the aggregated individual behavior and the interactions between agents. An end-to-end model, which can predict future trends by analyzing the patterns of knowledge propagation and the factors, which affect the rate and extent of knowledge flow, can be very useful when it comes to making decisions about policies and methods to promote the flow of knowledge.

Kobti et al. [40] introduce an agent-based model prototype for child vehicle safety injury prevention. This model is further enhanced by adding cultural algorithms [18] and social networks [19,20] aspects to it. These models aim to predict the factors responsible for the spread of knowledge related to child safety and the pattern/extent of the spread. The main drawback with these models was random initialization of the model and agent parameters. Ahmed et al. [6] introduce the idea of initializing these models by performing predictive data mining on a survey dataset related to child safety. This was the primary initial motivation for work presented in this thesis.

1.1 Current Research Motivation

One of the main issues with the present simulations in child safety is the validity of the model. There is no guarantee that the trends being shown by these simulations present an actual picture of what might happen in the real world. Major cause for this is high use of random parameters in these simulations to fill unknown values. Hence, an attempt is made to minimize this by using the values of parameters which are calculated after

analysis and mining of an actual survey data. The survey used here is the Canadian National Survey on Child Restraint Use 2010 [27], which was done in collaboration with the University of Windsor and AUTO21, Canada. Data pre-processing, regression analysis and mining is performed on the survey data in order to make a decision tree, which is then used to initialize the parameters in the agent-based model. This is an attempt to improve the quality and accuracy of the agent-based model when it comes to compare with real world data.

Moreover the simulations at present mostly revolve around homogenous agents. Heterogeneity of agents in these simulations has not been explored as much as it should have been. There are drivers around us with different age, gender, education level etc. Do these agent profile attributes like age, gender, education level etc. have anything to do with how they learn knowledge? Which intervention will yield better results: an intervention with young drivers or an intervention with older drivers? There has been no study, which can answer questions like these, taking into account the heterogeneity of agents to such an extent. This is an important aspect which plays an important and essential role in coining effective intervention policies.

1.2 Thesis Contribution

The aim of this research is to create an agent-based simulation on child vehicle safety based on an existing survey database, which performs close to real world and then create a framework through which we can test effect of different intervention policies on the

population using that simulation. The survey database is used to initialize different parameters in the agent-based simulation. Regression analysis and predictive data mining is performed on the survey database to extract these initialization parameters. Once the simulation is performing close to real world scenario, different intervention policies are tested on it. This testing is done by Brute force method and by using a Genetic algorithm.

Hence, the main goals of this study are:

- To create a close to real world agent-based simulation on child safety using regression analysis and predictive data mining on a survey database.
- To design a framework to test the effect and cost of an intervention policy on population using the agent-based simulation.
- To use exhaustive, or brute force, methods of analysis on intervention framework to determine general trends regarding performance of intervention policies based on different agent properties. This provides a basis for comparison for other modeling approaches.
- To use a Genetic algorithm to find the best intervention policy that can be performed under a given cost of intervention.

1.3 Thesis Outline

The main aim of this research is to create an agent-based simulation on child vehicle safety, based on an existing survey database, which performs close to real world and then create a framework through which we can test effects of different intervention policies on

population using that simulation. In order to discuss this, the thesis has been divided into the following chapters.

In Chapter 2, a literature review and survey is presented on Child Safety in vehicles, agent-based models on child safety, different agent-based simulations on patterns and prediction of knowledge flow and on the issue of validation of agent-based models.

Chapter 3 describes the survey database and different data processing, analysis and predictive data mining that were done on it. It also explains the formation of a decision tree based on the same processed database.

Chapter 4 describes formation of an agent-based model on child safety and different algorithms and techniques associated with it.

Chapter 5 describes the Intervention Policy framework and different methods that were used within the framework to explore the policies.

Chapter 6 presents the experiments that were done in the thesis and results produced by them, along with discussion of those results.

Finally, in the last chapter, the conclusions are presented and some potential future directions for this research direction have been suggested.

CHAPTER 2

LITERATURE REVIEW

This chapter includes a short report on previous works done in areas of initialization and validation of agent-based models. Some other agent-based models are discussed, which concentrate on patterns and prediction of knowledge flow, especially concentrating on factors that have major effects like heterogeneity of agents and different types of social networks. Then a small survey is presented on different works done in the field of child safety in vehicles and agent-based models developed on child safety. This survey also includes the terminologies related to these theories and models. It includes practical applications of these models in different fields, with focus on child safety measures.

2.1 Validation in Agent-Based models

Agent model validation has been a major issue in the area of social simulations, but yet there have not been many systematic considerations of whether different approaches to validation are appropriate for different approaches to modeling. Validation of models typically requires experts to look at the data, as errors and unwanted artifacts can appear in development of agent-based models. Some validation methods might be preferable to others when it comes to a particular style of agent-based models. Validation in agent-based models are broadly divided into following three categories [61]:

2.1.1 Empirical Validation

These validations are based on the comparison amongst the result obtained from the model and what we can observe in the real system. This gives a measurement of how good the model is in some given situations, but can't assure that it will prove with accurate results for situations which are different from those that can be observed in the real world. Moreover, just because the model gives the same results as the real world is no guarantee that the results have been obtained in the same way through the same processes.

2.1.2 Predictive Validation

This type of validation tries to give a proof that the results can be obtained through a model will have a validity in situations which are not directly observable in the real world. This is essential for purposes like "what-if" analysis and, in general, for the models that simulate non-repeatable phenomena like social and economic ones.

2.1.3 Structural Validation

This validation technique is concerned on the process by which the simulation results are obtained. A model can give results, which seem accurate, but are obtained through a totally different process than the real world. Hence the model should be examined and

inspected in order to guarantee that all the interacting parts are same as the corresponding real ones.

Windrum et al. [52] explain empirical validation procedures conditioned by their perspective as agent-based economic models. The discuss about a set of issues that are common to all models engaged in empirical validation giving rise to a novel taxonomy that captures the relevant dimensions along which agent-based models differ. They also explain three alternative methodological approaches being developed in empirical validation – indirect calibration [54], the Werker-Brenner approach to empirical calibration [55] and the history friendly approach [56].

Balci [53] presents guidelines for conducting verification, validation and accreditation (VV&A) of simulation models. Fifteen guideline principles are introduced to help researchers and practitioners comprehend what VV&A is all about. The activities under VV&A are described in modeling and simulation life cycle. The author also provides with taxonomy of 38 different V&V techniques for object oriented simulation models and 77 techniques for conventional simulation models. Baqueiro et al. [57] tackle the problem of standard verification and validation methodologies over agent-based modeling and simulation. Pure mathematical models deal with analytical equations only. The authors introduce integration of data mining with agent-based systems. They had technical difficulties to detect accurate and imperfect data in a given dataset.

Garcia et al. [58] research on validation process of marketing model along with

calibration, verification in the industry level, and harmonization. They tried to find out the best validation method, level of validation and how to learn which model was correct. A new calibration method is introduced, which is based on conjoint analysis that incorporate real world data into market based simulations. It is stated that conjoint data results are meaningful on an individual level and also on aggregate level, which is ideal for agent-based marketing models. Rand et al. [59] propose model validation by matching model components and process to real world, and by matching macro-level aggregate patterns, statistics and dynamics that were found across a variety of cases. They claim that macro measures they used provide useful information about the spatial patterns of real world. Sargent [60] performs data validation to develop theories, and mathematical and logical relationships in the model in order to create a conceptual model validation. Behavioral data is needed in the operational model validation. The theories and assumptions are tested using mathematical analysis and structural methods on data.

2.2 Agent-based Models on Knowledge Flow Patterns and Prediction

In this section, we will have a look on different agent-based models, which are used for prediction of knowledge spread and different factors that affect the spread. Everett Rogers [1] called this phenomenon of spread of knowledge in a society as 'Diffusion of Innovation' [2,3]. Different factors and elements of the models, which affect the phenomena of diffusion, are discussed below.

2.2.1 Heterogeneity of Agents

Agents can be usually of two types: homogenous and heterogeneous. Heterogeneous agents are those, who have varying degree of personal threshold and they are affected by word of mouth in different ways. Delre et al. [13] investigate how heterogeneity of agents effects the diffusion of innovation as shown in Figure 2.1. In the new proposed model, the consumer decides according to both his/her individual preferences and experienced social influences by other agents in the environment.

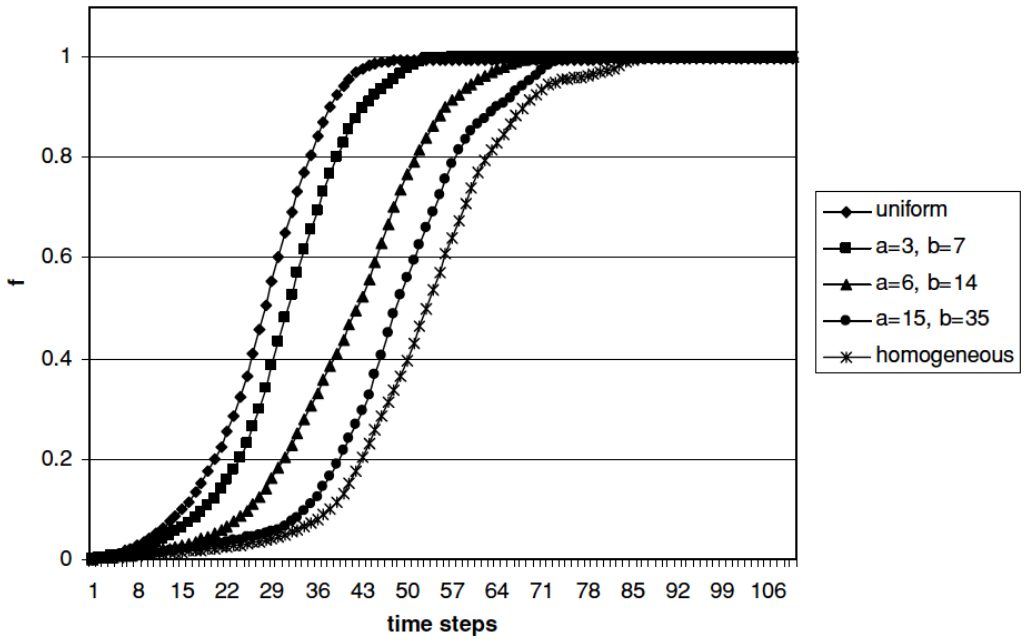

Figure 2.1 The S curves of diffusion varying with degree of heterogeneity [13]

Every agent communicates to its neighbors and diffusion happens through Word of mouth (WOM). Utility U of a product j depends on individual preference and social influence for a specific agent i. Agent adopts the product when $U_{i,j} > U_{i,j,min}$, where $U_{i,j}$

represents utility of product *j* for user *i* and *Ui,j,min* is the minimum utility for acceptance. It was observed that the speed of diffusion is low when personal threshold are high. Varying the heterogeneity in simulation resulted that more heterogeneous always causes a faster rate of diffusion. The authors claim that in more heterogeneous population, diffusion is better than homogeneous population, as critical mass is reached sooner.

Goldenberg et al. [15] investigate how the individual behaviors of adopters effect the collective diffusion of innovation. This is known as percolation model. This paper demonstrates how a microscopic presentation can be used for linking market level model to individual level behavior. It also allows examination of effect of heterogeneity in communication behavior of adopters on the aggregate adoption level. The percolation model has a critical percolation threshold *pc* such that for a given *Q* (quality of product), if $Q>pc$ an infinite cluster of neighboring buyers can be formed, while for $Q<pc$ all clusters of buyers are finite.

Alkemade and Castaldi [6] investigate whether a firm can learn about consumer characteristics given limited information and come up with a successful directed advertising strategy. The authors use the concepts from the literature on epidemics and herd behavior to study the problem of diffusion of innovation. A special genetic algorithm is used for the simulation based on the principle of "survival of the fittest". A population is randomly initialized with different strategies as genotypes. Now this population is improved in different generations by selection, recombination and mutation.

Hence better strategies are passed to next generation. Different diffusion dynamics are used by altering topology, advertising strategy and consumer characteristic.

It was seen that when using homogenous consumers, for random strategies, it is necessary for network of different agents to be connected for occurrence of cascading. It happens easily over random networks. With direct advertising, cascades are achieved easily on regular networks. When dealing with heterogeneous consumers, learned strategies outperform random ones in every aspect like size and speed of diffusion. The authors claim that firms can learn a direct advertising strategy taking into account both topology of social network and consumer characteristic. These outperform the random advertising strategies.

2.2.2 Structure of Social Network

The three main types of social networks discussed in this section are random network, highly clustered and scale free network. Abrahamson and Rosenkopf [4] were one of the first to state the effects of social network structures on diffusion process. They introduced the idea that each potential adopter experiences a different pressure for adoption, which depends on the social structure of the network and number of connection that adopter has, along with price, efficiency and legitimacy of the innovation.

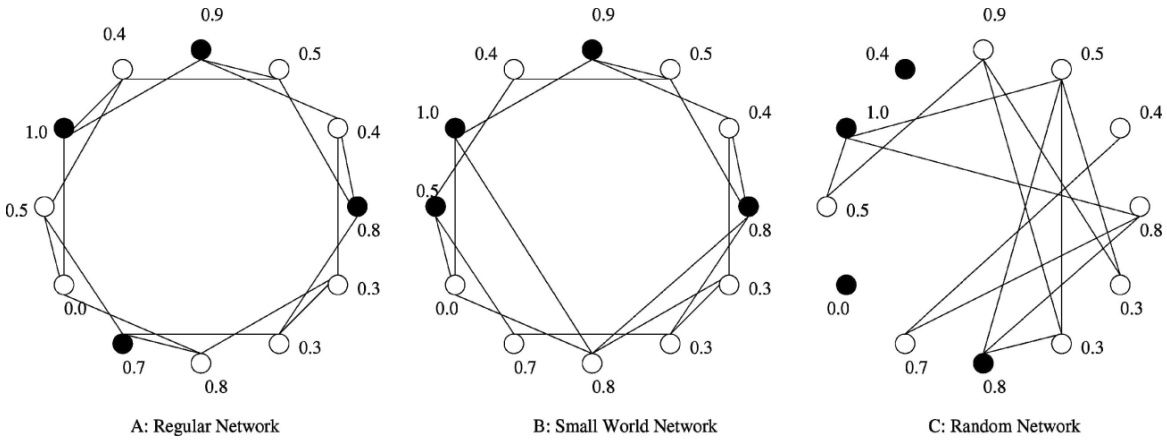

Figure 2.2 Different types of networks and threshold [6]

Three sets of simulations were performed. The first one tests propositions using a basic model of faddish diffusion. The second one explores the robustness of these findings assuming that every firm is not equally sensitive to information creating bandwagon pressure. The third set of simulation explores how these findings differ when model based on Learning is used rather than Fad theories [62]. The basic model simulation showed that an increase in network density increases the bandwagon pressure. Also the greater the number of pressure points and weaknesses at the boundary of a non-focal stratum, greater the adopters in it. Boundary pressure points and weaknesses have a greater effect on extent of diffusion than higher density network.

Delre et al. [13] relates degree of randomness in a network to innovation diffusion. Simulations are run with varying values of network randomness r [.0001– 1], alters L [1,2], weight of individual preference (y) Vs social influence (x) b [.4– 1], and personal threshold h [0,0.6]. Graph of Diffusion rate r Vs Randomness rare plotted for

every case.

When Randomness was varied, there was a maximum rate of diffusion found at r = 0.1. When compared against different values of alter, the trade-off was at L= 1.3.When varying value of weight b, it was found that randomness of network effects rate of diffusion more drastically when value of b is high. The author later claims that highly clustered networks support faster diffusion than random networks. Choi et al. [10] talk about network structure along with effects. The conditions of simulations are little different than as done by Delre et al. [13]. The results show that failed diffusions are more likely to happen, when network is highly random as shown in Figure 2.3. But surprisingly random links facilitates rapid diffusion process. Authors claim that presence of bridges (random network) reduces average social distance in a network and hence increasing speed of diffusion, but it might cause under-adoption. On the other hand, cliquish networks (highly clustered network) facilitate building up an early customer base, but it inhibits rapid diffusion. So the best strategy would be to work with a mix of both strategies. Kuandykov and Sokolov [19] compare random networks to scale free networks. Alkemade and Castaldi [6] discuss about different types of networks and threshold in theses networks, as shown in Figure 2.2.

