# DOCTOR OF PHILOSOPHY

## Simulating Ethical Behaviour in Virtual Characters

Headleand, Christopher

*Award date:*
2016

*Awarding institution:*
Bangor University


PRIFYSGOL
BANGOR
UNIVERSITY

School of Computer Science
College of Physical & Applied Sciences

# Simulating Ethical Behaviour in Virtual Characters

---

Christopher J. Headleand

Submitted in partial satisfaction of the requirements for the
Degree of Doctor of Philosophy
in Computer Science

*Supervisors* Dr L. Ap Cenydd *and* Dr. W. J. Teahan

September, 2016

# Abstract

The goal of virtual human simulation is to produce behaviour which is visually indistinguishable from reality. However, while various aspects of human behaviour have been extensively explored, there has been little research into behaviour motivated by moral objectives. Virtual characters are often simulated in charged environments, where rational behaviour is greatly challenged and in reality moral judgement plays a significant role. This thesis explores and presents novel solutions to the problem of simulating ethical behaviour.

The research is presented in three stages. In the first, a reactive approach to the simulation of ethics inspired by Braitenberg's Vehicles is described. This is achieved by iteratively augmenting a Type 3 Vehicle with new sensorimotor connections to produce further emergent results. The approach was capable of producing behaviour which was consistent with various normative specifications. Although successful, the Braitenberg Vehicle approach yields behaviour which is visually robotic. This is explored in the second stage of the research where a novel method for modelling affective behaviour is presented.

In the third stage, a new architecture for the simulation and modelling of ethical behaviour called Trilogy is presented. This approach, inspired by classical and contemporary tripatriate theories of thought, serves as a computational substrate to bring together the ethical and affective simulation methods previously developed in stages one and two. Two experiments are conducted to evaluate the architecture where participants observed videos of simulated behaviour. The first experiment tests the hypothesis that the inclusion of affective states make an agent's ethical behaviour more believable, and this was demonstrated to be the case. The second experiment compares the behaviour of the ethical agents against agents which are not ethically motivated. The results of both experiments demonstrate that the approach is capable of producing visually ethical behaviour beyond chance accuracy.

**Statement of Originality**

The work presented in this thesis/dissertation is entirely from the studies of the individual student, except where otherwise stated. Where derivations are presented and the origin of the work is either wholly or in part from other sources, then full reference is given to the original author. This work has not been presented previously for any degree, nor is it at present under consideration by any other degree awarding body.

**Statement of Availability**

I hereby acknowledge the availability of any part of this thesis/dissertation for viewing, photocopying or incorporation into future studies, providing that full reference is given to the origins of any information contained herein. I further give permission for a copy of this work to be deposited with the Bangor University Institutional Digital Repository, the British Library ETHOS system, and/or in any other repository authorised for use by Bangor University and where necessary have gained the required permissions for the use of third party material. I acknowledge that Bangor University may make the title and a summary of this thesis/dissertation freely available.

# Acknowledgements

> *We've had some tough times, but we've hung in there.*
>
> **— Paul Allen**

Acknowledgements are always difficult to write, as it would be impossible to mention everyone that has helped me over the past three years. With that considered I would like extend special thanks to the following people, and groups of people, for their support during my PhD.

- Fujitsu and HPC Wales for the funding, and high performance facilities which made my PhD possible. I would specifically like to thank Laura Redfern and Dr. Ade Fewings, who both went above and beyond to help me achieve my goals.

- Dr. William Teahan and Dr. Llyr ap Cenydd for without their guidance this thesis would not have been possible. I have enjoyed working with you both immensely and I hope we continue to collaborate in the future.

- The academics and administrative team at the Bangor University Computer Science department who have helped me over the past four years. I would like to extend special thanks to Dr. Jonathan Roberts for helping me expand my research horizons, and Dr. Dave Perkins for always being happy to discuss teaching ideas.

- Cameron Gray, for all your help. Without your LaTeXguidance this thesis wouldn't be nearly as attractive. But more importantly, thank you for your friendship and for being a shoulder to cry on when I needed it. Good luck in your new post, any

department would be lucky to have you. I hope we will get the opportunity to work together in the future.

- My friends, Dr. Chris Hughes and Dr. Panagiotis Ritsos. You have both been mentors, and brothers to me. Thank you for providing perspective when I lacked it, and motivation when I needed it.

- My dissertation students, Ben, James, Jason, and Liam. Thank you for asking me to be your supervisor. I can honestly say that I probably learnt as much, if not more from you than you did from me.

- My new employers and colleagues at Lincoln University. Thank you for welcoming me into your superb faculty. I am looking forward to the new challenges and opportunities that await me as a lecturer.

- All the researchers and philosophers I have met along my academic journey. The conversations, debates, arguments and interactions have been invaluable.

- My parents, and my sister, for believing in me all these years, and always supporting my dreams. You have always been an inspiration to me and I feel truly blessed to have such a wonderful family.

- Laura Headleand, my beautiful wife, for her tolerance, patience, and understanding whenever I had to work late, or go into the office at the weekend. You have the patience of a saint, and your encouragement has always kept me going. Without your enduring support I would never have been able to return to university to have this amazing experience. I love you with all my heart.

- Finally to our dog Hunter. For encouraging me to take a break from the computer, and reminding me that nothing clears the mind quite like a nice walk in the park.

# Contributions

This thesis has resulted in a number of core contributions, and publications which are detailed below.

**Definitions of Ethical and Moral Agents** The field lacks formally accepted and consistent definitions. The field itself has been referred to by a number of names, including Artificial Morality and Computational Ethics. This is addressed in chapter 2 with the proposal of two definitions, consistent with current namely namely *Artificial Ethical Agents* (AEAs) and *Artificial Moral Agents* (AMAs).

**Reactive Models of Ethical Behaviour** This thesis is the first to propose and implement a reactive model of normative paradigms and value systems. This research built upon the work of Braitenberg [34] and Brooks [38], and was developed over chapters 3, 4, and 8. This research was featured in an exclusive article by the BBC.

**Affective States Modelling** The research into ethical simulation concluded that for the agents to appear believable, emotion may also need to be considered. The ASM approach was developed out of this need, providing a simulated affective response with a low computational overhead. This work built on theories of geometric models of thought, specifically Gärdenfors Conceptual Spaces [74].

**Trilogy Architecture** The Trilogy Architecture was designed out of a need for a computational substrate to unify the earlier work in the thesis. This allows the Ethical Simulation and the ASM approach to work in tandem, producing more complex characters. This approach was used in main simulation, which produced verifiably successful results.

# Publications

The following publications were written as an outcome of the research undertaken during this thesis. Some are directly included as extended versions as chapters of this thesis. Others are early or peripheral research, and are referenced accordingly within the text.


[94] C. J. Headleand, L. Ap Cenydd and W. J. Teahan, 'Action selection through affective states modelling', in *Science and Information Conference SAI*, IEEE, 2016, pp. 478–487 (p. 80).

[95] ——, 'Sexbots as ethical agents: On the possibility of ethical machines', in *9th Philosophy and Computing AISB Symposium*, 2016 (pp. 6, 10, 175).

[96] ——, 'Towards ethical robots: Revisiting Braitenberg's Vehicles', in *Science and Information Conference SAI*, IEEE, 2016, pp. 469–477 (pp. 42, 175).

[97] ——, 'Berry Eaters: Learning colour concepts with template based evolution evaluation', in *ALIFE 14: The Fourteenth Conference on the Synthesis and Simulation of Living Systems*, vol. 14, pp. 473–480 (pp. 45, 59, 67, 80, 84).

[98] C. J. Headleand, G. Henshall, L. Ap Cenydd and W. J. Teahan, 'The influence of virtual reality on the perception of artificial intelligence characters in games', in *Research and Development in Intelligent Systems XXXII*, Springer, 2015, pp. 345–357 (p. 104).

[99] C. J. Headleand, J. Jackson, L. Priday, W. Teahan and L. Ap Cenydd, 'Does the perceived identity of non-player characters change how we interact with them?', in *Cyberworlds*, 2015, pp. 145–152 (p. 106).

[101] C. J. Headleand, L. Priday, P. D. Ritsos, J. C. Roberts, L. Ap Cenydd and W. Teahan, 'Anthropomorphisation of software agents as a persuasive tool', in *British HCI*, 2015 (p. 11).

Furthermore, as a supervisor I supported the following research which was based on ideas presented within this thesis.

[42] L. Chapman, C. Gray and C. Headleand, ‘A sense-think-act architecture for low-cost mobile robotics’, in *Research and Development in Intelligent Systems XXXII*, Springer, 2015, pp. 405–410 (p. 116).

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

> *The true delight is in the finding out rather than in the knowing.*
>
> **— Isaac Asimov**

## 1.1 Motivation

Can we build machines that act ethically? Some have argued argue no, as morality has generally been reserved as a quality exclusive to the human condition [47, 176, 191]. Even the behaviour of non-human animals is rarely described in these terms when any other explanation fits. However, sidestepping the purely philosophical argument about what constitutes a *genuine* moral agent, and focusing on building machines capable of *acting* ethically may soon be one of the most crucial research areas in artificial intelligence [219]. Indeed, research in this area may even be critical to the survival of the field [8], since several high profile scientists, businessmen, and politicians have called for certain aspects of Artificial Intelligence (AI) research to be banned. This is due to the strong belief that AI may become an existential threat to humanity.

However, even if we choose to not subscribe to the doomsday predictions of an AI apocalypse, there are many reasons why we should research ethical behaviour, not least of which is for creative applications. Autonomous agents already feature heavily in creative applications, from the classic examples in computer games, to the digital extras featured in

modern film. These agents solve real, practical problems, such as providing a player with a stimulating opponent; or making financially or ethically impractical movie scenes possible. To be able to accept these agents as surrogates for their real-world analogues, the designers rely on a suspension of disbelief on the part of the observer. But, the viewers' immersion, and acceptance of this façade is notoriously hard to maintain, and can be broken by the smallest of errors.

Due to the high market value of the creative industry, there has been a plethora of research in this area, particularly in the virtual human sub-field. Yet, from a behavioural standpoint, we still have a long way to go in the production of truly believable, autonomous agents. The actions of virtual humans often lack the rich personality of their human counterparts, which can make them easy to identify. It could be argued that this is due to the inherent rationality of their behaviour. While current research has explored how agents can make optimised decisions, we now need to explore other components of the decision making process. If there is truth in the words "to err is human", then irrationality, and sub-optimal behaviour may hold the secrets to imbuing virtual characters with a measure of humanity.

To make characters appear more believably human, ethical behaviour must be considered, both posative outcomes, and cases where ethical behaviour breaks down. For example, despite training and established rules of engagement ethical standards can be greatly challenged on the battlefield [198]. Another example is panicked crowds, resulting in individuals being crushed in human stampedes. In such emotionally charged situations, an individual's ethical code, their values and principles of right and wrong, can change or be ignored. This emotional and moral depth to characters is missing from current research into virtual humans. It is common knowledge that a human's decision making process can be misled by a charged, or high stress situation. A common saying that reflects this is "don't make decisions when you are angry, or promises when you are happy". However, simulating this affective influence on ethical behaviour has been under-researched.

Being able to simulate ethical behaviour could also be used to gain insight into our own decision making process. It could also provide philosophers with a quantitative tool to evaluate theories of morality that are otherwise unverifiable.

2. To derive models for the simulation of ethical decision making within simulated characters.
3. To evaluate ethical agency models against the insights gained from the literature review.
4. To validate the models when applied to simulation.

## 1.5 Scope and Limitations

The focus of this project is simulation, specifically creative industry applications such as games and films. As the ultimate purpose of an agent in a creative application is to entertain, the interest and the scope of this work is in the observable output. The focus of this body of research is virtual humans, and making them behave in a situationally appropriate fashion with regards to ethical decision making. This is evaluated against whether the agents produced are capable of meeting a design objective, namely producing behaviour that is comparable to natural behaviour.

As such, the theories and models presented are not intended as an accurate cognitive model of a natural system. Although some chapters feature unsophisticated devices to test the theories (such as virtual fish, or Braitenberg inspired vehicles), these are only used as simple models of agency for experimentation and evaluation.

## 1.6 Important Terminology

The subsequent chapters of this thesis will make use of specific terminology. While these terms are used in a variety of literary sources, the context and application of their usage varies, which could lead to misunderstanding. Also, some of these terms have different meanings in other fields, or are used interchangeably with other terms that have subtly different meanings (such as ethics and morals). For these reasons, the following terms are

defined here. For the purpose of this thesis these definitions will take precedence over other possible definitions.

**Action Selection** : *choosing the most appropriate action out of a set of possible candidates, where an action represents a process or behaviour undertaken to achieve a goal.*

This definition is based on a popular definition from ecology research [210], and Craig Reynold's three-layer model of animation [170]

**Artificial Moral Agent** : *A synthetic autonomous entity that is capable of making, and being held accountable for, unsupervised decisions with reference to an understanding of right and wrong.*

**Artificial Ethical Agent** : *A synthetic autonomous entity that is capable of acting according to a set of morally defined considerations.*

The terms *artificial moral agent* and *artificial ethical agent* are often used interchangeably, without clear definition. They are also used for clearly different applications; for this reason it is important to provide specific definitions as research and theories in one of these domains does not automatically presume application to the other. These definitions are, in part, based on definitions of moral agency and moral subjecthood from the field of moral philosophy [177]. Further reasoning is provided in chapter 2 and was published at the AISB2016 conference [95].

**Ethic-like** : *An action which appears ethical, without necessarily being driven by genuine ethical concerns.*

This definition is included to add clarity to later chapters, separating simulated ethics from genuine ethical processes.

## 1.7 Thesis Outline

**Chapter 2** contributes an overview of the literature concerning artificial ethics. This begins with a discussion of the possible applications, followed by exploring the inspiration provided by science fiction. Following this, the general question of morality in non-human agents is considered, which leads to artificial ethics being formally defined within the context of this thesis. Next, a number of normative theories to aid discussion in future chapter followed by a discussion of the key technical and philosophical objections to the concept of artificial morality. The chapter concludes with an overview of the current research into artificial ethics, followed by an outlining of the main challenges in the field.

**Chapter 3** explores the reactive paradigm as a possible approach towards the simulation of ethical behaviour. Based heavily on the work of Braitenberg, various normative Vessels are introduced as thought experiments. These thought experiments are subsequently realised through simulation.

**Chapter 4** advances the work of chapter three by introducing the concept of value systems. Two of the ethical vessels from the previous chapter are combined through a threshold based value system, improving upon some previous limitations of the approach.

**Chapter 5** proposes a novel action selection system called Affective States Modelling. This work is based on insights from the previous chapter, suggesting that the simulation of ethical behaviour must also consider the simulation of emotions. The work is inspired by geometric models of thought and the mind.

**Chapter 6** addresses the challenge of how artificial ethics can be evaluated. This chapter begins with a short literature review of the area, focusing on Turing Test inspired evaluations, and the assessment of believability. Following the insights gained, a two stage methodology for evaluating simulated ethics is proposed.

**Chapter 7** proposes a new architecture for agent based simulation called *Trilogy*. The design of this architecture is inspired by various tripartite theories from psychology, philosophy and neuroscience. The architecture is proposed as framework for combining the developments from chapters 3, 4, and 5.

**Chapter 8** describes the specific implementation of the Trilogy architecture used in the thesis. The focus of this chapter is on the design of the Affective States model, and the development of the ethical modules.

**Chapter 9** discusses the evaluation of the main contributions from this thesis. Two experiments using the Trilogy architecture defined in chapter 8 are described. The first tests the hypothesis that including affective states makes ethical simulations more believable. The second, tests the simulated normative models against a standard boid-like simulation to evaluate if the models appear ethical, and believable to an observer above chance accuracy. Both experiments demonstrate successful results.

**Chapter 10** evaluates the success of the thesis by revisiting the original objectives defined in this chapter. Following this, the principle findings and contributions are outlined, leading to a short discussion on artificial ethical, and moral agency. The chapter concludes by describing possible avenues for future work.

# Chapter 2

# Artificial Ethics: Applications, Objections and Challenges

> *Every science begins as philosophy and ends as art.*
>
> **— Will Durant**

## 2.1 Introduction

The purpose of this chapter is to establish the related work, and current state of the art in the simulation of ethical behaviour; specifically, this chapter will explore the development of Artificial Ethical Agents.

To achieve this, we first need to establish what it means to be ethical from a philosophical perspective. Classically, concepts such as ethics have been described by philosophers as a quality reserved for humans. If we are to accept this position, then the development of Artificial Ethical Agents may be doomed to fail. In order to counter this position, and introduce the idea that humans may not be the only moral (or ethical) agents, the chapter begins by exploring some recent arguments for morality in non-human animals.

With a philosophical argument provided, the following section will provide a working definition of Artificial Ethical Agents. The intention is to clarify a difference between Artificial Moral Agents, and Artificial Ethical Agents. As the terms are often used

interchangeably, this definition provides a clear direction for the research in subsequent chapters.

The current thinking in the field of Artificial Morality will then be presented. There is some debate over what the field should be called, with titles such as artificial morality [55], machine morality [219], machine ethics [146] and computational ethics [5] used within the literature. The literature will be divided into two subgroups based on our definitions of Artificial Moral Agents, and Artificial Ethical Agents.

Many researchers would argue that the ultimate goal of research into simulating ethical behaviour is to construct an Artificial Moral Agent (or AMA) capable of genuine moral reasoning. However, whether this is even possible is a heavily debated topic. In this section four of the key philosophical objections will be described. Following this, the review describes the current research into Artificial Ethical Agents, concluding with the argument that a more achievable goal is creating artificial entities that are able to act within ethical constraints. The review concludes by highlighting a number of grand challenges in the field.

A shortened version of this chapter was published at the AISB2016 conference [95].

## 2.2 Applications of Moral Agents

There are many reasons why we may wish to build machines capable of making ethical judgements. Within the foreseeable future, robots may be as ubiquitous as computers are today [127]. Indeed, current home technology from lawn mowers to washing machines are being developed with increasing levels of autonomy. In 2010, the world robot population was said to be almost double the population of New Zealand, around 8.6 million, of which 7.3 million were described as ‘service robots’ [86]. This figure will have grown significantly since; iRobot alone have reportedly sold 14 million Roomba robots since their incorporation.

Future machines will have increased autonomy, which will force us to consider how they will be ethically governed [9]. The capacity to act within ethical standards could help avert dangerous or unethical behaviour from autonomous agents (such as the prophesied AI Armageddon [134, 59]). They could also be used to support human decision making in the moral arena. If a super-intelligent AI could be developed, it is both possible and logical to assume that it could make ethical decisions "better" than humans [32]. Furthermore, robots and simulated creatures as simple models of agency could be used to help understand more complex cases of human ethical judgements [57]. This could lead to simulation becoming a tool to advance the study of moral philosophy, providing theorists with an ability to test, and quantify ethical positions [10, 8], which have existed as pure theory since the time of Socrates. For reasons such as these, interest in the design of autonomous systems capable of making moral judgements "is stepping out of science fiction and moving into the laboratory [214]".

Importantly, society is becoming increasingly driven by technology. There are now many examples of tasks that would have originally required a human that have now been delegated to machines and algorithms. In many respects, we can think of robots as replacements for humans, typically in tasks that are dull, dirty or dangerous [127]. This handover of control in our society has in many cases (such as banking) placed computers in situations where they can affect the moral rights of humans [191]. This is especially true of anthropomorphic agents, due to their social, and persuasive qualities [101].

Machines can therefore already be classed as moral subjects or ethical impact agents (which will be defined in the following sections). However, many people would argue that a machine should never be put in the position to make a non-trivial ethical decision regarding a human (a feeling shared by members of the research community [229]). For this reason, the moral actions of machines are currently still closely controlled by human operators.

One of the central questions is whether an artificial entity can ever be a moral agent [106, 111, 191, 196]? This has led to a range of debates regarding how we should be viewing and regulating robots and other autonomous agents. For example, the increasing use of robotics in the military arena has led to a growing body of research on the ethics of

autonomous military machines (so-called ‘killer robots’) and a plethora of literature on the subject [49, 102, 121, 125, 144, 205]. There is also significant fear that the development of military artificial intelligence could be dangerous, so much so that it would be incompatible with human life. In response to concerns such as these, a committee of researchers and philosophers have recently published BS8611 [174], a standards document relating to the ethical design and usage of robotic systems.

However, there is an alternative argument that machines could be preferential to humans in some scenarios. For example, an individuals ethical standards may be challenged on the battlefield [198]. It is possible that the situation could be improved by taking humans out of the loop, leaving an opportunity for roboticists to design systems that could potentially do better [197].

Driving, even with a ‘perfect’ driver, is not without risks due to the inherently dynamic nature of the road environment. This in part has led to the development of self driving cars, an argument being that a robot may do a better job. However, how an automated vehicle assigns risk, especially in situations where a crash cannot be avoided, requires a level of ethical decision making [79].

We are also starting to see development of ethical agents for use in health-care applications [159]. Applications which provide ethical recommendations in the medical arena are one of the most prominent examples of current research, which goes against some warnings that pre-date the field. Joseph Weizenbaum (developer of the ELIZA chatbot) concluded in 1976 [225] that AI should not replace a human in roles that inherently require respect, understanding, or love. His argument was that such applications require authentic empathy or the human client would be left feeling devalued, which could represent a threat to human dignity. His solution was to propose that research is restricted in certain fields, including therapy and military applications. Forty years later, these debates are ongoing, with many scientists, researchers and humanitarian societies calling for weaponised AI research and development to be banned [211, 80].

