[subfloat]font=footnotesize, labelformat=parens,labelsep=space, listofformat=subparens,subreffo
[subfloat]

UNIVERSITY OF SOUTHAMPTON
Faculty of Engineering and Physical Sciences
School of Electronics and Computer Science

A project report submitted for the award of
MEng Electronic Engineering with Artificial Intelligence

Supervisor: Mark Weal
Examiner: Yvonne Howard

A Framework for a Decision-Making Robot regarding Generalised Ethical Dilemmas using ML

by Adrian Azzarelli-Bonnardel

March 20, 2022

UNIVERSITY OF SOUTHAMPTON

FACULTY OF ENGINEERING AND PHYSICAL SCIENCES
SCHOOL OF ELECTRONICS AND COMPUTER SCIENCE

A project report submitted for the award of
MEng Electronic Engineering with Artificial Intelligence

by Adrian Azzarelli-Bonnardel

ABSTRACT

The level of philosophy currently implemented in *decision-making* (DM) applications is low stature. Where a technical application is flawless, the corresponding philosophical application may be flawed. This is most prominent when DM applications try to solve ethical dilemmas. Hence, the objective of this project is to develop a framework that provides reasonable solutions to generalised *ethical dilemmas*.

From our research we outlined two properties of current DM applications that need improvement: (1) the complexity of philosophy, and (2) the inclusion and diversity of multiple philosophies. As a result, we implemented three key DM perspectives (morality, social influence and legal influence) within our framework, developed and exemplar program and tested it against three separate ethical dilemmas, the outcomes of which were then evaluated by professionals and critiqued against our objective.

The evaluation found that the decisions made by each philosophical perspective was logical, thus the framework functioned appropriately, though further improvements could be made to the particular philosophies implemented. In conclusion, the project demonstrated the capability of advanced philosophical systems and their practicality within DM applications.

KEYWORDS

Artificial Intelligence, Machine Learning, Logistic Regression, Decision Making, Ethical Dilemmas, Metaethics

Contents

Acknowledgements

I would like to acknowledge Pamela Ugwudike and Atus Mariqueo-Russell for their contribution towards evaluating this project.

I would like to dedicate this project to a dear friend, James Bartle, for instigating my passion in philosophy with long debates followed by frustrating defeats.

Statement of Originality

I have read and understood the ECS Academic Integrity information and the University's Academic Integrity Guidance for Students.

I am aware that failure to act in accordance with the Regulations Governing Academic Integrity may lead to the imposition of penalties which, for the most serious cases, may include termination of programme.

I consent to the University copying and distributing any or all of my work in any form and using third parties (who may be based outside the EU/EEA) to verify whether my work contains plagiarised material, and for quality assurance purposes.

I have acknowledged all sources, and identified any content taken from elsewhere. I have not used any resources produced by anyone else.

I did all the work myself, or with my allocated group, and have not helped anyone else.

The material in the report is genuine, and I have included all my data/-code/designs.

The implementation of Logistic Regression in Chapter 5.1 is adapted from a coursework I carried out in COMP3223. Otherwise I have not submitted any part of this work for another assessment.

I have not submitted any part of this work for another assessment.

My work did not involve human participants, their cells or data, or animals.

Chapter 1

Introduction

1.1 Introduction

The common aim of most technological devices is to facilitate our lives, with the purpose of making decisions and carrying out actions that we either cannot or do not want to carry out. The development of these devices often focuses on some form of mathematical implementation, however, sometimes the aspects of our lives we wish to facilitate cannot be solved simply by using a mathematical approach. This is where a more *human*-based approach is necessary to solve our problems; an approach that incorporates an amalgamation of "thoughts" as well as human-like philosophies, perhaps where a device is given multiple options and has to subjectively choose one.

To exemplify this problem let us take an autonomous hoover - the *Roomba*. Its decisions do not require a complex level of human understanding, thus the device takes the mathematical approach to moving about, perhaps using sensors to measure distances from objects and then paving out a snake like path to cover all ground. Comparatively, an autonomous car fulfills a similar action of autonomous movement. Yet, due to the elevated risk to our wellbeings and greater human involvement the solution requires a higher level of human understanding, a level that cannot simply be handled by the mathematical approach. Therefore, the solution would additionally require the human-based approach.

In more recent years, we have furthered the capabilities of technologies such as *machine learning* (ML). Our ability to program devices for complex action/reaction is simplified, so emulating the human traits we endear for our systems is somewhat more possible. Nonetheless, the ML systems still rely on the mathematical approach, thus the requirement for more human traits is still left void, though now to a lower degree.

The problem we have described above is one of a philosophical nature. These systems missout on the *human experience*, so no way-of-thought is ever determined, no personal philosophies exist, therefore a system is unable to provide the "more human approach". Evidently, spending the time teaching a system to think a human way is currently out of our technological grasp, so the question we need to ask is: *how can we give a machine a way-of-thought without giving it the whole human experience?* One solution could be to directly program a system with an individual's or a group's philosophies. Alternatively, we could teach a system to act/react the way one or a group of individuals may. However, these solutions are both liable to in-group bias i.e. bias due to group-thinking. Otherwise, we could develop a philosophical framework which breaks down multiple philosophies from many individuals and groups and allows the system to amalgamate the different pieces to develop its own way-of-thought, thus reducing bias. This is the solution we aim to explore within this project.

As we mentioned previously, the rise of ML has made the mathematical approach more accessible. Though it would not contribute anything "human" we would nonetheless like to explore how it could be used alongside the philosophical framework, as certain statistical properties could be of use.

To accomplish this, we have set ourselves the task of developing a framework (and consequently an exemplar program) to facilitate the periods in our lives when we are faced with ethical dilemmas, situations classed as win-win/lose-lose scenarios. More precisely, our aim is to propose a framework that incorporates artificial intelligence methods to provide a plausible solution to a generalised ethical dilemma using a collection of philosophies. To achieve this, our framework will inspect several choices of *probable* outcomes, each relating to selected philosophical perspectives, namely Moral Ethics (*Methaethics*), Social Influence and Legal Influence. Thereafter amalgamating the choices, a final (single) solution will be drawn, thus the system will provide the client with a suitable choice to their dilemma.

In order to exhibit the functionality of the framework, an exemplar/exhibition program will be developed[1]; which we will evaluate using professional help and then draw conclusions as to the efficacy of our framework.

Presently, there exists little proposal regarding *multi-perspective* philosophical frameworks, where the viable research focuses more on the technical implementation of mathematical algorithms than philosophies/human DM complexities. We believe there should pertain an equilibrium between the two, an equilibrium not currently provided; we aim to propose a solution less weighted towards technology alongside a deeper philosophical base. To add to this, the philosophical framework being proposed will focus on generalised ethical dilemmas, again a subject with little-to-no research which succumbs to the same problem of equilibrium that multi-perspective frameworks do. Hence, arises the need for our project.

1.2 Background Literature

The two prevalent topics of research explored in this report consist of the ***subject of philosophy*** investigated in Chapter 3.1, which outlines the philosophical properties that our DM tool will incorporate, and ***technological modelling*** explored in Chapter 3.2, which underpins the DM techniques we have applied to our decision-making model.

Much of the (philosophical) literature around decision making focuses on the following branches of philosophy: (1)*Meta-ethics*, the study of morality, as Miner and Petocz (2003) outlines, there is a "need for training in moral philosophy"[2], (2)*Social influence*, described by Al-Ali et al. (2012) as the ethics that "change with the trends of society", and (3)*Legal influence*, which is acknowledged by Dennis et al. (2013) as important in any autonomous system, though can be disregarded when illegalities do not exist or cannot be avoided, or when other ethics take precedent. We have labelled these as the Metaethics, Sociology and Legality **pillars** to our DM model, where (2), Social Influence, holds the most weight, given a study from Nolan et al. (2008) supports *descriptive normative beliefs* (i.e.) what

[1] The program is not the focus of the project, nor are we proposing it as a product; it will be used as an evaluation tool for the framework

[2] In the context to training DM models

is classed as the social norm as the most influential behaviour[3] and (1), Metaethics, would hold the second rank, as we have reasoned with Dennis et al. (2013) behind the lowered importance of legal impact.

In addition to layering our model with philosophical reasoning, we need to apply a technical framework that supports it. There are several popular structures of DM models, such as *priority queues*[4], used in Islam and Rashid (2018) as a *heap* (tree-like structure), or *cost structures*[5], explored in Lee et al. (1999) as a *centralised system*[6]. There also exists purely learning structures (relating to ML), such as in Noothigattu et al. (2017), who uses a voting-based system to train decisions, and Kartal et al. (2016) who finds algorithms such as *support vector machines* (SVM) useful for *multi-criteria decision-making* (MCDM) tasks, with regards to automobiles.

Aside from structuring our framework, we need to hurdle two obstacles in order to develop an exemplar program. The first focuses on the translation of philosophical topics and contextual information (contextual to the dilemma being posed) into language that can be understood by a program. The second focuses on the reduction of bias when evaluating data, which is imperative for any human-oriented research.

Otherwise, the structure of this report is as follows. Chapter 3 will focus on the body of research. More precisely, Chapter 3.1 will focus on the philosophical reasoning implemented in our DM model, whilst in Chapter 3.2, we will apply DM and learning techniques in order to develop an exemplar program. In Chapter 4, we propose the finalised philosophical framework and in Chapter 5, we implement the framework in programmatic form. In Chapter 5.2, we illustrate the results from our testing, so that we can evaluate them and draw conclusions in Chapter 6.

[3] Interestingly the study exhibits peoples absent-thought towards the influence of normative action

[4] A *priority queue* is a selection sorting algorithm that assigns priorities to each outcome and decides on the highest-priority item

[5] A *cost structure* is a model that incorporates a cost-function with the purpose of finding the least-costly outcome

[6] A *centralised system* incorporates a central component that relays with external components individually. The alternative is a decentralised system, where all components relay with each other (similar to a fully-connected graph), though the central component still makes the primary decisions

1.3 Other Comments

The increased commercialisation of automotive cars has lead to a rise in research regarding ethical car-crashing systems, from which we have gathered there is very little consideration for multi-perspective ethical systems. We have found similar markets exist, for example in automated drone systems and further driver-less vehicles such as planes. The commonality between these areas of research is the focus on resolving a death-or-death outcome dilemma, so ensuring that these systems are programmed/taught appropriately should be the highest priority. Nevertheless we have found that this is not the case. Hence, the commercial need for this important in order to ensure systems are not programmed, developed and marketed with a poor ethical consideration.

To briefly comment on the impact of COVID-19, there is little to fret about the construction of our model, whether philosophical or technological, is not constrained by any physical instruments or facilities, nor is/will there be any need for person-to-person contact.

Chapter 2

Methodologies

2.1 Methodologies

Our initial methodology focuses on acquiring an *acceptable* level of the philosophy surrounding decision-making and ethical dilemmas. In order to accomplish this, we sought support from several peers (philosophy students) who assisted us in compiling a reading list consisting of books, papers and presentations.

The following methodology focused on the technological research to accompany the philosophical work and was intended to last several weeks. The plan focused on familiarising ourselves with a multitude of ML techniques that are relevant to DM) applications and outlining useful properties to understand what technical implementations are possible.

After the compilation of research and development of the framework, we designed a methodology to exemplify the framework's functionality. The method focused on developing an exemplar program, collecting data to train our program and testing. The development method focused on implementing the defined framework and subsequent ML and DM techniques in *Python* and researching available libraries; the data collection method looked at quantifying individual views programmatically using case studies and creating an online survey in order to gather data used to then train our model; and the testing method looked at program testing - given ethical dilemmas have no right/wrong property, we could only derive a testing method to ensure repeatability and reliability of the program.

The final methodology we needed to draw out regards the evaluation of the framework. Using the exemplar program, we planned to draw conclusions as to the reason behind the decisions made. Additionally, we envisioned outsourcing a more professional evaluation from volunteers with experience in philosophy, DM technologies and ethical dilemmas.

These methodologies are presented through a two-phase plan in the following chapter.