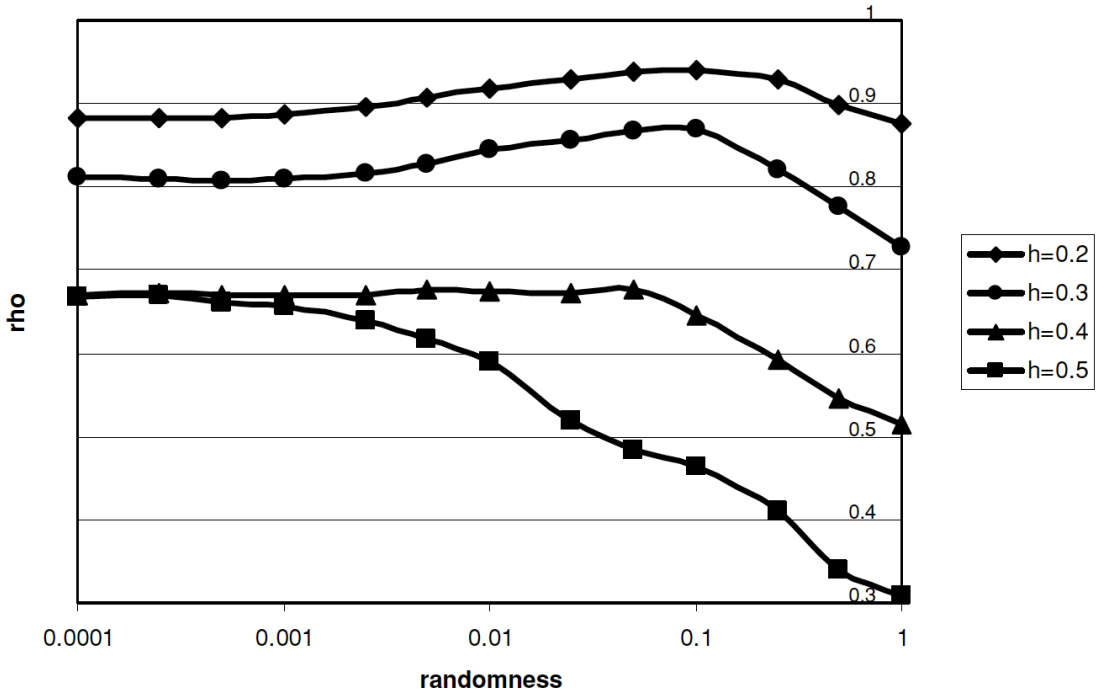

Figure 2.3 Speed of diffusion varying randomness and personal threshold (*h*) [13]

In the case of random networks, there are three cases: random network with each node having same number of neighbors, 3 clusters connected sequentially and absolutely symmetrical social network in which, all agents can establish links with other agents from their native or other clusters. Initial number of adopters for all simulations is 30. In case of scale free networks, most agents have few links (nodes) while some have lots of connection (hubs). The two cases for scale free networks are hubs and nodes as initial adopters.

The author states that diffusion was slower when the network was totally random compared to when random networks were divided into clusters. The diffusion flowed from adopter cluster to other clusters depending on the way they were connected.

In case of scale free networks [19], which are networks whose degree distribution follow a power law, diffusion in hubs as initial adopters case was way faster than the case, where nodes were the initial adopters. The reason for this observation is "information equality", where a network with higher information equality has higher diffusion of innovation. The authors claim that in random network, the diffusion of information is faster if network is split into clusters. Longer diffusion time in case of scale free network is related to lower information equality in comparison to random networks. Diffusion in scale free network is faster if initial adopters are hubs instead of nodes.

2.3 Child safety in Vehicles

Road crashes have been the leading cause of minor and fatal injuries amongst children in Canada, who are less than 14 years in age. Approximately 2 children die or are seriously injured everyday as a result of road crashes [41]. Different universities, non-profit organizations and government agencies have done numerous studies and surveys to figure out the reasons behind non-usage and misuse of proper child safety measures in vehicles. This section provides a small overview on these studies and their findings.

Apsler et al. [42] make an attempt to increase the usage of booster seats amongst low-income parents. A pre-test/post test design was conducted in daycare centers with post-test observation leading up to 8 weeks after the intervention. Parents participated in

an educational training and received free seats. Educational programs were provided to daycare staffs and children, and signs were put up in parking lot. This reduced the percentage of unrestrained children in vehicles from 56% to 26%. Ebel et al. [43] conducted a survey to measure booster seat usage and determine the factors predictive of proper child restraint and assess parental reasons for booster seat use and non-use. Cross-sectional, observational studies were done in Seattle, Washington and Portland, Oregon, where drivers were surveyed after picking up children from schools and daycare centers. Trained observers recorded child height, weight and age and directly observed restraint use. This was compared to recommended restrained method based on child's observed age, weight and height. Only 16.5% of children who should be in a booster seat were properly restrained compared with 80% of younger children, for whom, child safety seat was recommended. Relative to a 4-year-old child, a 6-year-old was half as likely to be in a proper booster seat. Many parents incorrectly believed that children are safe in a seatbelt and that they have outgrown the need of a special car seat.

Lee et al. [44] performed a study, which investigates child safety knowledge, the attitude and belief about booster seats in Latino parents. They also explore the effective strategies for message delivery in Latino community. Focus groups were conducted with Spanish speaking parents and information was collected through written survey and discussions. They found out that parents were widely misinformed about rules and guidelines for booster seat usage. Most of the participants did not own a booster seat. It was concluded that culture specific campaigns are needed to promote booster seat usage in Latino community. The guidelines should be preferably provided in Spanish.

Factors that influence use of booster seats in a multiethnic community are explored in Johnston et al. [45]. Three focus groups were conducted with low-income residents of central and southeast Seattle, Washington. Participants were especially sought from Somali, Vietnamese and African American communities. Recruitment of participants was done through posters, flyer and information booths at clinics, community centers etc. It was found out that participants expressed a lack of understanding about the working of booster seats in protecting child passengers, and how are they differ from a car seat. They attributed the lack of usage to ignorance or laziness among community members who do not value their children's life. They even expressed concerns regarding their own capability to practice usage of booster seats consistently. There were a lot of differences noted in different ethnic and linguistic groups. A need of education and training around booster seats and law requiring their use was identified.

Intervention studies about child safety in vehicles were done in Zaza et al. [46] and Pierce et al. [47]. A systematic development team reviewed scientific evidences of effectiveness of five interventions to increase child safety seat usage [46]. Community wide information plus enhanced enforcement campaigns and incentive plus educational programs had sufficient evidence of effectiveness. Education only programs aimed at parents, young children and healthcare professionals were seen as not being that effective comparatively. The main objective of [47] was to determine the knowledge level of head start providers, parents and students about booster seats. Booster seat usage before and after a combined educational program and booster seat giveaway was also observed.

2.4 Agent-based Models on Child Safety in Vehicles

In this section, the discussion is about work done in area of child vehicle safety using agent-based models. The general agent-based framework, which has been used for prediction of knowledge flow, child safety knowledge in this case, and the factors that affect the knowledge spread are explored.

2.4.1 Effects of Culture

Kobti et al. [18] discuss about modeling effects of social influence on driver behavior in applying child vehicle safety restraint. They use a cultural algorithm for the same. It enables drivers to learn from their individual driving experience with an option for immediate feedback from an expert intervention source following an accident. The cultural algorithm enables population level learning and captures dominant social beliefs among the drivers.

Situational knowledge is implemented in the belief space, which is based on top performing drivers. It was seen that in presence of a cultural belief system, the system that measures the correctness of use of child vehicle safety was positively influenced. But on the other hand, the population was more resilient to changes after an intervention. This portrayed that culture plays an important role when it comes to interventions and should be considered as a major factor by health practitioners. The introduction of cultural

framework is modeled to present a realistic reflection of the population model. It plays an important role in guiding the learning process of drivers after an intervention by health care practitioners.

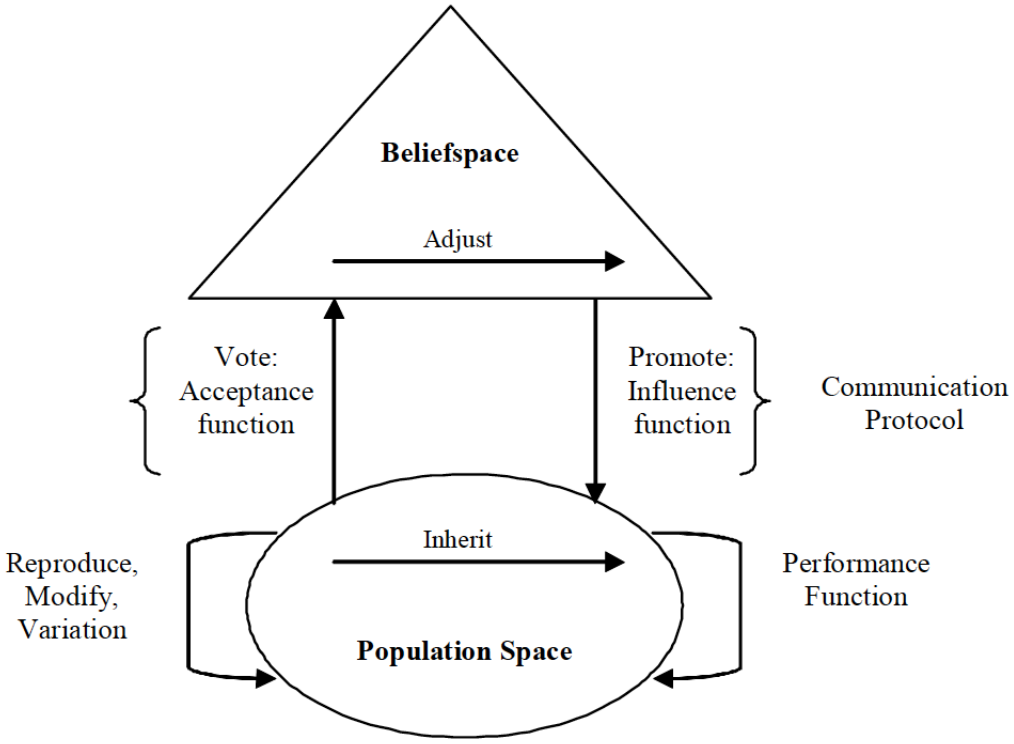

Figure 2.4 The Cultural Algorithm [18]

The cultural algorithm consists of a population and a belief space. The selected individuals from the population contribute to the knowledge in belief space depending on the acceptance function. The knowledge in the belief space is manipulated and changed based on individual experiences and their success or failures. The knowledge controls the evolution of the population using an influence function [Figure 2.4].

It was seen that learning from the expert source alone was most efficient. In the absence of cultural influence, the population demonstrated the most efficient use of child

safety measures [Figure 2.5]. The system resists change when cultural influences are present. The intervention methods both at population and individual levels were hindered by cultural buffer, which suggests that despite having some improvement, the system did not reach its full potential.

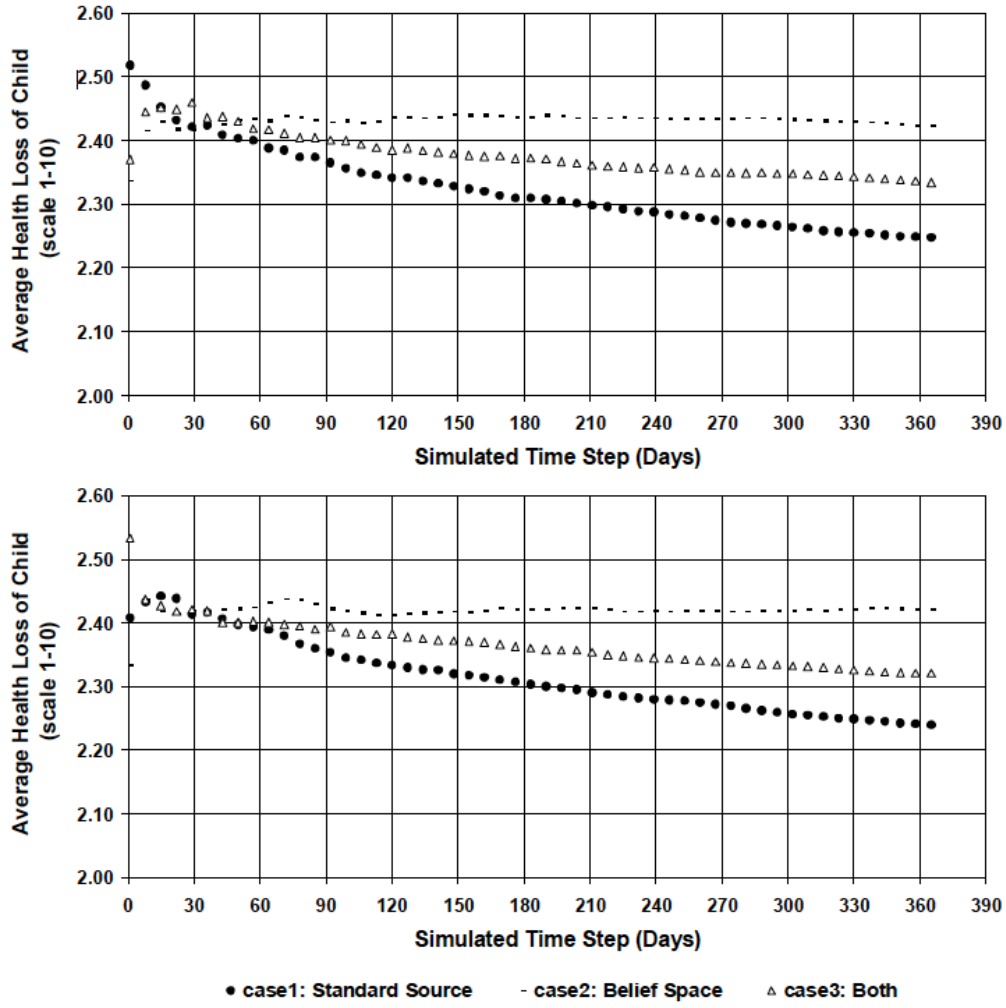

Figure 2.5 Average health loss in children in presence of social network (bottom) and its absence (top) [18]

2.4.2 Effects of Social Influence

In a framework which is socially motivated, the modeled agents or drivers are able to identify friendships and neighboring relations. In Kobti et al. [19], both positive and negative exemplars are used in the cultural algorithm to guide the belief at population level. Based on evolving individual experiences and changes in the belief system, both positive and negative exemplars influence the overall children population health and improved the possibility of drivers selecting the correct child seat.

The belief space is restricted to situational knowledge, where it encapsulates sets of best and worst examples taken from most influential individual experiences. Agents with positive experiences contribute to the good knowledge patterns and the one ones with bad experiences are used to prevent individuals from selecting failed strategies. Belief space is updated every 7 days, where population space is searched for top 2% of the best and worst drivers with best and worst performance, and the belief space is updated with their knowledge.

It was observed that the drivers were able to learn from both positive and negative experiences. Maintaining a set of worst patterns enabled the drivers to avoid the common mistakes and improve their performance. The negative pattern turns into a lesson, which need not be repeated by new drivers and hence contributing to learning process.

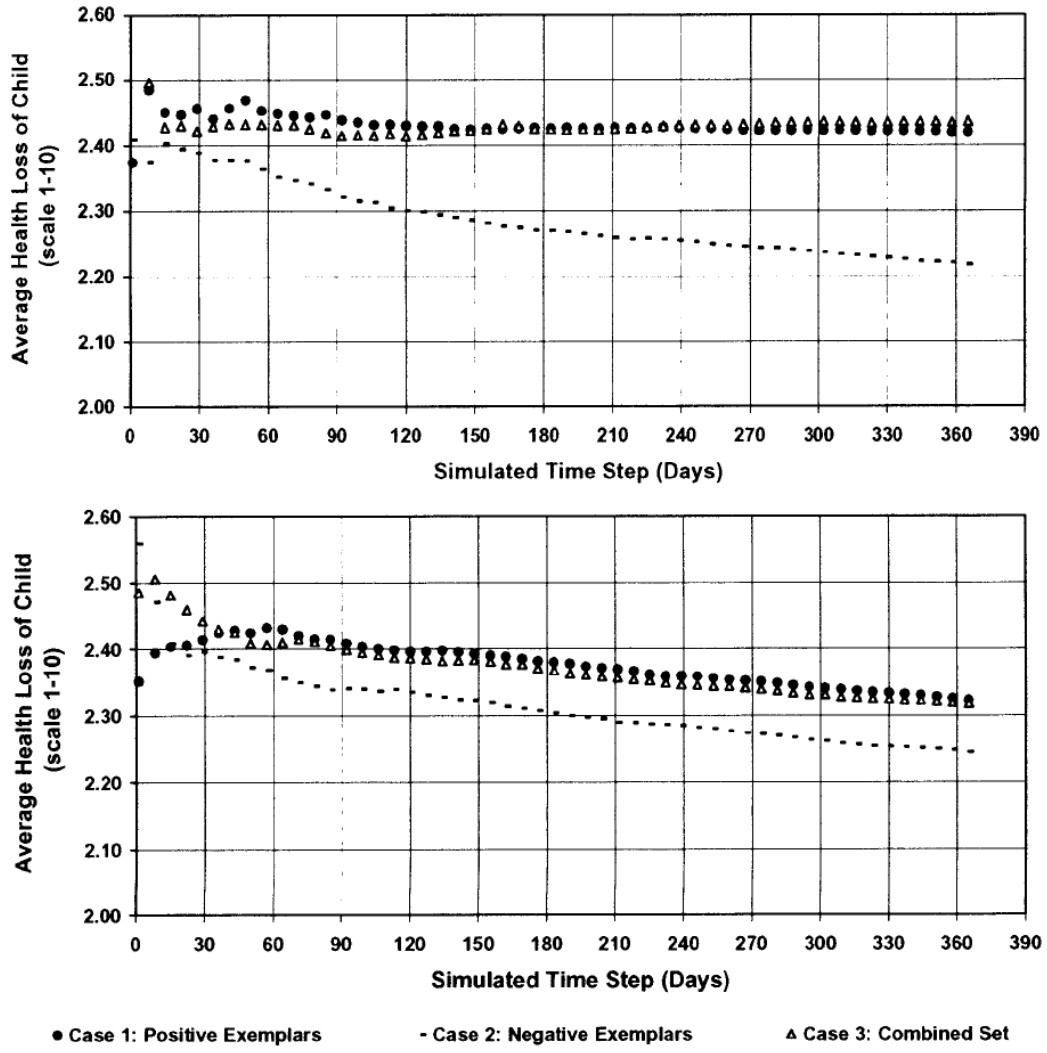

Figure 2.6 Average health loss in children in presence of belief space (top) and both belief space and intervention (bottom) [19]

2.4.3 Effects of Reputation

Modeling reputation of agents in a complex social simulation presents a significant challenge due to its distinct social nature. Kobti et al. [20] introduces a notion of reputation into child vehicle safety simulation. They hypothesize that selective

intervention criteria would achieve a better system convergence and introduces reputation as a variable for the same. They establish a generic reputation framework that tests alternate formalizations of reputation models.