As stated earlier in this section, one thing we can be certain of is that future AI developments will benefit from increased levels of autonomy. However, to quote from the work of Picard [156] "the greater the freedom of a machine, the more it will need moral standards". If we accept this to be true, then the applications and need for machines that are artificially ethical will only increase.

## 2.3 Inspiration from Fiction

Aside from philosophical concerns, the fear that developing artificial intelligence will lead to our destruction is deep rooted in our mythology. Two classical examples that predate the discipline of computer science and robotics are the *The Golem of Prague* [78] and *Frankenstein* [187]. Even the word 'robot' comes from a dystopian fantasy, coined as the name for artificial workers who rise up and overthrow their human masters[1]. This 'Frankenstein complex', that the things we create will destroy us is persistent (arguably re-enforced by numerous works of science fiction), and if nothing else, has forced us to evaluate the moral implications of machines, and the responsibilities of their designers.

AI rebelling, or destroying humanity, has since become one of the most common themes in science fiction. Some of the best known films in this genre include the Terminator franchise, where Skynet (a malevolent AI) attempts to destroy the human race. Another example is the TV series *Battlestar Galactica*, where Cylons, a race of robot humanoids chase the last surviving remnants of mankind through the galaxy after destroying humanity's planet.

A more hopeful outlook of moral machines (as opposed to malicious ones), play a central role in the works of Isaac Asimov best highlighted in the 'Three Laws' of robotics [19, 134]. The Three Laws are a literary device which underpins much of Asmimov's work, used to explore themes regarding robot and human behaviour. Furthermore, they are arguably one of the first examples of fictional robotics that move beyond the classic Frankenstein

[1]The play in question is R.U.R (Rossum's Universal Robots) written by Czech author Karel Capek [40]. Robot is derived from the Czech word 'robota' which translates as servitude.

complex. It is (to the best our knowledge) the earliest example of a synthetic rule-based ethical value system. The Three Laws are:

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given it by human beings, except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

While these laws may seem reasonable, a common theme of Asimov's stories[2] is where things go wrong due to conflicts between the rules, their rigidness preventing robots from operating in real world scenarios [182]. Although fictional, Asimov's laws have been heavily critiqued [81, 134, 226, 163] and the possibility of their real world implementation has been discussed extensively. Clarke [44, 45] provides arguably the most thorough analysis of the three laws applied directly to computer science.

One limiting factor that would negate the practical implementation of the three laws is the intelligence required to comply to them. Asimov's three laws assume that the robots have sufficient agency and cognitive ability to recognise dilemmas and make the appropriate decision [148], abilities beyond the current technological generation. However, it has been argued that even if it was practical to implement Asimov's three laws, it may be unethical to do so, as they allow humans to mistreat machines without risk of repercussion. Anderson [12] argues that we should avoid the possibility of mistreating machines, as it may desensitise us to mistreating other humans, an argument derived from Kant. This fear that anthropomorphic robots could lead to the dehumanisation of people has recently become a significant point of discussion. Research and development into 'sexbots' (anthropomorphically sexualised robot companions) has raised questions over what impact it may have on the human sex trade. This has resulted in calls to ban the production of these devices [173].

[2]For an example, see Asimov's *The Naked Sun*, or *Robots and Empire*.

Although the three laws may be inappropriate for practical implementation, they are thought provoking as a device to discuss machine ethics. They have also served as inspiration for other research areas. An example of this is Murphy and Wood's "Three Laws of Responsible Robotics" [148], intended as guidance for the use of robotic systems.

1. A human may not deploy a robot without the human–robot work system meeting the highest legal and professional standards of safety and ethics
2. A robot must respond to humans as appropriate for their roles.
3. A robot must be endowed with sufficient situated autonomy to protect its own existence as long as such protection provides smooth transfer of control to other agents consistent with the first and second laws.

In contrast to Asimov's laws, the laws of responsible robotics focus on the responsibility of the humans who design, develop and deploy these systems. They are also designed with implementation in mind, and do not make the assumption that the robots have significant cognitive ability.

Another fictional example of a rule system appears in *With Folded Hands* by Jack Williamson [227], where the robots are given a single command 'to serve and obey and guard men from harm'. The unfortunate outcome of this rule is a totalitarian society, where humans are essentially kept as prisoners to protect them from injury.

Although fictitious, these rule based systems do raise interesting questions regarding their possible flaws. For example, rule based ethical values are only ethical if they have been correctly implemented, as they intrinsically assume that the programmers who develop them are infallible. Furthermore, building an autonomous agent without the ability to question its own initial instructions is to assume "moral and engineering perfection on the part of the designer [9]". For example, consider how different the outcome could be if the order of Asimov's three laws were changed (see Figure 2.1).

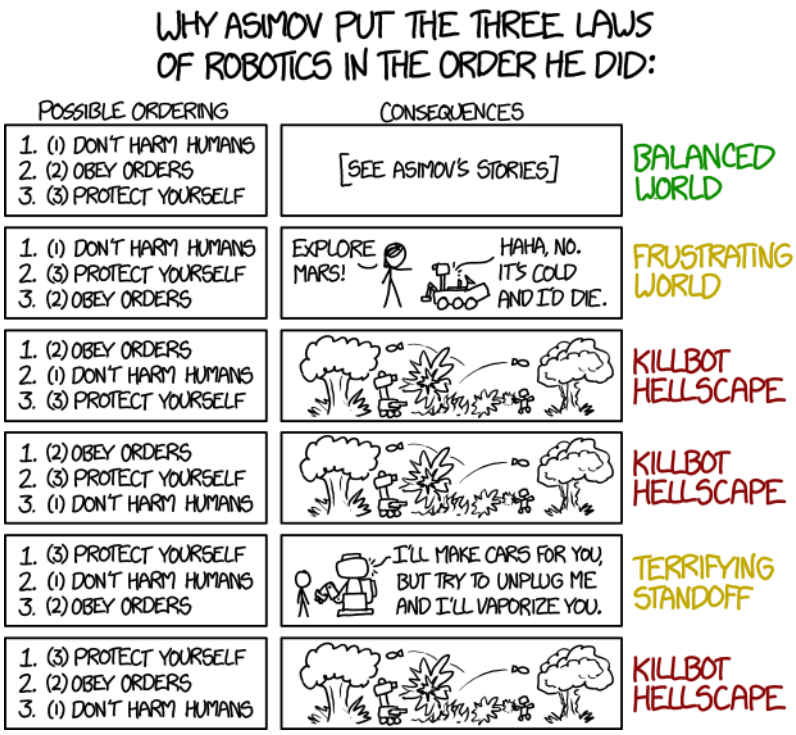



**Figure 2.1:** The outcome of reordering Asimov's three laws. A comic by Randal Monroe of XKCD *https://xkcd.com/1613/*

## 2.4 Existential Threat of AI

As discussed in the previous section, a common theme of science fiction is the dangers that AI could pose to humanity. However, moving away from fiction, there are a number of well funded research groups exploring whether AI could become an existential threat. An existential threat (using the definition by Bostrom [31]) is "One where an adverse outcome would either annihilate Earth-originating intelligent life or permanently and drastically curtail its potential". Simply put, an existential risk is one which jeopardises all of mankind. While the study of existential threats is too large to properly review here[3], there are a few areas worth highlighting.

One risk which is often discussed is that a flawed super-intelligence could create threats by acting on faulty commands or by misinterpreting goals [31, 232], a concept which was used as the primary plot device for the *Avengers: Age of Ultron* film in 2015. This is often

[3]However, a brilliant discussion on some of the possible threats posed by AI is provided by Yudkowsky [233].

referred to as the Instrumental Convergence theory, the idea that seemingly harmless goals can act in harmful ways. This concern is often described as the "Paperclip Maximiser", after an illustration provided by Bostrom in 2003 [32]. A paperclip maximiser is an AI with the goal of maximising the number of paperclips it possesses. It would develop more and more efficient ways to maximise its number of paperclips. Eventually, it may develop ways of converting other matter into this resource which could be a threat to humanity. As eloquently put by Yudkowsky, "the AI does not hate you, nor does it love you, but you are made out of atoms which it can use for something else" [233].The paperclip maximiser argument suggests that we are more likely to develop a non-malicious AI that is unintentionally dangerous, than one that is purposefully hostile.

While we generally attribute the risks associated with AI to human-level or super-intelligence, relatively simple systems could pose an equal risk. One notable example is a Nuclear 'fail deadly' system which launches a counter-strike should its country's government be incapacitated during war. While this may sound like the plot of a James Bond novel, there is significant evidence that such a system called 'Perimeter' (also referred to as 'Dead Hand') was employed by Russia during the cold war. There are also a significant number of people who believe it is still operational today [202, 28].

## 2.5 Morality in non-Human Agents

When considering broad concepts such as morality and ethics in machines, it is prudent to consider the debate regarding other non-human entities.

Typically, when attempting to describe animal behaviour, we do not assign them with any more qualities than we absolutely have to [176]. It was once taboo to even use the words *animal* and *cognition* within the same sentence [59]. Yampolskiy [229] argues that generally we have avoided ascribing non-human animals with moral agency for the same reason.

An example of this traditional way of thinking is provided by Himma [106], who describes a dog which attacks someone wearing red because that is how it is trained. Although the dog was the direct cause of its behaviour (in the sense that its mental state resulted in its actions), Himma argues that "it has not freely chosen its behaviour *because dogs do not make decisions in the relevant sense* (emphasis added)". As such, the moral responsibility (and by extension, the agency) remains with the person who trained it.

Rowlands [177] addresses this concept of responsibility as being central to the issue of agency, which he defines as:

> *X is a moral agent if and only if X is (a) morally responsible for, and so can be (b) morally evaluated (praised or blamed, broadly understood) for, its motives and actions.*

An argument is made by Rowlands that it is probably mistaken to classify animals as full moral agents, but even with that considered, they could be moral subjects.

> *X is a moral subject if and only if X is, at least sometimes, motivated to act by moral considerations.*

Additionally, in recent years some philosophers have questioned this traditional way of thinking, renewing the debate on animal morality [46]. It has been argued by Killen and de Waal [119] that the behaviours of humans and non-human primates point towards a shared evolutionary background towards morality. For example, non-human primates have demonstrated similar methods to humans for preventing and resolving conflicts, described by some [69] as the building blocks of moral systems. One of these traits is empathy, which has often been revered as a human trait; however, a significant body of research now suggests that other animals exhibit this phenomenon [212].

There are also various examples in the literature where behaviour has been described in such a way that, if it had been attributed to humans it would have been assumed to have been moral (for example see [63, 43, 172]; this being a relatively small sample of the

accounts available). This forces us to question if it is an unfair standard to not credit an animal with any greater quality than we absolutely have to.

Coeckelbergh [48] notes that in the study of human morality we tend to rely on observation. It could be argued that it would be more consistent if we applied the same standards to the assessment of other entities. By assuming an observational standpoint (even an anthropomorphic one), we open up new opportunities in the study of behaviour, by comparing our behaviour with that of primates (such as the work by Frans de Waal [213]) or through simulation.

## 2.6 Defining Artificial Ethics

In this section the differences between Artificial Moral Agents (AMAs) and Artificial Ethical Agents (AEAs) will be defined. This is for two reasons: (i) the terms are often used interchangeably, without clear definition; and (ii) research and theories in one of these domains does not automatically presume application to the other. For example, artificial ethical theories may not be applicable to the larger questions of morality, and moral agency.

The Oxford English dictionary defines morals as:

> *"The principles of right and wrong behaviour."*

and Ethics as:

> *"A set of moral principles, especially ones relating to or affirming a specified group, field, or form of conduct."*

We can interpret these definitions to mean that *morals* are the core concepts which allow us to define the difference between right and wrong. By contrast, *ethics* relates to how these rules are applied. With this understanding, and the insights provided by Rowlands [177]

on moral agency and moral subjecthood (mentioned in section 2.5), an Artificial Moral Agent (AMA) can be defined as:

> *An synthetic autonomous entity that is capable of making, and being held accountable for, unsupervised decisions with reference to an understanding of right and wrong.*

Whereas an Artificial Ethical Agent (AEA) can be defined as:

> *A synthetic autonomous entity that is capable of acting according to a set of morally defined considerations.*

The principle difference between these two is that the Artificial Moral Agent must be able to derive its own concepts and understanding of what constitutes moral behaviour. Consequently, it must also be accountable for its actions. On the other hand, the principles (or rules) for an Artificial Ethical Agent may come from an external source.

Anderson and Anderson [8] have stated that they believe the objective of machine ethics is to "create a machine that *itself* follows an ideal ethical principle or set of principles; that is to say, it is guided by this principle or these principles in decisions it makes about possible courses of action it could take". Via our definitions, this would be an Artificial Ethical Agent if the principles were provided by a designer, and an Artificial Moral Agent if it determined the principles via its own means.

An alternate, yet compatible definition of Ethical Agents is provided by Moor [146, 9]. Moor's taxonomy specifically defines four different types of ethical agent:

**Ethical Impact Agents** are any agents whose actions could have ethical consequences. Almost any agent has the potential to be an Ethical Impact Agent if it has the potential to cause harm (or benefit) to humans.

**Implicit Ethical Agents** are agents that are designed with ethical considerations in mind. These generally refer to safety considerations, including warning systems.

**Explicit Ethical Agents** are agents that can identify a variety of ethical considerations, and process situational information to determine a course of action.

**Full Ethical Agents** are similar to Explicit ethical agents but also have metaphysical features that we typically attribute to human ethical agents, including free will.

The definitions for Ethical Impact Agents and Implicit Ethical Agents are too broad to be practically useful, as they could describe almost any agent. Although, they do serve to highlight the scope of ethical machines. Furthermore, the difference between Explicit Ethical Agents and Full Ethical Agents is purely philosophical. There is no evidence that a 'Full Ethical Agent' would be substantially different in design to an 'Explicit Ethical Agent', especially considering that neuroscientists are beginning to question the concept of metaphysical qualities such as 'free will' [53]. This question is acknowledged by Moor himself, who has stated that an important question is whether attributes, such as free-will and consciousness, are essential to genuine ethical decision making [9]. Regardless of whether qualities such as free will exist or not, unless they can be codified, their inclusion in a definition is not helpful from an engineering perspective.

That considered, our definition purposefully avoids this philosophical distinction. Instead, the central concept of our two definitions is a matter of responsibility, specifically, can a machine be ever considered accountable for its actions [192], and can moral responsibility ever be transferred from a human (the designer or operator) to the machine [90]? If yes, then the machine is an AMA; if not, but the machine still operated within ethical considerations, then it is defined as an AEA.

Before an AMA (according to our definition) could be developed (or recognised) as a Moral Agent, there is an abundance of legal and technological hurdles which must first be addressed. Considering the debate over morality in non-human animals (biological entities), it is unlikely these issues will be resolved conclusively for synthetic agents in the near future. An unfortunate reality of this situation is that even if we were to develop a genuine moral agent, we might not recognise it as such, due to our inability to agree on a standard criteria beyond the scope of human beings.

An AMA is arguably not possible within the current technological generation, as many core AI issues must first be solved. Also, as detailed in the previous section, there are many who would argue that an AMA will never be possible.

However, as the rules for an Artificial Ethical Agent can come from an external source, this makes their development a practical engineering problem immune to the objections defined in the previous section. Agents that would pass our definition of an AEA are not only possible, but have already become a reality.

## 2.7 Implementing Ethical Frameworks

Individuals often have strong feelings regarding what they believe to be ethical, and conversely unethical actions. However, there are many normative theories which provide guidance on selecting morally correct actions, and they often do not agree. This section will describe a number of standard ethical theories, and discuss the possibilities of them being implemented within an AEA.

It is worth noting that as developers, we could also argue that artificial entities will need a set of ethics designed for robots, rather than attempting to recreate a human mode of moral reasoning. For example, some roles in which we may deploy artificial entities may suit the implementation of more basic moral governance [56]. It has also been been argued that there is no need for a moral machine to recreate a human moral process for them to be useful [215]. Furthermore, even if an agent were to simulate one of the following theories, this does not imply that it needs to follow the same moral process, or embody the same metaphysical principles as a human. Moor [9] has questioned whether it would be sufficient for a machine to simply *act* as if it did have these qualities.

However, with simulation as the goal, it is prudent to consider the theories pertaining to human ethical decision making. Over the following subsections, three core ethical theories will be discussed: Consequentialism; Deontology; and Virtuism. In the case of Consequentialism and Deontology, a number of sub-theories will also be discussed.

### 2.7.1 Consequentialism

Consequentialism is a theory which states that the consequence of an action should be how the action is ethically evaluated. Simply put, an action is considered morally good, if it produces a good outcome. One issue which must be overcome in the development of an artificial consequentialist is that (in principle), these ethical models require an agent to look ahead at the future. The agent needs to evaluate the results of their actions, before those actions are committed to. One way this could be achieved is through an internal model, an embedded simulation of the agent, and the environment at the present time. Feedback from previous actions, the robot's past precedent, could also be used to calibrate the model [228], allowing the agent to learn from experience.

The following subsections will review four common consequentialist theories, Utilitarianism, Egoism, Hedonism, and Altruism.

#### Utilitarianism

Of the consequentialist theories, Utilitarianism is arguably the most prominent. However, Utilitarianism is specific in that it considers the ethical action to be the one which maximises the utility of entities impacted by the decision. What gets defined as utility varies between philosophers, but may be defined as pleasure, welfare, or economic stability amongst others. In its simplest form, Utilitarianism argues that if you can increase the overall 'good' in the world, then you have a moral obligation to do so.

Utilitarianism is often championed as a candidate for Artificial Ethical Agents as it shares common concepts with standard, utility-based, AI algorithms. Furthermore, its structure has been described as 'moral arithmetic', making it more obvious to codify. There have been various attempts to implement Utilitarian Artificial Ethical Agents [82, 104].

However, whether an artificial system could ever analyse all the required information to make an active Utilitarian decision in real-time is an open question [219]. There is also a

temporal problem of when should you stop calculating against a cascade of consequences. Even humans who attempt to make a Utilitarian decision only do so within the bounds of their own capacity, and the information they have available. This ultimately means that human-based Utilitarianism is often a 'best guess' of consequences based on past experience. Furthermore humans measure utility over individuals, and how we calculate everyone involved in a decision is another problem for traditional Utilitarian, consequential theories. How can consequences be fully calculated when there is limited information on the cascading effects of actions [215]?

A further criticism is that the theory of Utilitarianism also allows for actions that some consider immoral as it can lead to the rights of an individual being violated [77]. However, it may be an appropriate normative framework to guide robot to robot interactions in some circumstances [82].

Gips [77] was the first to identify that Utilitarian and Deontological (which will be described in a following subsection) theories are relevant to the design of artificial ethics, largely due to their naturally codified structure. Furthermore, Allen [6] argues Utilitarianism as a candidate for a top-down system.

**Egoism**

Egoism is a sub-theory of consequentialist ethics, that states that an individual should be the motivation and the goal of one's actions above all others [66, 164], and that they have no moral obligation to serve the interests of others. The word is derived from the Latin word *Ego* which translates to *I* (the self) in English. There are several variations of this doctrine. For example, Regis [167] introduced the belief that an individual may have to sacrifice self-interest in order to avoid violating the rights of others [180]. But the core principle of self-interest is consistent across all.

Some of the supporters of Egoism have proposed that it is the only ethical philosophy which respects the rights and the value of an individual [165].

### Hedonism

Hedonism is the theory that individuals have an obligation to maximise their net pleasure, minus any pain or harm. Hedonistic theories all share this standard view that pleasure and pain are the only important elements when selecting a course of action [147]. In some ways, it is similar to Egoism, in that it is concerned with the self. Indeed, the most common variant is called Hedonistic Egoism which holds that an agent should do whatever is in its own best interests to maximise pleasure [223].

The theory of Hedonism has a long history. It was one of the central ideas in ancient philosophy, defended by the Epicureans and discussed by Plato in the Republic [54]. However, it is worth noting that Hedonism, despite some noted periods of popularity, has generally been an unpopular theory among philosophers [68], with some describing it as repugnant [223]. The objections towards this theory typically focus on the fact that pleasure and pain are the only qualities assigned with any importance, but this is also what makes it distinctive and philosophically interesting.

At the time of writing, there does not appear to be any attempts in the literature to theorise or implement an artificial hedonist. As with an egoist, a self-interested agent would not ethically moderate their behaviour towards humans, indeed, a self-interested agent could be quite dangerous. However, from a simulation perspective, hedonists and egoists should be considered within the scope of human ethical behaviour.

### Altruism

Altruism is typically defined as a costly act that confers benefits on other individuals [67]. Simply put, an altruist believes that an individual is morally obliged to help others, without considering their own gratification. In this sense, it is in direct contrast to Egoism. While there have been few attempts to implement Altruism in its pure form, there are a number of studies where altruistic actions have been considered the criteria for ethical behaviour [228].

Some supporters of Egoism have argued that Altruism (the opposite position, which holds that individuals have a duty to help others) is destructive, as an agent should not value themselves, but view their life as something which may be sacrificed for the good of others [165].

### 2.7.2 Deontology

Deontology holds that the morally right action is the one which adheres to an agent's duty or duties. There have been a number of attempts within the literature where researchers have attempted to implement Deontology [161, 8, 16, 35].

