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: Unknown

Decision-Making Assistant for Generalised Ethical Dilemmas

by Adrian Azzarelli

December 7, 2020

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

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

Within this progress-report I propose a philosophical framework and incorporate artificial intelligence (AI) techniques in determining a solution to an ethical dilemma, such as the *Trolley Problem* and the *Trapped Mining Crew*. In order to accomplish this I partitioned my framework into the three branches of philosophy that are key to any DM problem, Metaethics, Social influence and Legal influence. Within each category of my framework there exists several inter-connected processes that employ either multi-criteria decision-making or machine learning algorithms, with the purpose of making a decision which coheres with the principles of each category's philosophy. Thereafter, a final process works to make a reasonable choice between the three decisions suggested by the categories, which we take as our final decision.

The research explored within this report focuses on two key aspects of decision making, philosophical reasoning and technical application. Philosophical reasoning is a mature subject, thus much of the philosophy I have implemented dates back to Freudian times. Comparatively technical application is still in fruition, hence the lack of existing frameworks that surround ethical-dilemma decision making (EDDM), though there still exists a large body of research surrounding decision making as a whole. Together, the subjects have enabled me to apply the appropriate technical processes with regards to specific philosophies. The hope is that I will be able to implement the framework as a program and perform objective tests to determine the effectiveness of my framework.

Contents

Chapter 1

Introduction

1.1 Introduction

The goal of this project is to propose a philosophical framework that incorporates artificial intelligence techniques, with the aim of assisting a (human) client in making a decision in an ethical dilemma situation.

The two prevalent topics of research explored within this report consist of, the subject of philosophy investigated in Chapter 2.1, which outlines the philosophical properties that my decision-making (DM) tool will incorporate, and technological modelling explored in Chapter 2.2, which underpins the DM techniques I have applied to my decision-making model.

Much of the (philosophical) literature around decision making focuses on several 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"[1], (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 can not be avoided or when other ethics take precedent. I have labelled these (Metaethics, Sociology and Legality) as the **pillars** to my DM model, where (2) holds the most weight, given a study from Nolan et al. (2008) supports *descriptive*

[1]In the context of the impact of *Moral Theory* on decision making models

normative beliefs[2] as the most influential behaviour[3] and (1) would hold the second rank, as I have reasoned with Dennis et al. (2013) behind the lowered importance of legal impact.

In addition to layering my model with philosophical reasoning, I 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 artificial intelligence (AI)), 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 machine (SVM) algorithms useful for multi-criteria decision making tasks in cars.

Aside from structuring my framework, I need to hurdle two obstacles in order to develop an exemplar program. The first focuses on the translation of philosophical topics and contextual information[7] into language that can be understood by a program. The second focuses on the reduction of bias when evaluating testing data[8]. Even though these topics are not explored in this report, they are still worth mentioning as they will be my primary focus in the forthcoming months, as outlined in my Gantt Char, Appendix A.1, and in Chapter 4.2.

Otherwise, the structure of this report is as follows. Chapter 2 will focus on the body of research. More precisely Chapter 2.1 will focus on the philosophical reasoning implemented in my DM model, whilst in Chapter 2.2, I will apply DM and learning techniques in order to develop an exemplar program. In Chapter 3 I propose the finalised philosophical framework and

[2] *Descriptive normative beliefs* refer to what an individual think the social norm is

[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-cost 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

[7] *Contextual information* refers to information that can not be calculated using a deterministic approach, which allows a situation to be given context. Therefore acts as the source of information that need to be quantified to make a contextual-determination

[8] *Testing data* refers to the data used to test a machine learning model that has been trained using *training data*, by extension test data $\neq$ training data

in and Chapter 4 I aim to emphasise the work done and strictly layout the plan going forward.

1.2 Further Insight

The goal is to assist a client in choosing the solution from a list of reasonable solution, that will yield the *best*[9] result. More specifically the aim is to develop a philosophical framework with applied DM and ML techniques in order to assist a cleints decision making process in an ethical dilemma, and further illustrate it's workings by use of program. The plan outlined by the Gantt Chart follows the development of the philosophical model and a corresponding program. My research methodology can be found in Chapter 4.2.

To briefly comment on the impact of COVID-19: there is little to fret about the construction of my 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.