Reputation refers to trustworthiness of an agent in an artificial society. This framework allows external injection of knowledge, or intervention in health sciences, or a new strategy into the artificial society. They claim that a better performance can be achieved if agents could be carefully selected under some social criteria allowing efficient knowledge propagation in the society through the network. At each time step, the model is updated to reflect the changed reputation of agents. The algorithm, which is responsible for driving the logic of agent collaboration is also altered. Agents use the reputation to decide on the transfer and level of acceptance of the transferred knowledge.

The first reputation model assumes that reputation of an agent only depends on its degree of connectivity to the social network around it. The second model extends the previous models saying that reputation should also depend on quality of knowledge (QK) of the agent. The Reputation Index (RI) also depends on Income level (IL) and Education level (EL), which are more of agent properties rather than something that depends on the social network. The authors claim that from a network perspective, high degree nodes in a social network are not sufficient to be considered along in a reputation model, but rather a model rich with domain knowledge and agent characteristics would be more favorable.

2.4.4 Predictive Data Mining

Ahmed et al. [5] explore the use of predictive data mining, which aims at exploration of parameters that initialize the child safety model. They claim that existing data from surveys can be examined using data mining tools, exploring beyond basic statistics what parameters and values can be most relevant for a more realistic model run. The intent is to make the model replicate real world conditions as closely as possible, mimicking the survey data. This helps to discover patterns amongst drivers who have higher probability of improper usage of child car seats.

This framework uses predictive data mining technique to make predictions about values of data used in an agent-based model, using known results found from survey data. It focuses on predictive data mining technique using decision tree classification. A decision tree is a series of questions systematically arranged so that each question queries an attribute (e.g. age of the driver) and branches based on the value of the attribute. At leaves of the tree are placed predictions of the class variable (e.g. type of car seat used). The proposed architecture collects survey data from a database and generates a Decision Tree model on the fly. It also provides an Application Program Interface (API), which will be used by the Car seat model for initialization, prediction and validation. The system constitutes of three modules namely Data pre-processing module, Data mining module and API module. It highlights that data mining techniques can be used in agent-based models to overcome the gap between the real world and simulation.

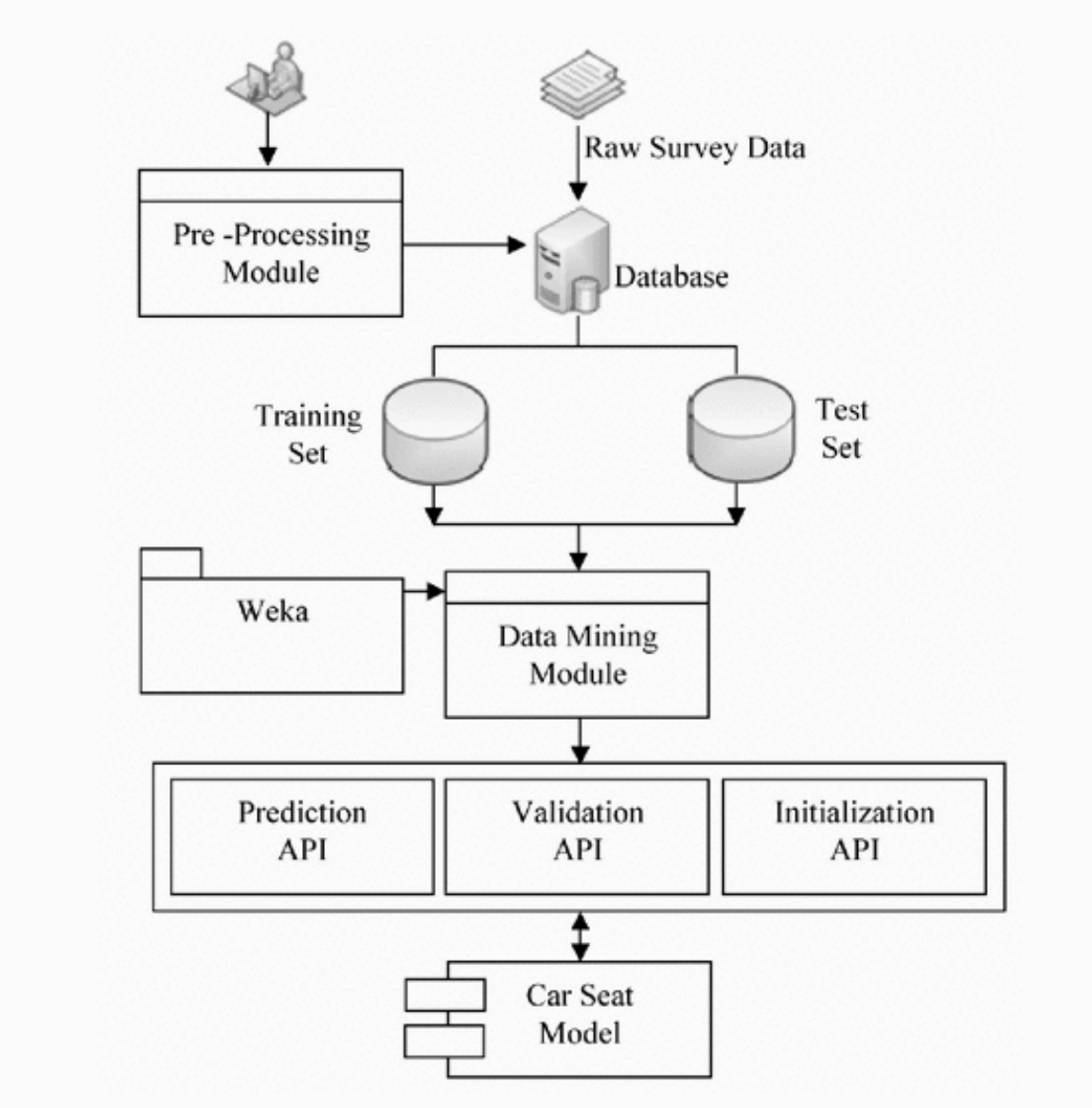

Figure 2.7 Predictive Data Mining: The Architecture [5]

2.5 Domain of Thesis

Figure 2.8 represents the domain of work done in our research study. It explains different concepts and theories that have been used to construct the whole framework. We have used Random network [19] to implement Social network. The heterogeneity of agents has

been extended to use of agents with different profiles in this research work. The initializations of model and agent parameters are being done through data mining of survey database and we are using empirical validation to validate our model. Then we implement the intervention policy testing framework, which is a totally new contribution by this thesis.

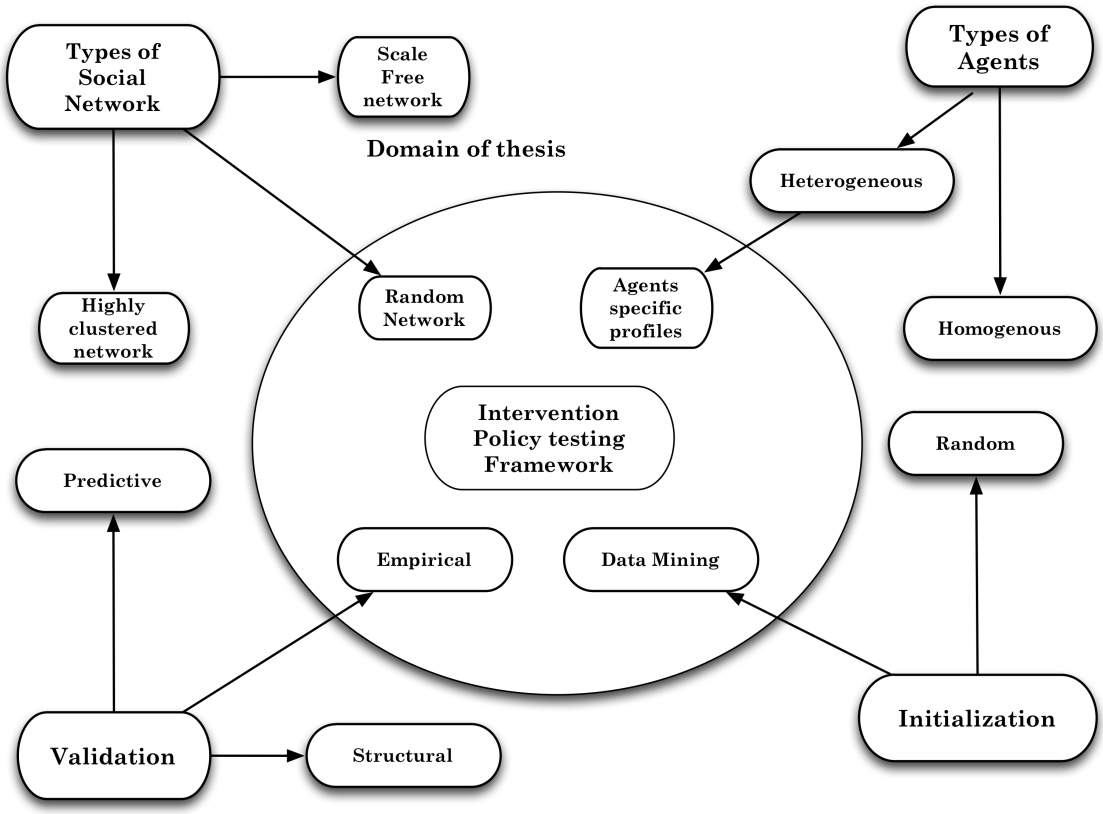

Figure 2.8 Domain of Thesis

CHAPTER 3
SOLUTION FRAMEWORK

The solution framework proposed in this thesis is divided into two major steps. Step one concerns with initialization of the agent-based model using predictive data mining on survey database, and validation of the model as shown in Figure 3.1. Step two involves using an intervention policy framework to test performance of different intervention strategies on the agent-based model. These intervention policies can be tested using brute force method or a genetic algorithm [Figure 3.2].

3.1 Initialization and Validation of Agent-Based Model

1. A survey database is created, which is based on a real world problem.
2. Predictive data mining is performed on the survey database.
3. An agent-based model is conceptualized and implemented based on the real world problem.
4. Agent profile parameters like age, gender etc. are initialized using the data from the database.
5. Agent behavior parameters are initialized by the mined data that we get after the predictive data mining.
6. The agent-based model is executed and final result is compared against the database for validation.

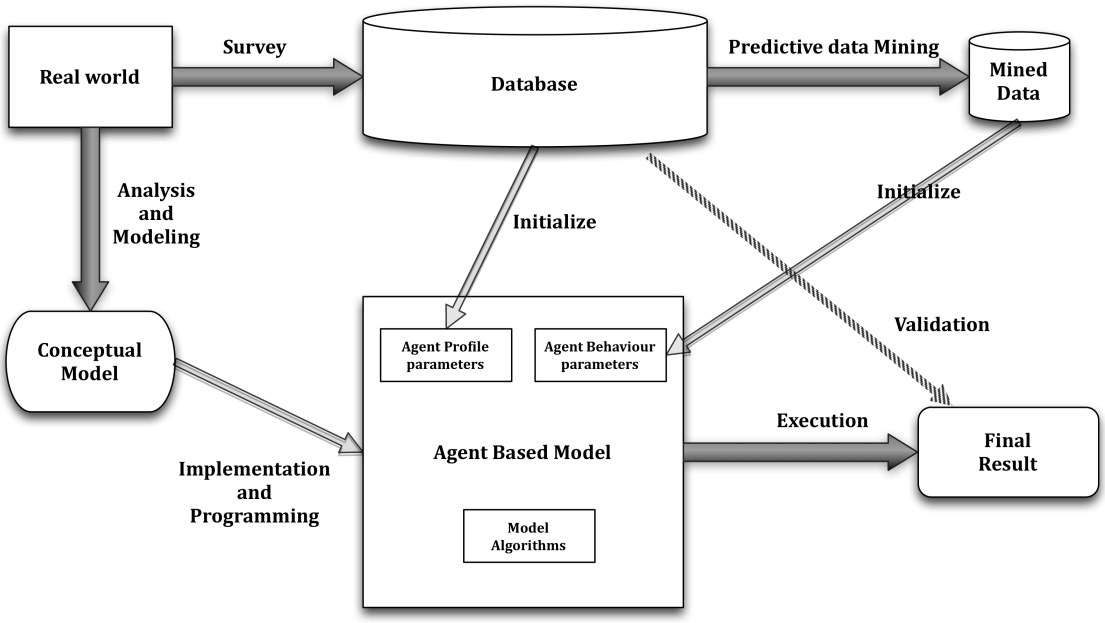

Figure 3.1 Framework for Validation of Agent-based Model

3.2 Intervention Policy Framework

1. Use the validated agent based model from section 3.1
2. Generate intervention policies using Intervention policy generator and test them on the agent based model.
3. Intervention policy generator can generate policies using brute force method or genetic algorithm.
4. The performance of an intervention policy can be tested by the final result it produces, when that policy is applied on the agent-based model.
5. These different policies can now be compared against each other using their performance and cost as a measure, to come up with best possible policies.

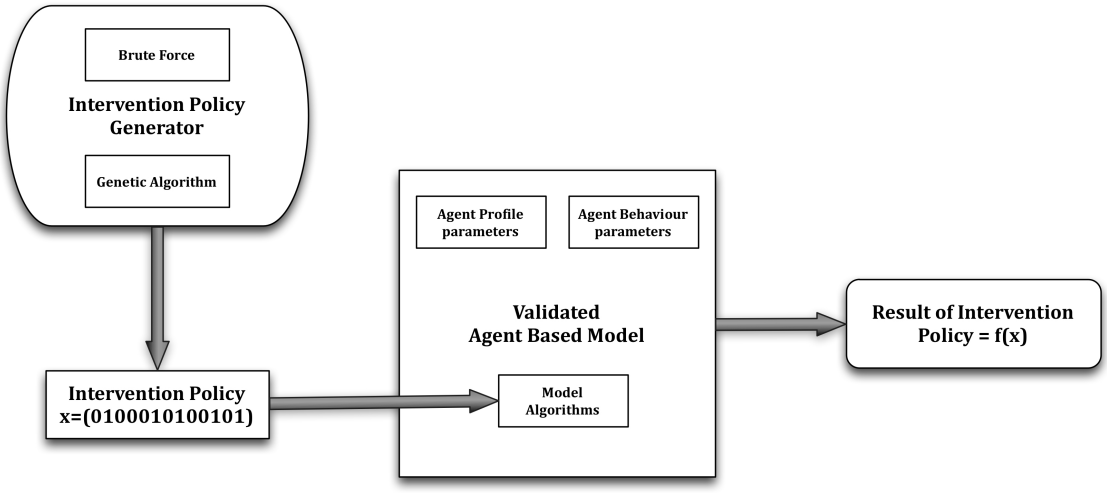

Figure 3.2 Intervention Policy Framework

In the next upcoming sections, we discuss implementation of this framework on child safety measures in vehicles.

CHAPTER 4

CHILD SAFETY SURVEY DATA ANALYSIS

The Canadian National Survey on Child Restraint use [29,41] was conducted by Transport Canada in partnership with Auto21 [30] and professors from Business and Statistics department at University of Windsor. This study was a follow up to the 2006 National child seat survey submitted to Transport Canada in 2007. In the previous technical report on Canadian National Survey on Child Restraint Use (2007), it was found that although most drivers used some type of safety restraint system, the rate of correct use of safety seats varied among different age groups. This survey was used as the base for construction and validation of the agent-based model.