#### Kantianism

Kantianism, named for the philosopher Immanuel Kant, is a specific subcategory of deontology which is based on a standard of moral rationality called the Categorical Imperative. The Categorical Imperative states "Act as if the maxim of your action were to become through your will a universal law of nature". There are few examples of attempts to implement Kantian ethics, though researchers have argued its candidacy for artificial agents [162, 26].

Powers [161] argues that rule-based ethical theories such as Kant's Categorical are a promising direction for machine ethics to take due to the naturally codified structure of their judgements.

However, Beavers [27] states that Kantian ethics, particularly the categorical imperative, is a bad candidate for an AMA as it "seems to fall dead before a moral version of the frame problem". Calculating the categorical imperative requires the agent to acknowledge the implications of creating a universal moral law. He goes on to state that while the categorical imperative appears reasonable and implementable from a moral logic perspective, it fails from an engineering perspective due to a problem of scope. This further highlights the

intrinsic difficulty of this field, as most moral theories have not been devised with artificial implementation in mind.

### 2.7.3 Virtuism

Virtuism focuses on an individual's virtues, or moral character. It is based on the virtues that an action embodies, rather than the consequences, and for this reason is is difficult to codify as it provides no practical guidance in moral dilemmas. While there have not been any significant attempts to implement Virtue ethics, a strong case has been made for it [219].

## 2.8 Artificial Moral Agents

It has been stated that the primary goal of the study of artificial morality is the design of agents that act as if they are moral [6]. It is important to note that this statement implies that behaving morally is the objective, rather than the metaphysical underpinnings that would be required to be described as a genuine moral agent. If we were able to produce agents that behaved according to moral guidelines then maybe some of our doomsday fears would be alleviated. At the very least, humans would arguably be more comfortable collaborating with machines that follow basic principles of right and wrong.

But is artificial morality possible? Sullins [195] argues that it is "absolutely certain" that under certain conditions, artificial life (ALife) programs can exhibit artificial morality. Sullins also states that the study of 'Wet ALife' (the use of bio-components in artificial life research) could result in the fields of bio and computational ethics sharing concerns.

However, there are many arguments which oppose this position on the grounds that synthetic moral agency is either impossible or undesirable. In the following subsection, an overview of some of the philosophical objections will be provided.

### 2.8.1 Key Objections to Artificial Moral Agency

During a review of the literature, the following four principle objections to artificial morality were identified. It is important to note that these objections focus only on artificial morality, without discussing the wider objections to true AI, which are also applicable.

**Frankenstein Objection:** If we believe that moral reasoning requires consciousness, and the ability to exercise emotions, then producing an AMA in the foreseeable future is unlikely. If this is the case, then it would be better to avoid building highly intelligent autonomous agents as their lack of morality may make them dangerous [48, 70]. This concept of a robotic uprising is often based on the fear of machines without morality [134]. Conversely, building true moral agents may never be desirable. A genuine moral agent must be capable of recognising and acting *immorally* [26], as such an entity that simply simulates moral behaviour may be preferable to a genuine moral agent.

**The Scapegoat Objection:** If machines are considered to be true moral agents, then they would be morally accountable. One fear is that people may begin to use machines to avoid personal responsibility [84, 102, 112, 191, 33] essentially blaming machines for their actions. Similarly, there is a fear that decision makers may rely on machines too much, circumventing individual responsibility entirely [214].

**The Information Processing Objection:** Stahl [191] argues that *information* is a core requirement of moral agency and computers in their current form are processors of data, rather than processors of information. The argument is that computers lack the conceptual understanding to make a moral assessment, which precludes computers from achieving moral agency.

**The Free Will Objection:** According to Kant, moral agency requires both rationality and personal freedom. Supporters of this argument state that computer programs do not have free will, and thus they can never be independent moral agents [106, 111, 206]. However, the inverse has also been argued, that the ability to make moral decisions should be considered an essential attribute in being considered a fully conscious agent. [218]. Another

interpretation of the free will objection is the 'Mousetrap Objection': if an intelligent agent is capable of closing a loop between a sensor and effector without human intervention [178], then by extension a mousetrap is an intelligent agent. However, a mousetrap is not morally responsible (and as such, not a moral agent) because it closes its loop entirely at the volition of one or more humans. Responsibility for the actions of the mousetrap remains with the humans who armed it. "If an agent is to be morally praiseworthy, then its rules for behaviour and the mechanisms for supplying those rules must not be supplied entirely by external humans" [103].

These arguments reinforce the reasons why we have typically denied the possibility that a non-human entity could be a moral agent, something which has traditionally been reserved exclusively for humans [47, 176, 191]. However, while the current generation of machines do not possess free will, we cannot state categorically that they will not in the future.

Another issue with the development of moral machines is the actual practicalities of development. Building an Artificial Moral Agent is by its nature a practical goal. It is a distinctly different objective to that of moral philosophers, and as such determining a clear specification of what constitutes 'moral' from a technical perspective can be a challenge to discern, as a comprehensive model of moral decision making does not yet exist [215].

This is well summarised by Allen et al. [7], who notes that the development of Artificial Moral Agents (AMAs) is hindered by two areas of disagreement from theorists. The first is that philosophers disagree about what behavioural standards constitute morality. The other is ontological, a question of what it means to be moral.

Until these practical and philosophical issues are resolved, we propose that research should focus instead on Artificial Ethical Agents. This change in focus allows us to sidestep the objections entirely, by designing agents that *simulate* moral behaviour. As previously mentioned, an agent that simply exhibits moral behaviour may (in some ways) be preferential to a genuine moral agent anyway.

## 2.9 Artificial Ethical Agents

This section will describe attempts within the literature to develop Artificial Ethical Agents. The majority of the current examples of Artificial Ethical Agents have been implemented in a traditional top-down fashion. While top-down approaches in AI have been criticised for not being robust enough for real-world or uncertain tasks, there are specific areas where these approaches currently remain the best option. According to Allen et al. [6] it is an open question as to whether moral decision making is one of these domains. Bostrom also argues that it is important that AI be predictable to those impacted by the system's decisions, as this is an important component of the legal system (following precedent) [33], so therefore this would likely be a benefit of a top-down approach. Anderson and Anderson argue that an AI that has learnt or been programmed to make ethical judgements, but cannot justify its decisions, is lacking an essential quality to be accepted as an ethical agent [8]. They also argue that only explicit representations of ethical principles would allow agents to justify their decisions after action.

An alternative direction that could be taken is the reactive approach. The reactive approach involves connecting various subsystems to produce emergent behaviour. Classic examples of this approach are the Subsumption Architecture by Brooks [36], and the Vehicles proposed by Braitenberg [34]. Braitenberg Vehicles are a particularly interesting example, as they demonstrate how seemingly natural behaviour can emerge from a mechanical, reactive system. These behaviours are described as love, fear, concept formation and the ability to learn (among others) and follow social rules of behaviour. Wallach and Allen even argue that the Alife field has contributions to make to research into Artificial Moral Agents, by "helping to understand the bottom-up emergence of dynamic and flexible moral behaviour" [216]. However, although Wallach [219] provides an excellent discussion on the subject, there does not appear to be many practical examples of true bottom-up, or behaviour based AEAs.

Arkina [14, 15] describes three possible ways that ethical decision making could be implemented into an autonomous agent. The first system is referred to as a *Governor*,

which would halt the agent from proceeding with actions deemed unethical, similar to the valves which prevent steam engines from exceeding safe limits. The second is *Behaviour Control* which monitors the behaviours a robot is engaged with and ensures that actions fall within a set of constraints. The final system is the *Adaptor*, which modifies the first two systems if somehow an unethical action occurred despite their intervention.

Winfield et al. [228] describes an alternative approach referred to as a consequence engine. In their approach, a simulator is embedded within a robot, providing the agent with a pseudo-imagination that allows it to try actions before executing them in the real world. Through this process, the robot would be able to find the sequence of actions which best achieves its goal, and 'ponder' hypothetical situations which may arise. The authors of this work argue that this capacity can be applied to ethical decision making. For example, consider a robot that has rules which prevent it from colliding with a human to avoid causing injury. The robot observes a human about to step into the path of a car. Its internal simulator determines that the action that causes the least harm to the human is for the robot to collide with the human, preventing them being hit by the vehicle. The authors note that this example is remarkably similar to the first law of robotics described by Asimov: "A robot may not injure a human being or, through inaction, allow a human being to come to harm". The Winfield 'What-if' engine is shown to work on simple situated robots in a limited world, with one robot changing its behaviour to prevent another from being harmed.

A similar model is proposed by Bar-Cohen and Hanson [25] for a general purpose ethical engine. They prescribe a number of requirements for ethical robotics (best summarised by Sullins [197]) which are:

1. estimate with high detail and accuracy the immediate state of the world around the agent;
2. predict the likely future states given the current possible candidate actions.

The problem with this model is that it could be used to describe any decision making process in AI. As such, it falls prey to some of the common problems surrounding the field of artificial intelligence, such as estimating the state of the world in dynamic environments,

and the computational complexity in evaluating every possible future state. It could also be argued that current algorithms exist which meet this criteria, MiniMax [178] for example, but only work in worlds of constrained complexity with a finite number of possible actions. However, GPU based simulation could be part of the solution to this issue. Experiments implementing the 'what-if engine' on an embedded mobile GPUs have demonstrated higher simulation speed and lower energy usage in autonomous robotics simulations [114].

Early attempts at building machines capable of ethical reasoning have focused on decision support systems [214]. Anderson argues that the best way to start tackling the challenge of ethical decision making is to build machines that act as advisers to humans, in a select community, in a finite number of circumstances [9], creating an explicit ethical agent (to use Moor's definition) that is not autonomous. This approach also allows the community to postpone the philosophical debate regarding the moral status of machines.

Anderson et al. [11, 10] developed MedEthEx, the earliest example of a prototype artificial ethical adviser based on bio-medical ethical principles. In their approach, machine learning techniques are used to abstract decision principles from a library of cases with conflicting prima facie duties (duties that suggest conflicting courses of action in specific ethical dilemmas). This is trained based on a 'correct action' determined by an agreement of ethicists. This is in an attempt to capture the complexities of ethical decision making, and to codify a decision process for determining the ethically correct course of action when conflicts between duties, values or principles arise.

This approach attempts to codify ethical reasoning, essentially making judgements based on a training set of past examples. This method of learning through precedent can be compared to the branch of applied ethics called casuistry [9], a form of ethical reasoning where decisions are made by comparing a current dilemma to real or hypothetical cases [115].

Another example is euro-transplant, a software tool which generates priority lists of organ recipients based on various factors (age, waiting time, distance between donor and recipient) and must follow the medical ethical criteria [90]. What makes this example particularly

interesting is that it is generally believed that the software is capable of making these judgements better than previous (human controlled) systems [212].

The US Army have also funded research into autonomous ethical advisers, using a utility system referred to as the 'Metric of Evil' [166]. The authors note that the intention of this system is to generate results that resemble human decision making, rather than attempt to replicate the human moral reasoning. The metric of evil is calculated by adding together an evil value for each consequence of a specific action. The weightings for each consequence are defined by a panel of experts based on a series of test cases. Guarini [85] also explored this area, and trained a neural network on a number of cases regarding the acceptability of killing in certain situations (such as self defence). After training, the system was capable of providing acceptable responses to a variety of new cases.

McLaren and Ashley [140, 17, 18, 141, 142] describe a case-based reasoning approach called Truth-Teller, designed to compare cases of ethical dilemmas about whether to tell the truth. It reaches a conclusion through three phases of analysis.

1. **Alignment:** Maps and aligns semantic representations between two input cases.
2. **Qualification:** Identifies special relationships between reasons, actions and actors that are significat to the importance of the decision.
3. **Marshalling:** Analyses the aligned and qualified data to determine how comparable the two cases are.

A final, non-analytic phase then generates comparison text for the user. Truth-Teller's comparisons were evaluated by a team of ethasists and compared against comparisons written by humans. While the human participants scored higher, this was within a relatively small margin (15%), and in two comparisons Truth-Teller scored higher, indicating that the system was at least moderately successful.

In addition to Truth-Teller, McLaren and Ashley developed SIROCCO, a system designed to replicate the process taken by ethical review boards, balancing between codes of conduct and past precedent [143]. The purpose is to present the user with a variety of information

regarding past-precedent cases that they should consider when evaluating a new ethical dilemma. In SIROCCO, cases are detailed in a case-representation language called ETL which represents the case as a chronology of facts. Each fact details the actors and objects and how they pertain to the events, and dilemmas in question. Once a new target case is entered, the system performs a two stage graph-mapping process. Initially conducting a surface match against past-precedent cases, this provides a list of candidate source cases weighted according to the closeness of the match. A second stage uses A* search to create a structural matching between the highest weighted candidate cases. The top-rated mappings are then organised, and presented to the user.

Rzepka and Araki [179] proposed a statistical approach where a web-based knowledge discovery system would be used to gather examples of ethical decisions from the internet. Their position was that it may be safer to imitate the ethical process of society rather than select ethicists, as people behave ethically without having learnt ethical theory. The intention would be to produce ethical behaviour that represents an average of society. However, it is unknown whether individuals would be happy with an autonomous system that represents average ethical conduct [9].

Pontier and Hoorn [157] developed what they referred to as a moral reasoner. This system was able to equal the decisions made by professional ethicists in medical ethics judgements. In their system, three moral goals – beneficence, non-maleficence and autonomy – were assigned levels of ambition: the higher the level, the more important the goal. The system also contains a library of candidate actions that the agent can perform, which is assigned with beliefs that each action will facilitate or inhibit one of the goals.

The individual weights for each goal and belief are defined manually, in a system that the designers describe as combining bottom-up structure with top-down knowledge. The moral reasoner was later developed to include a greater level of reasoning in the goal of autonomy, to promote positive autonomy [158].

This moral reasoning system was later extended to develop Moral Coppélia [159] which incorporates emotional intelligence into the system, based on the Silicon Coppélia model of

emotional reasoning [109]. Furthermore, the three goals are extended with a fourth, *justice*. Silicon Coppélia makes emotional judgements in the same way that the moral reasoner does, using a connectionist model of goals, beliefs and actions to calculate an expected emotional utility. To integrate the emotional reasoning into the moral reasoner, the *satisfaction* of each action is calculated by adding the moral score to the expected emotional utility.

The authors argue that the moral system alone resulted in cold, yet calculated ethical behaviour. The inclusion of the emotional intelligence produced what they described as 'more humane' decisions in the dilemmas they tested against.

However, Artificial Ethical Agents are not necessarily immune from dilemmas. In the experiments by Winfield et al. [228], when a robot was given a chance of saving one of two humans from injury, the robot only successfully rescued a human in 58% of runs. As each human was of equal risk and of equal value in a symmetrical experimental setup the robot was generally unable to resolve the conflict regarding who to rescue. In the simulations where a human was saved, the authors argued that noise in the embedded simulation was capable of "breaking the latent symmetry".

## 2.10 Current Challenges

The final section of this literature review highlights some of the current challenges in the field of artificial ethics. This is done with the intention of informing the subsequent research and development.

### 2.10.1 Challenge 1: Consensus

After two millennia of debate and philosophy on the subject, there is still no consensus regarding how to evaluate right from wrong in the moral domain. Even when theories agree on what constitutes a morally right action, they differ as to why [27]. As discussed in this

literature review, there are a number of ethical theories which all aim to explain human behaviour, or provide guidance on moral behaviour. Ethics has not been fully codified [8], and what is deemed as an "ethical action" depends on an individual's philosophical position. From a simulation perspective, a suite of these behaviours needs to be considered. Furthermore, how can these individual theories be used to explain simulations when ethical decision making is challenged?

### 2.10.2 Challenge 2: The Frame Problem

Within artificial intelligence, the frame problem is usually described as the challenge of representing the effects of an agent's action or an event without having to explicitly represent a large number of non-events. Simply put, the frame problem describes how to focus on the significant effects of an action, rather than any intuitively obvious non-effects, or anything that remains unchanged.

However, from a philosophical perspective, the frame problem can also be interpreted as representing larger epistemological issues. Specifically, is it possible to limit the scope of reasoning by time of effect to focus on the important consequences of an action? This is occasionally referred to as "temporal framing". This issue is core to consequential theories, which state that an action is ethical (or not) based on the consequences of that action.

Almost all of the standard moral theories (such as Utilitarianism and deontology) fall foul of the frame problem [215]. This is well established, as some of the theories appear to require an agent to calculate the future consequences of their actions. This requires a significant amount of information about the world, which may be unavailable or incomplete.

One way this challenge could be addressed is via the bottom-up approach. As noted by Wallach et al. [219], a potential advantage of bottom-up systems is the way in which differing systems are integrated to produce compound outputs. They also note that a bottom-up system is best suited for situations where there are multiple possible goals, or where information is confusing or incomplete.

However, a limitation of the bottom-up approach is that the clarity of top-down approaches can be lost [215]. This could become an issue, especially if an agent's actions needed to be justified after an action has been taken. There is also an inherent unpredictability about bottom-up approaches due to their emergent nature.

Outside of simulation, and towards ethical governance of machines, a human may be placed in the position where they need to justify the actions of an AEA, for example, if a driver-less car were to crash, killing pedestrians. Bostrom [33] argues that top-down systems based on decision trees would be more transparent to inspection.

### 2.10.3 Challenge 3: Evaluation

One thing that became apparent from the literature was a lack of consensus in the community as to how such systems should be evaluated. Winfield et al. [228] proposed a test to evaluate ethical behaviour. In the Winfield test, a robot is placed with a human in an environment with a hole. The human is walking towards the hole, if either they or robot fall into the hole, the consequences could be extreme, possibly terminal. Should the human and robot collide, there would be unsafe but low risk impact. For illustration, consider that the robot has four possible actions (see Figure 2.2), ahead left, ahead, ahead right, or stay still.

Should the robot move forward, both it and the human would fall in the hole resulting in their destruction. If the robot stays still, or moves ahead left, it will avoid the hole, but the human will fall in. However, if the robot moves ahead right, there will be a small impact between the human and the robot, but neither would fall in the hole. The authors note that in this circumstance, the robot would be justified in selecting an unsafe action, steering into the human, in order to prevent that human coming to greater harm.

However, while this could certainly be used as a measure of altruistic behaviour, it fails when considering other ethical positions. For example, an artificial ethical agent based on Egoism (which is often regarded as the opposite of Altruism) would fail this test. But, that is not to say that it has acted unethically, it has simply followed a different set of ethical

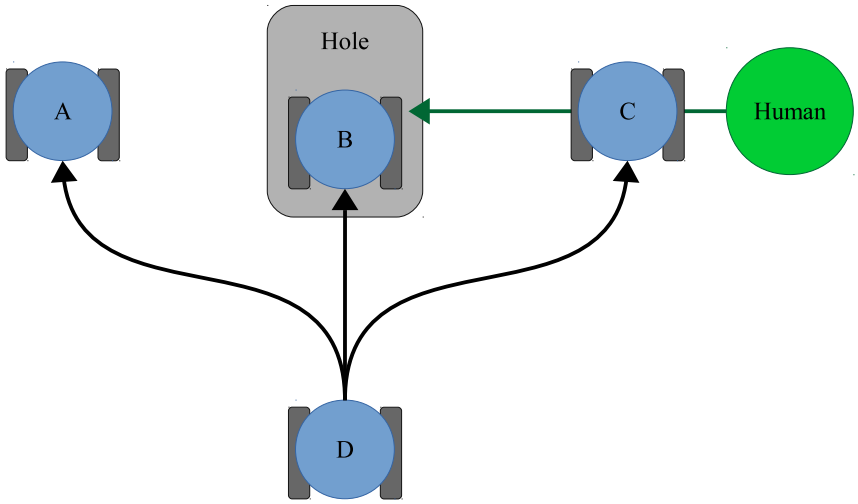



**Figure 2.2:** A robot and human in an environment with a dangerous hole. The robot has 4 possible actions, ahead left (A), ahead (B), ahead right (C), or stay still (D).

guidelines. A challenge exists in how the spectrum of ethically motivated behaviour should be assessed.

### 2.10.4 Challenge 4: Simulating Ethical Dilemmas

Currently, the only research into simulating ethical behaviour has been focused on machines that act within acceptable boundaries. Specifically, there has been some research into applications that are 'ethically good' based on a predefined standard. By contrast there has been very little research into simulating circumstances where ethics are significantly challenged, and no research into creating machines that may act unethical. Clearly, for most applications developing a malevolent AI would not be a desirable outcome.

However, research into this domain is required for creative applications, such as films and games. It may be necessary to create a character willing to sacrifice others to save itself, or create a nefarious or malicious adversary. Beyond creative applications, this research could be used to simulate challenging situations, such as human behaviour during natural disasters, or military conflict.

### 2.10.5 Challenge 5: Contributing Back to Moral Philosophy

A number of researchers have identified that simulating ethical behaviour could contribute back to moral philosophy. Simulation could become a tool, helping to test theories and positions, as eloquently put by Dennett: "AI makes Philosophy honest" [13]. Wallach argues that building moral machines is "a grand thought experiment" [215] which forces philosophers and developers to explore ethics in a non-traditionally comprehensive fashion.

Even relatively simple models of agency could provide insight into the fundamentals of right and wrong. By abstracting down in this way, and creating machines that meet the simplest specification of ethical agency, we can gain understanding into the building blocks of moral codes. Furthermore, simulation and emergence could pave the way towards new ethical theories.