A note on planning methodology: we began with an initial plan, labelled *Phase I*, which follows the research and design of our project including the submission of a mid-project *progress report*. Thereafter, we reviewed our progress along with the feedback from our report in order to outline the final plan, labelled *Phase II*, which follows the final stages of exemplar program development, data collection and framework evaluation. In addition, we added contingencies, whereby free-time was scheduled between Phase I and II so that any work from Phase I that was delayed could be completed during this period, otherwise the spare time could be used for preemptive work for Phase II. This would act as a contingency for the completion Phase I. Similarly, several weeks were left at the end of Phase II for the same reason.

2.2 Initial Plan, Phase I

This subsection outlines the plan prior to the mid-project review.

During the period of October-December 2020, we planned to spend a month beginning our research and as we gathered thought,s we would begin designing the framework within the following month, as outlined by the Gantt Chart in Figure 2.1. The research phase overlapped with the model design stage as it would allow us the time to confirm our initial research outcomes and perhaps instigate further conceptualisation of design. The progress report hand-in was set for 12 August 2020, so much of November and December was spent writing-up our research findings and framework design. We expected to receive feedback on the project during January, so no preemptive plan was made past this; which allowed us to keep the project open to large change given we were still uncertain at this stage how we would

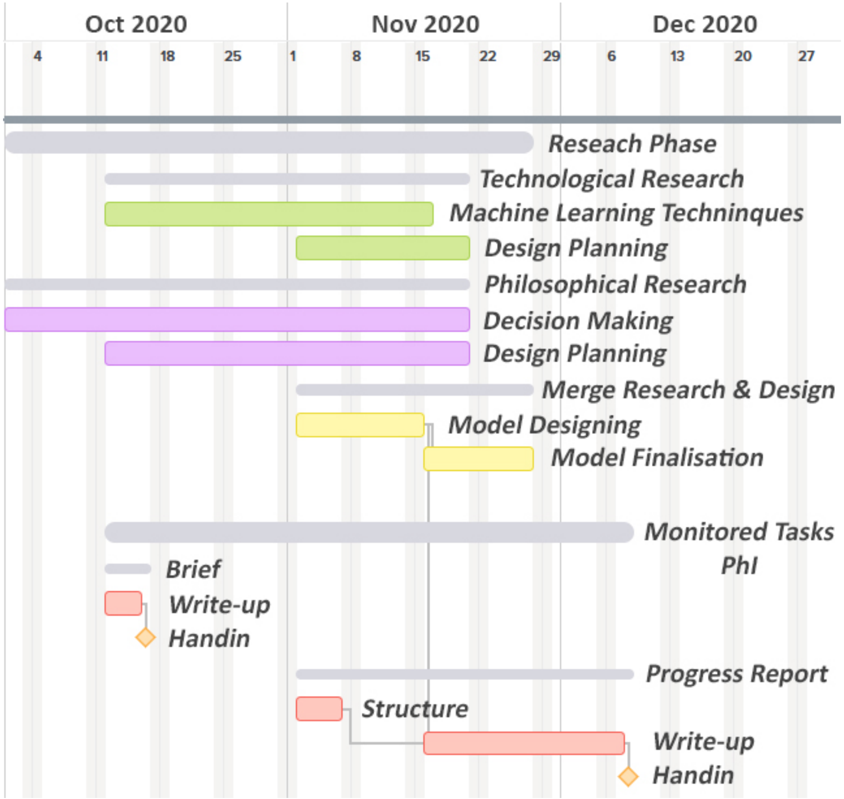


Figure 2.1: Gantt Chart for Phase I Plan, October - December 2020

exemplify the function of the framework. In addition, the spare time would acts as padding in the case where work may have been delayed.

2.3 Final Plan, Phase II

The report feedback focused on improving the writing style and content, therefore in planning for the following stage we did not place as much emphasis on immediate project completion, so to gather as much presentable content as possible. Though, as you can see from Figure 2.2, we stressed the significance of final report writing by allocating a larger portion of time to it.

Other than the external feedback, we decided to perform a personal review. Here, we weighed areas of the project to be completed that were more-or-less significant to the final outcome of the project. Domains such as exemplar program design and testing would require more time as the final evaluation of the framework hinged on outcomes from the program, therefore, more time was allocated to ensure simulations were appropriately

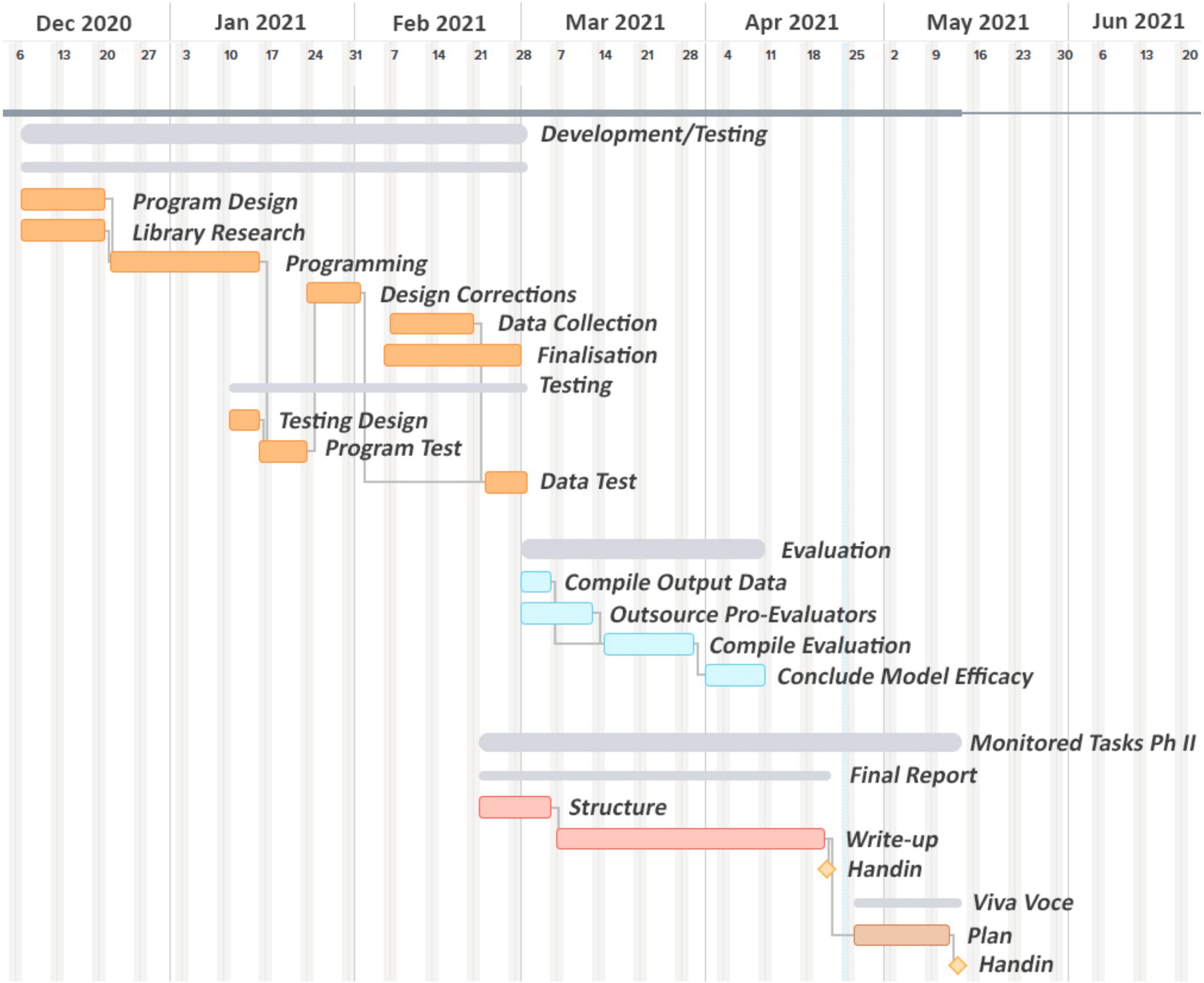


Figure 2.2: Gantt Chart for Phase II Plan, December 2020 - April 2021

completed. By comparison, the evaluation portion of the project was allocated less time as this required less overall effort. In addition, we permitted its completion to overlap with the report-writing portion of the project, as the majority of the report is independent of the final findings, allowing ourselves more time to make improvements on writing style. Similarly to Phase I, several weeks were left empty to act as contingency in the case where work was delayed.

The Gantt chart in Figure 2.2 illustrates the plan in more detail.

Chapter 3

Body of Research

3.1 Subject of Philosophy

3.1.1 Introduction to Philosophy

In order to introduce this section, we will cite from Joseph Raz, Raz et al. (2012), which outlines the philosophy on which we have based our model,

> *"All normative phenomena are normative in as much as, and because, they provide reasons or a partly constituted by reasons".*

This is to say that any action taken defined as *normative*[1] can only exist due to *reason*[2]. By extension, this section aims to apply *Reason* and *reason* to several ethical processes in order to build a reliable framework.

Furthering this, our model is based of three pillars of modern-day ethics, (1) Meta-ethics, (2) Sociology and (2) Legality, outlined in Chapter 1.1 and explored in the following section. Within these pillars, we have incorporated more unique ideologies that aim to represent their respective branches of philosophy in various ways. In addition, we have also implemented common

[1] *Normative* is defined as some actions/outcomes that are viewed by society as good/permissible and others as bad/impermissible

[2] Not to be confused with *Reason*. Raz et al. (2012) outlines, *reason* consists of situation $S(t)$ occurring at point t in time due to the occurrence of $S(t-p)$ at point $t-p$ in time, where $p > 0$. Whereas *Reason* is situation $s_x(t, c)$ occurring because of choice c, where $s_x \in S$ - set of possible scenarios, and $c \in C$ - set of possible choices.

ideologies present across the pillars, which we call **core philosophies** (also **cores**).

The core philosophies are based on ideologies taken from Wilson (1979). The main core is *compatibilism*, interpreted as the *compatibilist*'s stance on DM[3], which we have implemented by dividing the pillars into two groups we call *micro-frameworks*. The two groups correspond to: (1)Deterministic calculation which we have defined as deterministic reason, by which we mean the causation and occurrence of an outcome are inevitable, and (2)Reasoned calculation which we have defined as libertarian Reason, by which we mean the choice made was made from free will. The amalgamation of this philosophy is agreed upon by Friedrich Nietzsche, a prolific 19^{th} century philosopher who abode by the words *Amor Fati* (*a love for one's fate*), which suggests a deterministic life, yet is publicly understood as the choice to apply free will when necessary, otherwise love the fate your are given.

Another core considers *the Reasonable Man*'s[4] approach, which we have incorporated as the reduction of unreasonable emotional influence, an attribute partially backed by Gaudine and Thorne (2001), who states ”emotion is often considered a non-essential aspect to the ethical decision making process”. Our interpretation is that a reasonable solution would only incorporate emotion when it impacts a decision beneficially. For example, emotion in legal context is unreasonable, whereas emotion when proposed with a problem that incorporates a familial aspect is reasonable. This is implemented as the decision when to or not-to incorporate emotional influence within each pillar.

The last core theme is conceptualising common sense as intuition[5]; recognising when to use it and mimicking the process. As explained by the *recognition primed decision model*[6], discussed in Klein (1999), intuition

[3] *Compatibilism* is the idea that on occasion our actions are predetermined, otherwise we utilise *free-will*. Compatibilism as opposed to *Libertarianism*, the rational implementation of free-will, and *Determinism*, choice is predetermined

[4] *The Reasonable Man* is a person whom, no matter the education, background, environment, etc., will never fail to make the *most* reasonable decision. *Note:* This does not mean personal/emotion thoughts are removed, only that personal context is.

[5] Though Wilson (1979) explores intuition and common sense in separate chapters, they are linked through similar definitions, where intuition can be inferred as the subconscious influence of what *you* perceive as common sense. There also exist problems when modelling pure intuition so we have to appropriate it to common-sense in order to reduce complexities

[6] A DM model for complex situations, based on human thought processes

can lead to solutions that do not require weighing outcomes. Therefore, this core looks at understanding when a solution can be approached intuitively and what mechanisms are appropriate to each micro-framework in order to mimic intuition. For example, having to decide between killing a million people and a billion people is an intuitive problem, which does not need processes such as calculating legal implication. We acknowledge that there are potential dangers when incorporating common sense to simplify complex DM problems, as outlined in Goodwin (2009), consequently, the model only uses common sense intuition when faced with an solely intuitive problem, as per our previous example.