In this survey, participants were asked 9 questions related to Child safety measures in cars. This survey was done in 5 provinces of Canada. The questions asked in the survey were as follows:

1. What is the correct age to move a child from rear facing seat to a forward facing seat?
2. What is the correct weight to move a child from rear facing seat to a forward facing seat?
3. What is the correct height to move a child from rear facing seat to a forward facing seat?
4. What is the correct age to move a child from forward facing seat to a booster seat?

5. What is the correct weight to move a child from forward facing seat to a booster seat?
6. What is the correct height to move a child from forward facing seat to a booster seat?
7. What is the correct age to move a child from booster seat to a seat with seat belt?
8. What is the correct weight to move a child from booster seat to a seat with seat belt?
9. What is the correct height to move a child from booster seat to a seat with seat belt?

Apart from these questions, each participant was asked the following personal information. Except for age, all the questions were multiple-choice. The possible options for each question are mentioned below along with the question:

1. Age: Numeric Value

2. Gender: 1 = Male
   2 = Female

3. Marital Status: 1 = Single
   2 = Married/ Common law
   3 = Separated/ Divorced
   4 = Widowed

4. Ethnicity: 1 = Caucasian

   2 = Native Canadian

   3 = African Canadian

   4 = Asian

   5 = Arabic

   6 = Hispanic

   7 = East Indian

   8 = Other

5. Income Level: 1 = Under $20,000

   2 = $20,000 – 40,000

   3 = $40,000 – 60,000

   4 = $60,000 – 80,000

   5 = Over $80,000

6. Education level: 1 = Grade School

   2 = Some High school

   3 = High School Graduate

   4 = Some post-High School

   5 = College Diploma/ Certificate

   6 = University Degree

7. Population of city person lives in: 1 = Over 300,000

   2 = Between 100,000 – 300,000

   3 = Between 30,000 – 100,000

   4 = Between 1,000 – 30,000

   5 = Under 1,000

8. Was first driver training done in Canada: 1 = Yes

   2 = No

The response to each question was noted and kept for records. After the participants took the survey, an informative pamphlet was provided to each of them. These pamphlets contained the correct information about child safety measures in cars, including the correct answers to questions asked above. This stage is called as initial intervention in our study, when each participant is intervened/provided with knowledge about child safety.

After the first stage of questionnaire, the same survey is done again after 6 months, where same people who participated in stage one of the survey answer same questions for the second time. This gives us a quantitative measure of their knowledge about child safety measures in cars at two different times. This collected knowledge is put through the process of Data mining, Data Pre-processing, Regression analysis and Decision Tree formation, so that the agent-based model can be prepared and initialized with the processed data from the dataset. This will result in a better agent-based model,

which is closer to real world, when compared to simulations that use random values for initialization of agent parameters.

4.1 Data Mining/Pre-processing

After the surveys were done and the data was collected, came the stage of data cleaning, mining and pre-processing. This is a necessary step, as the data collected cannot be used in an agent-based model to initialize different parameters in its current form. This data has to go through a process of cleaning and pre-processing, so that it's fit to be used by the agent-based model for parameter initialization and other purposes. The actions taken to make the dataset capable of being used are explained below.

4.1.1 Data Cleaning

Data cleaning is the process of detection, correction and removal of corrupt records from a dataset, to get rid of all the dirty data and hence making it usable. All the entries in the provided dataset, which were not entered properly for every field were got rid off. All the fields should be properly entered for every person who took the survey; or-else the record is unusable for the agent-based model. 484 usable entries were left after getting rid of all the corrupt data. This meant that 484 participants took the survey properly and hence a maximum of 484 agents can be used in the agent-based model.

4.1.2 Data Pre-Processing

After the process of data cleaning, data pre-processing was performed on the dataset to make it usable with the agent-based model. The main objective of this process is to break and convert the dataset into a format, which can be parsed by our agent-based model as a .csv file, and can be used to automatically initialize different parameters in the simulation. The main pre-processing, that were performed on the dataset are explained below.

Pre-Processing on Age

In the given dataset, age was represented by a numeric value for e.g. 24. Since properties like age, marital status etc. are being used to create different agent profiles, using the actual numeric value of age for agents will result in numerous agent profiles, which might make the results less conclusive. For example, if age of all the participants ranges from 20 to 60, this will give 40 different agent profiles under age, which is a lot to handle for the agent-based model. Hence age is categorized into 4 groups. These groups are 20s(20-29), 30s(30-39), 40s(40-49) and 50s(50-59). The value of age in different records is changed accordingly. For example, 24 is replaced by 20, 36 is replaced by 30 etc. This gives just 4 different groups in our age field, which makes the job for framework easier.

Converting knowledge to bits

The designed framework deals with knowledge of the agents in a specific format. To have a quantitative measure for knowledge level of participants, the knowledge of every participant is needed in a bit format. As we see in section 4.1, there were 9 questions asked to every participant in the survey. Each of these questions has a correct answer, as stated below.

1. What is the correct age to move a child from rear facing seat to a forward facing seat? – 12 months
2. What is the correct weight to move a child from rear facing seat to a forward facing seat? – 26 inches
3. What is the correct height to move a child from rear facing seat to a forward facing seat? – 22 pounds
4. What is the correct age to move a child from forward facing seat to a booster seat? – 48 months
5. What is the correct weight to move a child from forward facing seat to a booster seat? – 40 inches
6. What is the correct height to move a child from forward facing seat to a booster seat? – 40 pounds
7. What is the correct age to move a child from booster seat to a seat with seat belt? – 96 months
8. What is the correct weight to move a child from booster seat to a seat with seat belt? – 57 inches

9. What is the correct height to move a child from booster seat to a seat with seat belt? – 80 pounds

For each of these questions, the answer given by the participant was either correct or incorrect. To convert these answers to bit format, "1" was assigned when the answer given was correct and "0" for every incorrect answer. This converts the knowledge of participants about child safety measures into a bit format, which is then easier to be dealt with while using the agent-based model.

Knowledge Level

Due to the pre-processing done in previous section, the knowledge of each and every participant is now converted into bit format. Since there were 9 questions asked in the original survey, the knowledge of each participant can be represented by a 9-bit array, where each bit represents a value, which tells us if the participant answered that particular question correctly or not. So the typical knowledge of a participant will look like below

Knowledge:	0	1	1	0	0	1	0	1	0

Each bit above represents if the participant gave the answer to the question associated with that bit correctly or not, depending on the value of the bit ("0" or "1"). This is called the knowledge array.

Now knowledge level of each participant can be derived from the above given knowledge array. Knowledge level is simply defined as the number of "1"s in the knowledge array. So for the knowledge array shown above, the knowledge level will be 4. It is to be noted that the knowledge level is a value between 0 and 9, 0 being the least possible knowledge level and 9 being the highest. So eventually, there are two knowledge levels for every participant; initial knowledge, which is the knowledge level on day 1 from the survey before the intervention stage and final knowledge, which is the knowledge level on day 180 of the survey. These are named K_i and K_f. The knowledge change K_c is defined as the difference between K_i and K_f. Hence

$$K_c = K_f - K_i$$

K_c can hold a numeric value between -9 to +9.

4.2 Regression Analysis

Regression analysis is a statistical analysis technique, which is used to estimate relationship between different variables. Analysis and modeling of relationships between a dependent variable and many independent variables can be done using this type of analysis. It lets us examine how the values of a dependent variable change when an independent variable varies. SPSS is used to perform this analysis, which is a statistical analysis software tool from IBM [31].

In the given survey, the agent profile i.e. age, gender, education level etc. are the independent variables and knowledge change K_c is the dependent variable. There is an assumption made that the knowledge change of the participants is dependent on their agent profile. Regression analysis is used to explore how properties of the participants affect their knowledge change and to what extent.

Some other data pre-processing is performed on the database before the start of regression analysis. To establish a proper relationship between dependent and independent variables, the distribution of data should be reasonable among different variables. Upon examination, it was seen that out of 484 entries, more than 90% have 'Caucasian' as their 'Ethnicity' and 'Married' as their 'Marital status'. Therefore it can be concluded that data distribution within these variables was not significant enough to be included in our regression analysis as an independent variable. The relationship between these variables and knowledge change might not be accurate due to lack of even distribution of data. Therefore, age, gender, income level, education level, driver training and population of city are used as independent variables and knowledge change is the dependent variable.

The accuracy of regression analysis is highly dependent on the number of probable values of dependent variable that are being predicted. Lesser the number of possible outcomes of dependent variables, the stronger will be the relationship between dependent and independent variables. The possible values of dependent variables K_c here

are 19 (from -9 to +9). We divide this knowledge change into three categories as show below:

Knowledge change = Decrease, if K_c is between -9 and -3

Knowledge change = Constant, if K_c is between -3 and +3

Knowledge change = Increase, if K_c is between +3 and +9

The possible outcomes of dependent variable, knowledge change, are reduced to 3 using the classification shown above. The numbers of possible outcomes for the independent variables are also reduced, as not every outcome has significant number of entries. After this process, the dataset takes the following structure:

1. Age: 20 = in 20s
   30 = in 30s
   40 = in 40s or greater than 40

2. Gender: 1 = Male
   2 = Female

3. Income Level: 1 = Under $20,000
   2 = $20,000 – 40,000
   3 = $40,000 – 60,000
   4 = $60,000 – 80,000

5 = Over $80,000

4. Education level: 1 = Grade school/ Some High school/ High school graduate

   2 = Some Post-High school

   3 = College Diploma/ Certificate

   4 = University Degree

5. Population of city person lives in: 1 = Over 300,000

   2 = Between 100,000 – 300,000

   3 = Between 30,000 – 100,000

   4 = Under 30,000

6. Was first driver training done in Canada: 1 = Yes

   2 = No

This new modified dataset now goes though the process of regression analysis, where age, gender, income level, city population, education level and country of driver training are the independent variables and modified knowledge change, as explained above is the dependent variable. Table 4.1 shows the result of regression analysis.

Coefficients					
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
	(Constant)	.348	.245		1.421	.156
	Parent Age	-.005	.005	-.048	-1.005	.546
	Parent Gender	.041	.080	.024	.515	.606
	Income Level	-.056	.028	-.102	-2.016	.044
	Driver Training	-.081	.074	-.053	-1.090	.276
	City Population	.067	.032	.098	2.105	.036
	Education Level	.022	.036	.032	.605	.315
Dependent Variable: Knowledge Change						

Table 4.1 Regression Analysis

In the above table, the field of importance is 'Sig.'. This field is an indicator of strength of relationship between the independent and dependent variable. The lower the value in this field, the stronger is the relationship. The 4 independent variables with the lowest value for the study i.e. City Population, Income Level, Education Level and Driver Training are chosen for further investigation. These variables have a strong effect individually on the knowledge change of the participants who took the survey according to the result of regression analysis.

4.3 Decision Tree

A decision tree is a tool that helps us in decision support, using a tree like model of decisions and their consequences. It even includes probability by which the outcomes occur, their resource cost, and utility. It helps in identifying strategies by which the set goal can be achieved. When used with data mining, it describes data but not decisions. The resulting tree can be used as input for decision-making, as in the case below.

A J48 pruned tree will be constructed here, using Weka data mining tool [32, 34]. Pruning is a process in machine learning by which the size of a decision tree can be reduced. This is done by removing sections of the tree which provide a little power of classification of instances. The goal of pruning a tree is to reduce complexity and have improved accuracy by removal of sections which are based on noisy data. J48 is an open source java implementation of the C4.5 algorithm of decision tree generation [33]. This implementation is done in the Weka data-mining tool [34], which will be used in this study. The nodes of the tree are different values of independent variables that we selected, and the leaves of the tree are the predicted knowledge change of the survey participant, based on the dataset. The tree generated by Weka is shown in Table 4.2

The prediction being performed here is if the knowledge of the agent will increase, be constant or decrease. This is done to increase accuracy of the decision tree. It can be seen

that given population of the city, income level, education level and country of driver training, we can predict the knowledge change K_c of an agent. The value in parentheses is a ratio of number of cases that follow the rule to those who don't. This decision tree can now be used as an input to the agent-based model and decide the learning rate LR of each agent-based on agent properties and the decision tree. The agents, whose predicted knowledge change is 'Increase' are given a highest learning rate and those with 'Decrease' are given the lowest.

```
J48 pruned tree
------------------

City Population = > 300,000
|   Education Level = University Degree: Increase (107.0/62.0)
|   Education Level = College Diploma: Constant (50.0/27.0)
|   Education Level = High School Graduate
|   |   Income Level = > $80,000: Constant (5.0/3.0)
|   |   Income Level = $60,000 to $80,000: Decrease (4.0/2.0)
|   |   Income Level = $40,000 to $60,000: Constant (4.0/2.0)
|   |   Income Level = < $20,000: Increase (18.0/9.0)
|   |   Income Level = $20,000 to $40,000: Increase (4.0/2.0)
|   Education Level = Some Post High School
|   |   Income Level = > $80,000: Constant (7.0/2.0)
|   |   Income Level = $60,000 to $80,000: Increase (4.0/1.0)
|   |   Income Level = $40,000 to $60,000: Increase (1.0)
|   |   Income Level = < $20,000: Decrease (2.0)
|   |   Income Level = $20,000 to $40,000: Constant (4.0/1.0)
City Population = 30,000 to 100,000
|   Income Level = > $80,000
|   |   Education Level = University Degree
|   |   |   Driver Training = Canada: Increase (22.0/10.0)
|   |   |   Driver Training = Outside Canada: Constant (7.0/3.0)
|   |   Education Level = College Diploma: Constant (9.0/3.0)
|   |   Education Level = High School Graduate: Increase (3.0)
|   |   Education Level = Some Post High School: Constant (4.0/2.0)
|   Income Level = $60,000 to $80,000
|   |   Education Level = University Degree
|   |   |   Driver Training = Canada: Constant (5.0/1.0)
|   |   |   Driver Training = Outside Canada: Increase (3.0/1.0)
|   |   Education Level = College Diploma: Constant (21.0/8.0)
|   |   Education Level = High School Graduate: Increase (6.0/2.0)
|   |   Education Level = Some Post High School: Increase (5.0/1.0)
|   Income Level = $40,000 to $60,000
|   |   Education Level = University Degree: Constant (9.0/5.0)
|   |   Education Level = College Diploma: Increase (7.0/2.0)
|   |   Education Level = High School Graduate: Constant (4.0)
|   |   Education Level = Some Post High School: Increase (5.0/2.0)
|   Income Level= < $20,000: Decrease (11.0/5.0)
|   Income Level = $20,000 to $40,000: Increase (13.0/7.0)
City Population = 1,000 to 30,000: Increase (61.0/29.0)
City Population = 100,000 to 300,000: Constant (79.0/43.0)
```

Table 4.2 Decision Tree for Learning rate of agents

CHAPTER 5

AGENT-BASED MODEL

Once the processes of data cleaning, pre-processing, regression analysis and decision tree formation were completed, the data was ready to be used in agent-based model. The multi-agent system was developed in a specific manner, so that the processed data from the survey can be used to initialize various parameters in the simulations, which normally would have been randomly initialized.

5.1 Repast

Repast stands for "The Recursive Porous Agent Simulation Toolkit". It's a widely used cross platform, open source and free agent-based modeling and simulation toolkit. David Sallach, Nick Collier, Tom Howe, Michael North and others developed Repast at University of Chicago. Currently Repast is being managed by "Repast Organization for Architecture and Development" (ROAD). Repast has been implemented in numerous languages like C++, Java, Python, .NET etc. The main features of repast are:

- Object oriented architecture
- Multi-platform
- Concurrent and discrete event scheduler
- Support for social networking tools
- In-built libraries for neural networks, genetic algorithms etc.
- Result logging and graphing tools

- Dynamic run time modification of agents and model is permissible

The version of Repast used to create the simulation here is RepastJ [35], which is the Java based version of Repast.

5.1.1 General Repast Setup

Repast works in a two-step process, namely simulation preparation and simulation running. Terminology for a single run of the simulation is a 'tick'. A simulation requires at least two classes, one for describing the agents in the model and other for describing the model itself. The model class 'AutoSimModel' inherits 'SimpleModelImpl' class from the Repast library, where latter overrides the methods provided by the former. There are methods, which are used to setup the simulation, and there are methods that are used to run the simulation. The two main parts of an agent-based model are Model class and Agent class.

The Agent class contains model specific information about the agents being used in the simulation. Model class creates agents using the agent class. Agent class consists of all the properties of the agents and get/set methods, which make these agent properties accessible to the Model class. The Model class has following parts

- Main Method: Creates instance of the model
- Variable for Model Infrastructure: These variables are the initial parameters for the model run. They also consist of variables that are responsible for size of environment, number of agents, number of time steps etc.