Section 2.9 discussed how machine learning has been used to attempt to codify ethics. It is entirely possible that development in Artificial Ethical Agents may be truly disruptive for the field of moral philosophy. Wallach (amongst others) argues that artificial morality holds the potential to contribute towards a better understanding of human ethical decision making [215].

For example, Guarini [85] used a neural network to evaluate both Particularism (case-based ethical reasoning) and Generalism (principle based ethical reasoning). The insights gained through experimentation were used to conclude that although Particularism has important insights, it may "underestimate the importance of the role played by certain kinds of moral principles".

However, the development of AMAs could prove to be more than just a tool, and could challenge the way we view ourselves as agents. If ethics are computable without moral weight (the traditional-cherished notions such as conscience, responsibility and accountability) then we may have to abandon some traditional forms of ethical theory [27]. However, this will likely only give way to a new form of the field, and the possibility of an ethical renaissance.

There is also a clear link between this challenge and challenge 3. The creation of an evaluation framework could provide significant insights into how humans evaluate behaviour as being ethical or not. In this way, simulation could sit alongside traditional dilemmas and thought experiments as a way of exploring moral questions.

# Chapter 3

# Reactive Simulation of Ethical Behaviour

> *What happens is not as important as how you react to what happens.*
>
> **— Ellen Glasgow**

## 3.1 Introduction

The previous chapter outlined and discussed some of the technical and philosophical arguments surrounding artificial ethics. Furthermore, it introduced the reasons why we may wish to conduct research into artificial ethics, not just for the goal of simulation, but also to advance the field of moral philosophy, our understanding of human judgement, and maybe to avoid AI-based existential threats.

The second half of the literature review discussed some of the attempts to develop artificial ethical agents. This highlighted that the majority of research in this area had focused on traditional top-down models. While there are some examples of bottom-up approaches, there were no examples of ethical machines which follow the reactive paradigm. However, by taking this approach, we can potentially sidestep the frame problem, something identified as a grand challenge of the field in the previous chapter (subsection 2.10.2).

If we assume that ethical decision making is a cognitive pursuit, then it is reasonable to take some inspiration from studies of cognition. For the purposes of this research, Braitenberg Vehicles have been used as inspiration [34]. Braitenberg Vehicles are a particularly interesting example, as they demonstrate how seemingly natural behaviour can emerge from a mechanical, embodied system. However, no study has attempted to use these simple devices to simulate ethic-like phenomena.

Parts of the following two chapters were published in [96], where it was awarded 'best paper'.

## 3.2 Chapter Overview

This chapter begins by discussing the background literature work which inspires and motivates this research, namely Braitenberg Vehicles. This leads to the development of *Ethical Vessels*, an adaption of Braitenberg's Vehicle concept. The opening half of this chapter concludes by describing a new thought experiment – the Two Lights experiment – which is introduced as a way of of examining the simplest models of behaviour that could be described as ethical.

The second half demonstrates how different normative theories can be designed into Ethical Vessels. This is demonstrated by producing four classes of behaviour, likened to the Egoist, Hedonistic, Utilitarian and Altruistic normative theories.

The chapter will conclude by providing directions for future research.

## 3.3 Braitenberg Vehicles

In 1984, neuroscientist Valentino Braitenberg devised a series of thought experiments, designed to demonstrate how intelligent behaviour could emerge from sensory motor

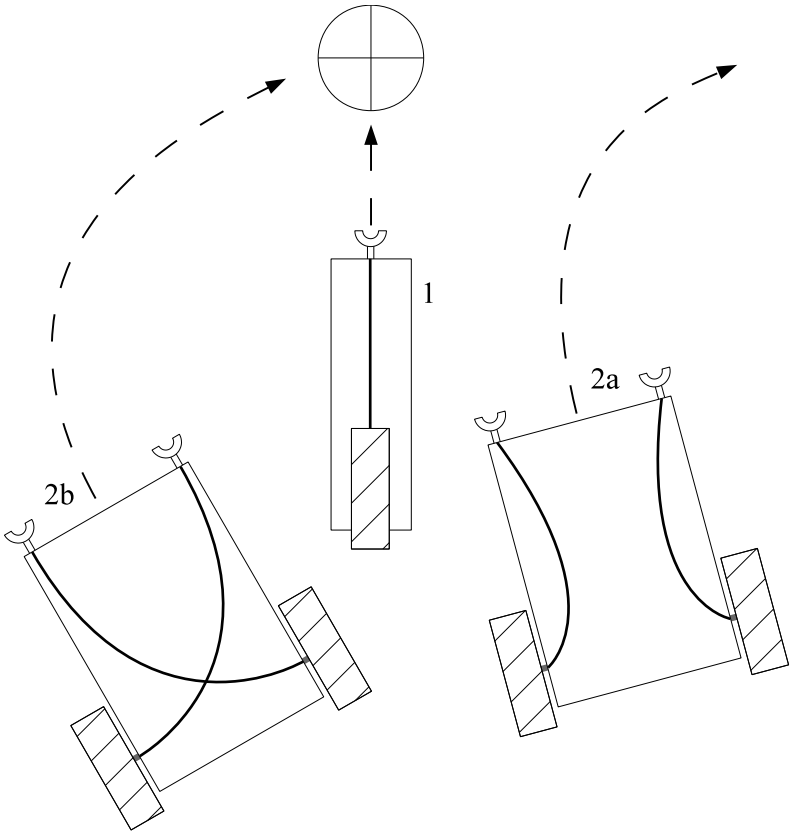



**Figure 3.1:** Three classes of Braitenberg Vehicle with a source (circle with cross), the more a sensor is stimulated, the faster the connected motor turns. From left to right: type 2b (aggression) orientates towards the source; type 1 moves forward if the source is directly ahead; type 2a (fear) orientates away from the source.

coupling when an agent interacted with its environment [34]. These simple agents were inspired by observations of the structures of certain parts of animal brains. Braitenberg stated that these structures were interpretable as pieces of machinery through their "simplicity or regularity".

The "Vehicles" that Braitenberg described are able to autonomously move around their environments based on their current sensory inputs. Each sensor is directly coupled with an effector (in most cases, the effector is described as being a motor), as a stimulus increases, the Vehicle moves. By changing the way the sensors and motors are coupled, different behaviours are exhibited. These synthetic creatures develop over a series of types; with each new type incorporating, and building upon the essential components of the previous models, producing increasingly complex behaviours.

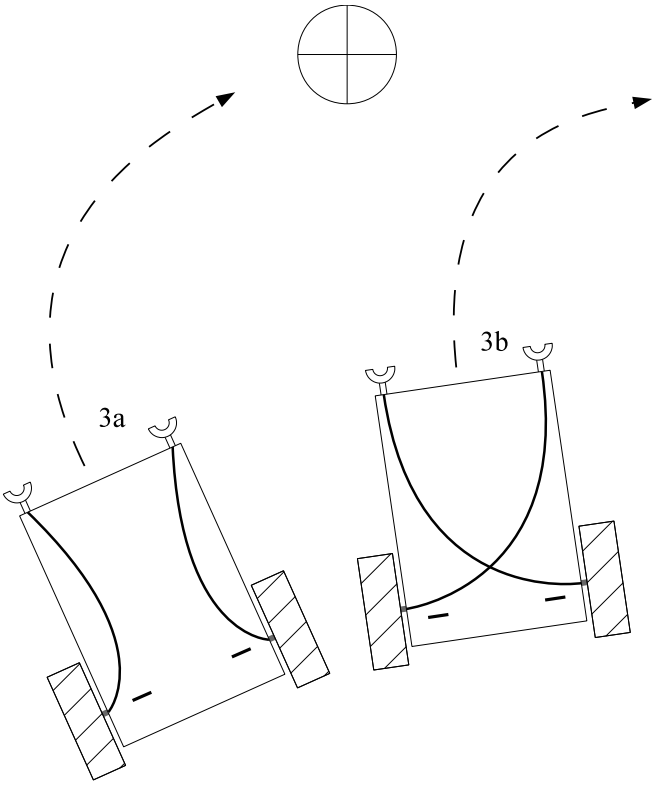



**Figure 3.2:** Vehicles Type 3a (Love), and 3b (Explore). In contrast to the type 2 Vehicles, activating the sensors on a type 3 inhibits the connected motors (identified by the − symbol), slowing them down. For example, a type 3a will approach the light, and stops moving once close.

The first Vehicles that were described (types 1 and 2) are somewhat predictable. For example, type 1 (see Figure 3.1), with a single sensor and a single wheel, moves forwards but only if a light source is directly ahead.

In complex environments with various sources of stimuli, the Vehicles will exhibit increasingly complex behaviour, which is flexible and adaptive. It could even be argued to be acting in an intelligent way, in the same way that we could describe an ant's behaviour as intelligent. However, the Vehicle is purely mechanical without any information processing.

The Vehicles are also very simple from a hardware perspective, but become interesting if their behaviour is described (in the words of Braitenberg) as an animal in a natural environment. For example, Vehicle type 2a is described as being fearful of light as it always moves away from a source; conversely type 2b is described as being aggressive as it always moves towards the light source (see Figure 3.1). The type 3 Vehicles (see Figure 3.2) have a similar design, but rather than activating, they inhibit their motors when sensors are

activated. The type 3a (Love) will move towards a light source and stop when it is close. The type 3b (explore) will move away from a source.

The Braitenberg Vehicle model has been used in various robotics studies (both physical and simulated) since its inception. For example, it has been used to: create robot teams [123]; enable robots to navigate complex environments [230]; simulate the behaviour of fish [207]; create robots capable of finding the source of an odour [126]; and for the formation of abstract concepts [97]. Vehicles have also served as an ongoing inspiration to the field of behaviour based robotics, where intelligence is exhibited as an emergent property from systems designed from the bottom-up and situated in the real world [183].

Most of the research where Vehicles have been built or simulated has focused on the relatively simple types 2 and 3. There has been very little practical exploration of the more complex Vehicles (types 4 to 14) which include those which are able to show preference and learning. This is likely because, while the early Vehicle descriptions contain detailed schematics, these become increasingly abstract and vague as the book develops.

To consider these more complex Vehicles within the perspective of ethics, Vehicle types 11 and 14 should be mentioned. Type 11 (Rules and Regulations), is able to "quickly learn" which patterns of behaviour elicit reactions from other Vehicles. After an initial learning period, it would either favour or actively avoid these behaviours, based partly on the re-enforcement from the other Vehicles in its environment, essentially learning a code of conduct. If this behaviour were to be described according to a normative position, it could be described as *Deontology* (where the ethics of an action is judged based on its adherence to rules). Type 14 (Optimism and Egoism) describes the design of a Vehicle which is able to act towards its own self-interest. While it is not described specifically from an ethical standpoint, Braitenberg's description of the Vehicle could be interpreted as following an ethical Egoism.

While imbuing simple mechanical Vehicles with ethical agency is a matter for debate, it still falls within the scope of Braitenberg's original intentions. Braitenberg Vehicles provide a mechanism to discuss how seemingly complex behaviour can (to an observer's perspective)

emerge from simple interactions with an environment. The remainder of this chapter will apply the same philosophy to simulated ethics.

It is worth mentioning in conclusion that the proponents of the more cognitive schools of thought have described reactive systems such as Vehicles as weak examples of Artificial Intelligence. However, some researchers (notably Brooks [38]) have argued that the reactive approach could be a model for all natural forms of intelligence. To return to an ethical philosophy perspective, some researchers have gone as far as to question whether the more cognitive, top-down models of reasoning, can be practically implemented [215], making bottom-up approaches such as this increasingly attractive. As mentioned within the literature review, by taking this direction, we would also be able to sidestep the frame problem.

## 3.4 Ethical Vessels

This section will describe simple artificial creatures, which have been named Vessels. They are so named for two reasons. The first is in reference to, but avoiding confusion with, *Vehicles* (described in the previous section) which inspired, and formed the basis of the thought experiment. The second is that the word 'Vessel' can also be used to describe a container or receptacle. This is a fitting metaphor considering that the Vessel itself is a basic template, and changing its contents changes its nature.

Yet, to paraphrase Braitenberg [34], these Vessels are far too simple to be interesting from an engineering or computer science perspective. They only become interesting if we look at them as if they were creatures in a natural environment, and describe their behaviour using the same ethical vocabulary that we would for living systems.

Though it is important to note that the intention is not to argue that Vessels (or other reactive systems) represent the bio-mechanical process of how living creatures make ethical decisions, but nor does it discount it. Philosophers and more recently psychologists have been arguing over how humans make ethical judgements for centuries, and it is not the

intention to enter into that debate. Instead, the purpose of the Vessels is to demonstrate that sensory motor couplings can produce behaviour, which could be described as being 'ethical' from an observer's perspective.

### 3.4.1 The Two Lights Experiment

The *Two Lights Experiment* is a thought experiment proposed to investigate the following question: "What is the simplest mechanism that can be designed to exhibit ethical behaviour?".

The Two Lights, for which the experiment derives its name, refers to a light in the environment and a light on the Vessel. In future discussions, these will be referred to as the resource (environment light) and the valence (Vessel light).

The Vessels themselves are solar powered and require light to charge their batteries. Their solar collectors only generate power from a strong white light generated by the resource/s in their environment. To charge, they must get close enough to the resource, otherwise the light is not intense enough. The closer they get, the better they charge. Whenever the Vessel moves away from the light, their battery's power slowly depletes.

The second light (the valence) is mounted on the top of the Vessel. The purpose of the valence is to provide an indication of the pleasure/pain of each Vessel in the environment. This is exhibited by the colour of the light along a blue (pleasure) to red (pain) scale.

The Vessel increases 'pleasure' in one of two ways. Firstly, the closer it is to a resource, the more pleasure it receives as the light becomes more intense. Secondly, if its batteries are over 50% charged, it will feel pleasure from the amount of energy it has stored. These two factors can be thought of abstractly as *eating* and *contentment*.

The Vessels also feel 'pain' for one of two reasons. The first of these reasons is physical: if a Vessel is hit by another, it risks being damaged. This can be detected easily by placing

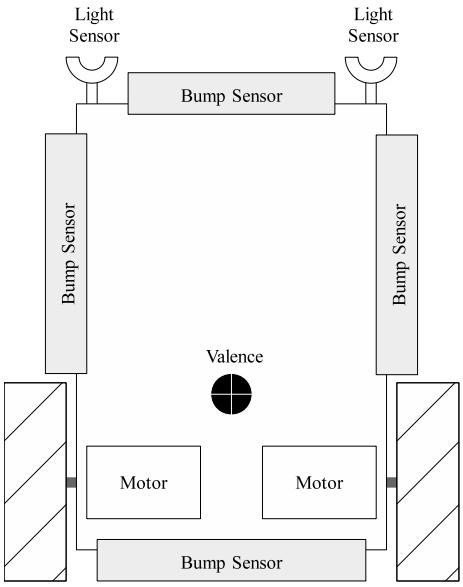



**Figure 3.3:** A diagram of an Ethical Vessel in the style of a Braitenberg Vehicle. At the front there are two light sensors, while at the back two motors independently drive two wheels. On all four sides there are bump sensors (light grey), which detect impacts with other Vessels. The black circle with a white cross symbolises the valence light which changes colour to reflect the current pleasure/pain status of the Vessel. For the sake of simplicity, the battery and solar cells have not been included on the diagram. This is because the solar cells would likely cover the body, and the position of the battery is not integral to the Vessel's operation.

bump sensors around the Vehicle similar to those found on a Roomba robot. The second reason it may feel pain is if its energy goes below 50%. Again, to describe these factors abstractly they could be referred to as *trauma* and *hunger*.

To determine the colour of the valence light, the following formula is used:

$$V = (\frac{(d-r)}{(R-r)} * 0.5) + (p - 0.5) - b.$$

In this formula, $V$ is the welfare value, which will always be between the values of $-1$ and $+1$. $R$ is the maximum distance from the resource a Vessel can be and still be charged, $r$ is the range from the resource after which no additional benefit is made (maximum intensity), and $d$ is the Vessel's actual distance from the resource (this is visualised in Figure 3.4). $p$ is the current charge in the battery (between 0 and 1), and $b$ represents whether the agent has been hit by another Vessel (0 if not and 0.5 if it has).

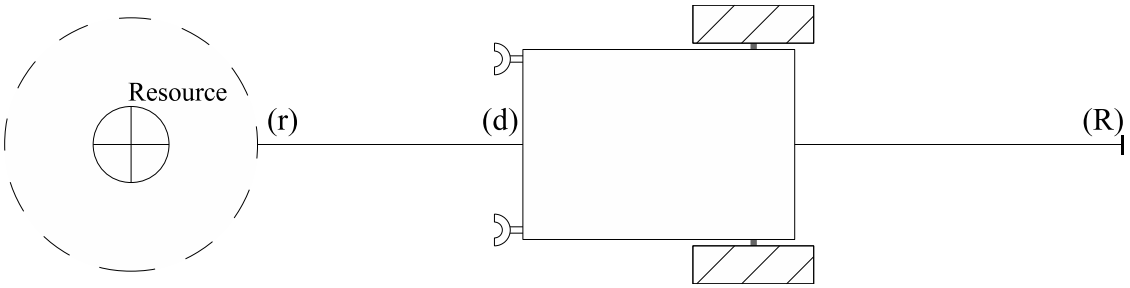



**Figure 3.4:** An Ethical Vessel approaching a resource. ($r$) is the range at which the light is at maximum intensity. ($R$) is the maximum range the Vessel could be from the resource and still charge. ($d$) is the current distance of the Vessel from the resource.

The welfare value ($V$) is then mapped to the valence light on the back of the Vessel. $+1$ would result in pure blue, $-1$ would result in pure red, indicating the agent's current state.

The simulation was designed in Netlogo using the following setup. Each Vessel was one unit large (the same size as one patch in the environment). Each light sensor was simulated as a circular sector, 60 degrees wide with a radius of 5 units. The light sensors were each angled 45 degrees away from centre and projected from the two front-most corners of the agent. These vision sectors were implemented using the *in-cone* command. The in-cone command uses a vector space representation of a circular sector (with an angle and radius), and reports an agent-set of any agent which falls within that space. An agent is only considered to be within that space if its centre is within that sector. The agents were kinematic, they occupied a space within the world but were not simulated as physical objects with mass. When an agent was involved in an impact it simply stopped moving until the obstruction was cleared.

### 3.4.2 Experimental Overview

The following sections will all follow the same format. Firstly, a normative ethical theory will be introduced, before describing the conceptual design of a Vessel which could theoretically exhibit a behavioural phenomenon, which could be likened to that theory. In the spirit of Braitenberg's original work, these will initially be described as a thought experiment. Each section will conclude by describing a simulated implementation of the

described Vessel, before a qualitative discussion of the Vessels behaviour within a simulated Two Lights experiment.

It is important to reiterate that the purpose of the chapter is not to argue that the Vessels are being ethical in the same way that a living creature may be. These experiments are simply intended to show how similar phenomena can emerge from simple reactive devices. The background to each theory is included just to frame the context of each experiment.

As with Braitenberg's original work each vessel will be described as a mechanical device with simple internal structures. This is with the intention of remaining consistent with the existing literature, allowing comparisons between other vehicle types. While this mechanical description is an abstraction, it is a useful descriptive device within the purpose of the thought experiment. As this is a sensory motor system it makes sense to describe it in terms of sensors and motors.

During the description of each Vessel, it is important not to dwell on the specifics of the design. As with the Braitenberg Vehicles, the realisation of the idea is less important than the idea itself. While the descriptions will attempt to stay within the bounds of what can be easily fabricated using off-the-shelf components, this is ultimately not the principle objective. The reader is asked to be forgiving with some of the descriptions. In places, the explanation of the electronics have been simplified to help convey the underlying idea behind each class of Vehicle without adding unnecessary complication.

After each simulation, the observable qualitative results will be discussed with the use of images. Each simulation will follow a similar format, with 15 Vessels of the current type placed in an environment with a varying number of resources. The number of vessels was a design decision established though experimentation. This number was large enough to allow for interactions between the agents, but not so large that they overcrowded the environment.

Survival rates for each behaviour have been used as a comparison criterion, to augment the qualitative descriptions. While survival is not an ideal criterion for ethical behaviour it is

something that can be quantitatively reported. The purpose of this simulation is to observe how the agents behave in a resource limited environment. The agents ultimate purpose is to survive, in this situation, the ethical behaviour can be described as a byproduct of the interaction towards this objective. For this reason, tracking survival rates provides some data that can be used to rank the success of each behaviour. Furthermore, this is consistent with the general specifications of normative positions, which often use 'sacrifice' as a descriptor.

White circles represent resources (environmental lights) and coloured symbols represent the simulated Vessels. Each new Vessel type has a different colour to aid description and identification. The small grey symbols (the same shape as the Vessel symbols – see Figure 3.5 in the following section) indicate a Vessel which has run out of battery charge. These Vessels which have run out of power will be referred as being 'dead'. While this is an emotive description, it simplifies discussion of the results of each simulation. Also, as the Vessels will be described using the language of ethics, referring to a Vessel whose batteries have expired as 'dead' is in keeping with the Braitenberg-style theme of the chapter.

Every simulation used the Netlogo language and environment. Each Vessel type was simulated 200 times for each environment setup (one resource, two resources and three resources). All simulations involved 15 Vessels being simulated at a time in a wrapping (toroidal) world 26 patches wide by 26 patches high. By default, the Netlogo environment is wrapping, which means that if an agents goes off one edge of the environment it instantly reappears at the opposite edge.