3.1.2 Meta-ethics (Moral Influence)

By definition, Meta-ethics is the study of moral thought, though within our model this concept is more appropriately linked to individual morality, and is based on Freud's division of mind into three: the *Id*, the primitive and instinctive characteristics of our personality, the *Ego*, how we currently view ourselves, and the *Super-Ego*, the ideal standard we set ourselves, explored in Freud (1923). Simply put (within our model), the Id represents instincts (related to the core of intuition), which will overwrite the other meta-ethical decisions, though only under particular circumstances; the Ego is represented by the depreciation/appreciation of ourselves after having taken the decision; the Super-Ego is represented by what we conceive as the *ego-ideal* - explored in J.C.Flugel (1945), "the ego-ideal is the first source from which the super-ego is derived". The Id will be present within the Deterministic frame, while the Ego and Super-Ego will work within the Reasoned frame, where the Ego will be the only area susceptible to emotional influence, under the constraint set by the core of reason. If the Super-Ego and Ego frame show disparity in choice, the Id frame will proceed to choose between the choices suggested by the Super-Ego and Ego frame.

3.1.3 Impact of Human Law

Dennis et al. (2013) reasons, "if necessary, disregard legal restrictions for ethical reasons". This is the stance we have taken with regards to the legal pillar. Though, only because it can be disregarded when concerning problems that operate outside the law, there still exists dilemmas where decisions can be swayed due to legal action. Note that this sub-section will *not* provide specific legal reasoning and advice. Instead, it aims to sway a decision dependant on the severity of legal implications.

It is important to understand the allocation of blame when regarding legalities as this can aid the conviction enourmously. Prentice and Koehler (2002) divides blame into two categories, omission bias, "a tendency to blame actions more than inactions", and a normality bias, "a tendency to react more strongly to bad outcomes that spring from abnormal rather than normal circumstance". We have taken Prentice and Koehler (2002)'s stance on blame and partitioned the legal pillar into two areas, *normality* (i.e. how usual is your choice), which will exist within the Reason frame, and *inaction* (i.e. has the neglect of other action led you to worse results), which will exist within the Deterministic frame. Within each of these micro-frameworks will lie a structure that dictates the criminal severity of a choice. Unfortunately, through reason[7], we must neglect emotion when determining legal impact, as Fiss (1991) explains, emotion is "inconsistent with the very norms that govern and legitimise the judicial power", enacting the core of reason. When the pillar faces disparity between choices ranked on the same scale, the theme of intuition will hold the decisive power, and given we appropriate intuition as common sense, we can incorporate this within the normality half of the micro-frameworks.

3.1.4 Social Influence

Normative social influence is "the influence of other people that leads us to conform in order to be liked and accepted by them", defined by Aronson et al. (2005). As outlined in Nolan et al. (2008) and Deutsche and Gerard (1955), it also poses the greatest behavioural change, despite its unknowing influence. Hence, we recognise it's importance to our model.

In order for our framework to conform to a sociological idea, we need to outline the prospective social beliefs that our framework may take. The most

[7] Referring to the core theme that is based on the Reasonable Man's approach

obvious social belief is that of *utilitarianism*[8], though in actuality this incorporates a large span of beliefs. Thus, the traditional definition of utilitarianism can be more accurately described as *preference utilitarianism*, which focuses on social utility, disregarding individual utility. Similarly, there exists a *hedonistic utilitarianism* which incorporates pleasure and pain, in place of good and bad - defined for both social and individual utility. There also exists a third approach, *ideal utilitarianism* which places less value on good/bad and more value on the mental states of intrinsic worth, perhaps taking the form of philosophical, scientific or artistic worth, as proposed in Moore (1912). Though, hedonistic and ideal utilitarianism are vulnerable, as explained by Sen and Williams (1982), who exemplifies the hedonistic approach as outdated and the idealistic approach as simply not respective of true group behaviour where individual lust for a certain mental state can easily take precedence over another. In addition, Sen and Williams (1982) illustrate that the preference approach consistently opposes *preference autonomy*, the ideas of good/bad based on an individual's desires; neglecting this from one's decision process removes the likelihood one will endanger the "greater good" of people due to a nefarious desire. This is the ideology we have decided to implement.

In order to cohere with this ideology, the pillar looks at defining what good can come out of the situation with regards to the largest population involved. Furthering this, mechanisms that define good are present in *both* the Deterministic and Reason frame. This permits the frame to apply our social ideology in multiple contexts. The Reason frame looks at choosing the greatest good for the greatest sum from the available outcome scenarios. Differently, the deterministic frame defines what society sees as the greatest good for the greatest sum, and determines which outcome is most-similar[9].

Unfortunately, emotion cannot be incorporated (core of reason) given our ideology removes personal utility as an influencer for DM, though it can influence our understanding of the problem. Therefore, emotion may warp the understanding of severity between choices in circumstances where individuals feel more emotionally attached. In context to the framework, the

[8] *Utilitarianism* is the belief than any action taken will be to maximise the greatest *good* and/or minimise the greatest *bad* for all, encapsulating social and individual utility

[9] This may seem like the implementation of Reason not determinism, given there exists a choice outcome (i.e. Reason) not a pre-determined outcome (i.e. determinism), however this is false. Understand that we aim to generate the predetermined outcome given societal beliefs and only choose the most similar available outcome, as the ideal may not be probable and cannot be considered

preference utilitarian approach is used to determine decisions for both Reason and Deterministic frames, where emotion can exaggerate the difference between decisions (with the aim of minimising uncertainty). The core of intuition is present when disparity exists between the choices made by the two micro-frameworks within the pillar, thus a socially focused intuitive choice is made.

3.2 Technological Research

3.2.1 Introduction to Technological Decision-Making

Technological modelling for ethical decision making is a field that has been amplified by the autonomous-revolution, more accurately, the rise of the autonomous vehicle, underpinned by Goodall (2014). Due to this, papers such as Islam and Rashid (2018) and Noothigattu et al. (2017) are proposed in order to solve a problem which bears a vast sum of solutions, which renders the field slave to the automated-vehicle DM problem, relinquishing a large gap for generalised problems. This is where our research enters the field, proposing a solution with regards to generalised ethical dilemmas.

In order to propose any technical solution we need to define our problem in a technical manner. Plainly, we aim to solve an MCDM problem, where the criteria we acknowledge is based on the pillars which we previously outlined in Chapter 3.1. Studies such as Jato-Espino et al. (2014) and Wang et al. (2009), who review MCDMs in high-pressure[10] environments, highlight the reliability that can be placed on widespread methods and the tendency to combine multiple methods. In addition, Zopounidis and Doumpos (2002), de Almeida et al. (2017) and Ananda and Herath (2009) provide comparative reviews of methods, with respect to other high-pressure environments, which enable us to outline which combination of MCDMs are appropriate for us to incorporate. In Triantaphyllou (2000), these methods are presented more elaborately, the weighted sum model (WSM) and the analytical hierarchy process (AHP) are outlined as the most reliable. Consequently, we have placed a primary focus on integrating these methods.

Otherwise, there is a large research base dedicated to learning (ML) algorithms. Papers such Busemeyer and Myung (1992) and Hill et al. (1994) propose variations of an artificial learning networks, which provide reliable though complex methods focused on improving precision. We do not consider incorporating methods for the purpose of precision; dilemmas are dilemmas because there is no precise answer. Classification methods would be an appropriate alternative. Kiang (2003), Harper (2005) and Hand and Henley (1997) compare classification methods and jointly conclude that logistic regression (LR) is among the most popular and highest-ranking

[10] Ethical dilemmas are high-pressure by nature are key, hence the necessity for reviews concerning high-pressure scenarios

method for classification problems. Consequently, we seek to incorporate LR within our framework.

Other than outlining the techniques we wish to include, we need to compare them in order to properly assign them to a pillar (and relative *sub*-pillars). Frutos-Pascual and Zapirain (2015) compares simple DM an ML methods and concludes that there is a significant advantage to classification technique, though it is outlined that MCDMs used in combination are also effective.

Within this section, we focus on applying several of the aforementioned techniques to each micro-framework. Weighted comparative techniques such as AHPs and WSMs will enable low-level decisions to be draw out, while the incorporation of LR will allow us to improve the certainty of decisions. The final application will be applying calculated scores from each pillar and subsequent micro-frameworks to weights associated with a single basis function, $\Phi(\cdot)$. This basis function will deliver a multivariate function whose coefficients will dictate which choice is finally made.

3.2.2 Decision Making and Machine Learning Methods Applied

The layout of our model incorporates a total of six micro-frameworks, which pertain to the three pillars outlined in Chapter 3.1.1 (i.e. two frameworks for each pillar). Each micro-framework will incorporate its own techniques. Recalling the pairs of micro-frames, one takes the Deterministic approach and the other which incorporates Reason. In this section, we aim to establish the technical application of the frameworks with reference to Chapter 3.1, where we proposed the philosophical ideologies we wish to incorporate. Note that this section will not incorporate any reasoning for technical implementations, this will be addressed in Chapter 4.

The first pillar we would like to consider is the meta-ethical pillar. In Chapter 3.1.2, we interpreted meta-ethics in terms of Freud's Id, Ego and Super-Ego, where the Id exists within the Deterministic frame (which we have called MD1) and the Ego and Super-Ego exist within the Reason frame (which we have called MR1). Given our interpretation relies of individual morals, we are unable to apply any learning methods to the model as this would corrupt the individuality we seek to incorporate. Therefore, this pillar utilises simpler DM techniques.

MD1 looks at incorporating the instantaneous choice made by the client, e.g. by incorporating a determination made by the client within the first 5 seconds of being presented with the problem and set of choices, the choice is denoted as $C_{X,id}$[11]. This allows us to implement intuition without having to program for it[12]. Differently, MR1 is divided into MR1a and MR1b, which focus on the Ego and Super-Ego respectively. MR1a employs a cost-structure using the AHP approach, which assigns a value for moral-wrongdoing to each choice, constrained by $0 < x < 1$, with a condition of proximity that states if the two most ideal ranks are within 0.1 score of each other then emotional influence takes control[13]. Here, the client-agent chooses the outcome that is most emotionally poignant. The choice made in MR1a is denoted by $C_{X,ego}$. Similarly to MD1, MR1b incorporates the instantaneous choice and is denoted by $C_{X,sego}$. Thereafter, a comparison between $C_{X,ego}$ and $C_{X,sego}$ is made and the condition $C_{X,ego} = C_{X,sego}$ needs to be satisfied in order to determine a choice for the meta-ethics pillar. If there is disparity, the decision made from MR1 is taken as the final choice[14]. This functionality is visually represented in Figure 3.1.

The second pillar we studied is the Legal pillar. In Chapter 3.1.3, we divided the pillar into two sections, normality and inaction which represent the Reason and deterministic frames, and labeled LR1 and LD1 respectively. To incorporate normality, we need to determine the norm choice and to accomplish this, we developed a ranking system that weighs choices depending on external data[15]; in scenarios where ranks receive the same weighting, an intuitive decision is made similar to that in MR1.