[9]The use of *best* suggests a layer of subjectivity to our model - this is a common flaw that many decision-making robots succumb to. I must recognise that it is an inevitable flaw within my model as my choice to (or not to) include certain philosophies means my robot will be constrained to *my* subjectivity, though I aim to generalise my solution so that the implication of my subjectivity in this manner is minimised

Chapter 2

Body of Research

2.1 Subject of Philosophy

2.1.1 Introduction to Philosophy

In order to introduce this section I will take a quote from Joseph Raz, Raz et al. (2012), which outlines the philosophy on which I have based my 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 *normativite*[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, my 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 I have incorporated more unique ideologies that aim to represent their respective branches of philosophy in various ways. In addition, I 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 I call **core philosophies** (also **cores**).

The core philosophies are based on ideologies taken from Wilson (1979). The main core is *compatabilism*, interpreted as the *compatibilist*'s stance on DM[3], which I have implemented by dividing the pillars into two groups I call *micro-frameworks*. The two groups correspond to: (1)Deterministic calculation which I have defined as deterministic reason[4] and (2)Reasoned calculation which I have defined as libertarian Reason[5].

Another core considers *the Reasonable Man*'s[6] approach, which I 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". My 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[7]; recognising when to use it and mimicking the process. As explained by the *recognition primed decision model*[8], discussed in Klein (1999), intuition can lead to solutions that don't 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 doesn't need processes such as calculating legal implication. I acknowledge that there are

[3] *Compatabilism* 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] By which I mean the causation and occurrence of an outcome are inevitable

[5] By which I mean the choice made was made from free will

[6] *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.

[7] 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

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

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 my previous example.

2.1.2 Meta-ethics (Moral Influence)

By definition, Meta-ethics is the study of moral thought, though within my 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 made a 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.

2.1.3 Impact of Human Law

Dennis et al. (2013) reasons, "if necessary, disregard legal restrictions for ethical reasons". This is the stance I 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/advice. Instead it aims to sway a decision dependant on the severity of legal implications.

It is important to understand allocation of blame when regarding legalities as this can aid 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". I 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 exists within the Reason frame, and *inaction* (i.e. has the neglect of other action led you to worse results), which exists 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[9], 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.

2.1.4 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, it's importance to our model.

In order for our framework to conform to a sociological idea we need to outline prospective social beliefs that our framework may take. The most obvious social belief is that of *utilitarianism*[10], 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 *hedonistic utilitarianism* which incorporates pleasure and pain, in place of good and bad - defined for both social and individual utility.

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

[10] *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

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[11], 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*[12];neglecting preference autonomy from ones decision process removes the likelihood one will endanger the "greater good" of people due to a nefarious desire. This is the ideology I 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[13].

Unfortunately, emotion can not 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 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

[11] The mental states can take the form of philosophical worth, scientific worth, artistic appreciation, etc.

[12] *Preference autonomy* is the decision what is good/bad for an individual based on *their* desires

[13] 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 I 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 can not be considered

two micro-frameworks within the pillar, thus a socially focused intuitive choice is made.

2.2 Technological Research

2.2.1 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 my research enters the field, proposing a solution with regards to generalised ethical dilemmas.

Though, in order to propose any technical solution we need to define our problem in a technical manner. Plainly, we aim to solve a multi-criteria decision-making (MCDM) problem, where the criteria we acknowledge is based on the pillars which we previously outlined in Chapter 2.1. Studies such as Jato-Espino et al. (2014) and Wang et al. (2009), who review MCDMs in high-pressure[14] environments, highlight the reliability that can be placed on wide-spread 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 me to outline which combination of MCDMs are appropriate for me 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, I 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. I have no desire to incorporate methods for the purpose of precision; dilemmas are dilemmas because there is not 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

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

regression (LR) is among the most popular and highest-ranking method for classification problems. Consequently, I seek to incorporate LR within my framework.

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

Within this section, I 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 me to improve the certainty of decisions. The final application will be applying calculated scores from each pillar ad 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.

2.2.2 Decision Making and Machine Learning Methods Applied

The layout of my model incorporates a total of six micro-frameworks, which pertain to the three pillars outlined in Chapter 2.1.1 (i.e. two frameworks for each pillar). Each micro-framework will incorporate its own techniques. Recall the pairs of micro-frames, one takes the Deterministic approach and the other which incorporates Reason. In this section I aim to establish the technical application of the frameworks with reference to Chapter 2.1, where I proposed the philosophical ideologies I wish to incorporate. Note that This section will not incorporate any reasoning for technical implementations, this will come in Chapter 3.