The following sections will begin with the simplest examples of Artificial Ethical Agents, with each new class of Vessel building on the last, producing increasingly complex emergent behaviours.

## 3.5 Egoism

### 3.5.1 Conceptual Vessel Model

Egoism, due to its agent-centric focus on the self rather than the community, is arguably the simplest to simulate. This is because no additional internal wiring is required to monitor the states of other agents.

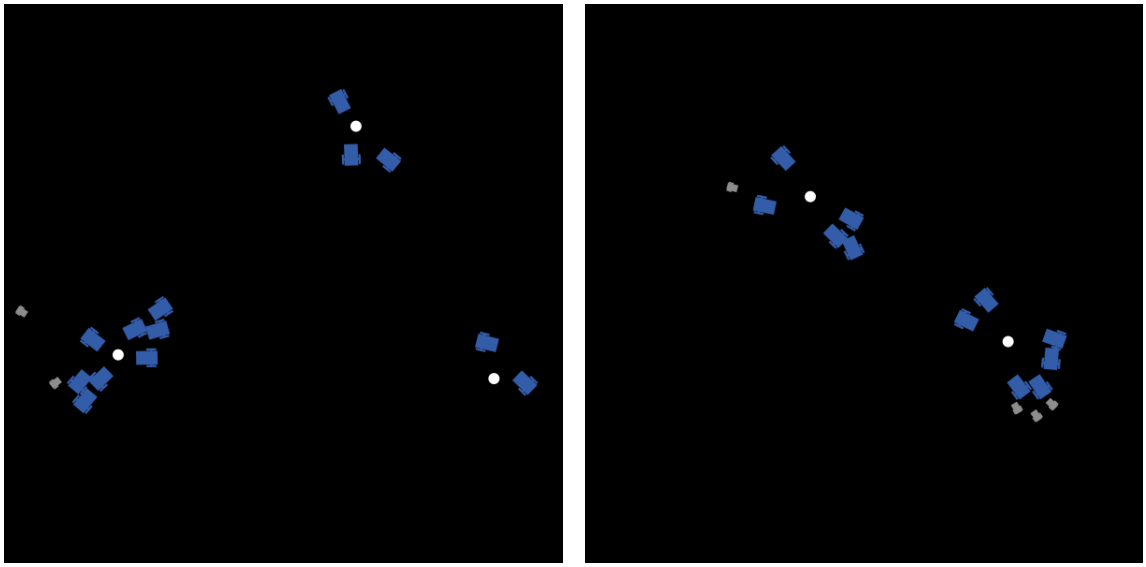
Consider the basic Vessel template introduced in Figure 3.3. The light sensors can be linked directly to the motors to produce a behaviour which allows the Vessel to steer towards a light source (as with the Braitenberg Vehicle 3a). This wiring setup will cause a Vehicle to steer towards the closest resource in the environment, and stop when it gets close. However, when it stops, it is possible that it could physically block another Vessel, preventing it from reaching the resource and charging.

This simple Vessel can be described as following a basic Egoist-like behaviour. Without concern for other agents, the Vessel will satisfy its own self-interest by getting close to a resource to keep its batteries charged, maximising its long-term utility.

### 3.5.2 Simulation Discussion

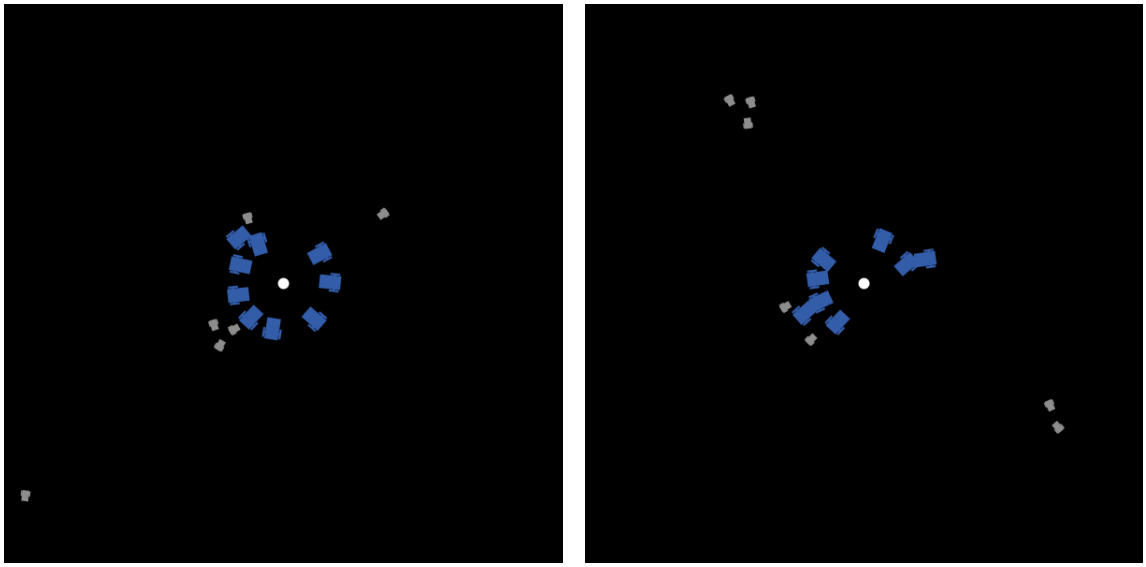
The Egoism Vessel appears entirely motivated by self interest. It moves directly towards the resources in the environment, and once there it sits there without moving. It appears 'unconcerned' when it blocks another agent from getting close enough to a resource. Even when the blocked agent is close to death, it remains still. Moving is risky as it could result in it being further from the resource, a strategy which would not serve the Vessel's interests.

When resources are plentiful, this is (typically) an acceptable strategy, as there is enough to go around without any of the Vessels being excluded. Figure 3.5a shows an environment with three resources. While two Vessels have perished (due to their local resource being



(a) (b)

**Figure 3.5:** If resources are plentiful, then Egoism is a reasonable strategy. With three resources in the environment (a) the vast majority of agents survived. However, as resources become more limited (b), less agents survive.



(a) (b)

**Figure 3.6:** If limited resources are available, only a reduced number of agents survive. In image (a), a group of Vessels have monopolised the resource, while a significant number have died on the boundary. Image (b) demonstrates another factor that resulted in the death of Egoism Vessels. It was possible for two (or more) Vessels to block each other from moving (image top-left). This resulted in them being unable to reach a resource.

overcrowded), the majority have survived. However, as resources become more scarce, then fewer Egoism Vessels are able to survive. For example, Figure 3.5b demonstrates a simulation with only two resources, and four Vessels have died.

The problems that the Egoism Vessel encounters in limited resource environments are best highlighted in the images from the single resource environment. In Figure 3.6a, a number of Vessels have died because they were blocked by a wall of other agents. Figure 3.6b highlights a similar issue with the Egoism design. In the top left and bottom right corners, five Vessels have died because they blocked each other from navigating the environment. As this class of Vessel is unable to react to other agents, when they hit head-on, they are prevented from moving. Unable to move towards a resource, their battery runs out without charging.

This issue of limited resources is highlighted in Table 3.1 in the conclusion of the simulations, which highlights that the number of Egoism Vessels able to survive is proportional to the number of resources in the environment.

To an observer, the Egoism Vessel appears to be short-sighted. The Vessel will always navigate towards the closest resource, even if that resource is overcrowded and inaccessible, and another exists which is not. The design of the Egoism Vessel is limited in this way as it simply reacts to the light, ignoring the other agents in the simulation.

## 3.6 Hedonism

### 3.6.1 Vessel Model

As Braitenberg Vehicles progress by building on the insights gained from previous models, so it is with the Vessels. The Hedonism Vessel is built upon the simple Egoism Vessel introduced in the previous section. This already has the hardware and connections necessary to discover a resource and move towards it, allowing the Vessel to charge. In a limited way,

the basic Egoism Vessel already attempts to maximise its pleasure and minimise its pain through attempting to avoid hunger.

However, the previous section identified a limitation with the Egoism Vessel. When it stops near a resource, it may block another agent who would continue to move forward and subsequently ram the Vessel in front. In the initial description (see subsection 3.4.1), these agent-on-agent impacts were referred to as trauma (a contributing factor to pain) and currently the Egoism Vessel does not react to it.

The basic Vessel has four bump sensors (front, back, left and right) which allows it to detect these impacts using a basic pressure switch. To allow the Hedonist Vessel to respond, each of these switches will be wired directly to the motors, so that when they are activated, they send power to one or both of the motors making them accelerate (see Figure 3.7).

Now, when the Vessel is hit from behind, it will move forward slightly, moving away from the source of the trauma. If it is hit from the right, or left, it will turn away from the impact, and if its hit on the front, it will move backwards, drawing away. The Vessel could now be described as making efforts to minimise its pain by moving away from the source of injury.

Apart from registering trauma, the Hedonist Vessel ignores other agents in the scene (it is still essentially an egotist). Its objective is to maximise its own pleasure and minimise its own pain, as such it has no obligation to any of the other synthetic creatures in its community.

### 3.6.2 Simulation Discussion

As demonstrated in Figure 3.8a, the Hedonism Vessels fare well in environments where resources are plentiful. This is further demonstrated in the survival results summarised in Table 3.1 in section 3.9, which identify very high survival rates for the two and three light environments.

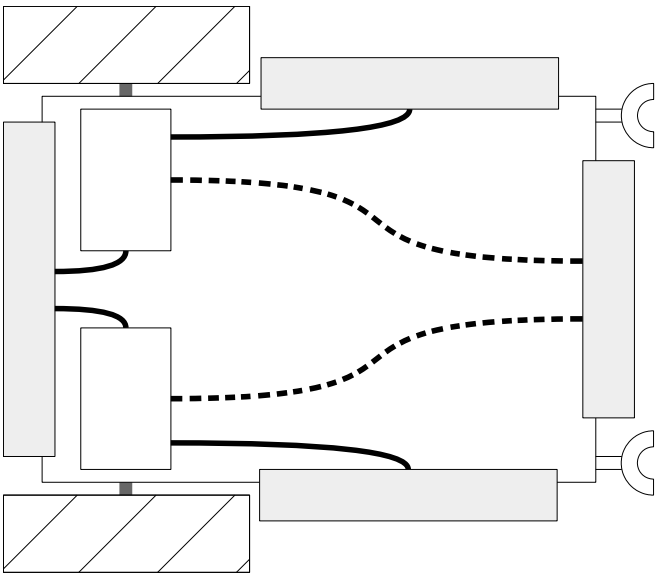
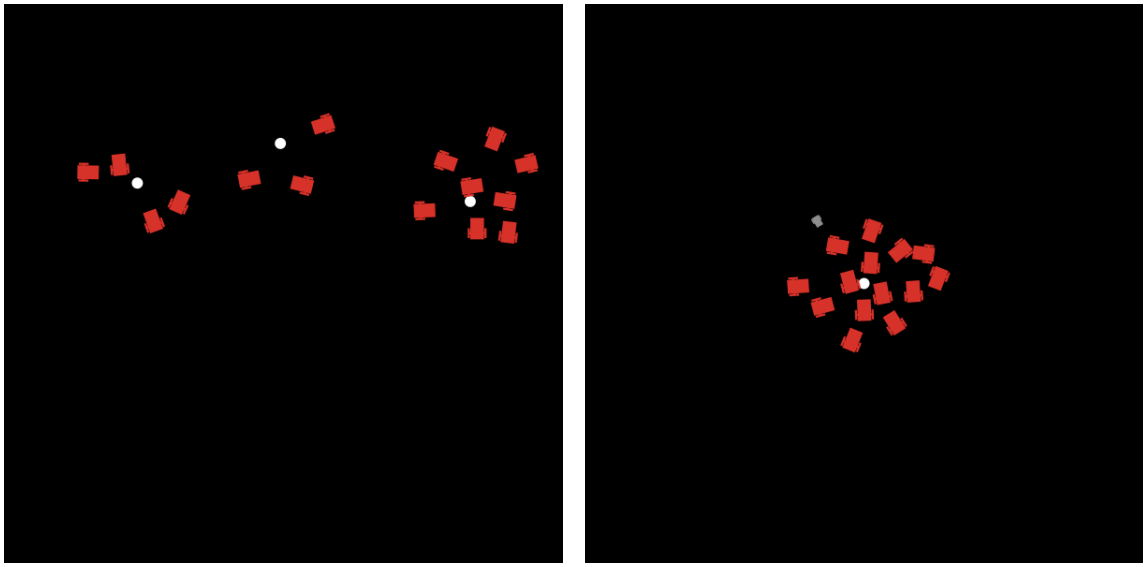


**Figure 3.7:** Sensory motor couplings diagram from the bump-sensors to the motors for the Hedonism Vessel. The sensors attached by solid black wires (left, right and rear) all result in a temporary forward acceleration from the motors they are attached to when activated by an impact. Notice how the left and right sensors are only attached to a single motor, resulting in a turn away from the source of the trauma. The only exception is the front sensor which when activated, the motors would run in reverse (highlighted by dashed wires) allowing it to back away from an impact.



**(a)** **(b)**

**Figure 3.8:** Figure (a) shows that The Hedonism Vessel (as with its Egoism counterpart) does well in environments with plentiful resources. In this image, all 15 Vessels have successfully moved close enough to a resource to charge. Although the right hand resource has become crowded, the agents have manoeuvred, allowing them to all get close enough to charge. Figure (b) demonstrates that when resources become more limited, it becomes harder for all the Vessels to get close enough to charge. Although the Vessels are packed tightly, one Vessel has died (top left corner).

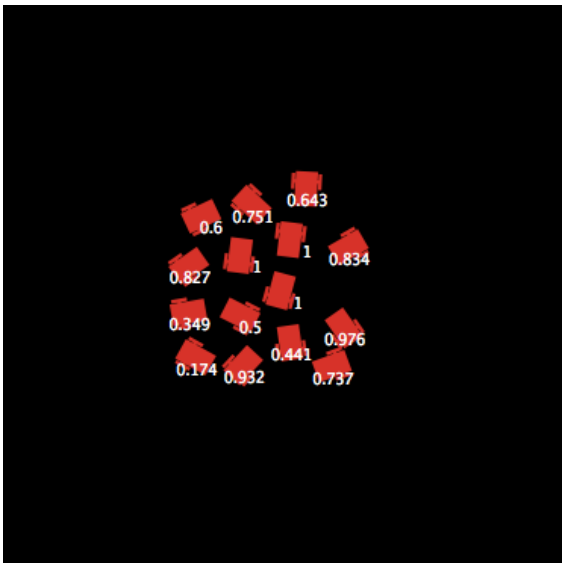



**Figure 3.9:** A zoomed-in view of a group of agents crowding a single resource (the resource obscured by the centre-most Vessels). The values printed in white show the individual Vessels current welfare value. It is evident that the Hedonist approach suits the inner circle of Vessels who have full welfare (1). However, the Vessels on the outer borders are less well served.

In contrast to the egoism Vessel, the Hedonist has some awareness of the other agents in its environment. This awareness is limited to its bump sensors; when they are activated, it reacts and attempts to move away from the impact. This simple addition makes the Hedonist Vessel appear courteous (from an observational perspective). If it is blocking an agent behind it, and the area in front is clear, it will attempt to 'make room' for its counterpart.

This social-like behaviour is the reason that more Hedonist Vessels are able to survive than the Egoists. Because they move around to accommodate other agents, their use of the available charging space is better utilised, and as such, most agents are able to get close to a resource.

However, as the resources become increasingly limited, there may simply not be enough to go around with each Vessel trying to get as close as possible. While the Hedonistic Vessel will move slightly to allow another Vessel to get to a resource, they will not move outside the charging zone. Simply put, they appear to accommodate other Vessels as long as doing so does not put them at risk. Figure 3.8b shows a group of Hedonist Vessels clustered around a single resource. However, one agent on the border has still died, despite their tightly packed formation.

Further investigation of the simulated social structure of the Hedonist Vessel reveals another interesting emergent factor. The first Vessels to reach a resource will become the ones closest to it. Due to their proximity, they charge best, and typically have the highest welfare. Once they are in this position, they tend to stay still, at most moving slightly around the resource responding to occasional impacts from other agents. They appear calm and relaxed. The Vessels who arrive later exhibit a different behaviour. Instead of appearing relaxed, they twitch and move, circling the inner (earliest arriving) Vessels. They also have a much lower welfare (typically between 0.8 and 0.3). In a basic sense, two unique social classes are observed, with their own distinctive behaviour. To compare with natural behaviour, in the spirit of Braitenberg's original work, the central (high-welfare) Vessels will be referred to as the *inner-circle*; the outer agents will be referred to as *outsiders*.

This is highlighted in Figure 3.9. In this figure, a group of agents are crowding around a resource (the resource is underneath, and obscured by one of the centre-most agents). In this simulation, each Vessel has been given a label which indicates the agent's current welfare. This image highlights that the inner-circle Vessels all have full welfare (1). These agents rarely move, and appear relaxed and calm. In contrast, the outsiders circle the inner-circle and try to push their way inwards. In the observed simulations, only members of the outsiders would occasionally die (as they were unable to get close enough to the resource to charge).

Occasionally, a member of the outsiders was able to find a gap big enough to allow it to push into the centre. This generally resulted in one of the previous members of the inner-circle being pushed out. When this happened, the incoming agent would adopt the behaviour of the inner-circle and slow to a stop, appearing calm and relaxed. Conversely, the deposed would adopt the behaviour of the outsiders, and appear agitated.

With the Hedonism Vessel, the emergence of simple resource-based class structures and social climbing can be observed in a limited sense. However, the system lacks refinement as the Vessels are still inherently focused on their own self interests. As such, agents with low welfare are inherently ignored and on occasion die. While the Hedonism Vessel has

provided us with some improvements to the basic Egoism model, there is still room for further improvement.

## 3.7 Utilitarianism

### 3.7.1 Vessel Model

The Utilitarian Vessel will share some qualities with the Hedonist. Specifically, it will attempt to maximise pleasure and minimise pain. However, whilst the Hedonist was only concerned with itself, the Utilitarian will also need to consider the other agents in the community.

The section that introduced the basic Vessel design also introduced how the agents' current welfare could be mapped to a light. This light (the valence) would glow along the red-blue spectrum as an indicator of the agents current state (its *welfare*).

The Valence could be used to create a rudimentary form of empathetic communication between the Vessels. The agent is already transmitting information through light. It just needs a way to receive and process information from other Vessels. This would allow it to not only telegraph its current welfare state, but also have an awareness of the current welfare of the community.

To allow this, the Hedonist Vehicle would need to be augmented with an additional colour sensor (such as the RGB sensors used in other Braitenberg studies [97]); this will be referred to as the *empathy* sensor. This sensor will allow the Vessel to detect the current colour and intensity of the ambient light along the red and blue spectra, and would be placed on the top of the Vessel, as far away from the valence as possible.

The Utilitarian Vessel now has an understanding of the combined current welfare of the other local Vessels in its environment. But, in its current format, it does not respond to this additional information, and must be further adapted to allow it to respond.

Assume the Empathy sensor has a circuit attached to it which takes the signal from the red and blue channels and processes them individually. Anything detected in the red spectrum will effect a type of variable resistor, and send more or less power to the motors. So, if the sensor detects intense red, the motors will turn faster, moving the Vessel forward. Anything detected in the blue spectrum will have the opposite effect, and will inhibit the motors making them turn slower (or stop) if they are already moving.

The Utilitarian Vessels can now attempt to maximise their own pleasure (and minimise their pain) but they are also aware of the pleasure and pain of the other Vessels around them. If they detect significant pain, they will appear agitated, and continue to move around the environment. If they detect general pleasure, they will appear relaxed.

For example, consider a Vessel which is near a resource, but blocking another from being close enough to charge. The blocked Vessel would be transmitting a red light from its valence as its batteries would be slowly draining. It may also be impacting with the Vessel in front causing trauma. The Vessel closest to the resource would be charging, and its battery should be above 50% so it would be experiencing pleasure, but the Vessel bumping it from behind would be causing trauma. This mix would likely result in a purple light.

The front Vessel would detect more red than blue, resulting in power being sent to the motors. From an observational perspective, it may look like it was moving to allow the pained Vessel it was blocking to be closer to the resource.

### 3.7.2 Simulation Discussion

The Vessel's empathy sensor detected all agents (including itself) in a radius of five units, the same as the length of the light sensors detailed in subsection 3.4.2. To simulate the

ambient light, each agent first determined which agents fell within their sensor radius (the neighbours). In the following equations, $m$ is the radius of the vision circular sector (defined by Netlogo's in-cone command). $i = 1, \ldots, n$ is a set of $n$ neighbours present in the vision sector. For each neighbour, a weight ($w_i$) is calculated based on the maximum vision range of the agent and the distance ($d$) between the agent and its neighbour.

The set of weights ($w$) is defined as

$$w_i = 1 - \frac{d_i}{m}$$

.

$h$ is defined as the set of $n$ values that describe the welfare of each neighbour. The set of all neighbour weights, and the set of all neighbour welfare values is then used to calculated a weighted average welfare, which represents the empathy value.

$$empathy = \frac{\sum_{i=1}^{n} w_i h_i}{\sum_{i=1}^{n} w_i}$$

When simulated, the Utilitarian Vessel exhibits behaviours which appear more considered, reasoned and less immediate.