- Repast template methods: These methods have to be defined in Model class for proper running of the simulation. These include
    - buildSchedule() : Defines which methods are to be run and when
    - buildDisplay() : Creates displays. We don't use this feature, as it doesn't work for batch runs
    - buildModel() : This is the main method that creates the model. All the agents and environment variables are created here and process of data collection happens here.
- Get/set Methods: These methods are used to change or retrieve the model infrastructure variables
- Interface Methods: These methods are part of SimpleModelImpl interface. These mostly concern with the initial parameters, name and setup of the simulation.
- Simulation specific methods: These are the methods, which are exclusive to a particular simulation. These define the logic and algorithms used in the simulation.

5.1.2 Repast Setup for AutoSimModel

There are 1000 agents in the simulation. 484 agents amongst these 1000 represent 484 participants from the survey. They are initiated with the same agent profile (age, gender, education level etc.) and initial knowledge K_i as in the survey. The agent profile, attributes and knowledge of the additional 516 agents are randomly decided, while maintaining the average knowledge of population before the survey. These agents are not

considered while calculating the average knowledge of the environment in later stages. The goal here is to be as close as possible to the real world, and hence the need of these dummy agents in the environment, as the people who took the survey interact with other people in real life, who have not been a part of the survey. Their use and importance in the simulation will be explained in detail in the coming sections. The simulation is run for 180 ticks, representing 180 days, as in the survey. The initialization process of all the agent parameters is explained in detail in next section.

5.2 Initialization of Agent parameters/attributes

Most of the agent parameters in the simulation can be directly initiated from the survey data. These parameters include agent age, gender, city population, income level, education level, country of driver training and initial agent knowledge K_i. The information about these parameters can be found in chapter 3. There are 4 other agent attributes, namely Learning rate, Knowledge deterioration rate, Accident rate and Reputation, which are to be initialized for every agent. These attributes play a very important role in the operation of the simulation. The initialization process of each of these is explained below.

5.2.1 Learning Rate

Learning rate L_r of an agent is the probability by which it acquires and remembers knowledge when provided to it. This knowledge can be given during an intervention, or

the knowledge from belief space or knowledge acquired during interaction in agent's social network. Learning rate is used in all these scenarios. The results of decision tree from section 4.3 are used to decide learning rate for agents. The decision tree predicts if the final knowledge level K_f of agents will decrease, be constant or increase after the 180 days period. It is assumed that if the final knowledge is being predicted to increase for an agent, that agent has a higher learning rate. So the learning rate of agents, after calibration of the model, are decided as shown below

L_r = .3, if prediction by decision tree is 'Increase'

L_r = .1, if prediction by decision tree is 'Constant'

L_r = 0, if prediction by decision tree is 'Decrease'

5.2.2 Knowledge Deterioration Rate

Knowledge deterioration rate K_{dr} is the rate by which an agent loses its knowledge of child safety measures in cars per day. This rate is different for every agent and is calculated by the formula below

$$K_{dr} = (K_i + (L_r * K_{int}) – K_f) / 180$$

Where K_i = Initial knowledge of agent on day 1 before intervention

K_f = Final knowledge of agent on day 180

K_{int} = Knowledge provided during intervention

L_r = Learning rate of the agent

5.2.3 Driving probability

Driving probability D_p is the probability of an agent driving a vehicle in a day. This has been derived from Canada Motor Vehicle Traffic Collision Statistic 2010 [36]. Its constant for every agent with a value of 0.3

5.2.4 Accident Rate

Accident rate A_r is the probability of an agent getting into an accident while it is driving. This has been derived from Canada Motor Vehicle Traffic Collision Statistic 2010 [36]. Its constant for every agent with a value of 0.007

5.2.5 Reputation

Reputation R is the probability by which an agent influences knowledge of other agents in its social network. It's a measure of his/her 'reputation' in the social network. Since the survey has no information about the social network aspect of the participants, the value of reputation is kept at a constant value of .4

5.3 Algorithms in the Simulation

An agent-based model is a collection of different algorithms running on agents at different specified time intervals. There are three basic algorithms running in the simulation, namely Basic Intervention Framework, Cultural Algorithm and Social Network. All three algorithms work in tandem to achieve the desired result of the simulation. Details of each of the algorithm are explained in detail below.

5.3.1 Basic Intervention Framework

Everyday, an agent decides to drive depending on their driving probability D_p. If they drive, they can get into an accident based on their accident rate A_r. Once in an accident, they have to go through an intervention about child safety measures, where they learn the corresponding correct knowledge K_{int} in accordance to their learning rate L_r.

5.3.2 Cultural Algorithm

Cultural algorithm is a branch of Evolutionary computing, which consists of a population and belief space [37,38]. Evolution takes place at both *cultural level* (belief space) and at *population level* (for each individual). Belief space is a cultural knowledge, which is shared amongst all the agents in the population. Selected elite individuals contribute to cultural knowledge by means of an acceptance function. This knowledge manages the

evolution of population based on an influence function, thereby sharing it with all the agents in the population.

Cultural algorithm is used in the simulation to spread the common knowledge about child safety in cars amongst all the agents. There is a belief space, which is updated by the average knowledge of the best drivers using an acceptance function. This belief space is updated weekly with the average knowledge of the top 5% drivers with the best knowledge in the population. Everyday, a collection of randomly selected agents update their knowledge from knowledge in belief space K_{belief} using an influence function. The agents learn this knowledge in accordance to their learning rate L_r [Figure 4.1].

5.3.3 Social Network

Every agent has its own social network and there is a reputation R associated with every agent. This reputation gives us a measure of influence that a particular agent has on other agents in its social network. Everyday, a randomly selected collection of agents reach out in their social network and update their knowledge depending on the knowledge of other agents. Agents in their social network are influenced based on their reputation R. The agent collects the knowledge from the social network but only updates it if there is a $2/3^{rd}$ majority amongst the agents in its social network. The agents learn this knowledge in accordance to their learning rate L_r [Figure 4.1].

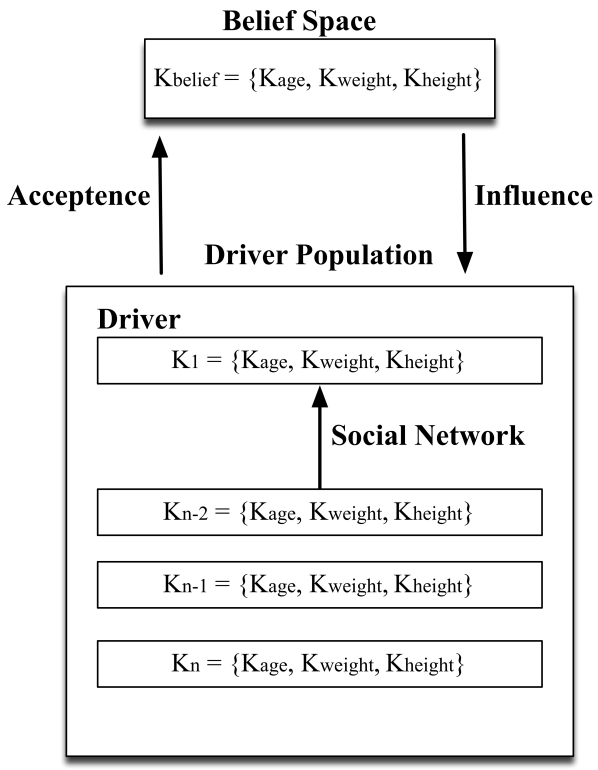

Figure 5.1 Cultural Algorithm and Social Network

5.4 Flow of Simulation

The algorithms mentioned above all work in harmony along with each other in the agent-based model for 180 ticks, which represents 180 days as in the survey. The aim of the simulation is to be as close as possible to the real world. If the final average knowledge of all the 484 agents after 180 days is close to that as in the survey, then this claim can be confirmed. We don't take knowledge of other 516 random agents into account, as they were not a part of the initial survey and there is no method of validating their knowledge. The flow of the simulation, along with initialization and its working is explained in the algorithm below

5.4.1 Algorithm

1. Create the simulation environment
2. Create 1000 agents
3. Initialize 484 agents with agent profiles and knowledge from the survey database. Initialize their agent attributes (L_r, D_p, Ar, R) as explained in previous sections.
4. Initialize rest of 516 agents with random agent profiles, agent attributes and knowledge.
5. Randomly create social network for every agent.
6. Make every agent out of 484 agents in the survey go through an intervention and inject them with the intervention knowledge K_{int} on day 1 of the simulation. Agents learn this knowledge based on their learning rate L_r.
7. Calculate initial belief space knowledge K_{belief} of the whole population. This knowledge is the average knowledge of top 5% of all the drivers.
8. Execute the following steps everyday for 180 days (1 day = 1 tick of agent-based model):
    a. Reduce the knowledge of agents based on their individual knowledge deterioration rate K_{dr}.
    b. Every agent decides to drive or not based on its driving probability D_p. If they are driving, they might get into an accident based on their accident rate A_r. If in an accident, they go through an intervention where they are injected with knowledge about car safety K_{int}. Agent learns this knowledge based on on their learning rate L_r.

c. Update knowledge of a collection of randomly chosen agents using knowledge from belief space. They learn the belief space knowledge K_{belief} depending on their individual learning rate L_r.

d. Update knowledge of a collection of randomly chosen agents from their social network. An agent contacts agents in its social network and inquires about their knowledge. It then updates its knowledge based on learning rate L_r if it gets a $2/3^{rd}$ majority about the knowledge in the social network.

9. Update the belief space knowledge K_{belief} every 7 days of the simulation.
10. If the number of days is less than 180, go to step 8.
11. At day 180, calculate the average knowledge level of all the 484 agents in the simulation, who were a part of the survey. Compare this average knowledge to average knowledge from the survey on day 180.

5.4.2 *Flowchart of Agent-Based Model*

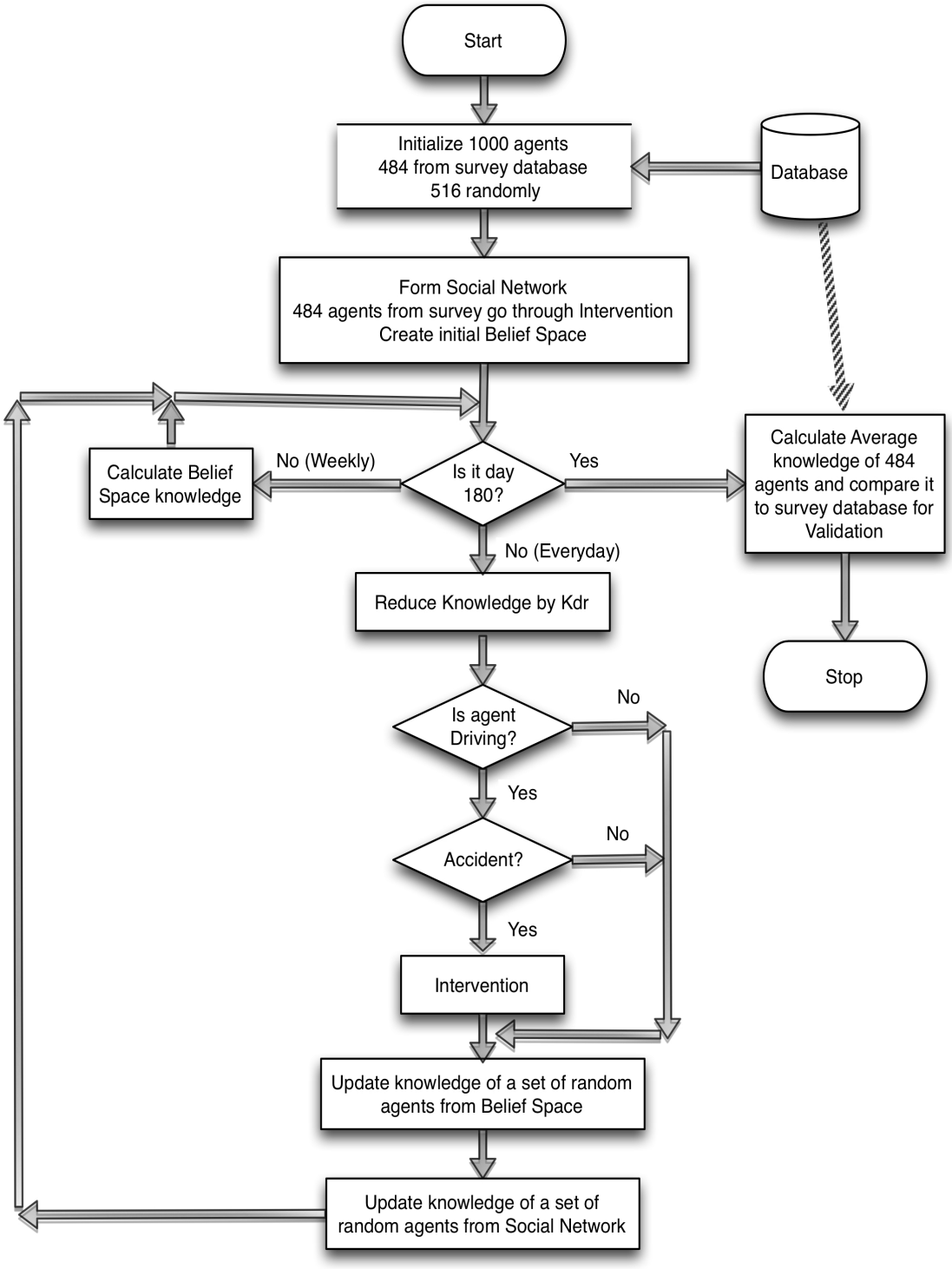

Figure 5.2 Flowchart of Agent-based Model

CHAPTER 6

INTERVENTION POLICY FRAMEWORK

In sections 3 and 4 above, a simulation on child safety in vehicle was created, which is close to real world scenario. This was done using a survey database as a basis for creating the simulation and performing regression analysis and decision tree algorithms on it. This processing of database helped with initialization of different parameters of the agent-based model, which decide the final outcome and result of the simulation.

Now, this simulation can be used to measure performance of different intervention policies, which can be implemented on the population in order to increase the awareness about child safety. Many of these policies have been discussed in [46,47]. An intervention is a policy implemented by government or a similar organization to educate people about child safety in vehicles. These interventions policies are costly to implement and the cost depends on number of interventions being performed. In upcoming sections, we discuss about different intervention policies, which can be implemented in our simulation and different methods of finding the best intervention policies.

6.1 Intervention Policy

An Intervention Policy is defined as the methodology of performing an intervention and deciding the subset of population that will be a part of that intervention. Cost of an

intervention policy will depend on number of people who are included in that intervention. As discussed in section 3.2, the population in database has following properties

1. Age: 20 = in 20s

   30 = in 30s

   40 = in 40s or greater than 40

2. Gender: 1 = Male

   2 = Female

3. Income Level: 1 = Under $20,000

   2 = $20,000 – 40,000

   3 = $40,000 – 60,000

   4 = $60,000 – 80,000

   5 = Over $80,000

4. Education level: 1 = Grade school/ Some High school/ High school graduate

   2 = Some Post-High school

   3 = College Diploma/ Certificate

   4 = University Degree

5. Population of city person lives in: 1 = Over 300,000

   2 = Between 100,000 – 300,000

3 = Between 30,000 – 100,000

4 = Under 30,000

6. Was first driver training done in Canada: 1 = Yes

2 = No

The selection of people who have to go through the intervention can be based on these properties. Each intervention policy can either include or exclude people from specific categories. This will decide the number of people being intervened by that specific intervention policy and also cost of that intervention policy. The intervention has to be repeated after a specified number of days. Hence each intervention policy has three parts.

1. Number of days after which the process of intervention is repeated
2. Different profiles of people who are being included in the policy
3. Cost of Intervention policy

The number of days after which the interventions are repeated is fixed at 20 for all the experiments in the thesis, but this can be easily changed. The intervention policy is simulated in the agent-based model and the model is then run for 180 days. The average final knowledge K_{favg} of the population is calculated on day 180 and is used as the performance measure for that intervention policy. The cost for each intervention policy is also calculated, which is basically the number of individual interventions that happened

during 180 days period. An example of an intervention policy based on agent properties is given below:

Age<20,30> *// Include people in age group 20s (20-29) and 30s (30-39)*

Gender <Male, Female> // *Include both Males and Females*

Training <In Canada, Outside Canada> // *Include people trained in and outside Canada*

Income level < 20000-40000,40000-60000,over 80000> // *Include people from these income groups*

City Population < under 30000,30000-100000,100000-300000> // *Include people from cities of these population level*

Education Level <High School Grad or under, Some post High School, College Diploma, University Degree> // *Include people with these education level*

Intervention Policy

It should be noted that the logical operation within different options of same property is OR and the logical operation between different properties is AND. Hence the policy above can be represented logically as

{Age: 20 OR 30} AND {Gender: Male OR Female} AND {Training: In Canada OR Outside Canada} AND {Income Level: 20000-40000 OR 40000-60000 OR Over 80000} AND {City Population: under 30000 OR 30000-100000 OR 100000-300000} AND {Education Level: High School Grad or under OR Some post High school OR College Diploma OR University Degree}

Intervention Policy as combination of Logical Operations

This intervention policy can be encoded in the simulation using a simple bit string, where each bit takes a value of 0 or 1 depending on whether that option/property is being included in the intervention policy or not. The intervention policy above can be encoded in bit string as follows.