The first pillar I would like to consider is the meta-ethical pillar. In Chapter 2.1.2, I interpreted meta-ethics in terms of Freud's Id, Ego and Super-Ego, where the Id exists within the Deterministic frame (I have called MD1) and the Ego and Super-Ego exist within the Reason frame (I have called MR1). Given my 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 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 $C_{X,id}$[15]. This allows me to implement intuition without having to program for it[16]. 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[17]. Here, the client-agent chooses the outcome that is most emotionally poignant. The choice made in MR1a is denoted by $C_{X,ego}$. Similar to MD1, MR1b incorporates 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[18]. This functionality is visually represented in Figure 2.1

The second pillar I will look at is the Legal pillar. In Chapter 2.1.3, I 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 I need to determine the norm choice and to accomplish this I developed a ranking system that weighs choices dependant on external data[19]; 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, I 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 $D =$

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

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

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

[18]This is the intervention of intuition given uncertainty

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

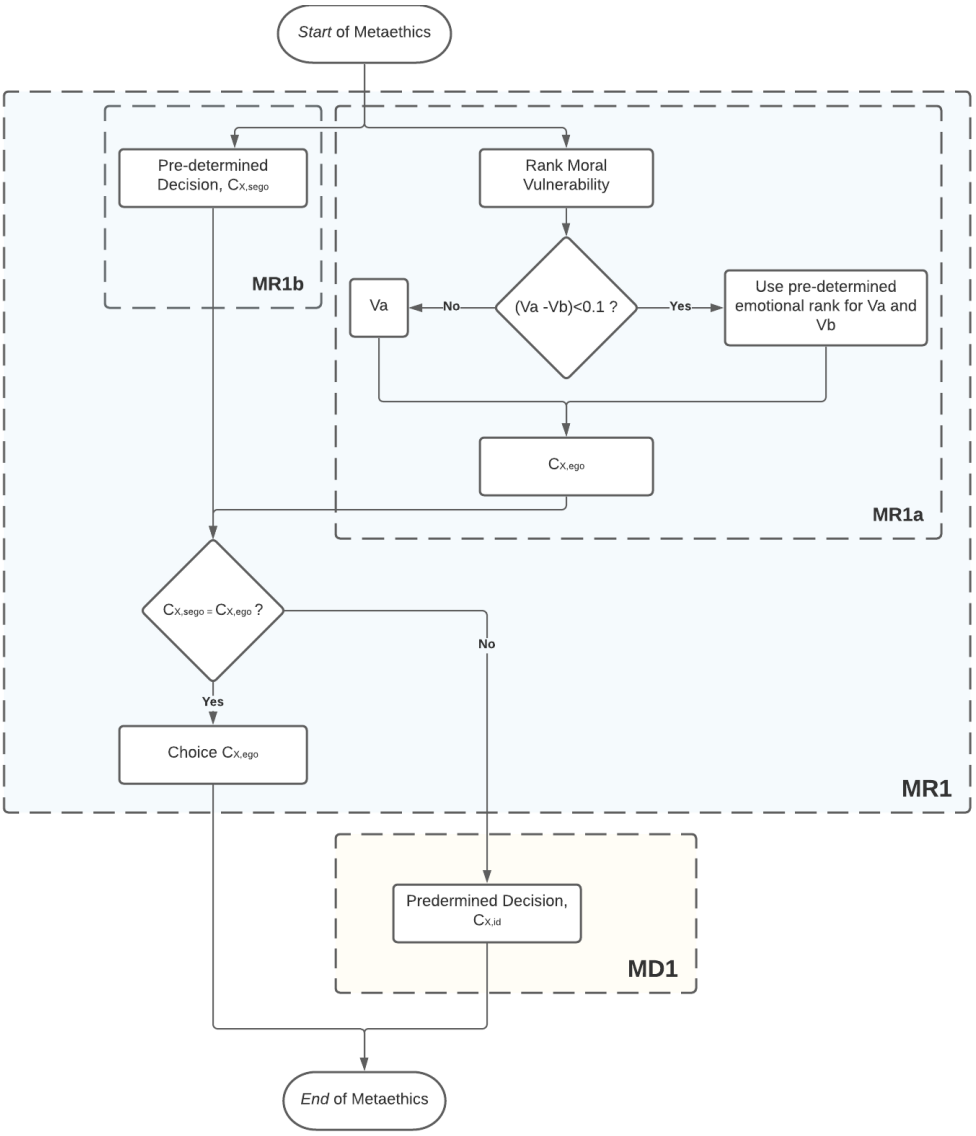

FIGURE 2.1: Decision Model of the Metaethics Pillar

$\{D_{C1}, D_{C2}, \ldots, D_{CN}\}$, for $n = 1, 2, \ldots, N$. Furthering this, a set of weights, $W = \{W_{C1}, W_{C2}, \ldots, W_{CN}\}$, are calculated for the respective choices using Equation 2.1, 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.

$$W_{Cn} = 2D_{Cn} - \sum_{i=1}^{N} D_{Ci} \tag{2.1}$$

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 2.2 illustrates the mechanisms of the legal micro-frameworks.

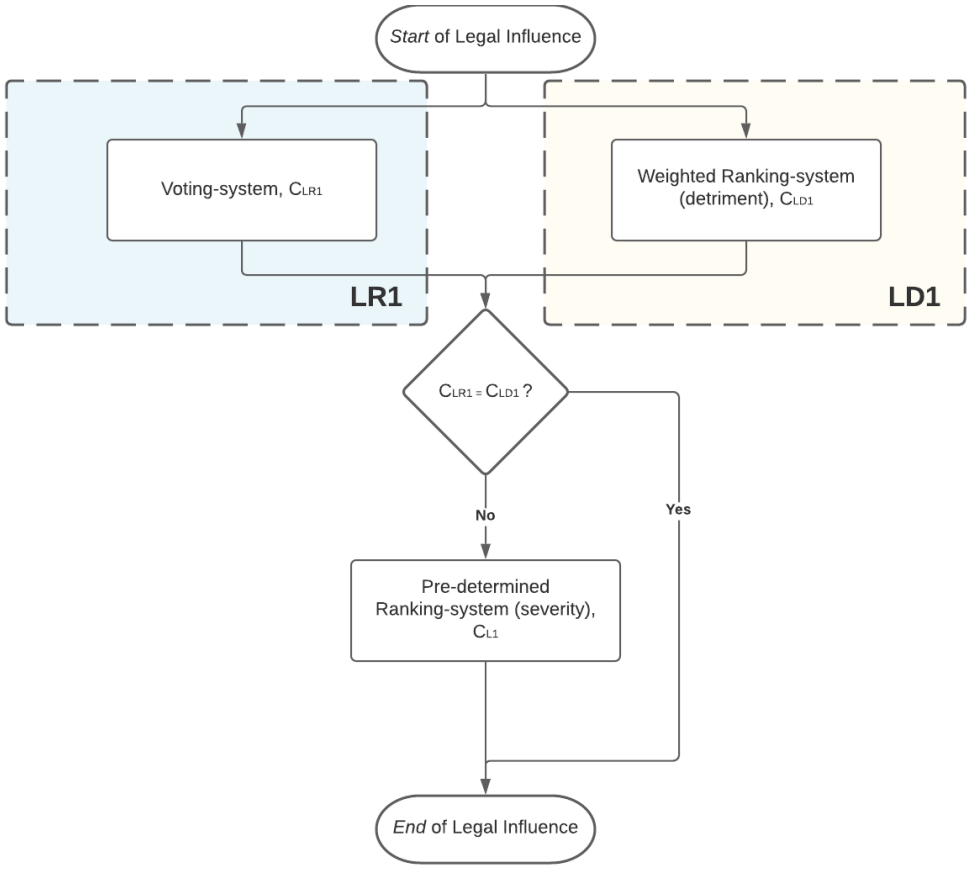

FIGURE 2.2: Decision Model of the Legal Pillar

The final pillar I need to delineate as a frame is the Social pillar. In Chapter 2.1.4, I 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 I have imposed the LR method for classification. More accurately said, I have incorporated a learned-weight logistic-regression algorithm (LWLR), defined in Chapter 4 of Friedman et al. (2001).