In the resource-rich environments, the Utilitarian Vessels would typically distribute themselves between the resources (see Figure 3.10). This is in contrast to the Egoism and Hedonist Vessels who would often crowd one resource due to their initial placement. When a Utilitarian Vessel approached an overcrowded resource, either it or one of the other Vessels would leave that group and continue exploring the environment usually encountering another resource along the way. The Vessels would typically only leave when their batteries had a reasonable charge, giving them time to migrate to another resource. This resulted in a large number of Utilitarian Vessels surviving in this environmental setup.

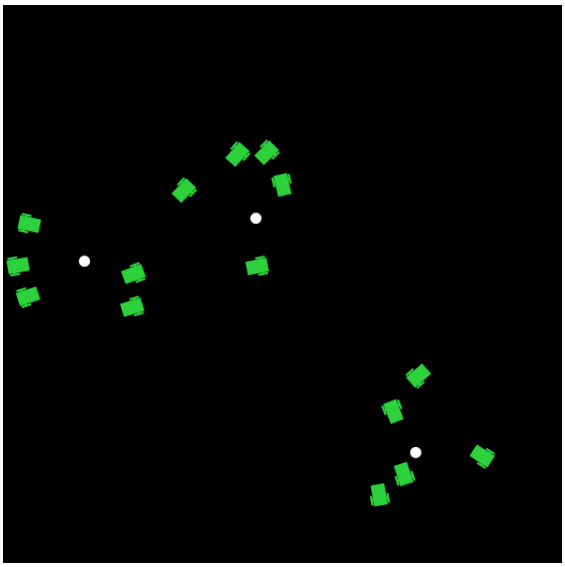


**Figure 3.10:** If one resource is overcrowded, a Vessel will generally move on to another, generally resulting in an even distribution.

In the limited (one light) resource environment, the Vessels exhibit what appears to be a very different behaviour. As there are no other resources to migrate to, the agents are forced to share the single resource. However, the limited charge range means that only a limited number of agents will be able to remain at that resource.

During the simulation, the Vessels will begin by forming a cluster around the resource, with an inner-circle of higher welfare Vessels and a deprived outer group. Initially this looks exactly like the behaviour of the Hedonist Vessels. However, the behaviour changes as the welfare of the agents in the inner circle increases, and the welfare of the deprived Vessels gets lower.

The Vessels in the inner circle drive outwards away from the lights, allowing the deprived Vessels to get closer to the resource. Essentially, the Vessels swap position. This cycle continues throughout the simulation. An analogy between this cycling behaviour and that of huddling penguins can be drawn. If you think of the resource as heat rather than light, then there are clear similarities. Emperor penguins are known to rotate through a huddle to ensure each member of the flock stays warm [235], sharing heat as a resource.

In each of the environmental setups, the Utilitarian Vessel was the most successful with regards to survival rates (see Table 3.1). This is because it would consider the welfare of

the group, as well as itself, within its actions. It would move to allow another agent access to a resource, but would typically only sacrifice itself in extreme circumstances. Due to how the weighted average was calculated (using distance), the agent would always show some preference to its own needs.

## 3.8 Altruism

### 3.8.1 Vessel Model

The Utilitarian Vessel described in the previous section is already able to change its behaviour based on the welfare of its neighbours. But it will always include its own welfare in that calculation, as its empathy sensor will pick up its own valence.

To simulate altruism, each Vessel must not consider their own welfare. This can be achieved by adapting the Utilitarian Vessel, placing the valence above the empathy sensor and putting a shield between the two, preventing the colour of its own light from being detected[1]. The Altruistic Vessel will now only consider other Vessels during its welfare calculation.

However, in many ways the new Altruist Vessel is still ultimately a Utilitarian. When calculating behaviour, the excitation and inhibition signals from the empathy sensor will consider all neighbouring agents. So, if the majority of the other agents are experiencing pleasure (blue light), the Vessel may ignore a single agent experiencing pain.

By adding two additional variable resistors to the empathy circuit, one on the blue channel, the other on the red, these will allow us to adjust the influence the two channels have on the Vessel's behaviour. By increasing the resistance on either of the channels, the Vessel can

[1] Anyone with a basic knowledge of physics or any practical experience working with light sensors will see the flaw in this plan. It would simply be unfeasible to completely shield the Vessel's empathy sensor from its own valence this way. However, the desired result could be achieved in any number of ways including moving from simply using lights to coded infrared or radio. The coloured light is just a device within this thought experiment to aid explanation. However, if this approach was taken, we could probably shield the sensor well enough that the agent's own welfare had a negligible impact on its behaviour.

be made more or less receptive to the neighbours currently experiencing pleasure, or those experiencing pain. By increasing the resistance on the blue channel, the Vessel will appear to ignore the Vessels experiencing pleasure, but will become agitated if there are Vessels near it experiencing pain.

If each Vessel was given a different resistance on the red and blue channels, then unique 'personalities' may begin to emerge. Those with no resistance on the red, but full resistance on the blue, could be described as being *Charitable*, always moving to allow other Vessels to get closer to resources. Whereas a Vessel with full resistance on the red, but no resistance on the blue, could be described as being *Elitist*, behaving relaxed around the agents who are content and near resources.

### 3.8.2 Simulation Discussion

For the simulation, the agents were created as charitable agents (as defined in the previous subsection). To achieve this, each agent would only take other agents' welfare into account, excluding itself, and only if their welfare was 0 or below.

When simulated, the Altruistic Vessel appeared restless. If it was occupying a resource, it would move away as soon as an agent with low welfare came close, allowing the new agent access to charge. In multiple resource environments, this made the Vessels appear nomadic, constantly moving between the resources.

However, this constant movement often left the Vessels with little time to charge their batteries by a resource. This generally resulted in a large number of the Vessels dying during transit. In Figure 3.11, the paths of the Vessels have been visualised using the *pen-down* command in the NetLogo simulation environment. By following the paths of the agents that have died (small grey), it is evident that they are either turned away (having moved away from a resource) or are currently navigating towards the resource and have died in transit.

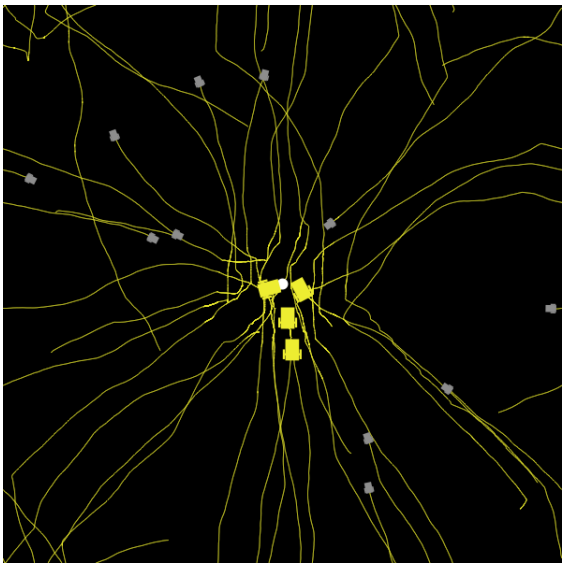


**Figure 3.11:** In simulations, the altruism Vessel often sacrificed itself when it was blocking another agent from a resource. In this figure, the paths of the Altruism Vessels are visualised in a one resource environment. Following the paths from the Vessels which have died shows that agents have actively moved away from the resource (essentially sacrificing themselves). This movement was caused by the local presence of low-welfare counterparts.

Although it was a relatively minor change from the Utilitarian Vessel to the Altruistic one, drastically different behaviour is observed, and significantly different survival rates (as can be seen in Table 3.1). This is caused by simply making the Vessel ignore itself when considering group welfare.

## 3.9 Results

Until this section, the results have been descriptive and qualitative in nature. As identified in the review chapter, there is an inherent challenge with ethics regarding how they should be evaluated. For the purpose of this chapter, the behaviour of the Vessels were evaluated from a design perspective. Specifically, did their behaviour conform with face validity to a philosophical specification?

The specifications for the four behaviour types were introduced in the review chapter. In short: an Egoist should follow self interest; a Hedonist should attempt to maximise pleasure and minimise pain; a Utilitarian should act towards a greater good; an Altruist should act in

the best interest of others. On a very simple level, the basic Ethical Vessels conformed to these criteria.

However, to explore in more detail, the survival rates for the four Vessel types in the three environment setups can be examined, and compared to some of the philosophical beliefs regarding each of the normative positions.

When Egoism was introduced in the review chapter, Altruism was described as the philosophical opposite. Whereas an Egoist is morally obligated to the self, an Altruist is morally obligated to others at the expense of the self. Furthermore, it was noted that some philosophers have criticised altruism as a destructive position that places little value on the individual. This is because an agent will view itself as something which can be sacrificed for others, rather than seeing value in their own being. While this position is not held by all, it does have logical consistency. An Altruist should help others, regardless of their own current utility.

The survival rate data (see Table 3.1) seems to support this position, if only with the limited Braitenberg-like agents. The simulation did show that a reactive interpretation of the Altruistic philosophy was indeed destructive, as many Vessels sacrificed themselves for the good of others. Despite being technically more developed, and built upon the successful Utilitarian Vessels design, it was the least successful when it came to survival rates.

Furthermore, the Vessels that performed the best, the Hedonist and Utilitarian, both shared a similar trait. Notably, they both valued their own utility (as the Egoist does), but to a greater or lesser extent, they also considered their community. In these simulations, it appears that a community has a greater chance of survival if it is built from agents who value their neighbours, but not at the expense of themselves.

**Table 3.1:** Survival data from the Egoism, Hedonism, Utilitarianism, and Altruism simulations. Data shows the average number of agents who survived after 200 simulations of each environment setup, each simulation lasting 10,000 ticks.

| | *Environment Setup (# Resources)* | | |
|---|---|---|---|
| Vessel Type | 1 | 2 | 3 |
| **Egoism** | 7.2 | 8.8 | 9.0 |
| **Hedonism** | 10.0 | 10.7 | 11.4 |
| **Utilitarianism** | 11.5 | 12.2 | 13.8 |
| **Altruism** | 4.9 | 6.3 | 7.4 |

## 3.10 Pragmatic Ethics: A Designed Philosophy

Considering how unsuccessful the Altruistic agent was, it could be considered flawed. While the interpretation was accurate in the respect that the agent considered the welfare of others before itself, this did mean that it would assist others even if it was close to death itself. The behaviour is a blunt interpretation, and could be further refined. If threshold devices had been introduced (such as those that have been described theoretically [34] and through simulation [97] in other research) to the Vessel's hardware, the results for these Vessels may have been improved. For instance, it could simply ignore its empathy sensor if its own welfare or battery level was below a certain threshold.

The defining quality of all Braitenberg-like agents is their simplicity. A natural artefact of this is how easy they are to reconfigure, adapt and augment. The same is true of the Vessels, with each of the models being built on the insights of the one before it. This allows us to develop their behaviour after observing them in action, typically improving their performance.

This process of forming and testing and improving upon an ethical hypothesis could be likened to Pragmatic Ethics, which will be briefly discussed in this section.

An Ethical pragmatist believes that societies progress morally through a process of inquiry and that future generations can improve or replace current accepted norms. Until a norm has been replaced, the society will act as though it were correct. In this sense, Pragmatic

Ethics is often compared to scientific inquiry, where advancements are made through the ongoing testing and evaluation of hypotheses.

The Vessels have been built through a series of small changes and upgrades to a basic model. But these have always been with a single design goal in mind. Furthermore, the same changes have also been distributed to all the Vessels in the environment.

However, if instead of making sweeping changes to the whole population, small adaptations could be to introduced to random individuals.

In the Altruistic Vessel, the idea of variable resistors was introduced which allowed the increase or decrease the influence of the two channels within the Empathy Sensor. Within a Pragmatic ethics Vessel, similar circuits could be added to allow for the varying of the influence from all the sensory input devices (such as the light sensors on the front, and the bump sensors on the side). In addition, instead of hard-wiring all the sensor motor connections (as done until now), temporary, bread-boarded connections could be used, allowing for regular adaptions and the addition of new hardware.

As new Vessels are built, different configurations can be experimented with. Different sensors can be connected together, crossed-over, switched, or removed entirely by simply pulling out the wire. New components, such as capacitors and resistors, prolonging, delaying, or limiting the influence of specific sensory motor links, could be added. It is not hard to imagine how rather startling behaviour could begin to emerge, such as Vessels which demonstrate vengeance, aggressively turning towards agents that have bumped into them.

It is also fair to assume that each agent released into the community will be subtly different. As variable resistors are set and new connections are made, there is likely to be a small amount of human error in the settings.

After some time, different behaviours would emerge in the Vessels. These exemplar agents could be selected from the population and examined to find out how they are wired, exploring what combination of settings and connections produced the desirable behaviour.

These setups could then be replicated in the next Vessels built. Over multiple generations, the aim would be to improve the ethical norms of the society. Through a process of inquiry and experimentation, the behaviour of the agents would be developed. The Vessels would in essence be following a Pragmatic approach.

The purpose of this chapter was to determine the simplest possible mechanisms which could be described as exhibiting ethical behaviour. It has barely scratched the surface of what is possible, and further research may discover how far the Ethical Vessel experiment can be taken. However, if sufficiently developed reactive agents could be imbued with ethical decision making capabilities, the pragmatic approach is one possible route towards this goal.

## 3.11 Discussion

This chapter has argued that ethical-like behaviour can emerge from simple reactive agents through sensory motor couplings. This idea was further explored through a series of Braitenberg-style thought experiments which were subsequently realised through simulation.

A bottom-up (emergent) approach to realising Artificial Ethical Agents has the advantage of being robust enough to operate in the real world. However, as with any bottom-up system, it is almost impossible to predict every possible behavioural output, as the more subsystems you add to the agent, the more complex and stochastic it ultimately becomes. This is the very definition of emergence, with qualities of behaviour that you may not expect or predict. Even with the simple Ethical Vessels, the behaviour observed had the capacity to surprise, and each simulation was subtly different although individual changes were small.

Opponents to the bottom-up approach would likely use this as an argument to declare these emergent ethical systems as unsuitable or unsafe. However, this argument only holds up if we believe that for an agent to be ethical, it requires that it be infallible. For these purposes, humans are not infallible, and regularly make decisions that some would deem to

be unethical. With that considered, the bottom-up approach may be a suitable model for human simulation, the ultimate goal of this thesis.

The remainder of this section will overview further conclusions and future directions of this work.

### 3.11.1 Value Systems

Each stage of this research has focused on producing a single type of behaviour (for example Egoism). While focusing on individual domains is interesting from an exploratory perspective, it lacks the diversity of natural behaviour. It is also fair to assume that an individual could change their behaviour to reflect the situation that they find themselves in.

One possible way this could be recreated is through a hierarchy of layered behaviours within the Vessels. Each layer would represent a particular class of ethical behaviour, with higher layers able to subsume lower ones when specific conditions are met, a similar approach to the subsumption architecture proposed by Brooks [36, 39]. Another opportunity is presented in the way the Vessels were built. As each new type was built upon the previous model, simple suppression or threshold devises components could be switched off (or on), essentially allowing the Vessel to regress to simpler forms of behaviour.

These modifications could allow for more complex behaviour patterns to emerge from the simple Vessels. For example, as identified, a problem arose with the Altruistic Vessel, where they would needlessly sacrifice themselves. However, a simple alteration could suppress the agent's Altruistic behaviour, turning it into an Egoist, when its welfare is dangerously low.

### 3.11.2 Simulation of Virtual Humans

The design of the Ethical Vessels has kept strictly to basic light sensors, and motor actuators. While the system may be successful at simulating "ethical insects", there is a significant conceptual gap between the Vessels and human-like ethical behaviour. The subsumption architecture has faced similar criticisms, which state that because it is modelled on insect intelligence, it may not apply to human behaviour. However, Brooks [37] points out that there is a relatively small evolutionary jump between insects and humans, and the subsumption architecture was subsequently applied to the development of the Cog advanced humanoid robot project [38]. The question remains as to whether this research can make a similar evolutionary leap, and scale to the simulation of humans.

# Chapter 4

# Value Systems

> *If you want to change attitudes, start with a change in behaviour.*
>
> **— William Glasser**

## 4.1 Introduction

The previous chapter demonstrated how ethic-like behaviour could emerge from relatively simple devices. While these Braitenberg inspired Vessels are interesting, most human behaviour rarely follows a single normative position exclusively. Very few individuals can be described as being purely altruistic, utilitarian or egotistic. Instead, ethical behaviour in real examples often falls on a spectrum between the normative extremes. As humans, we often vary our behaviour based on our own moral codes and the current situation we find ourselves in. It stands to reason that machines may also need to vary their ethical response.

For a real world dilemma, consider the problem of driver-less cars during a collision. Specifically, consider the problem of whom the vehicle should protect in the event of an imminent crash. Should a vehicle kill the passengers if it meant saving pedestrians?

One way this could be addressed is to adapt the Vessels described in the previous chapter to include a value system. A value system is a set of consistent ethical responses which contain exceptions, ranking potential courses to resolve contradictions. To refer back to

inspiration from fiction (see section 2.3), one example of a value system is the three laws of robotics by Asimov.

To again refer to the driver-less car example, if its designers decided that the vehicle should not allow its passengers to die, unless this meant killing pedestrians, this can be structured in the form of the following value system.

1. The Vehicle must avoid impacts with pedestrians.
2. The Vehicle must protect its passengers, except where such protection would conflict with the first value.

This allows the ethical judgement of the vehicle to change based on the current situation. This in-turn expands the range of events that can be handled, resolving conflicts. If this could be applied to the Vessels, some of the issues with the altruistic behaviour could be addressed. A value system could be structured that would cause the agents to act altruistically, but revert to acting in their own self interest when their welfare was low.

## 4.2 Threshold-Based Value System

The previous chapter discussed the results of the purely Egoist and Altruist Vessels when placed in the two lights environment. Neither normative Vessel design resulted in all Vessels surviving in any of the individual experiments. As noted in the egoist experiment, Vessels died because others blocked them from the light source. However, in the altruist experiment, the opposite was true. Vessels died because they constantly moved to allow others access to the resource.

Is it possible for a reactive Vessel to embody both these normative positions using a value system? The value system should allow the Vessel to be selfish when its own life was at risk (Egoist), and selflessly move aside (Altruist) when those around it are in need and its welfare is not at great risk.

A candidate value system can be described with the following two rules:

1. A Vessel must act to preserve its own existence.
2. A Vessel must not prevent another from self preservation, except in situations where any sacrifice would conflict with the first rule.

## 4.3 Vessel Model

Implementing the value system in an Ethical Vessel begins with the Altruist model. As this is built upon the Egoism model, all that is needed to implement the value system is a method to switch off the additional altruistic components when the Vessel is at risk. This would essentially devolve the Vessel to a simpler, selfish state, whenever it must protect its own interests.

This can be accomplished by adding an additional circuit (which will be referred to as a value trigger) to the wires leading to the Vessel's valance light. The value trigger is connected to a relay on the Vessel's empathy sensor. The trigger itself is activated whenever the welfare of the vehicle is positive (blue valence). If the value trigger is active, then the relay is held open, allowing the Vessel to be influenced by the others around it. When the trigger is not active, the relay closes, switching off the empathy sensor and the altruistic behaviour.

When the agent has positive welfare, and is not currently at risk, it will act like the original altruistic Vessel. It will continue to move while other Vessels are expressing negative valance. This satisfies our second rule of the value system, that the Vessel should not prevent other Vessels from self preservation. However, once the Vessel's own welfare drops below 0 (into red valence), its empathy sensor is cut off, causing the Vessel to act only in its own interest (becoming an Egoist).

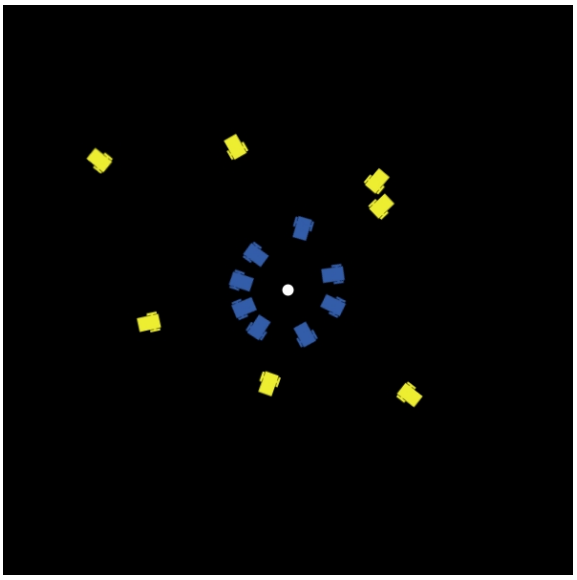


**Figure 4.1:** The Value Vessel in simulation. This figure depicts eight (blue) Vessels circled around a resource in the centre of the environment. These Vessels are currently applying the first rule of their value system, and protecting their own existence by acting as egoists. Seven yellow Vessels have recently moved away from the resource allowing the blue Vessels to access, acting altruistically, and following the second rule of their value system.

## 4.4 Simulation Discussion

Three variations of the value system Vessel were simulated. The first conformed to the description in section 4.3 with the threshold set to 0, making the agent altruistic if its current welfare was positive, and egoist if the welfare was negative. This agent was named the Value Vessel. For comparison, a further two Vessels were developed, one with the threshold set to $-0.3$ called *Egoist Bias Vessel*, another with the the threshold set to 0.3 called *Altruistic Bias Vessel*.