In order to assimilate the inaction frame, we utilised the WSM method by assigning each choice, C_n, a value that represents the legal detriment of making that decision, D_{Cn}, which forms the set of detriments $D = \{D_{C1}, D_{C2}, \ldots, D_{CN}\}$, for $n = 1, 2, \ldots, N$. Furthering this, a set of weights,

[11] This is where the degree of reliance on the client exists, hence the assisted side to our proposed DM framework

[12] Referring to a footnote we made previously where we mentioned the difficulty with programming for intuition

[13] This was determined as a proximity of 0.1 is small enough to be swayed by emotional influence, though is subject to change

[14] This is the intervention of intuition given uncertainty

[15] This data is taken from a sample of people's choice given this/similar situations

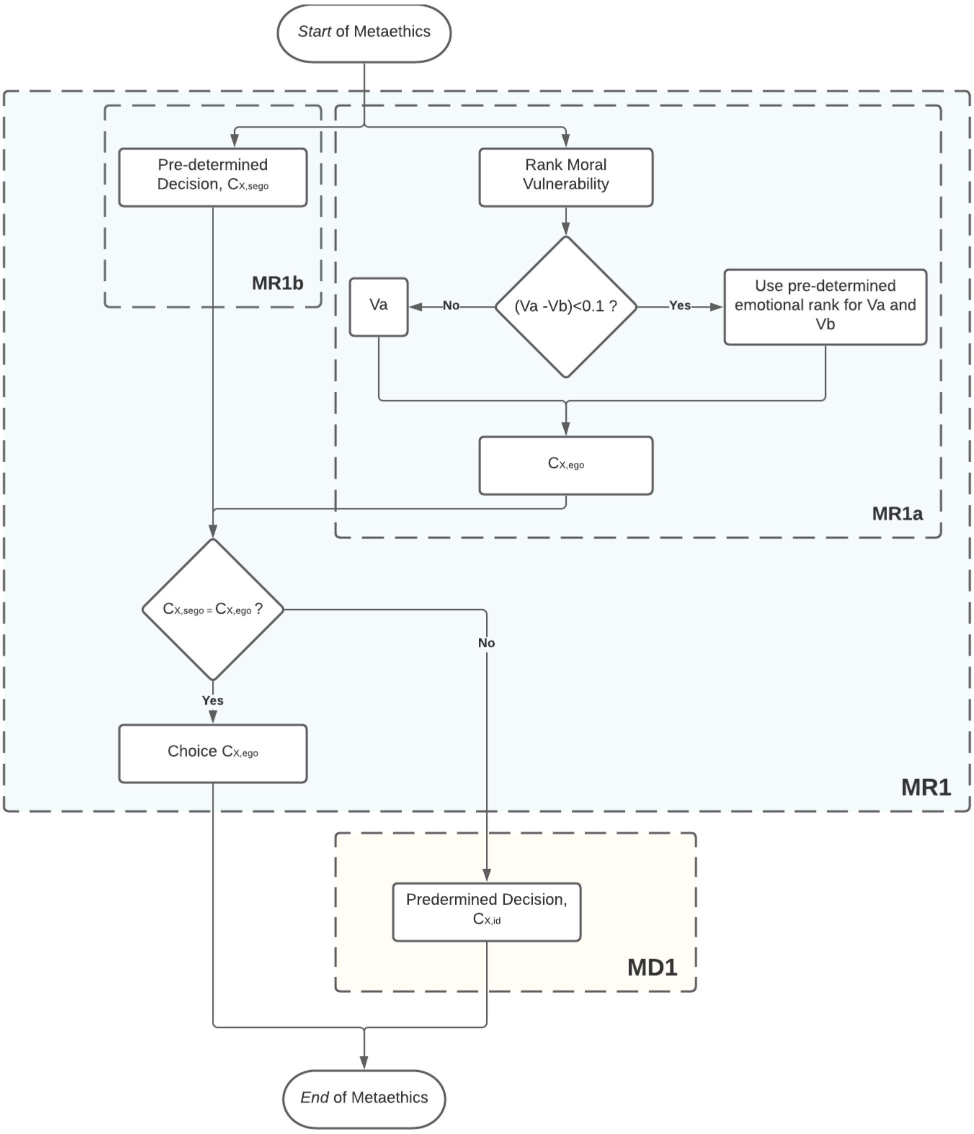


FIGURE 3.1: Decision Model of the Metaethics Pillar

$W = \{W_{C1}, W_{C2}, \ldots, W_{CN}\}$, are calculated for the respective choices using Equation 3.1, which determines a weight-value associated with the inaction of a choice C_n. Values of W are then applied to the WSM approach in order to find the optimal decision.

$$W_{Cn} = 2D_{Cn} - \sum_{i=1}^{N} D_{Ci} \tag{3.1}$$

Thereafter, values for LR1 and LD1 are calculated then compared. If $C_{LD1} = C_{LR1}$ then C_{LD1} is passed as the final legal choice, if $C_{LD1} \neq C_{LR1}$ then a set of ranks that evaluate the severity of choices is used, and the least sever choice from C_{LD1} and C_{LR1} is passed as the legal choice. Figure 3.2 illustrates the mechanisms of the legal micro-frameworks.

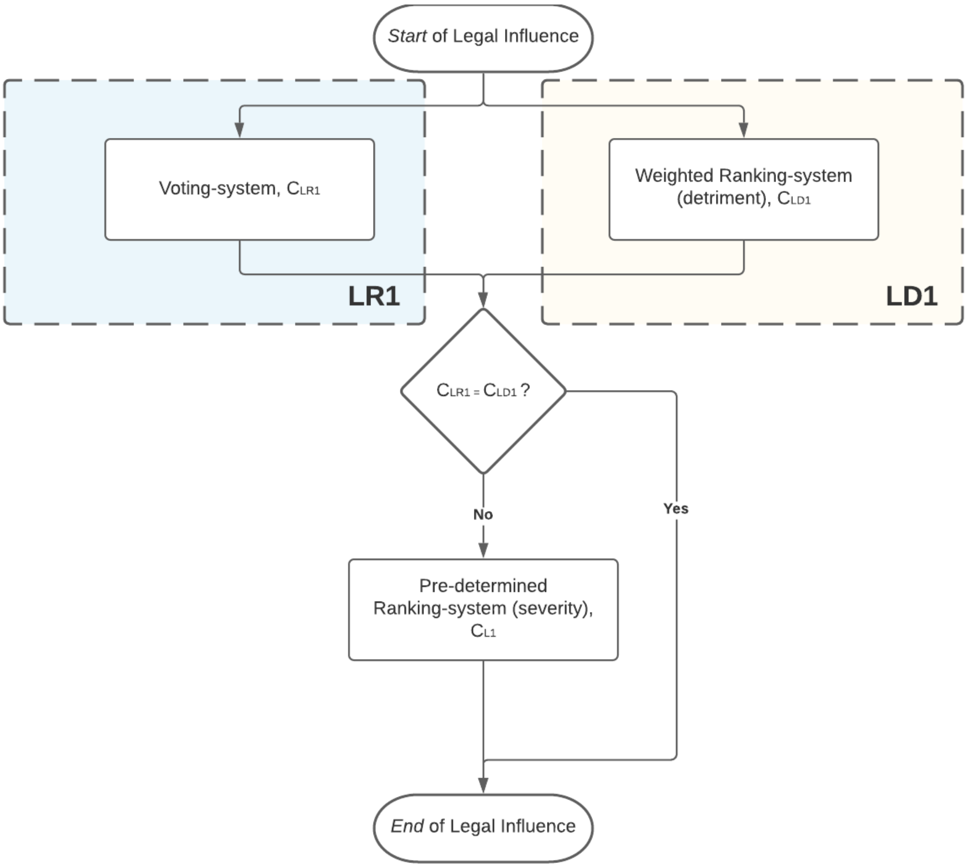


FIGURE 3.2: Decision Model of the Legal Pillar

The final pillar we need to delineate as a frame is the Social pillar. In Chapter 3.1.4, we defined the implementation of the greatest good for the greatest sum within respect to the Reason and Deterministic frames, taking the preference utilitarian approach. Going by technical application, a voting-based structure best suits the Reason frame, where the final choice is defined as C_{SR1}. Whereas for the deterministic frame we have imposed the learned-weights logistic regression (LWLR) (this is where we fit the aforementioned LR) method for classification, as defined in Chapter 4 of Friedman et al. (2001).

Considering the problem $\Phi(C, X)$ of multi-variate classification solved using LWLR, where the aim is to classify the point X using a set of class-related points, C. In our problem, any point can be defined by 2-Dimensional vector, $\begin{pmatrix} x_1 & x_2 \end{pmatrix}$, where x_1 represents the greatest good and x_2 represents the greatest sum of people and are constrained by, $0 < x_1, x_2 < 1$. X represents the ideal outcome, (i.e.) the greatest good for the greatest sum of people, therefore is evaluated as $\left(\begin{smallmatrix}1 & 1\end{smallmatrix}\right)$.

The problem is defined as

$$\begin{aligned} &\Phi(C, X) \quad \text{where} \\ C =& \{C_1, \dots, C_N\} \\ C_n =& \{\overline{x_1}, \dots, \overline{x_{N_n}}\} \\ \overline{x_i} =& \left(\begin{smallmatrix}x_1 & x_2\end{smallmatrix}\right) \\ X =& \left(\begin{smallmatrix}1 & 1\end{smallmatrix}\right) \end{aligned}$$

where N represents the number of classes, C_n is a set of points which represent class n, N_n represents the size of class n and $\overline{x_i}$ represents a point in pertaining to class n.

To briefly explain our LWLR implementation, we consider that Logistic-regression is a probability-dependant classification method where the probability of a point x pertaining to a class k, is found using the soft-max regression function described by Equation 3.2, where $s_k(x)$ is the projection of the point x onto learned weights for class k, ω_k, thus $s_i(x) = \omega_k^T \cdot x$. Furthering this, LWLR focuses on presenting certainties, this is mathematically represented in Equation 3.3, which equates probabilities in order to calculate the certainty of pertaining to class k. Which is further simplified in Equation 3.4. Using $\omega_k^T x - \omega_K^T x$, we can derive objective values of certainty which will allow us to classify point x, thus classify the point $X = \left(\begin{smallmatrix}1 & 1\end{smallmatrix}\right)$. Equations 3.2 and 3.3 were taken from Equation 4.17 in Friedman et al. (2001).

$$P(k = K|x = X) = \frac{exp(s_k(x))}{\sum_i exp(s_i(x))} \tag{3.2}$$

$$log\left(\frac{P(k = k|x = X)}{P(k = K|x = X)}\right) = \beta_{10} + \beta_k^T x \tag{3.3}$$

$$\ln \frac{\frac{e^{s_k(x)}}{\sum_i e^{s_i(x)}}}{\frac{e^{s_K(x)}}{\sum_i e^{s_i(x)}}} = \ln \frac{e^{s_k(x)}}{e^{s_K(x)}} = s_k(x) - s_K(x) = \omega_k^T x - \omega_K^T x \tag{3.4}$$

We must recognise this as an important implementation, as without it there is a high likelihood of contention between classes. This is due to the nature of ethical dilemma DM, where by it is already difficult to decisively classify potential outcomes as humans, therefore classifying programatically would be equally, if not more difficult. The final choice made will be labeled C_{SD1}. Figure 3.3 illustrates the sociological pillar, where a decisive cost-structure is used to differentiate between choices C_{SD1} and C_{SR1} in cases when they do not satisfy equality, in a similar fasion to the legal frame.

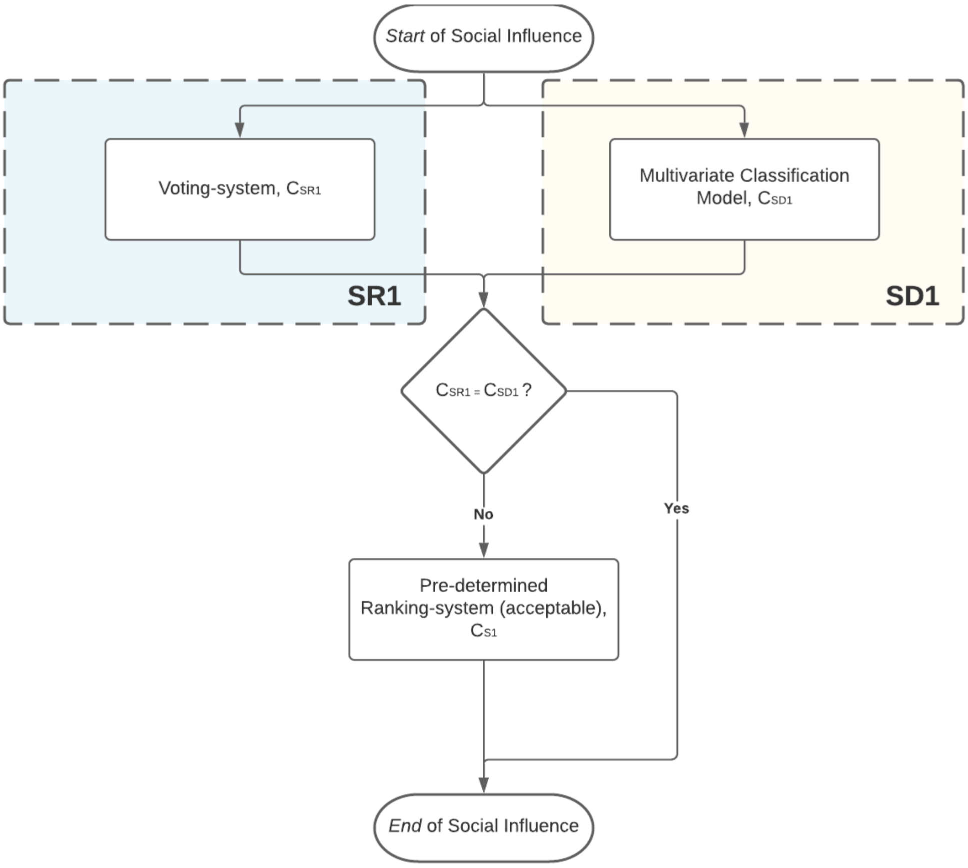

Start of Social Influence
Voting-system, CSR1
Multivariate Classification Model, CSD1
SR1
SD1
CSR1 = CSD1 ?
No
Yes
Pre-determined Ranking-system (acceptable), CS1
End of Social Influence

FIGURE 3.3: Decision Model of the Social Pillar

Chapter 4

Proposed Framework and Justification

4.1 Proposal

The philosophical framework being proposed is illustrated in Figure 4.1. The final function described by $f(C_M, C_S, C_L)$ represents Equation 4.1, where α, β, γ are predefined constants which aim to amplify, reduce or neglect choices from either pillar.