Age			**Gender**		**Primary Driver Training**	
20	30	40	Male	Female	In Canada	Outside Canada
1	1	0	1	1	1	1

Income Level				
Under $20000	$20000-$40000	$40000-$60000	$60000-$80000	Above $80000
0	1	1	0	1

Population of City			
Over 300,000	100,000 - 300,000	30,000 – 100,000	Under 30,000
0	1	1	1

Education Level			
Grade School, Some High School, High School Graduate	Some Post High School	College Diploma/ Certificate	University Degree
1	1	1	1

Intervention Policy represented in bit string

The final bit string that represents the above intervention policy has a length of 20 bit and is represented as follows

1	1	0	1	1	1	1	0	1	1	0	1	0	1	1	1	1	1	1	1

Final bit string representing Intervention Policy

When simulation is initialized with intervention policy represented by bit string shown above, it gives us an Average final knowledge K_{favg}, which is performance measure of

that policy. It also gives the cost of the intervention policy, which is basically the number of individual interventions performed if done every 20 days within a 180 day time period on the people who were a part of that intervention policy.

In the next sections, we discuss about methods, which were used to test these intervention policies and methods to come up with best intervention policy within a given cost.

6.2 Brute force Method

Brute force method is also known as proof by exhaustion, proof by cases or perfect induction [48]. It's a type of mathematical proof, in which the statement to be proved is split into a finite number of cases and each and every case is examined. It involves systematically enumerating all possible outcomes of a problem and checking each one of them.

A 20-bit string, as shown above represents an intervention policy. Each bit can hold a value of either 0 or 1. The total number of combinations possible for intervention bit string are 2^20= 1048576. This is the total number of possible intervention polices, although many of them might not produce any results. All these possible intervention policies can be brute-forced on the simulation one by one, resulting in 1048576 different simulation runs, which will result in the same number of Average final knowledge K_{favg} and cost of intervention policy. All this can be documented for further analysis through

which effects of including people of different agent properties in intervention policy can be analyzed based on the final average knowledge.

6.3 Genetic Algorithm

Genetic algorithm [50] is a class of evolutionary algorithm, which generates solutions to optimization problems using techniques, which are inspired by process of natural evolution and selection. This search heuristic is mostly used to generate solutions for optimization and search problems. A population of candidate solutions, known as individuals, is evolved towards a better solution for an optimization problem. Each candidate has a set of properties, known as its chromosome, which is mutated and altered throughout different evolving generations. Traditionally, solutions are represented as a binary string of 0 and 1, but other representations are possible too.

The evolution starts from a generation consisting of a population of randomly generated individuals. Fitness of each individual in the generation is calculated. Fitness is a measure of performance of an individual towards the optimization problem being solved. The more fit individuals are stochastically selected from the population and these individuals go through a process of crossover, based on the crossover probability of the algorithm. Crossover is a genetic process in which, two parent genes create child genes. An example of process of crossover is shown below

Parent Gene 1

a	b	c	d	e	f	g	h	i	j

Parent Gene 2

0	1	2	3	4	5	6	7	8	9

Child Gene 1

a	b	c	d	e	5	6	7	8	9

Child Gene 2

0	1	2	3	4	f	g	h	i	j

Figure 6.1 Process of Crossover

After crossover, these children genes go through process of mutation, where the bits of these genes are altered depending on algorithm's mutation probability. This whole process is repeated till the formation of a new generation of individuals. Then fitness of individuals in this new generation is calculated. Thereafter, the whole iterative process is repeated till we reach a satisfactory fitness level or for a maximum number of generations.

This method of reaching a solution using genetic algorithm is used to find the best intervention policy under a given cost. It might not always be possible to brute force the policies if total number of policies is very large. Use of a genetic algorithm is preferred in those cases. Since intervention policies are represented by bit string, the genetic algorithm can be used for the same easily. The whole process is discussed in detail in the next chapter.

CHAPTER 7

EXPERIMENTS AND RESULTS

7.1 Agent-Based Model

The model was run to simulate 180 days, in which different aspects and algorithms used in our multi-agent system and the final result from the original survey were compared against each other. In Figure 6.1, CA represents Cultural Algorithm, SN represents Social Network and INT represents basic Intervention framework. 4 different runs were performed using different combinations of algorithms discussed in section 4.3 against each other and compared their performance. These are:

- Cultural algorithm, Social network and Intervention framework: On
- Cultural algorithm and Intervention framework: On; Social network: Off
- Social network and Intervention framework: On; Cultural algorithm: Off
- Intervention framework: On; Social network and Cultural algorithm: Off

As we see in Figure 6.1, the best result is displayed when everything is kept on. This means that the average knowledge is highest when cultural algorithm, social network and intervention all work together. This is closely followed by the run in which only social network is off. The poorest performing run is when just intervention and social network were kept on. The run with just intervention framework shows an improvement over the former.

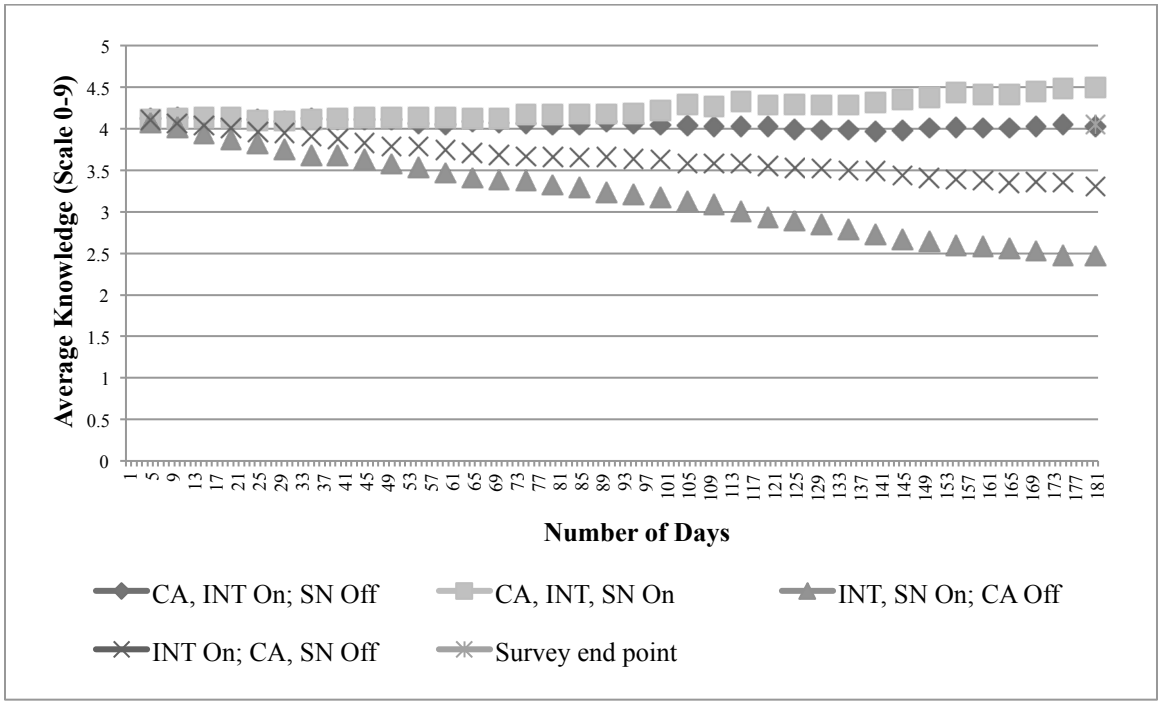

Figure 7.1 Average knowledge of Child Safety over 180 days

The experiments suggest that maximum diffusion of knowledge is achieved when cultural algorithm, social network and intervention all work together. Social network has a marginal effect on the increase of average knowledge, as the knowledge being spread through the social network might not always be the correct knowledge. In fact, in absence of a cultural algorithm, a social network might prove to be harmful for spread of correct knowledge. We can infer that social network performs better for spread of knowledge when a considerable number of people already have the correct knowledge, or else it might backfire and spread incorrect knowledge. When compared to the average final knowledge from the survey at day 180, as we can see from Figure 6.1, the closest performance is given by simulation in which social network was off, cultural algorithm

and intervention framework were on. This indicates that information exchange due to social network rarely happened amongst the people who took the survey. This represents a simulation, which gives us a very close picture of what happened in real world during those 180 days.

We can conclude that belief space and cultural learning play a big role in the spread of knowledge. In other observations, we see that during the simulations with cultural algorithm, the belief space quickly reached a constant value and rarely changed in later stages of the simulation, suggesting that most of the knowledge gained was amongst people with lower knowledge level and people with higher knowledge level didn't improve their knowledge much. It was also observed that when social network was kept on, a large number of populations ended up having the same exact knowledge, indicating a mass convergence of knowledge.

7.2 Intervention Policy: Brute Force Method

Brute forcing of all the possible intervention policies on the simulation was done on a 16-core system. The whole problem was divided into 16 smaller denominations, which can be all executed in parallel. Windows Powershell [49], a task based command line and scripting language was used for the same. Below is the technical specification of the software and hardware used for brute force.

Operating System: Windows Server 2008

Processor: Intel Xeon E5520 @2.27 GHz (16 CPUs)

Memory: 24566MB RAM

Netbeans IDE 6.9.1

Windows Powershell

The result of the brute force is stored in a .csv (comma separated value) file. It's observed that out of possible 1048576 intervention policies, there were 394151 cases where at least one intervention was performed, as not all policies resulted in actual interventions due to their logical nature. This gives us with 394151 different intervention policies and their corresponding Average final knowledge and cost. Further analysis is performed on this .csv file to examine the effect of including different agent properties in intervention policy on average final knowledge of the whole population. It should be re-instated that these analysis results are based on results given by the agent-based simulation and might need further explanation/validation by field experts.