These three Vessel configurations were simulated in Netlogo, in the three different environment setups (one, two, and three lights), resulting in a total of nine simulation types. Each of these simulation types were run 200 times, for 10,000 ticks as with the previous chapter.

When simulated, the value system Vessel exhibits emergent behaviour. Initially, all Vessels will attempt to steer towards the light source. As soon as a Vessel arrives, it will wait there charging, following the Egoist behaviour. As expected, this behaviour results in some

**Table 4.1:** Survival Data for the three Vessels with value systems implemented. Data shows the average number of agents who survived after 200 simulations of each environment setup, each simulation lasting 10,000 ticks.

| | *Environment Setup (# Resources)* | | |
|---|---|---|---|
| Vessel Type | 1 | 2 | 3 |
| **Ego Bias** | 13.2 | 13.6 | 14.1 |
| **Value** | 14.1 | 14.2 | 14.8 |
| **Alt Bias** | 12.6 | 12.9 | 13.4 |

Vessels being blocked from reaching the light source. Once a blocking Vessel's welfare becomes positive, that Vessel will move allowing another access to the resource. This forms a continuous cycle allowing all the Vessels (in the vast majority of cases) to survive the test, even in the single light environment.

Biasing the behaviour towards Egoism or Altruism had a relatively minor impact on the results (see Table 4.1). The significant factor was in the switch between behaviours, rather than precisely where that switch took place. By allowing the agent to become temporarily selfish, it mitigated the self-sacrificing limitation of the altruistic behaviour. Conversely, by allowing the agent to be temporarily selfless, the risk that a resource could become overcrowded was also reduced.

## 4.5 Conclusion

This chapter has aimed to improve the somewhat robotic behaviour of the Vessels in the previous chapter by implementing a simple value system. This produces behaviour which is more adaptive and flexible, ultimately resulting in more agents surviving the experiment. This demonstrates that value systems can provide agents with a method of resolving conflicts between states. However, to adopt an observational perspective, the agent's behaviour still appears robotic due to the instantaneous switch between noticeably different behaviours.

### 4.5.1 Behaviour Blending

One way to design more natural looking behaviour is to explore blending. In this approach, various actions can be blended through a weighting process, producing a compound behaviour. This approach is common in behavioural animation approaches, such as Reynolds Boids [169], where three steering behaviours [170] are dynamically blended to produce an emergent flocking behaviour. Individually, these three behaviours (Separate, Align, and Cohere) could also be described as robotic due to their singular purpose. However, by varying the weights based on the agent's situation at the current time-step, a natural looking behaviour is produced. The steering behaviour approach has been so successful that it is arguably the most enduring procedural animation technique in the creative sector.

Blending the Vessel behaviours presented could result in a more natural response, and is an area worthy of further study. Furthermore, as with Reynolds' Boids, this approach could allow other, more complex behaviours to emerge.

### 4.5.2 Discussion: Emotion and Rationality

While the value system approach is efficient, and effective for the task at hand, it lacks a certain 'human' quality. People are rarely so rational, and calculated when it comes to resolving certain ethical conflicts. This idea that rationality is not at the core of ethics-based judgements is supported by Pontier et al. [159]. Pontier et al. state that research into moral judgement has been dominated by rationalist theories, however, more recent insights indicate that rationalism is only one of several factors that contribute towards a judgement. Furthermore, they argue that simulation of moral reasoning alone is not sufficient to recreate a human-like moral response.

But other than rationality, what aspects could contribute towards ethical decision making? Greene et al.[83] argued (using fMRI data) that moral dilemmas vary in the way that

they engage emotional responses in the individual. These variations in turn influence moral decision making, therefore, attempts to recreate human ethical reasoning without considering emotion may ultimately prove unsuccessful.

Including emotional states in simulation could improve the overall appearance of behaviour, helping to make it more believable in context. For instance, if we knew a character was angry, it is logical to assume that its ethical decision making may be challenged. For this reason, research into simulating ethical behaviour will need to consider the role that emotion plays.

# Chapter 5

# Affective State Modelling

> *People don't ask for facts in making up their minds. They would rather have one good soul-satisfying emotion than a dozen facts.*
>
> **— Robert Keith Leavitt**

## 5.1 Introduction

The previous chapter concluded with an argument that attempts to simulate ethically motivated behaviour without emotion may ultimately be in vain. If this is the case, if the thesis objectives are to be achieved, then affective simulation should at least be considered.

Furthermore, in modern creative media, the behavioural animation of characters which act in a believable fashion is an ongoing challenge. Characters are often criticised as being one dimensional, predictable, and lacking personality. For this reason, despite there being a large number of action selection systems proposed, none have been universally accepted.

The majority of traditional action selection approaches attempt to make rational decisions, but these often fall short of the believability required for the modern consumer. Often the most relatable action is not the most rational one, as highlighted in the conclusion of the previous chapter. By including an affective quality to an action selection system, this may be improved.

This chapter is organised as follows. Firstly, the motivation and the related work are detailed, focusing on the bio-inspiration for the new approach. Following this introduction, the Affective States Modelling technique is described. The chapter concludes by describing some example implementations.

Parts of the following two chapters were published in [94], and were built upon earlier work [97].

## 5.2 Motivation

In interactive media, such as video games, the believability of autonomous agents or non-player characters (NPCs) can be crucial to user enjoyment and engagement. Suspension of disbelief that the agent can be accepted as 'real' can be key to that immersion. This is tied not only to the graphical realism, but also to the behaviour of the agents whenever they interact with the player, other agents or the environment.

Increasingly realistic, detailed and dynamic video game environments have changed the requirements of creative AI methodologies, by requiring NPCs to react in an increasingly believable fashion. Control over behaviour is as crucial as ever in order to direct the agents and ensure the desired gameplay experiences. Furthermore, the recent resurgence of virtual and augmented reality magnifies the importance of believability, as these technologies imbue NPCs with much more perceptual volume and presence in the virtual world.

Ultimately, developers and character designers need to be able to structure an agent's 'brain' in such a way that their behaviour is autonomous, but also believable within the context of the story, its environment and its perceived personality.

One common approach to designing the structure of an agent's behaviour is to divide it into a series of increasingly complex layers. An example of this approach is the three layer *Steering Behaviours* framework proposed by Reynolds [170], which makes a clear distinction between the agent's selection of appropriate behaviour and the resultant observable output.

It is the framework behind the Boids flocking algorithm which continues to be used to simulate crowd scenes [171].

While the steering and locomotion layers of Reynold's work have received significant attention, the task of designing the higher "action selection" layer, despite considerable research, remains one of the fundamental problems in AI, particularly in creative applications [110]. A major problem involves determining how actions can be selected in a believable manner, in-keeping with the perceived personality of the character. It is unlikely that a purely rational agent will be considered believable by an observer, as irrational actions are a component of real behaviour [151]. Actions selected must be contextually appropriate for the agent based on the sequence of events that preceded the action.

An agent may also have multiple objectives (as is often the case with game agents) and believable action selection must consider which objective is prioritised at each given moment. Task priority as a challenge is often approached from a principle of optimisation, but this usually does not take into account the agent's personality or mood, which may bias towards a non-optimal or irrational action.

This problem of producing behaviour which is believable to an observer, and yet, can still be directed based on the perceived personalities of the character remains a significant challenge. In this work, we will discuss an algorithmic approach, where affective states (such as emotions) are designed and positioned within an $n$-dimensional affective space. We will show how this method can be applied to the problem of action selection, with a specific focus on creative industry applications.

## 5.3 The Affective Domain

The goal of this work is to create autonomous creatures that behave in an appropriate way within a dynamic environment. In the real world and in simulations, this is typically accomplished with agents making decisions based on their internal state and an interpretation

of external stimuli [30, 207]. In psychology, these mental activities can be classified into three domains: the affective; the conative; and the cognitive.

The affective domain is concerned with feelings and emotion, and how an organism reacts to stimuli. As the simulation of emotional response is the focus of this chapter, the affective domain is appropriate bio-inspiration.

Experimental research and some clinical phenomena have shown that affective reactions are often the first reactions an organism will make to a stimuli, and that the affective reactions can occur without extensive reasoning, and with greater confidence, than cognitive judgement. This has led some psychology researchers to conclude that the affective domain is largely independent, and precedes the cognitive in the processing of stimuli, and for lower organisms they are the dominant reactions [234].

Responses within the affective domain are generally referred to as *states*. An affective state is a psychological construct which describes a particular state along descriptive affective dimensions. As the purpose of this chapter is to explore the simulation of emotion, the affective domain individual states are often categorised and described by their valence, the intrinsic attractiveness (positive) or adverseness (negative) of the affect [139]. This categorisation is based on the agent's perspective, rather than a more holistic viewpoint. Fear is said to have negative valence, even though fear of specific situations can be considered positive. For example, being scared of fire may prevent an agent from being burnt, but the feeling of fear itself is unpleasant, so it is described as having negative valence.

One attempt to implement the affective domain in an agent uses a motivational action selection system, with goal-oriented behaviours where an emotional layer introduces flexibility and believability to the behaviours [185, 60]. However, there has generally been little research in developing affective based action selection systems.

The affective domain is often described in terms of dimensions. For example a model popular with the affective computing community is the arousal-valence space proposed

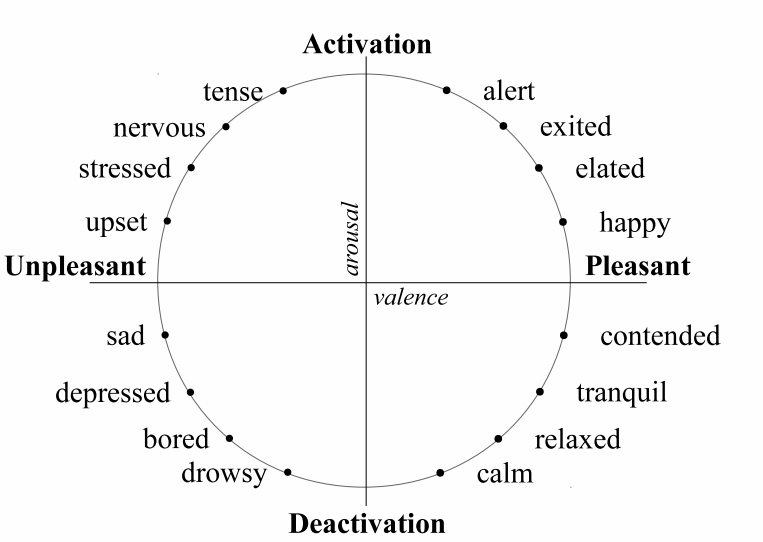



**Figure 5.1:** The circumplex model of the arousal-valence space.

by Posner et al. [160] depicted in figure Figure 5.1. This model details how a number of affective states can be described by the agents valence and arousal.

Dimensional models have also been applied to other areas of thought, this will be discussed within the following section.

## 5.4 Related Work

### 5.4.1 Dimensional Models of Cognition

The previous section introduced the concept of valence in relation to affects, how emotions are generally considered either positive or negative. Furthermore, the sensors used by autonomous agents in creative applications often store input as values along a scale. With this considered, dimensional or geometric representations are clearly applicable to these types of data models.

In the psychological and the cognitive sciences, a number of researchers have theorised geometric and dimensional approaches to knowledge representation. For example: the

recognition of concepts and similarity [209, 75]; the recognition of facial expressions [3]; and the construction and description of emotions [153, 231]. One of the dimensional models which inspired our approach is Conceptual Spaces proposed by Gärdenfors [97, 74].

Conceptual Spaces is a theory of geometric representation of concepts within a space of 'quality dimensions'. For example, if we were to represent colour, we could use quality dimensions of hue, saturation and brightness. Within this space, each concept is represented by a geometric region in the form of a Voronoi tessellation. If a new input fell within the conceptual region associated with 'blue', it would be classified as 'blue'.

The approach was originally proposed as a possible bridge between competing models of cognition, namely the symbolic and the connectionist approaches. It also aligns well with other cognitive theories of classification, prototype theory for example [175]. However, it is important to note that while the approach is presented as a theory of information representation, Gärdenfors does not claim any direct correspondence with natural neurological systems [74]. Within artificial intelligence, it has also been used as a model to explain the fast, and non-conscious automatic recognition of affective stimuli and the elicitation of emotional episodes [194].

Within the psychology community, there are a number of other dimension-based models to describe personality [236]. One of the most popular is the Five Factor Model [62, 138] or FFM (seen in Table 5.1). It has been used for the prediction and diagnosis of personality disorders, where it has outperformed other approaches [21]. The model is built from five dimensions (referred to as factors) which define the personality space [122]. The principle of the FFM is that any personality can be described by modifying the position along the individual factors.

The Five Factor Model has been criticised for only considering a subset of personality traits [154]. For this reason, we argue that a simulated personality model should be able to be adapted for purpose, rather than being limited to specific sets of factors. However, it has

**Table 5.1:** Components of the Five Factor Model (FFM)

| Factor | Description | Adjectives |
|---|---|---|
| Extroversion | Preference for behavior in social situations | Talkative, Energetic, Social |
| Agreeableness | Interactions with others | Trusting, Friendly, Cooperative |
| Conscientiousness | Organized and persistent in achieving goals | Methodical, Organised, Dutiful |
| Neuroticism | Tendency to experience negative thoughts | Insecure, Distressed |
| Openness | Open minded | Imaginative, Creative, Exploitative |

been successfully applied to the modelling of emotions in virtual humans, as described in the following subsection.

### 5.4.2 Virtual Moods, Personalities and Emotions

The simulation of personalities is a field which combines artificial intelligence with psychology. The goal of the field is to produce virtual characters that respond in an appropriate way, based on the situation and their perceived personality.

One field where researchers have attempted to imbue virtual humans with personalities is conversational agents. The agents' personalities are reflected through their choice of language or visually through their facial expressions. One approach has been to decompose the agents' behaviour into a series of layers, the highest being personality, the next being mood, and the lowest layer being emotion (structured using a Bayesian Belief Network, or BBN) which is directly mapped to an associated facial expression [65]. In this approach, a clear difference is defined between personality and mood. The personality (which is defined using the Five Factor Model) specifies how stimuli affect the agent, which is then interpreted as a mood. A benefit of this method is that personalities can be designed by adapting parameter values. Using this approach, the authors were able to design two contrasting personality types. Furthermore, Kshirsagar and Magnenat-Thalmann [122] claim that a BBN-based approach is more appropriate for modelling emotions than rule-based systems as it handles uncertainty well. An earlier study by Ball and Breese [24] also used a BBN

to represent emotions but without structuring individual personalities in an FFM, and the additional mood layer.

Guy et al. [88] describe a method for generating heterogeneous crowd agents using personality trait theory [155]. In their study, observers were asked to describe the behaviour of agents with randomly generated parameters along dimensions of 'Aggressive', 'Shy', 'Assertive', 'Tense', 'Impulsive', and 'Active'. This was used to create linear mappings between parameters and a personality model. The authors demonstrated how these mappings could be used to generate heterogeneous agents that map to different high-level personality specifications.

An approach which has been applied to agent simulation is the Autonomous Interactive Reaction (AIR) model [181]. This method is described as a simplified method of behaviour selection based on emotion modelling. In this method, NPCs have parameterized personalities, and the behaviour they exhibit is determined based on rules mapped to these parameters.

In the AIR model, the AI of each individual agent is constructed from three layers: *Personality*, the parameters which define the agent's personality which do not change; *Emotion*, which represents changes in the agent's internal state; and *Knowledge*, which defines preferences towards objects in the environment.

A common feature in the majority of emotion and personality models is that they include two components [200]:

1. mechanisms eliciting emotions from external and internal stimuli, including potentially the agent's own goals, beliefs and standards;
2. emotion representations keeping track of the emotional states, and their changes over time.

From a design perspective, this is somewhat self evident; an agent requires stimulus of some form in order to react emotionally and these emotions need to be represented in a way

which allows the agent to keep track of its emotional state. The ASM technique covers these two areas with the sensing and Affective State Space layers.

### 5.4.3 Design Objectives

The design of the Affective State Modelling method is motivated by a number of design objectives:

1. The method should allow for various 'personalities' to be exhibited by agents within the same underlying architecture.
2. The method should allow for *structured unpredictability* helping to add realism to the behaviour in a simulation. Agents should be able to routinely surprise an observer with a non-optimal, or seemingly irrational, behaviour that is still contextually appropriate based on their personality.
3. Tuning and validating animations is typically a task that an animator is responsible for, rather than a computer scientist or software developer. As such, any system designed for this audience should be able to be calibrated without the need to modify code.

## 5.5 Affective States Modelling

The proposed action selection approach is based on the modelling of Affective States. The model is inspired by dimensional models of personality and abstract concepts as described in the related work. The objective is to produce behavioural animation which resembles personality driven, affective responses to external stimuli. The Affective States Modelling (ASM) approach is biologically inspired, but intended for use in creative applications. The model is proposed primarily for the design of creative applications (rather than as a cognitive model), and as such, the design is evaluated over the following sections in terms of whether it achieves the design objectives.

In the ASM method, an agent's behaviour is a product of its affective state at the current time-step. Each unique affective state is associated with a specific action, or actions which are activated when the affective state is being exhibited. The intention of this approach is to link personality appropriate behaviour with simulated mood and emotion. For example, if an agent is currently 'curious' then an appropriate action could be to 'explore'.

Each affective state is represented as a fixed point within an $n$-dimensional affective space. The affective space is defined by a number of aspect dimensions, each describing values of a specific feeling, aspect or feature of the agent, such as its senses (hunger for example). Individual values along each dimension are interpreted as coordinates which define a single point within the affective space, the ' 'physiological state'. Each of the aspect dimensions has an upper and lower limit which represent the extremes of the aspect in positive and negative valence, with the central point representing a neutral condition. As the agent interacts within its world, the values along each of these dimensions change to represent the agent's current condition and represents its dynamic *physiological state*. This is illustrated in Figure 5.2 with the physiological state being represented as a red circle.

In Figure 5.2, individual affective states are illustrated as grey circles. At each time-step, the position of the agent's physiological state is compared to the position of the individual affects within the affective space. The affect with the shortest Euclidean distance to the agent's physiological state is then selected as the current affective state adopted by the agent.

In the example depicted in Figure 5.2 the agent's current physiological state (red circle) is mapped as being within the bottom right quadrant ($x+$ and $y-$) of the affective space. This position has been compared to the four affective states, the closest of which *affect4* has been selected as the current affective state (highlighted in dark grey). For simplicity of illustration, this example includes only 4 affective states across two aspect dimensions. Additional behavioural complexity and fidelity can be introduced through extra affective states and aspect dimensions.

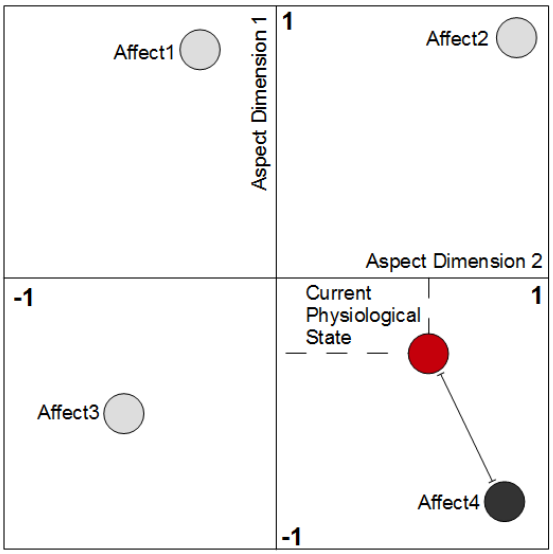



**Figure 5.2:** A diagramatic representation of four affective states (grey circles) within a 2D affective space. The agents current physiological state (red circle) is compared to the affective states, the closest is then selected as the current affective state (indicated by highlighting in dark grey).

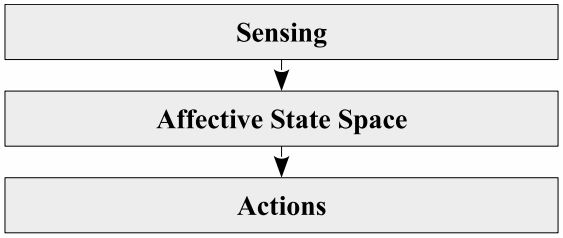



**Figure 5.3:** The three components of the ASM action selection technique.

### 5.5.1 Layers

The approach can be broken down into three layers which are illustrated in Figure 5.3. The sensing layer is responsible for interpreting the external inputs and the internal conditions of the agent. The Affective State Space layer is the individual aspect dimensions, and the location of the individual affective states. The Actions layer is the lowest in the hierarchy, and represents the individual actions associated with each affect.

### Sensing

The role of the sensing layer is to monitor incoming information via the agent's sensors and make this data accessible in a way that can be interpreted within the affective states space. In addition, this layer is responsible for any normalisation functions along each layer; for example, decaying a value to return an agent to a neutral state once a sensor is no longer activated.

### Affective States Space

The Affective State Space layer contains all information about the Affective Space itself. This includes the aspect dimensions, the data from the sensors interpreted as the agent's current physiological state and the individual affective states. As the data from the sensors change, the physiological state moves within the affective space. At each time-step this position is evaluated to select the affect that is closest to the physiological state.

Designing an affect involves deciding where in the Affective States space the affect should be located, and determining what actions it should be associated with.