Consider 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 $\begin{pmatrix} 1 & 1 \end{pmatrix}$.

The problem is defined as

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

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 my LWLR implementation: Logistic-regression is a probability-dependant classification algorithm where the probability of a point x pertaining to a class, k, is found using the softmax regression function described by Equation 2.2, where $s_k(x)$ is the score given to the point x using learned weights for class k, ω_k, hence $s_i(x) = \omega_k^T x$. Furthering this, LWLR focuses on presenting certainties, this is mathematically represented in Equation 2.3, which equates probabilities in order to calculate the certainty of pertaining to class k. Which is further simplified in Equation 2.4. Using $\omega_k^T x - \omega_K^T x$, I can derive objective values of certainty which will allow me to classify point x, thus classify the point $X = \begin{pmatrix} 1 & 1 \end{pmatrix}$. Equations 2.2 and 2.3 were taken from Equation 4.17 in Friedman et al. (2001).

$$P(k = K | x = X) = \frac{exp(s_k(x))}{\sum_i exp(s_i(x))} \tag{2.2}$$

$$log\Big(\frac{P(k = k | x = X)}{P(k = K | x = X)}\Big) = \beta_{10} + \beta_k^T x \tag{2.3}$$

$$\ln \frac{\frac{e^{s_k(x)}}{\sum_i e^{s_i(x)}}}{\frac{e^{s_K(x)}}{\sum_i e^{s_i(x)}}} = \ln \frac{e^{s_k(x)}}{e^{s_K(x)}} = s_k(x) - s_K(x) = \omega_k^T x - \omega_K^T x \tag{2.4}$$

Additionally, I need to incorporate social emotion, as I mentioned at the end of Chapter 2.1.4. This will be in the form of a transformation matrix, T, which will translate the means of class distributions, μ_n, away from the

mean of means[20], μ, using Equation 2.5, along a direction-vector formed between μ and μ_n. T will translate μ_n by factor E. This will allow for larger class separation, resulting in more certain classification.

$$\mu_{n,new} = T\mu_{n,old} \tag{2.5}$$

I 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 EDDM, where by it is already difficult to decisively classify potential outcomes as humans, therefore classifying programatically would be equally if not more difficult. Increasing separation between class means will allow us to more decisively classify X. The final choice made will be labeled C_{SD1}. Figure 2.3 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, like with the legal frame.