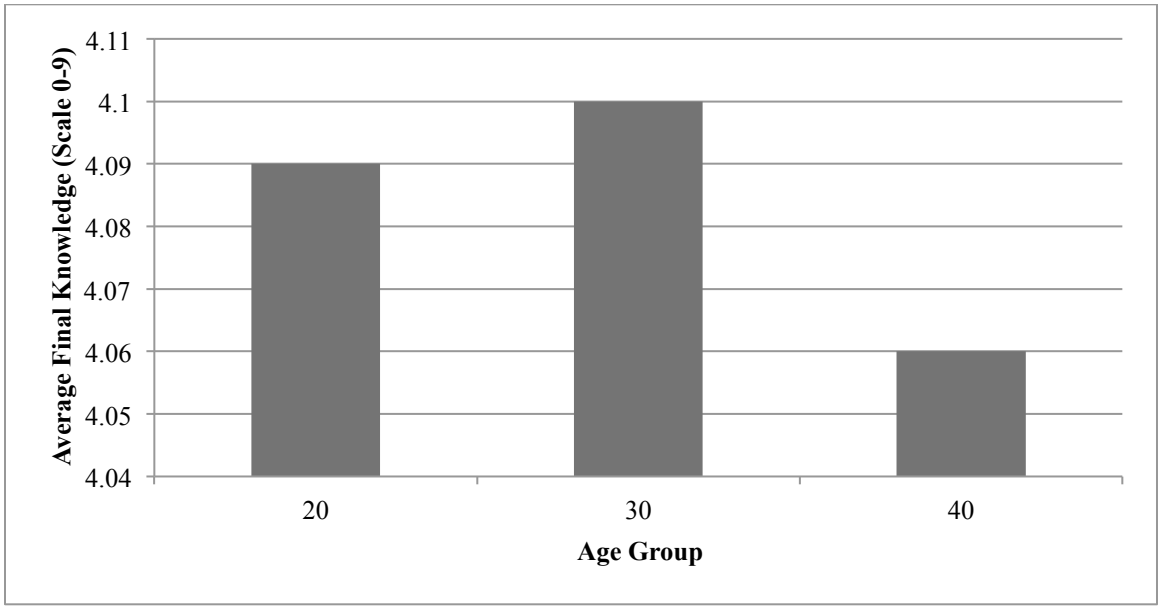

Figure 7.2 Comparison of Average Final Knowledge in different Age groups

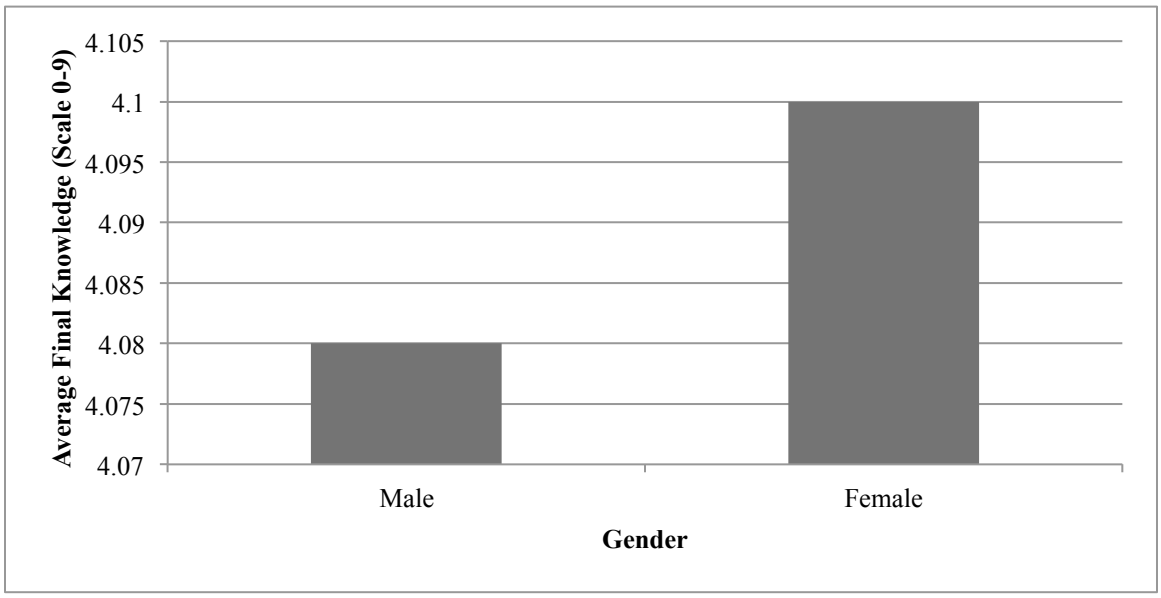

Figure 7.3 Comparison of Average Final Knowledge in different Gender groups

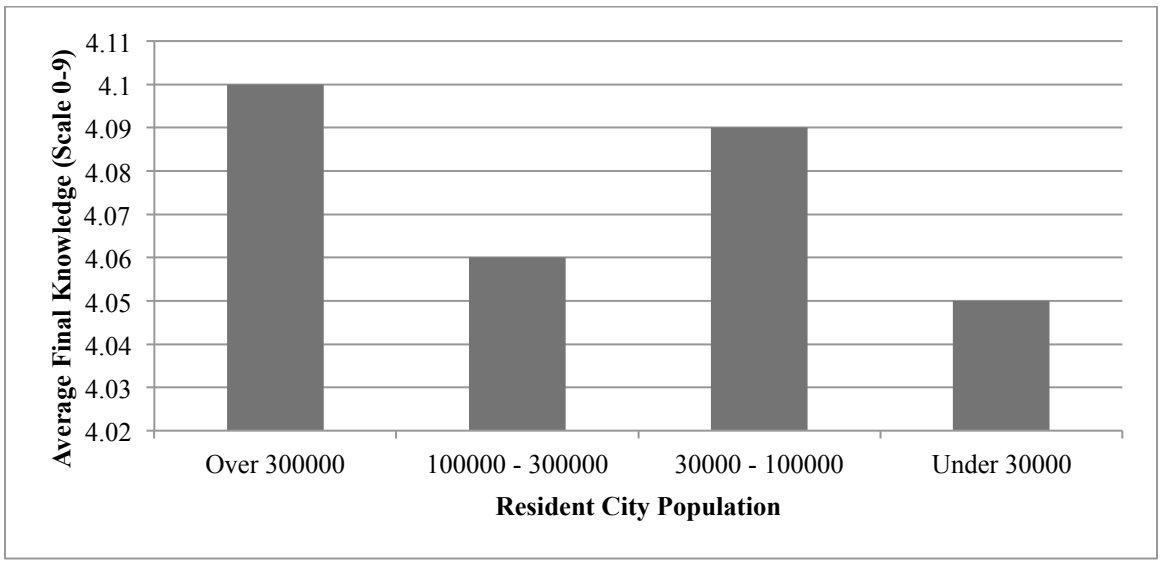

Figure 7.4 Comparison of Average Final Knowledge in different City Population

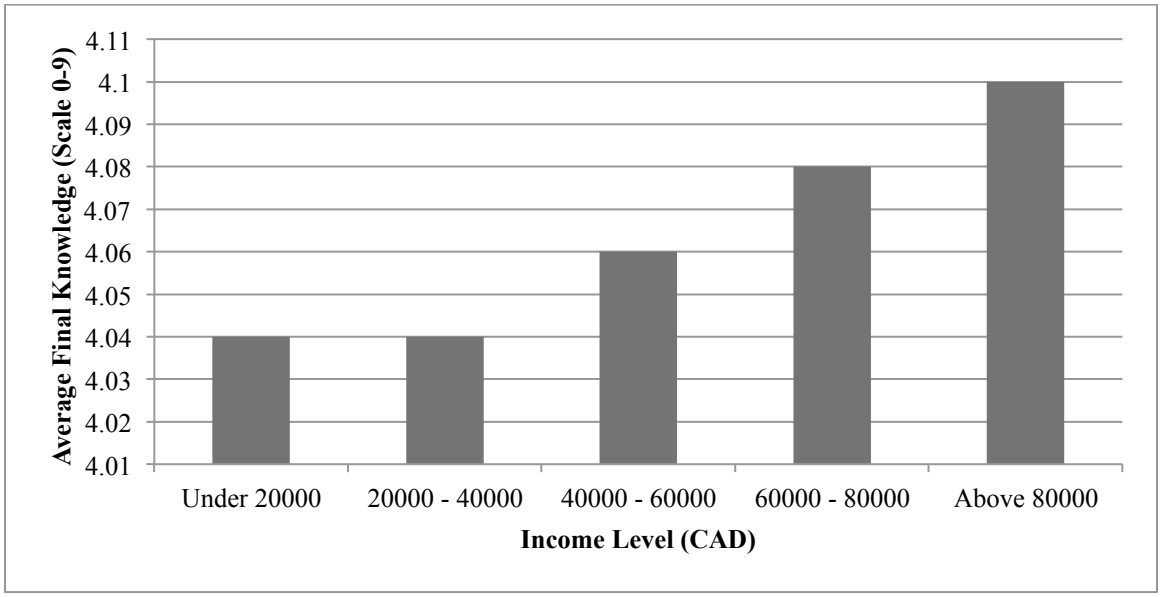

Figure 7.5 Comparison of Average Final Knowledge in different Income Levels

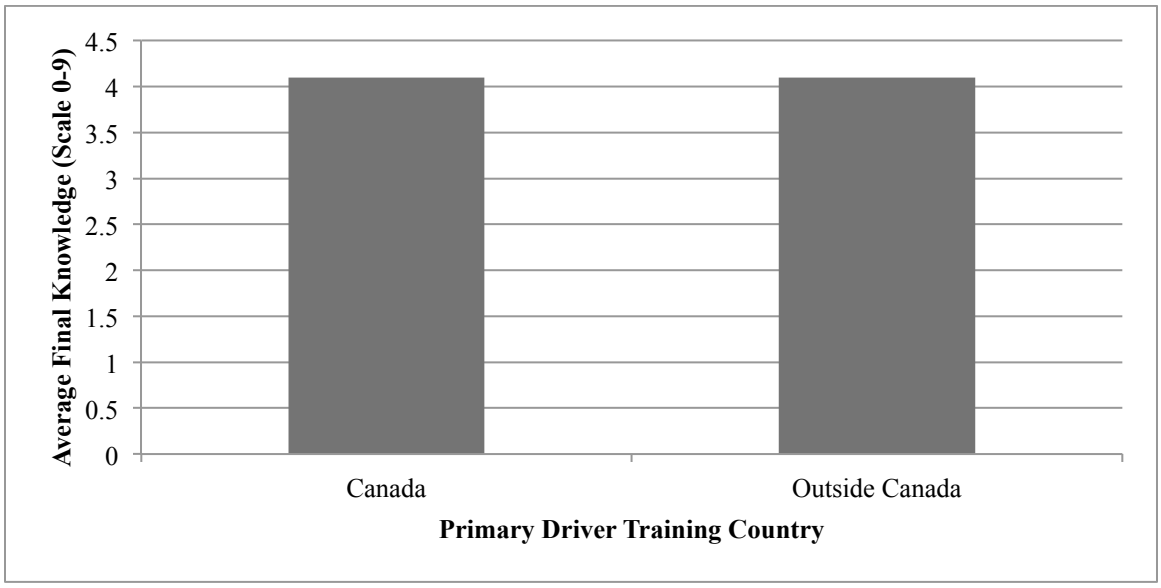

Figure 7.6 Comparison of Average Final Knowledge in different Countries of Primary driver training

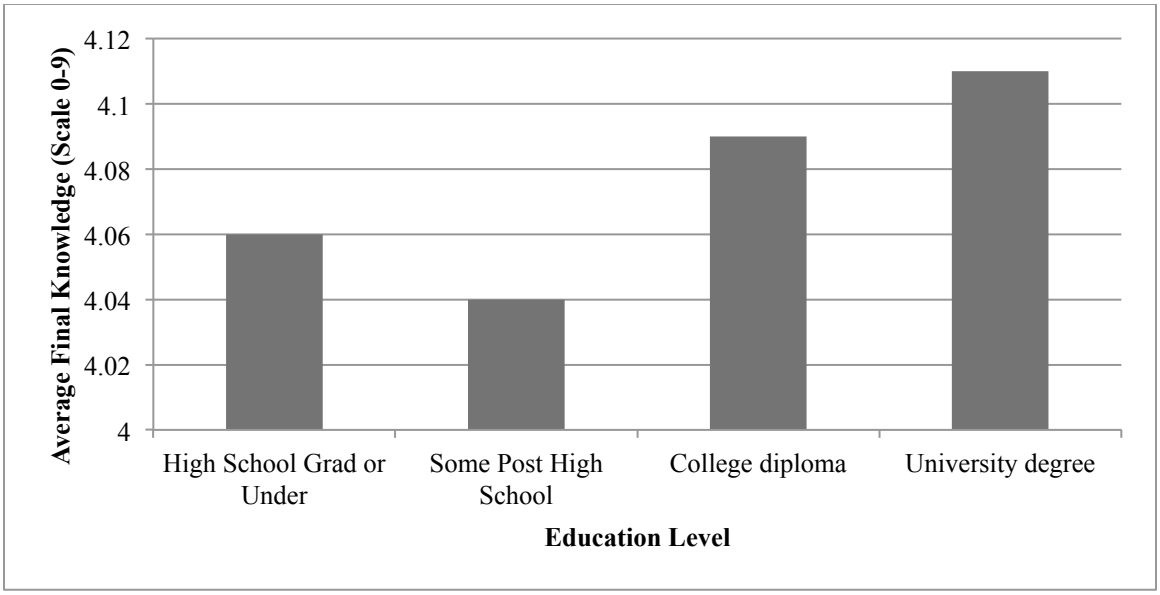

Figure 7.7 Comparison of Average Final Knowledge in different Education Level

Upon further analysis of the results produced by brute forcing of intervention policies, it is observed that policies in which agents of age group 30-39 are included produce better results than the policies, where agents from age groups of 20-29, and 40 above are included [Figure 6.2]. This indicates that agents in the age group of 30s learn and retain more knowledge during intervention that other age groups. Similarly, interventions on females work better than those on males [Figure 6.3]. Comparison of interventions on agents from cities of different population level is shown in Figure 6.4.

In figure 6.5, it can be seen that interventions yield better results when they are performed on agents who have higher income level than compared to agents with lower income level. Also, the responsiveness to these interventions increases with increase in education level of the agents [Figure 6.7]. There was no such difference seen when it came to country of primary driver training of these agents [Figure 6.6].

7.3 Intervention Policy: Genetic Algorithm

In section 5.3, it was discussed how genetic algorithm can be used to find best or a set of best intervention policies given a maximum cost. Government agencies and policy makers usually have budgetary restrictions while making these intervention policies. The absolute best policy would be obviously to perform an intervention on the whole population frequently, but this will require a lot of individual interventions and cost of implementing this policy would be really high. Genetic algorithm can help in finding the best intervention policy, which will give best results, under a specific budgetary restriction; cost of intervention in this case. The experiments were performed under following conditions.

- Crossover Probability: Low (.2), Medium (.5), High (.8)
- Mutation Probability: Low(.01), Medium (.05), High (.10)
- Maximum cost of intervention policy: 1500
- Number of individuals in each generation: 10
- Number of generations: 50
- Selection method for parent genes: Roulette wheel method [51]

Roulette wheel selection method [51] is a way of selection of parent genes for crossover and other genetic process, so that the next generation genes can be created. This method works on basic principle of a roulette wheel. The better the fitness of a specific gene is, the larger area it is assigned on a roulette wheel. Hence its probability of being selected is higher than that of genes that have a lower fitness, but still the selection

is not guaranteed. The fitness used in this genetic algorithm is the Average final knowledge K_{favg} of the policy described by that individual gene. Higher the K_{favg} of a policy, higher the fitness of that gene is. If the cost of the policy is above 1500, it's fitness given a penalty based on a penalty function.

The aim here was to come up with intervention policies, which will produce best results under the intervention cost of 1500. Using the .csv file created by brute force method, the best intervention policy and the associated Average final knowledge K_{bf} for this condition can be easily extracted. Therefore, the performance of genetic algorithm can be measure using this Average final knowledge K_{bf} as benchmark. The results of the experiment is documented in table 6.1

Crossover Probability Mutation Probability	Low: .20	Medium: .50	High: .80
Low: .01	90.90%	89.30%	88.21%
Medium: .05	92.08%	93.61%	96.66%
High: .10	95.97%	96.30%	96.64%

Table 7.1: Sensitization Table for Genetic Algorithm

As evident from the table above, different runs were done of genetic algorithm using different combinations of Crossover probability and Mutation probability for better results. An average of 10 runs was taken for better consistency. The percentages in the table indicate how close was the average of 10 runs to K_{bf}. The best result (96.66%) was given by the genetic algorithm, when Mutation probability was .05 and Crossover probability was .80. This means that this genetic algorithm, after 50 generations, gives us

intervention policies, using which results in an Average final knowledge, which is 96.66% of K_{bf}. The list of 10 best policies from this genetic algorithm is explained in Appendix B. For purpose of comparison, the best 10 policies under the cost of 1500, when using brute force method, are explained in Appendix A.

7.4 Explanation of methodology and results on an abstract level

There was a short discussion about 'Diffusion of Innovation' [1-3] in section 2.3 of this thesis. There is a wider application of the work done in this thesis on an abstract level when it comes to theory of Diffusion of Innovation. Diffusion of Innovation is a theory that seeks to explain how, why and at what rate do new ideas and knowledge spread through cultures. It also explores the factors that affect these patterns and extent of knowledge flow and tries to predict the same. Comparing this theory to work done in this research study, innovation can be compared to knowledge about child safety in vehicles; the interventions can be compared to different advertising methods, which are used to promote the innovations. It is evident that given another similar data set in some field of Innovation diffusion; a similar model and framework can be created using the methodology discussed in this research. Hence, although the possibilities were limited in this research work due to limited nature of available dataset, the scope of application of used methodology is quite broad by making minimal changes to it.

CHAPTER 8

CONCLUSION AND FUTURE WORK

Our work was motivated by an ongoing societal challenge, namely, improving child safety in vehicles. Part of this challenge involves changing behaviours regarding proper usage of safety technologies, such as child safety restraints. In order to produce changes requires interventions whose design and implementation is complex and may be well-served using agent based modeling approaches.

In this thesis, a method of creating a close-to-real-world scenario agent-based model on child safety in vehicles using a survey database was developed. In chapter 2, we reviewed research done in the field of child safety in vehicles, and on knowledge flow patterns and prediction using multi-agent systems. In chapter 3, discussion was on different types of data cleaning and pre-processing that were performed on the survey database. Use of regression analysis to determine driver characteristics that affected knowledge change and decision tree formation on those characteristics was also described in the same chapter. In chapter 4, we created a framework to test different intervention policies on this agent-based simulation. These intervention policies were based on different characteristic and properties of the population who took the initial survey. Two methods were used to test these intervention policies. We used an exhaustive, or brute force, approach to test all the possible combinations of intervention policies and document the final results along with cost of performing each intervention

policy. We also used genetic algorithm as a method to find the best intervention policies that can be put into action, given a limitation on maximum cost of the policy.

The results from the experiments give us an insight on many aspects of child safety measures in vehicles. These include the following. (a) The agent-based model shows that belief space and cultural learning play a big role in the spread of knowledge. (b) We also infer that social network performs better for spread of knowledge when a considerable number of people already have the correct knowledge; however, it might backfire and spread incorrect knowledge under certain circumstances. (c) Through analysis of results produced by brute force method of different intervention policies, it was seen that interventions works best on population with high income and knowledge, as they learn and retain more knowledge during interventions. (d) Also, younger age group population and females will respond better to interventions. (e)On average, interventions done in bigger cities will yield better results that those done in smaller cities. It should be noted that these results are based on the results produced by agent-based simulation and might require validation and explanation by field experts. By using genetic algorithm, we can quickly find a list of best possible intervention policies under a given cost that can be implemented on the population.

A future extension of this work would involve implementing the same methodology on a different database that is related to knowledge/innovation flow in a population. A similar framework can be developed for intervention or marketing to promote the innovation or knowledge. More work is required on making social networks

more realistic and surveys can be designed in ways that provide computer scientists with information about social network of the agents to work with. This would provide more capability for validation of the proposed methodology and establish the correctness of the framework used.

REFERENCES/BIBLIOGRAPHY

1. Rogers, E.,: Adoption and Diffusion of Innovations, 4th Edition, Free press New York (1962)
2. Arndt, Johan.:Role of Product-Related Conversations in the Diffusion of a New Product. Journal of Marketing Research, 291-5. (1967)
3. Bass, Frank M.: A New Product Growth Model for Consumer Durables. Management Science, 215-27. (1969)
4. Abrahamson, E., Rosenkopf, L.: Social Network Effects on the Extent ofInnovation Diffusion: A Computer Simulation.Organization Science, pp.289-309. (1997)
5. Ahmed, S., Kobti, Z., Kent, R.: Predictive Data Mining Driven Architecture to Guide Car Seat Model Parameter Initialization. Intelligent DecisionTechnologies, pp. 789-797. (2011)
6. Alkemade, F., Castaldi, C.: Strategies for the Diffusion of Innovations on Social Networks. Computational Economics 25(1), 3-23. (2005)
7. Baqueiro, O., Wang, Y., McBurney, P., Coenen, F.: Integrating Data Mining and Agent-based Modeling and Simulation. Advances in Data Mining. Applications and Theoretical Aspects, pp. 220-231. (2009)
8. Bohlmann, J., Calantone, R., Zhao, M.: The Effects of Market Network Heterogeneity on Innovation Diffusion: An Agent-Based Modeling Approach. Journal of Product Innovation Management 27(5), 741-760. (2010)

9. Cantono, S., Silverberg, G.: A Percolation Model of Eco-Innovation Diffusion: The Relationship between Diffusion, Learning Economies and Subsidies. Technological Forecasting and Social Change 76(4), 487-496. (2009)
10. Choi, H., Kim, S., Lee, J.: Role of Network Structure and Network Effects in Diffusion of Innovations. Industrial Marketing Management 39(1), 170-177. (2010)
11. Delre, S., Jager, W., Bijmolt, T., Janssen, M.: Targeting and Timing Promotional Activities: An Agent-Based Model for the Takeoff of New Products. Journal of business research 60(8), 826-835. (2007)
12. Delre, S., Jager, W., Bijmolt, T., Janssen, M.: Will it spread or not? The Effects of Social Influences and Network Topology on Innovation Diffusion. Journal of Product Innovation Management 27(2), 267-282. (2010)
13. Delre, S., Jager, W., Janssen, M.: Diffusion Dynamics in Small-World Networks with Heterogeneous Consumers. Computational & Mathematical Organization Theory 13(2), 185-202. (2007)
14. Derolan, F.: Formation of Social Networks and Diffusion of Innovations. Research Policy 31(5), 835-846. (2002)
15. Goldenberg, J., Libai, B., Solomon, S., Jan, N., Stauffer, D.: Marketing Percolation. Physica A: Statistical Mechanics and its Applications 284(1), 335-347. (2000)
16. Hassan, S., Antunes, L., Pavon, J.: Mentat: A Data-Driven Agent-Based Simulation of Social Values Evolution. Multi-Agent-Based Simulation X pp. 135-146. (2010)