Note: C_M, C_S, C_L do not represent numerical values, so the final form of the function will be a linear multivariate, with a maximum of three variables. The variable (i.e. choice) with the largest coefficient is the ultimate choice.

$$\Phi(C_M, C_S, C_l) = \alpha C_M + \beta C_S + \gamma C_L \tag{4.1}$$

The technical model is designed to exemplify, test and therefore determine the pros and cons that come with the proposed framework. The individual models to these were laid-out in Chapter 3.2.2.

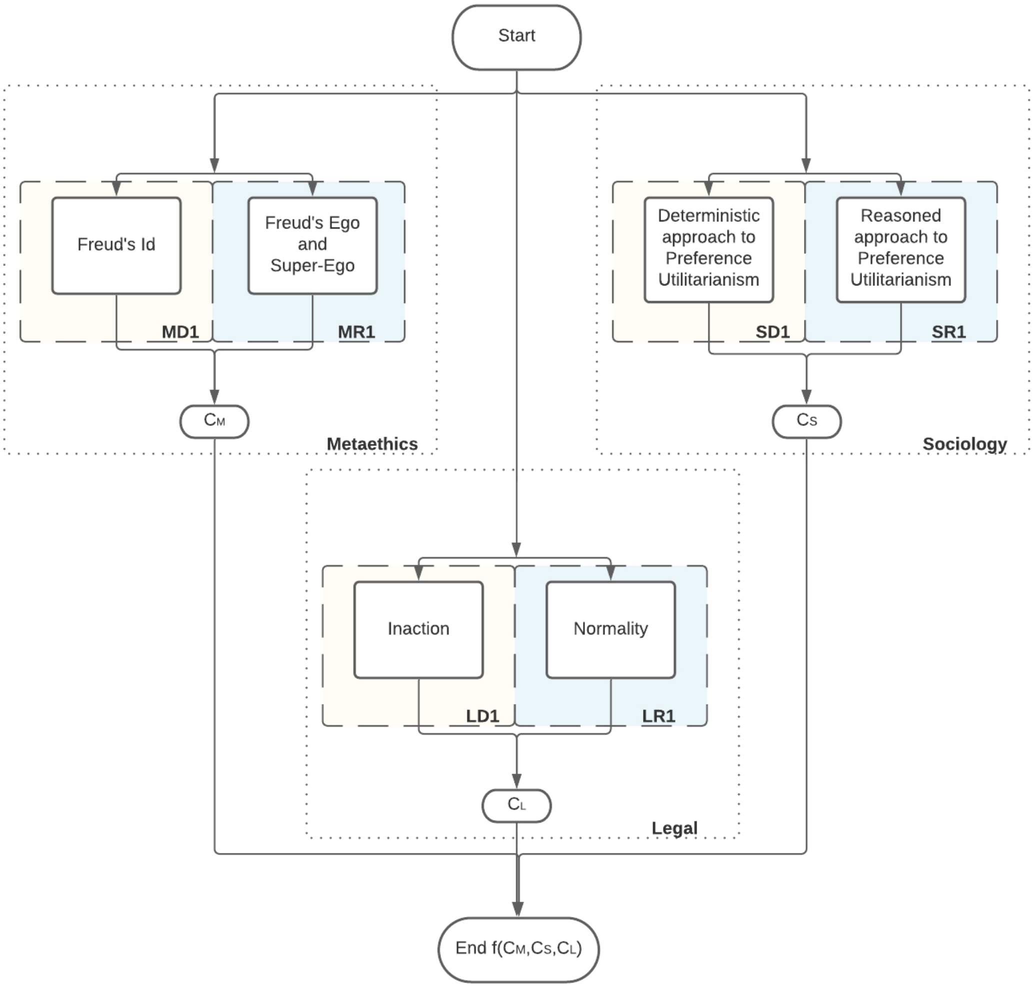

Start
Freud's Id
Freud's Ego and Super-Ego
MD1
MR1
CM
Metaethics
Deterministic approach to Preference Utilitarianism
Reasoned approach to Preference Utilitarianism
SD1
SR1
CS
Sociology
Inaction
Normality
LD1
LR1
CL
Legal
End f(CM,CS,CL)

Figure 4.1: Framework flow-graph

4.2 Technical Justification

Now we have applied technical application to the proposed philosophical framework, we need to cover the justification for the use of DM and learning algorithms.

Within the meta-ethical pillar, we made the case that learning techniques are unwise given our interpretation of meta-ethics (moral theory) relies heavily on individuality. This is where client-agent interaction is necessary in order to assist in determining a decision. Thereafter, we made the decision to avoid using complex tools such as *personality tests* to map a client's morals as we thought there were more efficient methods at our disposal. For MD1 and MR1b, the implementation of instantaneous choice allowed us to incorporate two instinctive aspect of a client's moral code. The Id, which is itself a characterisation of our instincts, and the Super-Ego, which is the intrinsic property of our moral standards (therefore should come instinctively to us). In addition, it allowed us to implement the core of intuition. Alternatively, MR1a utilised a cost-structure determined by the client. The Ego is not thought of instinctively, and in lieu of our prior decisions, it was necessary that a part of the moral mechanism was not instinctive.

The Legal pillar is the simplest implementation of the three pillars as it holds the least significance.The LD1 frame incorporates normality, so a simple voting-system determined from external data (a group of people, not including the client) seemed most fit. Similarly, the LR1 frame incorporated a voting-system, though done so to stage a WSM structure which would allow the program to determine a choice due to inaction. We felt a candid voting-system that asked individuals to vote on inaction could leave the ranking vulnerable to ulterior interpretation[1], hence the structure of the WSM.

The Social pillar is the only pillar that incorporates a learning-based algorithm as it is the only pillar that can be influenced by large sums of people. Due to the nature of determinism, we struggled to conceive an approach that utilises a choice between final outcomes, given this concept coincides with free will and not determinism. To work around this, we implemented a classification model where the point we aim to classify is predetermined. Hence, no choice is made; it is simply ”classified”[2]. Otherwise, we implemented a voting-system for LR1 determined by external data, which seemed sensible given the requirement needed to satisfy Reason.

The use of algorithms such as AHP, WSM and LR arose from the research mentioned in Chapter 3.2.1. Whereas the idea to use LWLR instead of a simple LR implementation came about from tests we had personally executed prior to conceptualisation project, though this choice is echoed in Friedman et al. (2001).

[1] Though inaction seems like a straightforward concept, once applied to complex problems it could be exposed to misinterpretations, e.g. take inaction in a scenario where there a more than five choices of outcome. How would you quantify inaction?

[2] Play on words

Chapter 5

Epitomising the Framework

It should be noted that the construction of the ***exemplar program*** *is not the focus of project, thus should only be viewed as a method of evaluating the framework.*

5.1 Program Discussion

To epitomise the on-goings of our framework and objectively evaluate its efficacy, we have chosen to apply our framework programmatically. This will consist of the design and development of an exemplar program, the collection, filtration and generation of training data and a post-simulation evaluation.

5.1.1 Key Implementations

Using the flow charts developed in Chapter 3.2, we can more easily understand what needs to be implemented in the main program body and how it can be accomplished. Considering the flow charts, we have separated the program into three separate modules, where each module will focus on a different pillar of the framework, namely the legal, meta-ethics and social-influence modules.

The legal module is split into three sub-modules labeled, *inaction* - referring the micro-frame LD1, *normality* - referring to the micro-frame LR1, and *severity* - referring to the process which determines a final outcome for the legal frame. In finality, the main legal module takes the outputs from each sub-module and applies conditional statements in order to make a final choice for the legal frame, as depicted in Figure 3.2.

Similarly, the meta-ethics module is split into three sub-modules, each sub-module representing one of Freud's Id, Ego or Super-Ego. The main meta-ethics module amalgamates the choices from each and makes a resulting choice for the meta-ethics frame, as depicted in Figure 3.1.

Again, the social-influence module is split into three sub-modules, representing the reason and deterministic frame, as well as a visualisation sub-module for the LWLR implementation. The visualisation tool will grant us the opportunity to comment on the efficacy of the program as an evaluation method for the framework, which is to say a poorly distributed plot would suggest a high margin of error in either the distribution of data or data collection. This is discussed in detail later on. The particular LWLR program has been adapted from a separate project[1], where we evaluated the effectiveness of several LWLR classification techniques. Specifically for this project, we have determined that the LWLR technique was the most appropriate and should prove the most effective classification method. To conclude, the main social-influence module takes the choices made by the two sub-modules and if necessary applies a choice discriminator in order to derive a final choice for the social-influence frame, as depicted in Figure 3.3.

Let us briefly mention how the visualisation tool will be implemented. Using the LWLR implementation mentioned previously, we have the opportunity to call a group of different optimal weights (this is a function command that comes with the implementation). This group consists of the best sets of weights for projections in descending order and as we aim to visualise our data using a 2-Dimensional plot, we will take the top two sets of weights. Thereafter, we will map our data onto both sets of weights and plot the projections against each other. This will form our visual graphic, where training-data will be colour-coded by its respective class and the test-data point, [1,1] as described in Section 3.2.2, will be outlined with a particular shape. In addition, we will return the values of certainty associated with our test-point to the user-agent, thus allowing us to determine the effectiveness of the LWLR algorithm. For example, poor sets of certainties, and

[1] Coursework for COMP3223

weak distribution of classes, could signify the program has difficulty differentiating classes, which would suggest LWLR was the wrong classification method to use.

Separate from the main body, we need to recognise three other key implementations: the structure of the input data, the filtration of real data, and possibly the generation of new data. The first, data-structuring, is essential for any implementation to be trained; the second, data filtration, will allow the program to filter out incomplete data-points; and the third, generating new data, could be used to expand our data set if the data collection method yields a poor sum of data.

The data-structure follows; the data-set will take the format of a Python-dictionary, where each data-point contains several subsets of values specific to each program module mentioned above; each set will be indexed in an appropriate manner for ease-of-access. If a data-subset is incomplete it will be flagged by the letter *'X'*.

From this, the data-filtration implementation will remove any data-points which contain incomplete subsets using the flag. The resulting list of complete data will form a temporary data-set.

If the resulting data-set is low in quantity, the data-generation implementation could be used. A small data-set would deprecate the reliability of the ML methods used within the main-body program, thus there may be a need to expand our data set - data-generation being the method we would use to accomplish this. This entails generating N new points per complete data-point (taken from the temporary data-set), where new points would be taken randomly from a normal-distribution of the complete data-points. The resulting data-set of N-points being our final input data-set.

If the resulting data-set is of appropriate quantity the temporary set will be used as our final input data-set.

Within the main-body program, the input data will be distributed to their respective modules. After the program has completed its DM processes, a comprehensive output will outline the final choice, as well as the choices made in each sub-module. This is done so the user-agent understands where each choice has been determined within our framework, so allows the user-agent to apply reason to each decision.