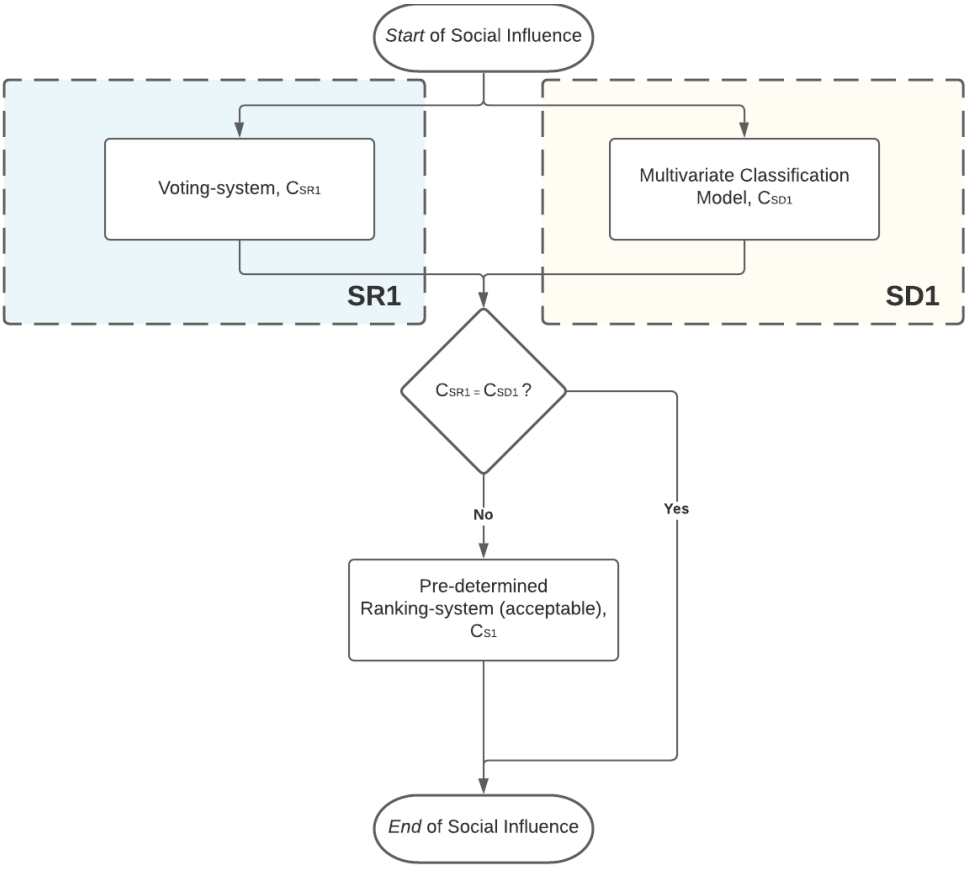

FIGURE 2.3: Decision Model of the Social Pillar

[20]This is a point that represents the mean of the class-means

Chapter 3

Proposed Model and Justification

3.1 Proposal

The philosophical framework being proposed is illustrated in Figure 3.1. The final function described by $f(C_M, C_S, C_L)$ represents Equation 3.1, 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 don't 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.

$$\Phi(C_M, C_S, C_l) = \alpha C_M + \beta C_S + \gamma C_L \tag{3.1}$$

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 2.2.2.