17. Hassan, S., Gutierrez, C., Arroyo, J.: Re-thinking Modeling: A Call for the use of Data Mining in Data-Driven Social Simulation. Workshop W31 Social Simulation of the 21st International Joint Conference on Artificial Intelligence, Pasadena, California (2009)
18. Kobti, Z., Snowdon, A., Rahaman, S., Dunlop, T., Kent, R.: A Cultural Algorithm to Guide Driver Learning in Applying Child Vehicle Safety Restraint. Evolutionary Computation 2006. CEC 2006. IEEE Congress on. pp. 1111-1118. IEEE (2006)
19. Kobti, Z., Snowdon, A., Rahaman, S., Dunlop, T., Kent, R.: Modeling the Effects of Social Influence on Driver Behavior in Applying Child Vehicle Safety Restraint. Systems, Man and Cybernetics, 2006. SMC' 06. IEEE International Conference on. pp. 145-150 IEEE(2006)
20. Kobti, Z., Snowdon, A., Rahaman, S., Kent, R.: A Reputation Model Framework for Artificial Societies: A Case Study in Child Vehicle Safety Simulation. Advances in Artificial Intelligence. pp. 185-190 (2008)
21. Kuandykov, L., Sokolov, M.: Impact of Social Neighborhood on Diffusion of Innovation S-Curve. Decision Support Systems 48(4), 531-535. (2010)
22. Kurahashi, S., Saito, M.: Word-of-Mouth Effects on Social Networks. Knowledge-Based and Intelligent Information and Engineering Systems pp. 356-365. (2011)
23. Mahajan, V., Muller, E., Wind, Y.: New-Product Diffusion Models, vol. 11. Springer (2000)

24. Remondino, M.: Diffusion of Innovation in a Social Environment: A Multi-agent-based Model. Computer Modeling and Simulation, 2008. UKSIM 2008. Tenth International Conference on. pp. 573-578. IEEE (2008)
25. Remondino, M., Correndo, G.: Data Mining Applied to Agent-based Simulation. Proceedings of the 19th European Conference on Modeling and Simulation, Riga, Latvia (2005)
26. Sargent, R.G.: Verification and Validation of Simulation Models. Proceedings of the 37th Conference on Winter Simulation. pp. 130-143. (2005)
27. Schramm, M., Trainor, K., Shanker, M., Hu, M.: An Agent-Based Diffusion Model with Consumer and Brand Agents. Decision Support Systems 50(1), 234-242. (2010)
28. Thiriot, S., Kant, J.: Using Associative Networks to Represent Adopters` Beliefs in a Multi-Agent Model of Innovation Diffusion. Advances in Complex Systems 11(2), 261-272. (2008)
29. Canadian National Survey on Child Restraint Use 2010, Site: http://www.tc.gc.ca/eng/roadsafety/resources-researchstats-child-restraint-survey-2010-1207.html
30. Auto 21, Canada's automotive R&D program, Site: https://www.auto21.ca/en/
31. IBM SPSS Software. Site: http://www-01.ibm.com/software/analytics/spss/
32. http://wiki.pentaho.com/display/DATAMINING/J48
33. Ruggieri, S.: Efficient C4.5. In: IEEE Transactions on Knowledge and Data Engineering, pp. 438-444 IEEE (2002)

34. Weka 3: Data Mining Software in Java. Site: http://www.cs.waikato.ac.nz/ml/weka/
35. The Repast Suite. Site: http://repast.sourceforge.net
36. Canada Motor Vehicle Traffic Collision Statistic 2010, Site: http://www.tc.gc.ca/eng/roadsafety/tp-1317.htm
37. Holland, J.: Adaptation in Natural and Artificial Systems. In: The University of Michigan Press, Ann Arbor (1975)
38. Reynolds, R.: An Adaptive Computer Model of the Evolution of Agriculture for Hunter-gatherers in the Valley of Oaxaca. In: Ph.D. diss., Univ. of Michigan (1979)
39. World Report on Road traffic Injury Prevention, Site: http://www.who.int/violence_injury_prevention/publications/road_traffic/world_report/summary_en_rev.pdf
40. Kobti, Z., Snowdon, A.W., Kent, R.D., Rahaman, S.: A multi-agent model prototype for child vehicle safety injury prevention. In: Agent 2005 Conference on Generative Social Processes, Models and Mechanisms, Chicago, Illinois, USA, October 13-15, pp. 271–294 (2005)
41. Snowdon, A. W., Hussein, A., High, L., Stamler, L., Millar-Polgar, J., Patrick, L., & Ahmed, E.: The effectiveness of a multimedia intervention on parents' knowledge and use of vehicle safety systems for children. Journal of pediatric nursing, 23(2), 126-139. (2008)

42. Apsler, R., Formica, S. W., Rosenthal, A. F., & Robinson, K.: Increases in booster seat use among children of low income families and variation with age. Injury Prevention, 9(4), 322-325 (2003)
43. Ebel, B. E., Koepsell, T. D., Bennett, E. E., & Rivara, F. P.: Too small for a seatbelt: predictors of booster seat use by child passengers. Pediatrics, 111(4), e323-e327 (2003)
44. Lee, J. W., Fitzgerald, K., & Ebel, B. E.: Lessons for increasing awareness and use of booster seats in a Latino community. Injury prevention, 9(3), 268-269 (2003)
45. Johnston, B. D., Bennett, E., Quan, L., Gonzalez-Walker, D., Crispin, B., & Ebel, B.: Factors influencing booster seat use in a multiethnic community: lessons for program implementation. Health promotion practice, 10(3), 411-418 (2009)
46. Zaza, S., Sleet, D. A., Thompson, R. S., Sosin, D. M., & Bolen, J. C.: Reviews of evidence regarding interventions to increase use of child safety seats. American journal of preventive medicine, 21(4), 31-47 (2001
47. Pierce, S. E., Mundt, M. P., Peterson, N. M., & Katcher, M. L.: Improving awareness and use of booster seats in Head Start families. Wisconsin Medical Journal, 104(1), 46-48 (2005)
48. Paar, C., & Pelzl, J.: Understanding cryptography: a textbook for students and practitioners. Springer (2010)
49. Windows Powershell, Site: http://msdn.microsoft.com/en-us/library/windows/desktop/dd835506(v=vs.85).aspx

50. Goldberg, D. E: Genetic algorithms in search, optimization, and machine learning. Addison-Wesley Professional. (1989)
51. Zhang, L., Chang, H., & Xu, R.: Equal-Width Partitioning Roulette Wheel Selection in Genetic Algorithm. In Technologies and Applications of Artificial Intelligence (TAAI), 2012 Conference on (pp. 62-67). IEEE. (2012)
52. Windrum, P., Fagiolo, G., & Moneta, A.: Empirical validation of agent-based models: Alternatives and prospects. Journal of Artificial Societies and Social Simulation, 10(2), (2007)
53. Balci, O.: Verification validation and accreditation of simulation models. In Proceedings of the 29th conference on Winter simulation (pp. 135-141). IEEE Computer Society, (1997)
54. Dosi, G., Fagiolo, G., & Roventini, A.: An evolutionary model of endogenous business cycles. Computational Economics, 27(1), 3-34, (2006)
55. Werker, C., & Brenner, T: Empirical calibration of simulation models. Max-Planck-Inst. for Research into Economic Systems, (2004).
56. Malerba, F., Nelson, R., Orsenigo, L., & Winter, S.: 'History-friendly'models of industry evolution: the computer industry. Industrial and Corporate Change, 8(1), 3-40, (1999)
57. Baqueiro, O., Wang, Y. J., McBurney, P., & Coenen, F.: Integrating data mining and agent-based modeling and simulation. In Advances in Data Mining. Applications and Theoretical Aspects (pp. 220-231). Springer Berlin Heidelberg, (2009).

58. Garcia, R., Rummel, P., & Hauser, J.: Validating agent-based marketing models through conjoint analysis. Journal of Business Research, 60(8), 848-857, (2007).
59. Rand, W., Brown, D. G., Page, S. E., Riolo, R., Fernandez, L. E., & Zellner, M.: Statistical validation of spatial patterns in agent-based models. In Proceedings of agent-based simulation (Vol. 4). (2003).
60. Sargent, R. G.: Verification and validation of simulation models. In Proceedings of the 37th conference on Winter simulation (pp. 130-143). Winter Simulation Conference. (2005).
61. Remondino, M., & Correndo, G.: Mabs validation through repeated execution and data mining analysis. International Journal of Simulation: Systems, Science & Technology, 7(6) (2006)
62. Smith, J.: Learning styles: Fashion fad or lever for change? The application of learning style theory to inclusive curriculum delivery. Innovations in Education and Teaching International, 39(1), 63-70, (2002)

APPENDICES

Appendix A: Best Intervention Policies (Brute Force Method)

Below are the 10 best intervention policies by Brute force method. The number of days between each intervention is 20 days and maximum cost of intervention is 1500 units. *Cost of intervention* is defined as number of interventions that occur during a period of 180 days under that specific policy. *Predicted Final Average knowledge* is the predicted final average knowledge, by the simulation, using that specific policy. The explanation of a specific policy is given below:

Example of an Intervention policy

Age<20,30> *// Include people in age group 20s (20-29) and 30s (30-39)*

Gender <Male, Female> // *Include both Males and Females*

Training <In Canada, Outside Canada> // *Include people trained in and outside Canada*

Income level < 20000-40000,40000-60000,over 80000> // *Include people from these income groups*

City Population < under 30000,30000-100000,100000-300000> // *Include people from cities of these population level*

Education Level <High School Grad or under, Some post High School, College Diploma, University Degree> // *Include people with these education level*

10 Best policies

1. Age <20,30>
   Gender <Male, Female>

Training <In Canada >

Income level <20000-40000,40000-60000,60000-80000,over 80000>

City Population <under 30000,30000-100000,100000-300000, over 300000>

Education Level <College Diploma, University Degree>

Predicted Final Average knowledge through simulation: 5.768

Cost of Intervention: 1386

2. Age <30>

   Gender <Male, Female>

   Training <In Canada, Outside Canada>

   Income level <under 20000, 20000-40000,40000-60000,60000-80000,over 80000>

   City Population <30000-100000, over 300000>

   Education Level <Some post High School, College Diploma, University Degree>

   Predicted Final Average knowledge through simulation: 5.760

   Cost of Intervention: 1494

3. Age <20,30>

   Gender <Male, Female>

   Training <In Canada, Outside Canada>

   Income level < 20000-40000,40000-60000,over 80000>

   City Population < under 30000,30000-100000,100000-300000>

   Education Level <High School Grad or under, Some post High School, College Diploma, University Degree>

   Predicted Final Average knowledge through simulation: 5.752

   Cost of Intervention: 1467

4. Age <30,40>

   Gender <Male, Female>

   Training <In Canada, Outside Canada>

   Income level <under 20000, 20000-40000,40000-60000,60000-80000>

City Population <under 30000,30000-100000,100000-300000, over 300000>

Education Level <College Diploma, University Degree>

Predicted Final Average knowledge through simulation: 5.714

Cost of Intervention: 1341

5. Age <30,40>

   Gender <Male, Female>

   Training <In Canada, Outside Canada>

   Income level <under 20000, 20000-40000,40000-60000,60000-80000,over 80000>

   City Population <under 30000, over 300000>

   Education Level <College Diploma, University Degree>

   Predicted Final Average knowledge through simulation: 5.698

   Cost of Intervention: 1404

6. Age <30,40>

   Gender <Female>

   Training <In Canada, Outside Canada>

   Income level <under 20000, 20000-40000,40000-60000,over 80000>

   City Population < under 30000,30000-100000,100000-300000, over 300000>

   Education Level <High School Grad or under, Some post High School, University Degree>

   Predicted Final Average knowledge through simulation: 5.696

   Cost of Intervention: 1341

7. Age <20,30,40>

   Gender <Male, Female>

   Training <In Canada, Outside Canada>

   Income level <under 20000, 40000-60000,60000-80000,over 80000>

   City Population <100000-300000, over 300000>

   Education Level <Some post High School, University Degree>

Predicted Final Average knowledge through simulation: 5.690

Cost of Intervention: 1386

8. Age <30,40>

   Gender <Female>

   Training <In Canada, Outside Canada>

   Income level <under 20000, 20000-40000, over 80000>

   City Population < under 30000,30000-100000,100000-300000, over 300000>

   Education Level <High School Grad or under, Some post High School, College Diploma, University Degree>

   Predicted Final Average knowledge through simulation: 5.683

   Cost of Intervention: 1458

9. Age <20,30>

   Gender <Male, Female>

   Training <In Canada>

   Income level <20000-40000,40000-60000, over 80000>

   City Population < under 30000,30000-100000,100000-300000, over 300000>

   Education Level <High School Grad or under, Some post High School, College Diploma, University Degree>

   Predicted Final Average knowledge through simulation: 5.683

   Cost of Intervention: 1458

10. Age <30,40>

    Gender <Male, Female>

    Training <In Canada, Outside Canada>

    Income level <under 20000,40000-60000, over 80000>

    City Population <30000-100000, over 300000>

Education Level <High School Grad or under, Some post High School, College Diploma, University Degree>

Predicted Final Average knowledge through simulation: 5.681

Cost of Intervention: 1386

Appendix B: Best Intervention Policies (Genetic Algorithm)

Below are the 10 best intervention policies by Genetic Algorithm method. The number of days between each intervention is 20 days and maximum cost of intervention is 1500 units. *Cost of intervention* is defined as number of interventions that occur during a period of 180 days under that specific policy. *Predicted Final Average knowledge* is the predicted final average knowledge, by the simulation, using that specific policy.

1. Age <30,40>
   Gender <Female>
   Training <In Canada, Outside Canada>
   Income level <under 20000, 20000-40000,40000-60000, over 80000>
   City Population < under 30000,30000-100000,100000-300000, over 300000>
   Education Level <High School Grad or under, Some post High School, University Degree>
   Predicted Final Average knowledge through simulation: 5.696
   Cost of Intervention: 1341

2. Age <20,30,40>
   Gender <Male, Female>
   Training <In Canada, Outside Canada>
   Income level <under 20000, 40000-60000,60000-80000,over 80000>
   City Population <100000-300000, over 300000>
   Education Level <Some post High School, University Degree>
   Predicted Final Average knowledge through simulation: 5.690
   Cost of Intervention: 1386

3. Age <20,30>

Gender <Female>

Training <In Canada, Outside Canada>

Income level <20000-40000,60000-80000,over 80000>

City Population < under 30000,30000-100000, over 300000>

Education Level <Some post High School, College Diploma, University Degree>

Predicted Final Average knowledge through simulation: 5.654

Cost of Intervention: 1377

4. Age <20,30,40>

   Gender <Male, Female>

   Training <In Canada, Outside Canada>

   Income level <under 20000,60000-80000,over 80000>

   City Population < under 30000,30000-100000,100000-300000, over 300000>

   Education Level <University Degree>

   Predicted Final Average knowledge through simulation: 5.642

   Cost of Intervention: 1494

5. Age <30,40>

   Gender <Male, Female>

   Training <In Canada, Outside Canada>

   Income level <20000-40000,60000-80000,over 80000>

   City Population < under 30000, over 300000>

   Education Level <High School Grad or under, Some post High School, College Diploma, University Degree>

   Predicted Final Average knowledge through simulation: 5.636

   Cost of Intervention: 1440

6. Age <30,40>

   Gender <Male, Female>

   Training <In Canada, Outside Canada>

   Income level <over 80000>

City Population < under 30000,30000-100000,100000-300000, over 300000>
Education Level <High School Grad or under, Some post High School, College Diploma, University Degree>
Predicted Final Average knowledge through simulation: 5.634
Cost of Intervention: 1485

7. Age <20,30,40>
   Gender <Female>
   Training <In Canada, Outside Canada>
   Income level <20000-40000,40000-60000,60000-80000,over 80000>
   City Population <30000-100000, over 300000>
   Education Level <College Diploma, University Degree>
   Predicted Final Average knowledge through simulation: 5.626
   Cost of Intervention: 1395

8. Age <20,30,40>
   Gender <Female>
   Training <In Canada, Outside Canada>
   Income level <under 20000, 40000-60000,60000-80000,over 80000>
   City Population <30000-100000, over 300000>
   Education Level <High School Grad or under, Some post High School, University Degree>
   Predicted Final Average knowledge through simulation: 5.621
   Cost of Intervention: 1458

9. Age <20,30,40>
   Gender <Female>
   Training <In Canada>
   Income level < 40000-60000,60000-80000,over 80000>
   City Population < under 30000,30000-100000, over 300000>

Education Level <High School Grad or under, Some post High School, College Diploma, University Degree>

Predicted Final Average knowledge through simulation: 5.613

Cost of Intervention: 1386

10. Age <20,30,40>

    Gender <Male, Female>

    Training <In Canada, Outside Canada>

    Income level <under 20000, 20000-40000,60000-80000,over 80000>

    City Population <100000-300000, over 300000>

    Education Level <High School Grad or under, University Degree>

    Predicted Final Average knowledge through simulation: 5.613

    Cost of Intervention: 1395

VITA AUCTORIS

NAME:	Ritwick Gupta
PLACE OF BIRTH:	Varanasi, India
YEAR OF BIRTH:	1985
EDUCATION:	Sunbeam English School, Varanasi, India, 2004
	University of Pune, B.E.., Pune, India, 2009
	University of Windsor, M.Sc., Windsor, ON, 2013