5.1.2 Data Collection Discussion

To generate results from the program we need data to train the LR and input into the other *artificial intelligence* (AI) methods used. To accomplish this, we will utilise the questionnaire method as our primary source of data. The secondary source of data being the data-generation process mentioned in the previous chapter

The questionnaire method, though proven easy to use, contains some flaws. For example, a subject could exhibit *Interpretative Bias*, however this could be overcome through careful elaboration on questions and case-studies, ensuring little is left to interpretation. *Personal Bias* can also be of issue, though given our framework relies on personally-subjective data we can accept it. Otherwise, more isolated problems may lie in the type of questionnaire being posed.

The particular style of questionnaire we have chosen is a situation judgement test (SJT). SJTs are a common methods of measuring subjective decision making, though as described in Lievens et al. (2008), SJTs can be prone to faking or coaching. As there is little-to-no incentive for this behaviour within our questionnaire, we can choose ignore it.

The format of the SJT is as follows, three case studies will be laid-out each posing a different ethical dilemma and a set of probable outcomes will be provided for the subjects to choose/rank/other depending on the type of question being answered. For example, a question may ask to rank the probable outcomes with regards to their legal-normality. Further, the first case study will focus on the infamous *Trolley/Train Problem*[2], where possible outcomes will be simply to pull the lever or ignore the lever. The second case study depicts the subject as a chief financial office (CFO) whom has to allocate monetary resources to settle debts and/or train post-graduate employees and/or employees bonuses, where a combination of either three can be chosen/ranked/other; totalling seven possible outcomes[3]. The third case study places the subject in a casino with the certainty of always winning, where the possible outcomes rely on whether the subject wants to include a combination of friends and/or strangers in their winnings, with a

[2] Explanation video: https://www.youtube.com/watch?v=bOpf6KcWYyw

[3] (1) Settle debts,(2) Train postgraduate (PG) employees, (3) Employee bonus, (4) 1&2, (5) 1&3, (6) 2&3, (7) 1&2&3

total of four possible outcomes[4]. The case-studies depict classic lose-lose, win-loss, win-win ethical dilemmas, respectively, and ensure that a spectrum of properties are varied, such as class size, where each class represents a different outcome, and problem-type.

Separately, each set of questions will propose different views for the subject to take into consideration and will ask for their opinion in light of this new perspective. For example, Q1 could ask a subject to apply a certain legal perspective, whereas Q2 could ask to apply a societal perspective. The actual breakdown of the question set follows: Q1 focuses on applying the Freudian perspectives, where parts a, b and c represent the application of the Id, Ego and Super-Ego, respectively; Q2 aims at assessing the legal implications, where parts a and b represent the normativity and inaction perspectives, respectively. Thus, Q3 assesses the moral views of the subject (for the purpose of developing a social view), where a and b quantify the subject's view on utilitarianism and c simply asks for their view without any applied perspective.

As we mentioned previously, the secondary source of data is a generative method based on the Normal distribution of complete data found in the primary data-set. This data-source will take the form of technical implementation, as outlined in Section 5.1.1. We should also acknowledge that the generation of fake-data based on real-data will result in a less representative distribution, as fake-data will only be a variation on the real-data. However, it will also provide a larger data-set. This is a benefit to our program as larger data sets provide stronger mean values and reduce the impact of anomalous data which would otherwise skew the classification, thus providing a smaller margin of error.

[4] (1) Only you win, (2) Subject and the subject's friends win, (3) You and strangers win, (4) Subject, the subject's friends and strangers win

5.2 Results

5.2.1 Data Collection

The data-collection resulted in fourteen total data-sets where only six are complete sets. Consequently, we utilised the data-generation method to produce six-hundred new data-points.

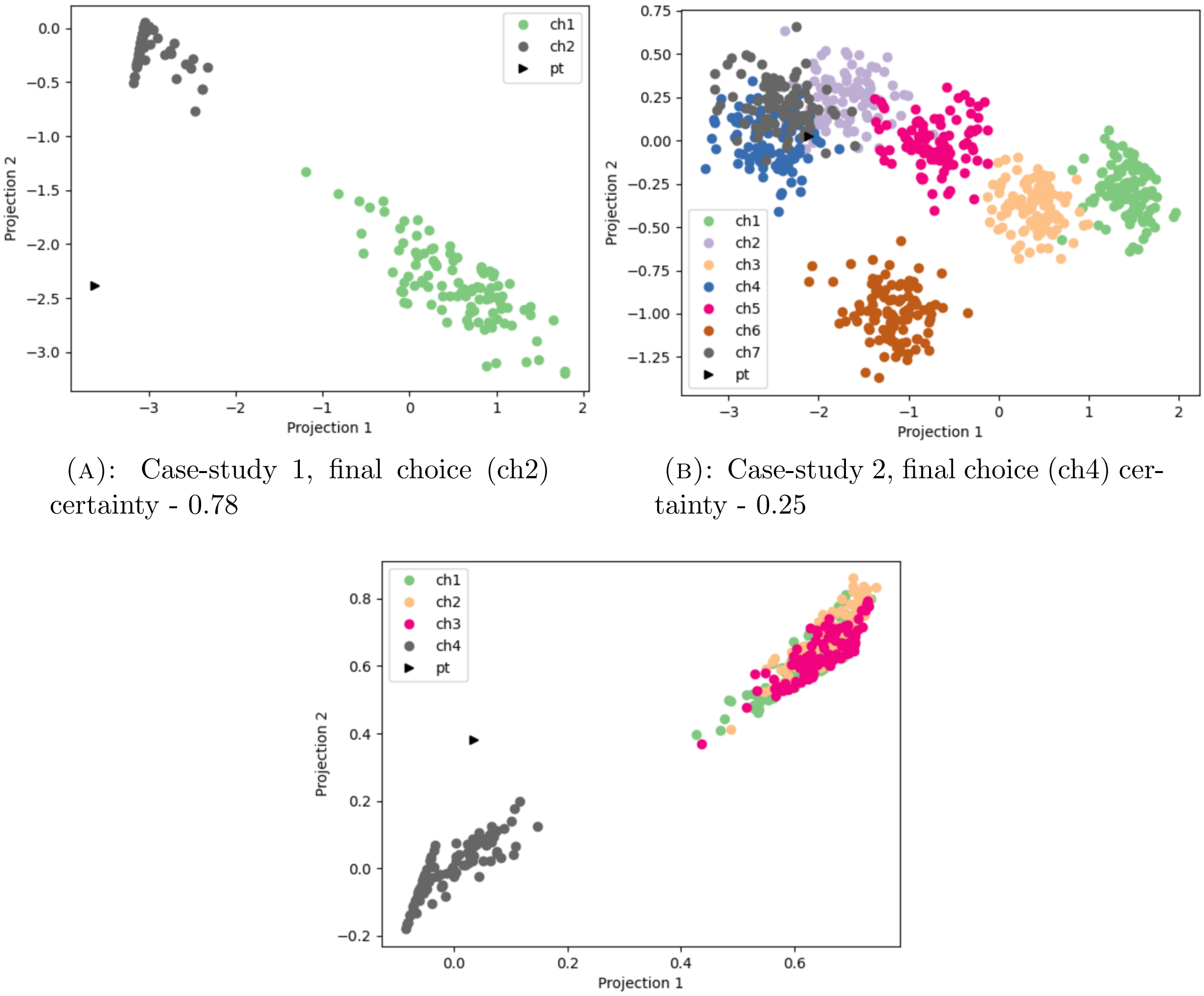
5.2.2 Emulation Results

Using the visualisation implementation mentioned in Chapter 5.1.1, we mapped three 2-Dimensional LR distributions, presented in Figure 5.1. The three sub-figures illustrate the projections attributed to each case-study's data-set, where each class describes the possible choices that exist within the case-study. The point *pt* depicts the position of the *ideal* choice. From Figure 5.1, we can see data-distribution within each class is strong, however actual class distribution is weak. For example in Figure 5.1(c), classes pertaining to choices 2, 3 and 4 all overlap, which would make it difficult for the program to differentiate between them, thus resulting in smaller LWLR certainties. Similarly, in Figure 5.1(b) with classes ch 2, 4 and 7, which is fortified by the final choice having a low certainty of 0.25. We believe this is a result of the data-collection, where participants deemed some choices to be of similar-value and in cases where *many* choices are required the LWLR performs poorly. Evidently, our reason for using LWLR may have been misplaced, where instead of aiming to increase probabilities, we should have focused on increasing class separation. A reasonable alternative ML classification method would have been linear discriminant analysis (LDA); an algorithm that primarily focuses on increasing class separation.

As a result of eight program-runs, we found complete coherency in choices being made in every area of the program. This signifies there is certainty in choices made by the framework.

Finally, Figures 5.2, 5.3 and 5.4 all illustrate the outcomes from case-study 1, 2 and 3, respectively depicted as 2, 7 and 1. With regards to case study 1, the final decision is reasoned to be socially and morally just, however faces more strenuous legal consequences. With regards to case study 2, the final decision is deemed just in all cases. With regards to case study

3, the final choice is legally and socially just, however there is clear moral injustice. Given the context of each case, we believe the reasoning behind each final choices is both reasonable and logical.



(A): Case-study 1, final choice (ch2) certainty - 0.78

(B): Case-study 2, final choice (ch4) certainty - 0.25

(C): Case-study 3, final choice (ch4) certainty - 0.75

FIGURE 5.1: Logistic-Regression projections for each case study: (a) Train/Trolley Problem, (b) Financial Allocation as CFO, (c) Casino Wins; For each case, points are classed by each possible choice 'ch$n$', where n=0,1,..,,N, where N represents the number of possible outcomes, as described in Section 5.1.2; Point *'pt'* represents the *ideal* choice; Projection 1,2 are the LR projections of the data-set, and p, onto the ideal weights of the LR problem

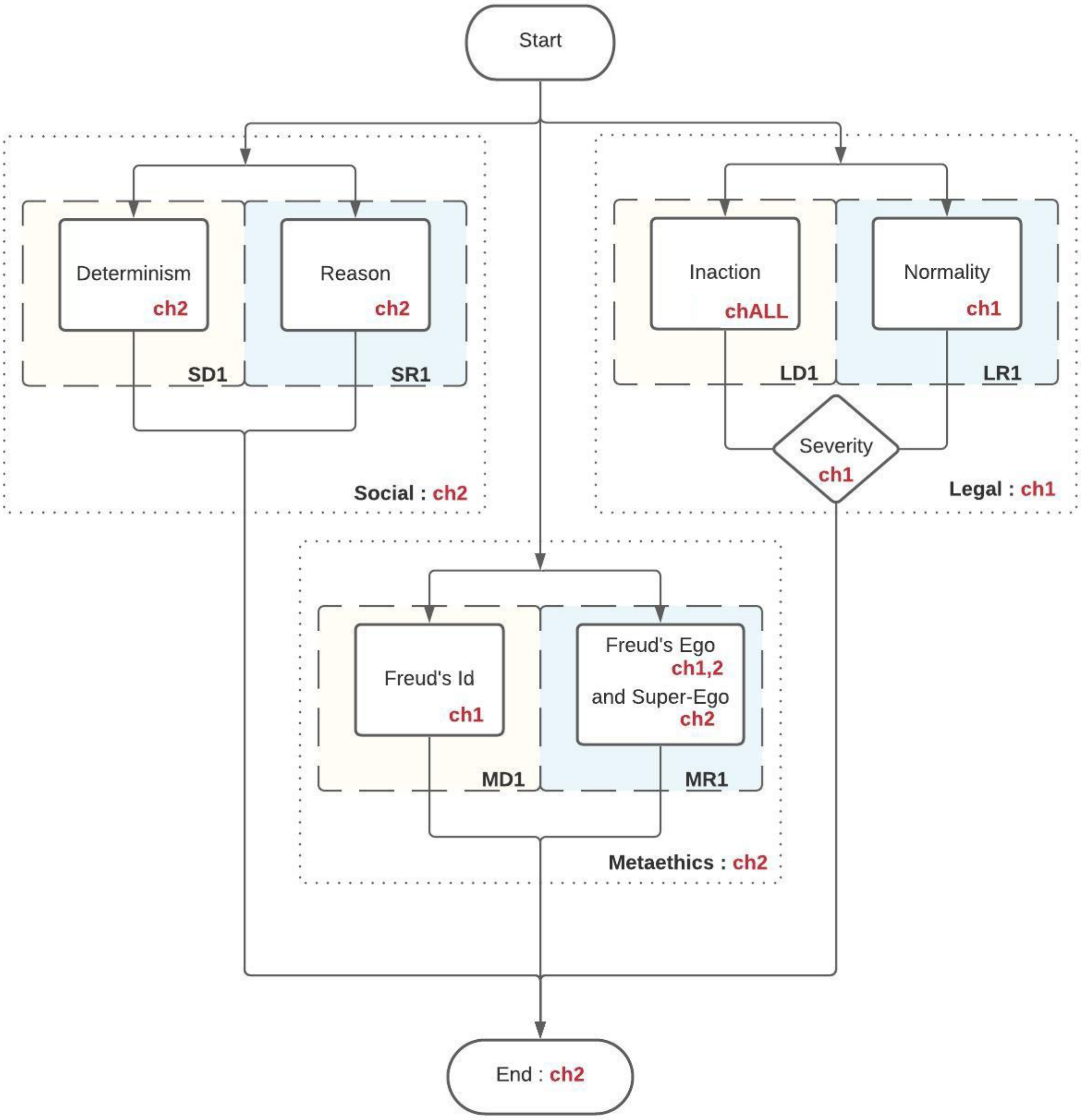


FIGURE 5.2: The choices made in each micro-frame for Case Study 1, where ch2 is to pull the lever