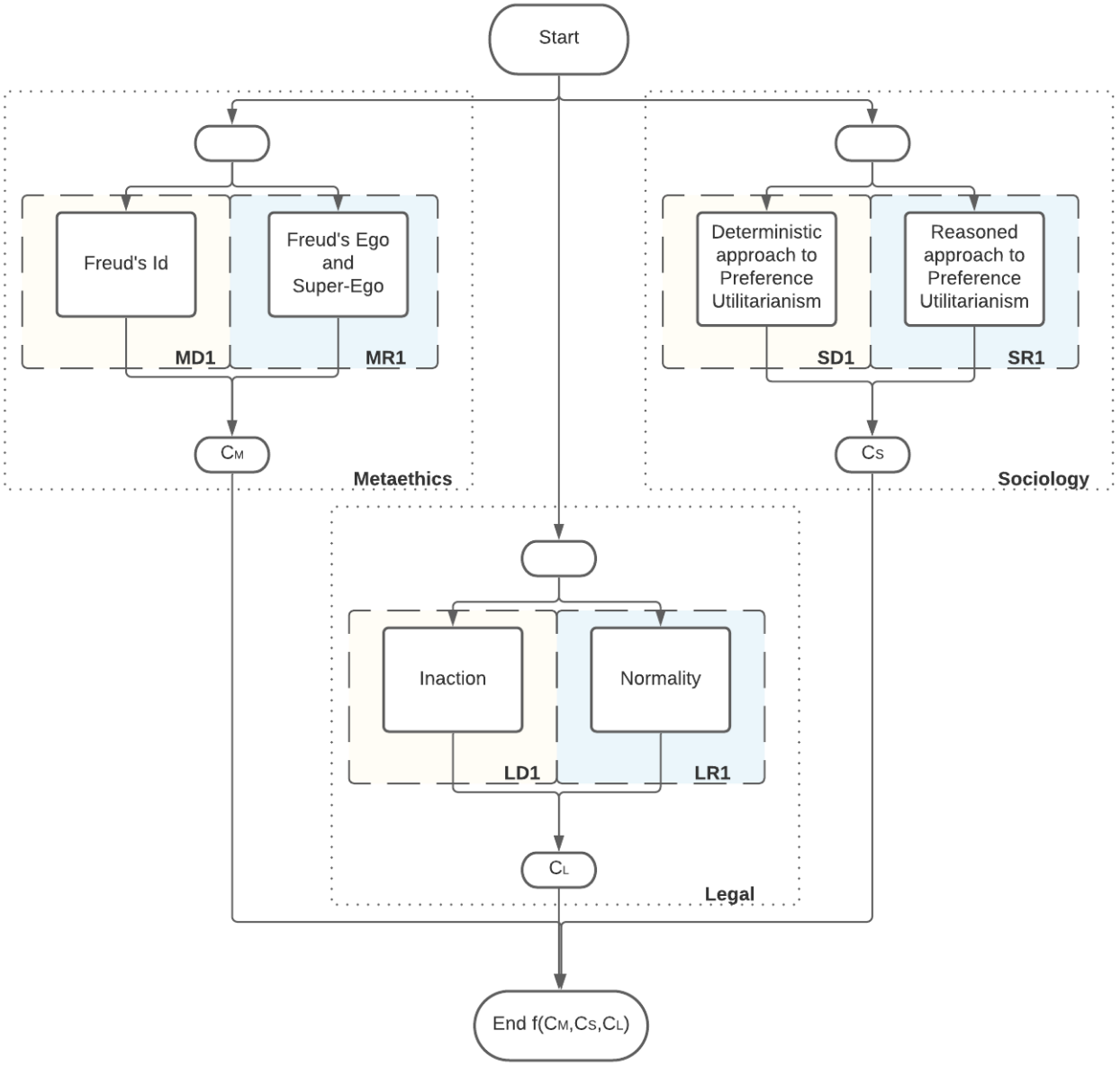

FIGURE 3.1: Decision Model of the Social Pillar

3.2 Technical Justification

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

Within the meta-ethical pillar, I made the case that learning techniques are unwise given my 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, I made the decision

to avoid using complex tools such as *personality tests* to map a clients morals as I thought there were more efficient methods at my disposal. For MD1 and MR1b, the implementation of instantaneous choice allowed me 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 me to implement o 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 wasn't 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. I 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 I struggled to conceive an approach that utilises a choice between final outcomes, given this concept coincides with free-will not determinism. To work around this, I implemented a classification model where the point I aim to classify is predetermined. Hence, no choice is made it is simply "classified"[2]. Otherwise, I implemented 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 2.2.1. Whereas the idea to use LWLR instead of a simple LR implementation came about from tests I had personally executed

[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

prior to conceptualisation project, though this choice is echoed in Friedman et al. (2001), Chapter 4.

Chapter 4

Account of Work

4.1 Current Account

Over this last term I have devoted my time to in depth research in the domains of philosophical ethics (with a hard focus on reasoning) and technological application of decision making models (including both ML and deterministic algorithms). The prior being a subject I had very little understanding of once I started this project; the latter being a subject I had deep understanding of.

My initial methodologies focused on acquiring enough philosophical understanding in order to develop the primary (only philosophical) framework. Using advice I was given from several peers (philosophy students), I began with books, different to research papers, books provided me with established, well known, concepts that I could build my own thoughts around. Which clearly influenced me given a majority of my philosophical implementations are based on books, and then backed by research. Regarding the technical research, I had to first familiarise myself with current technological implementations. This allowed me to build a solid understanding of what algorithms I could implement. My final methodology within the research phase of my project counted on appropriately assigning technical implementation to the philosophical frame.

In addition to compiling research I have been able to develop a framework which I believe to be a reliable decision making assistant. Though I initially set out to fully automate the decision framework, I discovered the difficulty

in completing such a task in the time-frame allocated for this project (one academic year, October-April), so I divulged and developed a framework for assisted decision making. Having now reviewed it within this report, I am confident of its functionality.

4.2 Future Plan

Having now developed a framework, I aim to build a program that illustrates it's function. In order for this to be possible I will need to outline a fair input-method which will allow me to train my model's social influence (and remove bias) as well as apply external data to the AHP and WSM structures that exist within the other micro-frameworks. My methodology for data-collection focuses on developing a set of case studies that the a random group of people can evaluate (i.e. rank, provide values for, etc.), and further randomly pick one individual to act as my client-agent, i.e. to assist the decision regarding the Meta-ethical pillar of my framework. To ensure the program works, and works effectively, I will need to design a reasonable testing method and apply it to the program. A testing method is going to be difficult to evaluate as with ethical dilemmas solutions are not so binary (right-wrong). Currently the aim is to employ professionals within the domain of philosophy and psychology to assess the decision (accompanied by reasoning) which the program suggests to the client. This assessment will allow me to conclude on the true effectiveness of the framework.

The final methodology I need to incorporate revolves around removing bias from data-collection. This is important, especially in cases where I require data for the legal frame, as it will be necessary to remove any information that has been skewed by emotion. Potential bias could fall under the category of racism, sexism, ableism, ageism, though a potential solution would be retracting any information which could lead to such bias, e.g. "Person 1, 2, ..." a opposed to "Greg, Karen, ...".

Please refer to Appendix A.1 to further understand the breakdown of these methodologies and the time frames I have given myself.

Appendix A

Gantt Chart

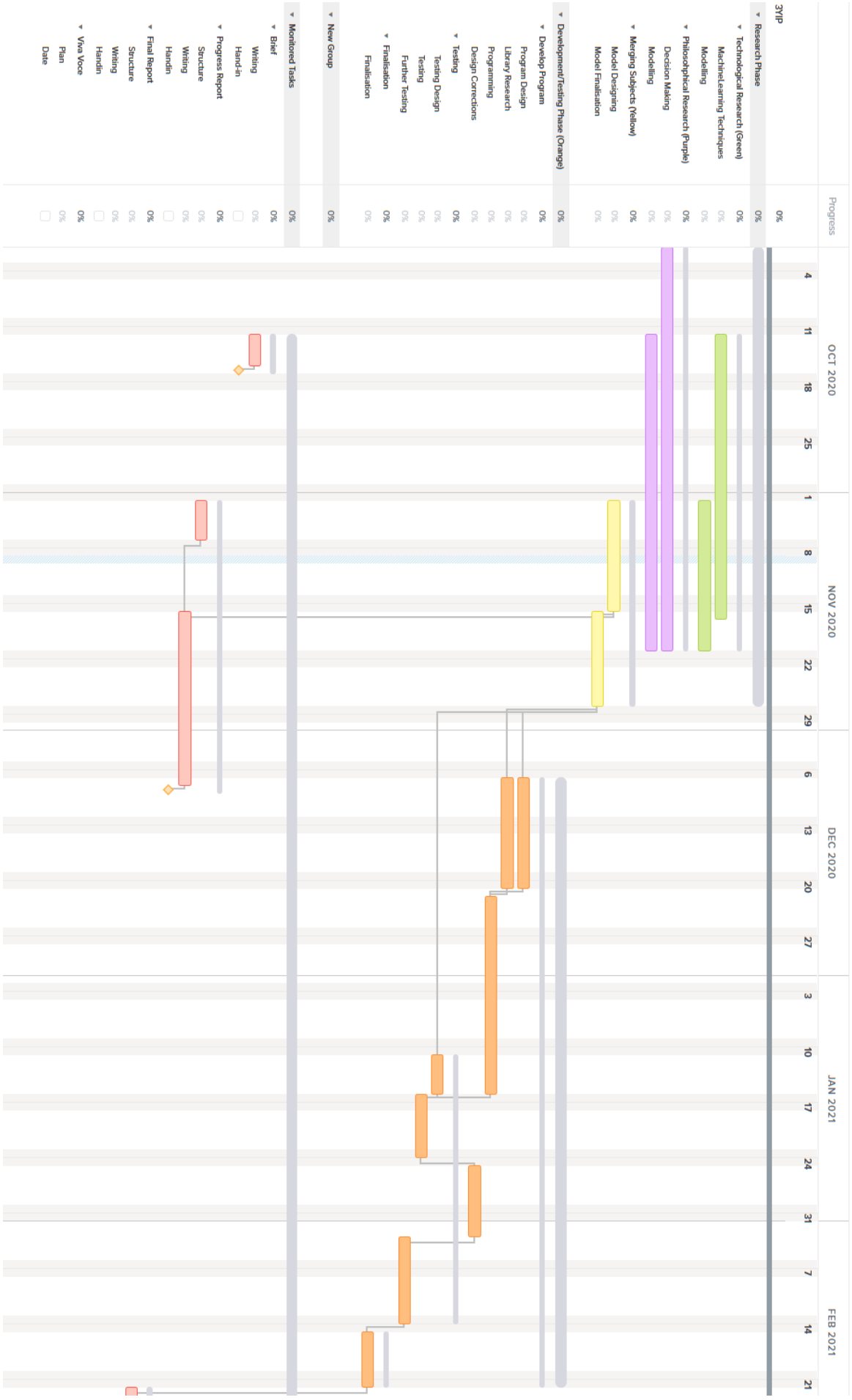

FIGURE A.1: GanttChart

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.

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.

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.

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.