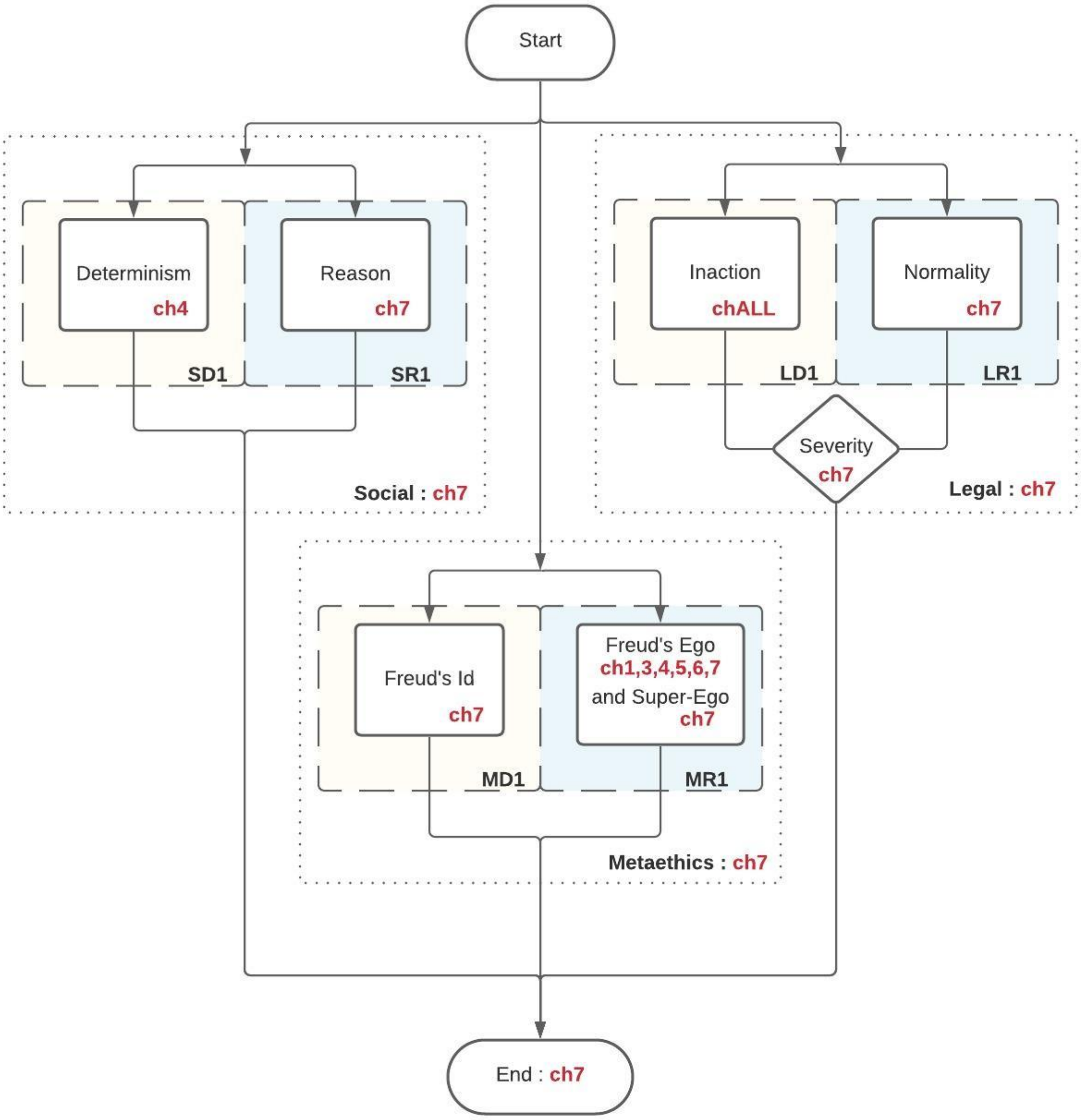


FIGURE 5.3: The choices made in each micro-frame for Case Study 2, where ch7 is to allocate resources to all faculties

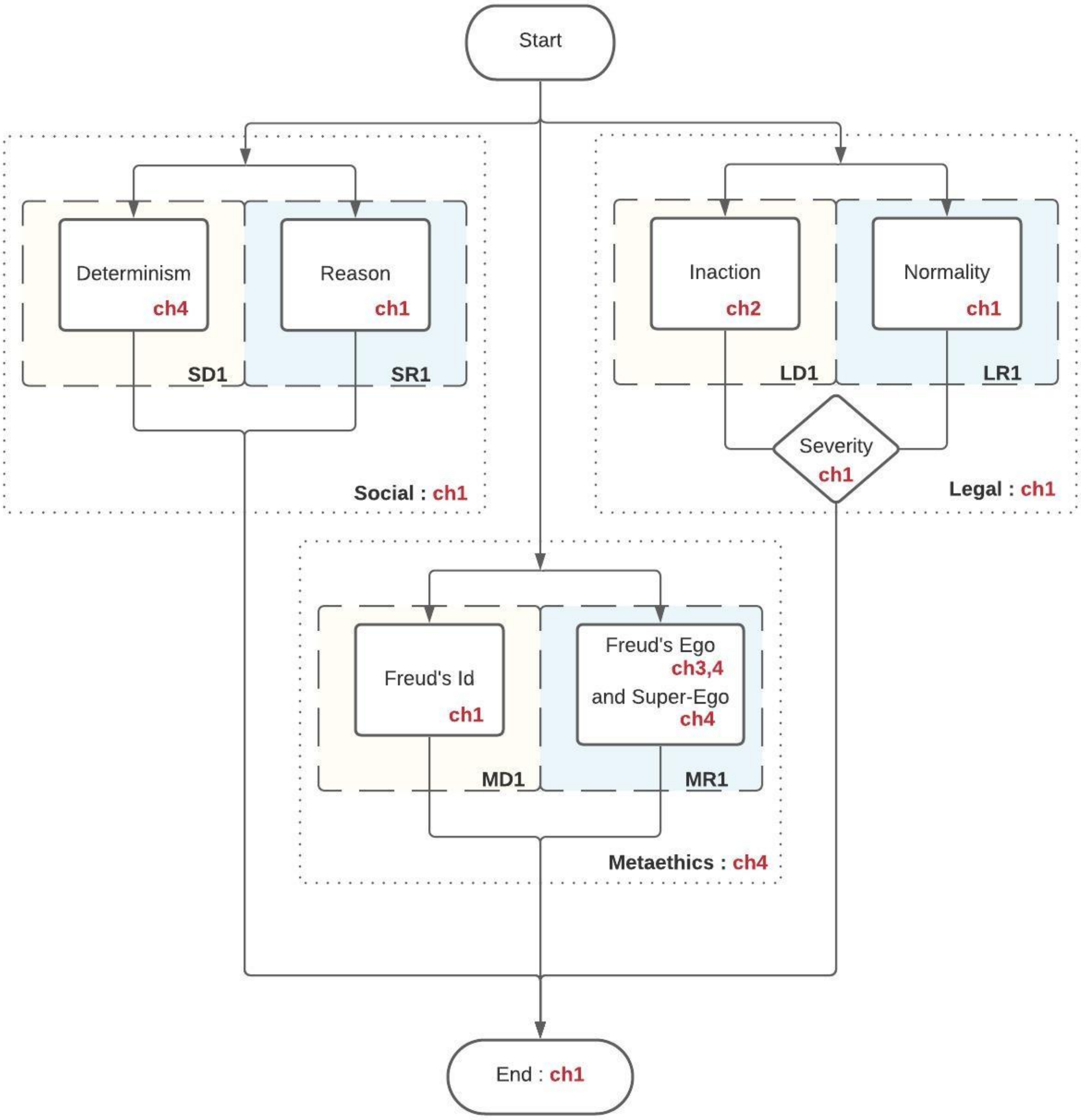


Figure 5.4: The choices made in each micro-frame for Case Study 3, where ch1 is where you are the only winner

Chapter 6

Model Evaluation and Conclusions

6.1 Model Evaluation

6.1.1 Evaluation Introduction, Format and Participation

Solutions to ethical dilemmas can be tricky to evaluate. By definition, dilemmas have no ”correct” (by extension ”incorrect”) solution(s). Consequently, there is no course of action the Author could take to personally, objectively, evaluate the outcomes of the program, hence the model itself. Instead, a professional who has dealt with ethical dilemmas could evaluate the outcomes with less subjectivity. We acknowledge there is still a margin of subjectivity involved, however this would be on a *strictly professional* basis, reducing the likelihood of any personal bias and the impact of subjectivity on the evaluation. Additionally, a review could provide us with professional criticism directed towards our framework, so extending the scope of the evaluation.

Participants willing to evaluate the project would have to meet the following criteria: (1) hold a qualification rated level 7 or above[1] in fields

[1] Further information on qualification levels: https://www.gov.uk/what-different-qualification-levels-mean/list-of-qualification-levels

relevant to the project, (2) maintain a post which involves either the action or understanding of critical decision making, (3) have a comprehensive understanding of philosophical principals tied to ethical dilemma decision making.

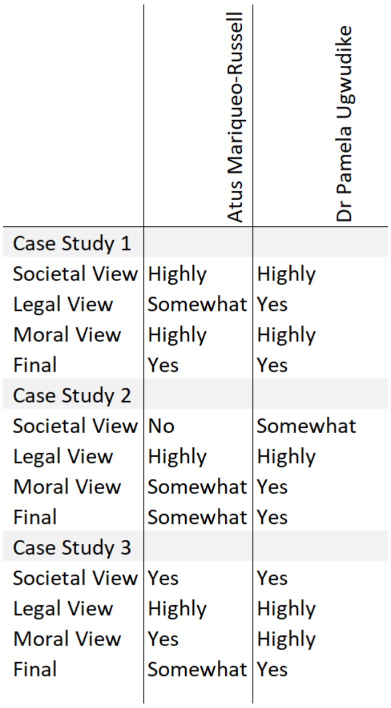
The professional evaluation will be held as an interview-type study, where we will ask participants to evaluate the choices made by the program regarding each study, as well as evaluate the framework design, jeered towards how appropriate the philosophies are for generalised problems.

As a result, we recruited two professionals: **Atus Mariqueo-Russell**, who holds an MPhilStud in Philosophy and is completing a PhD at the University of Southampton in *desire theories of welfare* with reference to catastrophic risk avoidance; and **Dr Pamela Ugwudike**, Associate Professor in Criminology at the University of Southampton, currently involved in several research projects that focus on ethical implementation of AI in the criminal justice system. We should affirm our confidence in their ability to subjectively evaluate the program and framework as both participants meet our evaluation requirements and have shown an impressive level of apprehension and analytical skills during their interviews.

6.1.2 Evaluation

As previously mentioned the evaluations was split into two areas, program evaluation and framework evaluation. Consequently, the evaluation is presented in this order.

The program was evaluated using the outcomes from the three case studies, where each participant was asked: (1) how ”agreeable” (”Not”/”Somewhat”/ ”Yes”/”Highly”) they thought the outcomes are in each micro-frame, and (2) how appropriate (”Not”/”Somewhat”/”Yes”/”Highly”) the final choice is, with open commentary. As illustrated in Figure 6.1 the participants found the answers from case 1 extremely agreeable with no critique, whereas with case 2 some disagreement was voiced. Mr Mariqueo-Russell explained that the choice made for the social-view sub-model was ”counter intuitive” to the perspective applied and a choice of 2/4/10 (Figure 5.3) would be less-so. In addition, Mr Mariqueo-Russell found that although the outcomes for Freud’s Id and Super-Ego ”made sense”, the outcome for Freud’s Ego was less representative, thus deeming the moral-view (metaethical) sub-model only "Somewhat Agreeable". With regards to case study 3, again Mr Mariqueo-Russell was "surprised" over the social-view sub-module outcomes, although in this case he voiced understanding as to why those particular choices were made. On the other hand, Dr Ugwudike found the criticised program models adequate, explaining that though some choices may not be *ideal* they are still representative of the philosophical perspectives presented. The overall program choices were rated a minimum of "Somewhat Agreeable", though both participants sharing feelings of satisfaction over all choices.



	Atus Mariqueo-Russell	Dr Pamela Ugwudike
Case Study 1		
Societal View	Highly	Highly
Legal View	Somewhat	Yes
Moral View	Highly	Highly
Final	Yes	Yes
Case Study 2		
Societal View	No	Somewhat
Legal View	Highly	Highly
Moral View	Somewhat	Yes
Final	Somewhat	Yes
Case Study 3		
Societal View	Yes	Yes
Legal View	Highly	Highly
Moral View	Yes	Highly
Final	Somewhat	Yes

FIGURE 6.1: Participants' answers to how "agreeable" they believe the outcomes of each micro-frame and whole framework are for each case study

Following this, participants were asked to evaluate the framework design which consisted of: (1) rating how "appropriate" the implemented philosophies are regarding ethical dilemmas ("Not"/"Somewhat"/"Yes"/"Highly"), (2) stating any potential improvements or changes to the philosophies or technical implementation, (3) discussing whether the *current* framework could be applicable to any area of life (personal/business/legal/etc.), and (4) discussing whether an *improved* framework could be applicable to any area of life.

Dr Ugwudike found the philosophies presented appropriate for the application but showed uncertainty in using Freud's psychoanalytic Personality Theory (Id/Ego/Super-Ego). Further, she believes that more philosophical perspectives could be useful, though only if they relate to the social and legal sub-frames. Despite Dr Ugwudike voicing her admiration over the framework, she stated that she would only use the current framework in minor personal matters, where no emotional, physical or other harm could be caused. Though she stated that even if the framework was improved upon it should still not have a large role in society.

Similarly, Mr Mariqueo-Russell found little use in the metaethical reasoning applied within the framework. Instead, he suggested the use of Kantian and Virtue ethics alongside the existing Utilitarian ethics (this grouping in summed up in Repetti), stressing that definitions to be taken from the *Stanford Encyclopedia of Philosophy*[2]. In addition, he advanced his satisfaction with the legal micro-frame stating it "made perfect sense".

Contrasting with Dr Ugwudike's comments on the framework's use, Mr Mariqueo-Russell found that such a tool should not be used to *make decisions*, even if improved upon. However, he stated that an improved tool could be used to *make suggestions*, emphasising that some level of human intervention should exist to monitor suggested choices, using the outcomes from case study 2 to validate his point.

6.2 Conclusion

We began this project with the intention introducing philosophically-advanced DM technology. In doing so, we spent several months researching and designing a novel framework, another several months reviewing and improving it, followed by the programmatical implementation of said framework and a thorough professional evaluation of the implementation. As a result, we successfully developed a framework that provides a higher state of philosophical-thought than any current framework. This is echoed by the praise received during the professional evaluation, focused towards the technical implementation of determinism, under the social influence micro-frame, and the overall implementation of legal-influence.

[2] https://plato.stanford.edu/index.html

In spite of the framework's accomplishments, it became evident that several aspects of the framework need to advance before any suitable application can be drawn out. Namely, discarding the meta-ethical micro-frame while expanding on the philosophies used within the social-influence micro-frame. And though the program (and the adjacent outcomes) seemed to represent the framework *appropriately*, there were still apparent flaws in the data-collection, as outlined in Chapter 5.2.1.

Consequently from the evaluation, more research into the best (more suited) philosophical attributes and an expansion on implemented attributes is necessary before *serious* use of the framework can be considered, as echoed by the evaluators. However, from the evaluations, it is clear that programming for philosophies is a possibility and there exists a potential to develop similar tools for similar problems.

In reflection to the comments made in Chapter 1, exemplifying a more balanced relationship between AI and philosophy by ”setting ourselves the task...to facilitate periods in our lives when we are faced with ethical dilemmas” (Chapter 1.1) may have complicated our aims. A softer approach may have been better suited to a *third year individual research project*. Though, given our framework's accomplishments in an area where many AI-only projects have floundered, solving ethical dilemmas, we can be certain in stating that a ”more balanced relationship” is possible and potentially more beneficial. It will be interesting to see how the extortion of DM philosophies evolves over the coming years. We are particularly interested to see what philosophies will become prominent points of debate within this field and what type of equilibrium(s) could persist between AI and philosophy. It would be fascinating to do

As a closing remark, it will be exciting to see how society reacts to the inevitable commercialisation of similarly styled applications. As discussed in Christensen (2013), it is often difficult to imagine what consumers really want and how they will react, and though both evaluators felt hesitant towards the application, neither had experienced a similar tool. Perhaps these perceptions will change with habits, perhaps not, nevertheless it will be interesting to follow the discourse.

Bibliography

B Al-Ali, MMA Ul Haq, AA Al-Rebh, M Al-Qahtani, and T Al-Qurashi. Decision making in ethical dilemma. In *2012 Proceedings of PICMET '12: Technology Management for Emerging Technologies*, pages 589–599, 2012.

Jayanath Ananda and Gamini Herath. A critical review of multi-criteria decision making methods with special reference to forest management and planning. *Ecological economics*, 68(10):2535–2548, 2009.

E Aronson, TD Wilson, and AM Akert. *Social Philosophy (5th Edition)*. Pertinence Hall, Upper Saddle River, NJ, United States, 2005.

Jerome R Busemeyer and In Jae Myung. An adaptive approach to human decision making: Learning theory, decision theory, and human performance. *Journal of Experimental Psychology: General*, 121(2):177, 1992.

Clayton M Christensen. *The innovator's dilemma: when new technologies cause great firms to fail*. Harvard Business Review Press, 2013.

Adiel Teixeira de Almeida, Marcelo Hazin Alencar, Thalles Vitelli Garcez, and Rodrigo José Pires Ferreira. A systematic literature review of multicriteria and multi-objective models applied in risk management. *IMA Journal of Management Mathematics*, 28(2):153–184, 2017.

L Dennis, M Fisher, M Slavkovik, and MP Webster. Ethical choice in unforeseen circumstances. In *TAROS*, 2013.

M Deutsche and HB Gerard. A study of normative and informational social influences upon individual judgment. 51:629–636, 1955.

Owen M. Fiss. Reason in all its splendor. 1991.

Sigmund Freud. *Das Ich und das Es (The Ego and the Id)*. Internationaler Psycho- analytischer Verlag (Vienna), W. W. Norton Company, Vienna, Austria, 1923.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

Maite Frutos-Pascual and Begoñya García Zapirain. Review of the use of ai techniques in serious games: Decision making and machine learning. *IEEE Transactions on Computational Intelligence and AI in Games*, 9 (2):133–152, 2015.

Alice Gaudine and Linda Thorne. Emotion and ethical decision-making in organizations. *Journal of Business Ethics*, 31:175–187, 06 2001.

NJ Goodall. Ethical decision making during automated vehicle crashes. *Transportation Research Record*, 2424(1):58–65, 2014.

Paul Goodwin. Common sense and hard decision analysis: why might they conflict? *Management Decision - MANAGE DECISION*, 47:427–440, 04 2009.

David J Hand and William E Henley. Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 160(3):523–541, 1997.

Paul R Harper. A review and comparison of classification algorithms for medical decision making. *Health Policy*, 71(3):315–331, 2005.

Tim Hill, Leorey Marquez, Marcus O'Connor, and William Remus. Artificial neural network models for forecasting and decision making. *International journal of forecasting*, 10(1):5–15, 1994.

M. A. Islam and S. I. Rashid. Algorithm for ethical decision making at times of accidents for autonomous vehicles. In *2018 4th International Conference on Electrical Engineering and Information Communication Technology (iCEEiCT)*, pages 438–442, 2018.

Daniel Jato-Espino, Elena Castillo-Lopez, Jorge Rodriguez-Hernandez, and Juan Carlos Canteras-Jordana. A review of application of multi-criteria decision making methods in construction. *Automation in Construction*, 45:151–162, 2014.

J.C.Flugel. *Mans, Morals and Society.* Pelican Books, 219 N Cortez St, Prescott, AZ 86301, United States, 1945.

Hasan Kartal, Asil Oztekin, Angappa Gunasekaran, and Ferhan Cebi. An integrated decision analytic framework of machine learning with multi-criteria decision making for multi-attribute inventory classification. *Computers Industrial Engineering*, 101:599 – 613, 2016. ISSN 0360-8352.

Melody Y Kiang. A comparative assessment of classification methods. *Decision support systems*, 35(4):441–454, 2003.

Gary Klein. *Sources of Power, How people make decisions.* The MIT Press, Cambridge, Massachusetts, 02142, 1999.

T. S. Lee, S. Ghosh, and A. Nerode. A mathematical framework for asynchronous, distributed, decision-making systems with semi-autonomous entities: algorithm synthesis, simulation, and evaluation. In *Proceedings. Fourth International Symposium on Autonomous Decentralized Systems. - Integration of Heterogeneous Systems -*, pages 206–212, 1999.

Filip Lievens, Helga Peeters, and Eveline Schollaert. Situational judgment tests: A review of recent research. *Personnel Review*, 2008.

M Miner and A Petocz. Moral theory in ethical decision making: Problems, clarifications and recommendations from a psychological perspective. *Journal of Business Ethics*, 42:11–25, 01 2003.

GE Moore. *Ethics.* Williams Norgate, London, UK, 1912.

JM Nolan, P Wesley-Schultz, RB Cialdini, NJ Goldstein, and V Griskevicius. Normative social influence is underdetected. 34:913–923, 05 2008.

R Noothigattu, S Gaikwad, E Awad, S Dsouza, I Rahwan, P Ravikumar, and A Proccia. A voting-based system for ethical decision making. pages 1–7, 09 2017.

Robert A Prentice and Jonathan J Koehler. A normality bias in legal decision making. *Cornell L. Rev.*, 88:583, 2002.

Joseph Raz, Simon Blackburn, James Dreier, David Enoch, Stephen Finlay, Anti Kauppinen, James Lenman, Mark Schroeder, Jussi Suikkanen, and Julie Tannenbaum. *Oxford Studies in Metaethics.* Oxford University Press, Great Clarendon Street, Oxford OX26DP, United Kingdom, 2012.

Riccardo C. Repetti. Utilitarianism kant and virtue theory.

A Sen and B Williams. *Utilitarianism and beyond*. Cambridge University Press, Cambridge, UK, 1982.

Evangelos Triantaphyllou. Multi-criteria decision making methods. In *Multi-criteria decision making methods: A comparative study*, pages 5–21. Springer, 2000.

Jiang-Jiang Wang, You-Yin Jing, Chun-Fa Zhang, and Jun-Hong Zhao. Review on multi-criteria decision analysis aid in sustainable energy decision-making. *Renewable and sustainable energy reviews*, 13(9):2263–2278, 2009.

Edgar Wilson. *The Mental as Physical*. Routledge Kegan Paul Ltd., 29 Store Street, London, WC1E7DD, 1979.

Constantin Zopounidis and Michael Doumpos. Multi-criteria decision aid in financial decision making: methodologies and literature review. *Journal of Multi-Criteria Decision Analysis*, 11(4-5):167–186, 2002.