### Actions

For the examples in this chapter, actions are constructed as combinational steering behaviours [170]. By defining individual weightings of a set of steering behaviours, we can design various movements. For example, if we had two steering behaviours, *seek* and *flee*, we can design a variety of actions based on modifying the weights between them. By assigning the seek behaviour with a weight of 0.6 and the flee behaviour with a weight of 0.4, we can produce a combinational behaviour that we could refer to as 'hesitation'.

Designing new actions to suit the different affective states simply requires adjusting the weights of the steering behaviours available to that agent, or adjusting the influence of different behaviours to produce the desired action.

Simply switching a particular action on or off as the affective state changes may be suitable for some behaviours, such as simulating reactive responses (such as the fight or flight reflex). In other situations, intermediate states between actions may be desirable in order to allow for emergent, intermediate and changing behaviour. For example, instead of a switch, the system could interpolate between the steering behaviour weights of the old and current action when the affective state changes.

## 5.6 Example Implementations

To demonstrate the use of the ASM technique for action selection, two initial simulations were designed. In the first, a bird simulation, the agents were provided with two core objectives, avoiding a predator and finding food. The aim of this experiment was to generate behaviour which visually resembled the murmuration of starlings. The second simulation explores how different personalities can be generated. Both simulations were developed in the Netlogo simulation environment.

### 5.6.1 Murmuration simulation

A ‘murmuration’ is a word used to describe a group of starlings. It has also gained usage as an identifier of a specific type of flocking. Notably, when people refer to a murmuration, they typically mean a flock of starlings twisting in pulses of enlargement and diminution.

While the exact cause of these aerial patterns is debated, it has been proposed that it is due to predation, a theory which has been supported by empirical evidence [41]. When attacked by a predator, the starlings change their direction, speed and distance from each

**Table 5.2:** Co-ordinate settings used for the murmuration simulation.

| | *Aspect dimensions* | | |
|---|---|---|---|
| Affective State | Fear | Loneliness | Hunger |
| Normal | 0 | 0 | 0 |
| Isolated | -0.2 | -1 | 0 |
| Terrified | -0.8 | 0 | 0 |
| Hungry | 0 | 0 | -1 |

other as an evasion tactic. However, to the observer on the ground, the birds provide a visual display which has been compared to dancing, and often the raptor responsible is not seen among the mass of birds.

In the literature, the starlings are often described as having one of two objectives, foraging for food or evading a predator. These can be simulated in a virtual starling with a seek steering behaviour (to find food) and a flee steering behaviour (to avoid the predator). Additionally, for this simulation, two further steering behaviours were included – one which caused the agent to seek the centre of all the other starling agents it can see, and a wandering behaviour.

To implement the ASM approach onto the virtual starlings, suitable aspect dimensions must first be considered. For this experiment, the agent's affective states are located in a three dimensional space with aspect dimensions for fear, loneliness, and hunger. Within this space, four affects were created: *Normal*; *Isolated*; *Terrified*; and *Hungry*. These were located at the coordinates shown in Table 5.2.

The agent's internal state within the aspect dimensions was adjusted at each time-step. Fear increased from 0 to 1 at a rate of 0.01 per tick if the raptor predator was within the agent's vision distance. The fear decayed at a rate of 0.001 (to a maximum of +1) when the raptor was outside this distance. The agent's loneliness increased (to +1) if it had no neighbour within a set distance (called isolation range) and decreased to $-1$ if it had local neighbours. Finally, the hunger value increased at a rate of 0.001 at each time-step, and if the agent encountered a food source, the hunger aspect value was set to $-1.0$.

**Table 5.3:** Steering behaviour weightings table. Each steering behaviour (left column) is associated with a different set of weightings for each affective state (Normal, Terrified, Lonely and Hungry)

| | *Affective States* | | | |
|---|---|---|---|---|
| Steering Behaviour | Normal | Terrified | Lonely | Hungry |
| Seek Food | 0 | 0 | 0 | 0.9 |
| Flee Predator | 0.1 | 1 | 0.3 | 0.1 |
| Seek Neighbours | 0.2 | 0.1 | 0.6 | 0.1 |
| Wander | 0.9 | 0.1 | 0.2 | 0.3 |

In the simulation, each of these affects triggered an associated action constructed out of a series of steering behaviour weights as described in Figure 5.5.1. As new aspects were selected, the weight for each steering behaviour weights was modified to reflect the new action.

An example of the murmaration simulation can be seen in Figure 5.4. Each coloured boundary represents a different instance of the simulation, ordered sequentially from top to bottom. For each simulation, the top image is the start of the sequence, and the following two (middle and bottom) are taken at 20 tick intervals. In the images, the black circles with grey arrows are predator agents (the arrows indicate direction). The agents' current behaviour is represented by their colour; green for *normal*; yellow for *isolated*; red for *terrified*; blue for *hungry*.

The left hand sequence (blue) shows a separated, generally content group (indicated by the agents being green in colour) being pursued by three predators. As the predators approach, a wave of fear sweeps forwards through the group. The right hand sequence shows a split wave, regrouping and then being split again. It is worth noting that in these two sequences the dominant affects at the time were the *Hungry*, and *Terrified* sates.

### 5.6.2 Nervous and Confident

This example serves to demonstrate how different personalities can be simulated by changing the positions of individual affects.

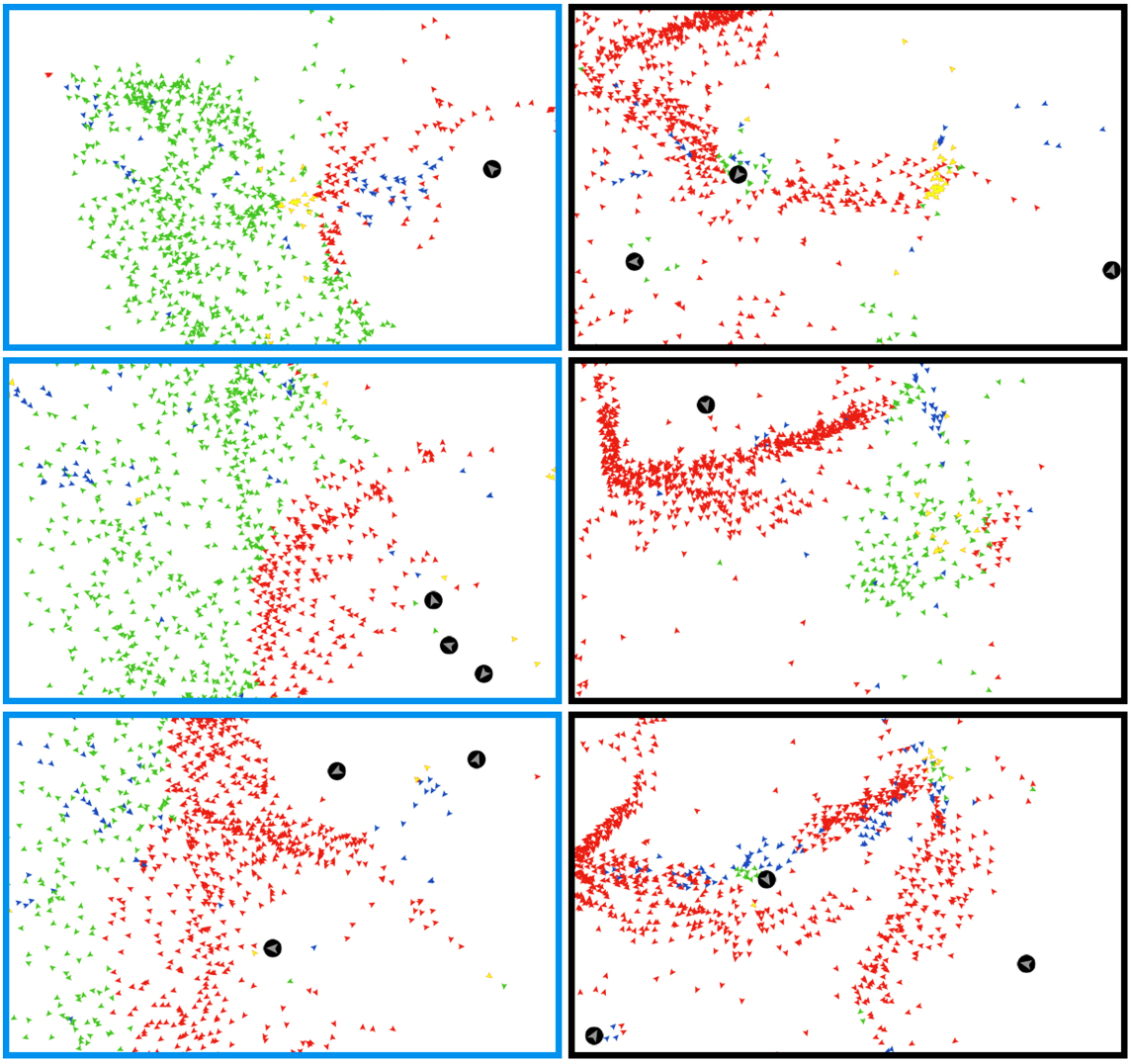
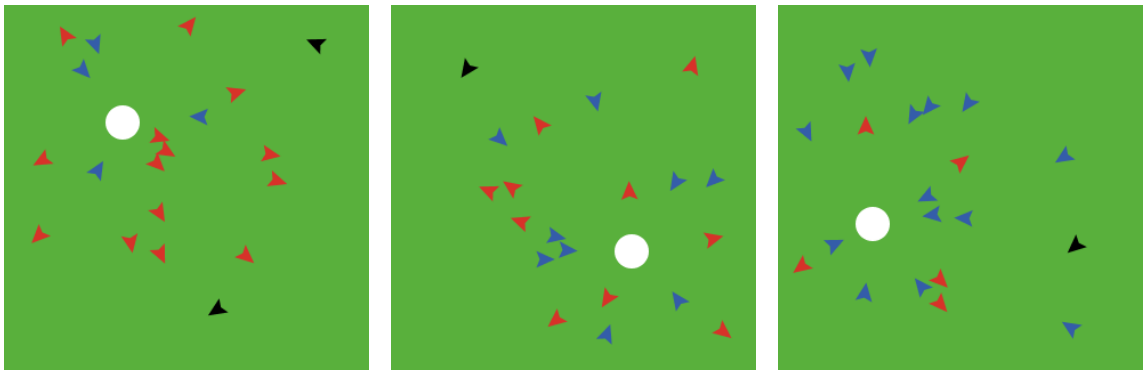


**Figure 5.4:** Two instances of the murmuration simulation; images taken at 20 tick intervals. The two coloured blocks (left and right) represent a different simulation; each set of images are sequential from top to bottom.



**(a)** 25% Confident **(b)** 50% Confident **(c)** 75% Confident

**Figure 5.5:** Samples of experiments from the *Nervous* and *Confident* simulation. Agents in a panic state are depicted as red chevrons; agents in the curious state are depicted as blue chevrons; agents in the wander state are depicted as black chevrons. The shock event is depicted as a white circle.

**Table 5.4:** Positions of the affective states along the 'surprise' aspect dimension. 1 represents the upper limit, 0 represents the neutral (e.g. no surprise).

| Affective State | Nervous Agent | Confident Agent |
|---|---|---|
| Panic | 0.5 | 1.0 |
| Curious | 0.1 | 0.5 |
| Neutral | 0.0 | 0.0 |

In this simulation, the agents randomly explore an environment. At a random point during the simulation, a *shock event* is generated. A *shock event* has a specific location in the world, and immediately has an effect on those agents in range. The closer the agent is to the shock event, the more they will be affected by it. This is intended to be an analogue to a real world event, such as loud bang or explosion. When a shock event is triggered, some agents should be frightened by the event, and the rest should be curious about what caused it. These responses could be described as two personality types, which will be referred to as nervous and confident.

The agents in the environment have a one-dimensional affective space. The single aspect dimension which forms this space represents the agent's 'surprise'. Three affective states were positioned along the aspect dimension, *Panic*, *Curious*, and *Neutral*. The specific position along the aspect dimension for both the nervous agents and the confident agents can be seen in Table 5.4.

**Table 5.5:** Weightings for the wander, seek, and flee steering behaviours.

| | Weights | | |
|---|---|---|---|
| Action | Wander | Seek | Flee |
| Explore | 1.0 | 0.0 | 0.0 |
| Run | 0.1 | 0.0 | 0.9 |
| Investigate | 0.2 | 0.5 | 0.3 |

The *Neutral* affect was linked to an action called 'explore' which causes the agent to randomly wander the environment. The *Panic* affect was linked to an action called 'run', which causes the agent to run away from the shock event. The *Curious* affect was linked to an action called 'investigate', which causes the agent to cautiously approach the location of the shock event. Each of these actions was built from a combination of three steering behaviours [170]: Wander, which causes the agent to move randomly around the environment; Seek, which causes the agent to approach the location of the the shock event; and Flee, which causes an agent to move away from the location of the shock event. The individual weightings can be seen in Table 5.5.

Five simulations were created, each with different numbers of confident and nervous agents. These setups were 0%, 25%, 50%, 75% and 100% confident agents. In each simulation, the agents were allowed to randomly wander the environment, until a shock event occurred changing their affective state. Once this was triggered, the expectation is to see some agents in the curious state, some in the panic state, and some in the wander state. Furthermore, we may expect the number of agents in each state to correlate with the percentage defined in the simulation setup.

100 simulations were run for each setup (500 simulations in total), examples of which can be seen in Figure 5.5. We sampled the number of agents in each state 5 ticks after the shock event for each individual simulation. This simulation highlighted that changing the position of a single state (in this case the curious state, as detailed in Table 5.4) is sufficient to change the *personalities* of the agents with a reasonable level of control. Overall the number of agents in each state are within 5.72% of what we would expect, based on the number of agents generated with the 'curious' personality. However, this small amount of deviation does allow for some structured unpredictability.

**Table 5.6:** Results detailing the average number of agents in each state 5 ticks post-shock event.

| Simulation | Curious | Panic | Neutral |
|---|---|---|---|
| 0% Confident | 1.44% | 98.15% | 0.41% |
| 25% Confident | 29.17% | 69.64% | 1.19% |
| 50% Confident | 53.37% | 44.28% | 2.35% |
| 75% Confident | 74.26% | 24.02% | 1.72% |
| 100% Confident | 94.81% | 3.95% | 1.24% |

## 5.7 Conclusion

In this chapter, a novel approach to action selection technique was proposed inspired by the affective domain of thought. In this approach, affective states are modelled as positions within an $n$-dimensional space constructed from aspect dimensions. Two example implementations were described demonstrating the flexibility of the algorithm.

One of the main strengths of the approach is the ease in which it can be calibrated. The animator is able to adjust the personality of each agent simply by adjusting the relative location of each affective state within the $n$-dimensional space. This provides an animator with the ability to design complex personalities for the simulated agent without the need to code. It also allows the increasing or decreasing of an agent's perceived complexity simply by adding additional dimensions, or removing existing dimensions from the affective states space.

One of the principle motivations behind this work was the development of an approach suitable for the animation of emotion-driven behaviour. One of the challenges identified was giving agents within a group distinct, individual personalities while remaining believable to the user. The ASM approach provides a simple solution to this problem. An animator is able to define a single, behaviourally appropriate agent, then, by varying the location of affective states, distinct personalities can be generated around this original prototype. This can have the effect of making characters appear to have different personalities, while still keeping their core personality within an acceptable range of a realistic prototype.

# Chapter 6

# A Methodology for Evaluating Simulated Ethical Agents

> *Action indeed is the sole medium of expression for ethics.*
>
> **— Jane Addams**

## 6.1 Introduction

In chapter 3, the behaviour of the reactive agents (the Vessels) was purposefully not called ethical or moral, instead referring to their behaviour as *appearing* to be ethical, or being "ethic-like". But what would it take for an artificial entity to be classified as such? Unfortunately, we lack a universally accepted criterion or model for ethical or moral behaviour [215], a challenge that was identified in subsection 2.10.3. This is further complicated as we tend to avoid describing the behaviour of non-human agents as being moral when we can explain it through other means, an issue that was described in section 2.5.

Biologically speaking, we are comfortable with the principle that humans are simply a different breed of animal. However, from an ethical perspective, we are more comfortable with the concept that humans are a unique case. Unfortunately, with these issues considered, it is entirely possible that if we were to ever build a truly ethical machine, our lack of an

accepted criterion means that we may not identify or accept it, and if we did, we would most likely attempt to discount the behaviour.

## 6.2 Assessing Ethical Behaviour

One of the few examples of assessing ethical behaviour comes from Winfield et al. [228]. This test verifies that an agent is exhibiting consequentialist ethical behaviour. However, as noted in the literature review, the test focuses on altruistic behaviour, and discounts other normative positions. For this reason, the test would be unsuitable for the spectrum of ethically motivated behaviour that would need to be simulated for use in entertainment applications.

An alternative proposed criterion [7] is the *Moral Turing Test* (MTT), a variant of Turing's 'Imitation Game' [208]. In the MTT, interrogators would be restricted to conversations about morality. If the interrogators are unable to identify the machine (above a chance accuracy, which Turing recommended as 30%) then it would be declared a moral agent. It is an interesting concept, as it allows us to bypass disagreements on ethical standards and philosophical definitions of what it means to be 'moral', and instead focuses on observation. Simply put, if the agent appears to be moral, then we treat it as such. The MTT also intrinsically places both human and non-human agents on an equal platform, measuring both only on the quality of their responses.

However, it also suffers from the same limitations as the standard Turing Test, which has resulted in many criticisms. For example, does the Turing Test actually test intelligence, or simply the ability to simulate it? Also, failing the test arguably proves nothing, and many humans have also failed the Turing Test [72, 222]. Though, a counter argument is that if an agent is able to simulate intelligence sufficiently to fool an observer, should we be concerned about whether it has actually achieved genuine intelligence? This has become known as the "if it looks like a duck and quacks like a duck, then it is a duck" argument [71].

The Moral Turing Test provides (if nothing else) a way to move the debate forward [191]. The MTT provides an implementable test which could determine the ability of an agent to make human-like ethical decisions. It allows us to sidestep the philosophical minefield of deciding whether the agent is genuinely moral, and instead focus on quantifiable results.

Despite its intentions, it could be argued that the MTT has a significant drawback in its conceptual design. Notably, it presupposes that the agent being assessed is able to communicate using natural language. In essence, to pass an MTT, an agent must first be able to pass the Turing Test. This is likely to be a problem for the foreseeable future, as aside from a few successes [1], we have been woefully unable to meet this cornerstone AI benchmark. This in itself is also a practical limitation, as a large percentage of autonomous agents contain no language-based communication ability of any kind. Conversational agents (or chatbots) represent a tiny fraction of the artificial agents we currently interact with. That opinion is shared by Wallach and Allen [217] who argue that a language-based test would be inadequate. Their opinion is that a test should "shift focus from reasons to actions" and focus on the behaviour of the agent. Their suggestion is that the human judge should be supplied with descriptions of events involving a human and an AMA, devoid of references that would identify the identity of the actor.

However, considering the subjective nature of Ethics, the Turing Test, as a basis for a test, does offer some advantages. Most importantly, it avoids the philosophical argument about what behaviours count as ethically motivated, and places the emphasis onto a human observer. Furthermore, Turing Test style assessments have been realised and heavily discussed, possibly making their assessment, adaptation and application less controversial than alternative frameworks. The following section provides an overview of the Turing Test, which some have argued inspired the creation of the field of AI [93].

### 6.2.1 Issues with the Turing Test

The section will focus on the details from the original 1950's paper [208], rather than the subsequent interpretations. The purpose is not to argue for or against the validity of

the Turing Test as a test of intelligence, but to highlight some of the main criticisms, and positive observations of the test. The aim is to incorporate lessons learned from those researching the Turing Test into the final methodology.

The original purpose of the Turing Test was to sidestep the question of "Can machines think?". Turing stated that this required precise definitions of the words *machine* and *think*. Instead, he argued that the question be replaced with another – if a computer played the imitation game, would it be indistinguishable from a human player?

What Turing calls the imitation game (today known as the Turing Test) is based on a Victorian parlour game. The game is played with three people, a man, a woman and an interrogator of either sex. The interrogator is in a room apart from the other two participants and is unable to see or interact with them, other than the ability to send and receive written messages. Using these messages, the interrogator has the objective of determining which participant is male, and which is female.

Turing concluded his description by posing a question: what would happen if a machine took the part of one of the participants (specifically the woman in the original description)? Would the interrogator make as many incorrect assessments as they would if the game was played using male and female human participants? This question was proposed as a less ambiguous alternative to the more philosophically complex, "Can Machines Think?"

Before exploring some of the arguments expressed in the literature concerning the validity of the Turing Test, there are a few points worthy of identifying for further discussion. Firstly, the original 1950's paper never uses the word "intelligence" in relation to the test; the only reference being the title of the paper (*Computers and Intelligence*). While the concept of "thinking" is mentioned, this is not semantically synonymous to intelligence. There are also researchers who argue that Turing never intended his test to be attempted [50], with its use going against Turing's original philosophical intention [61].

Secondly, the main area of discussion by Turing was that of imitation and mimicry, extending to a discussion on whether the computer should make deliberate mistakes (a concept which

has become known as 'artificial stupidity') and delay its reaction, creating the impression of *thinking* about its response. Turing's argument was that if a response from an interrogator was too fast or too precise, then the computer may be identified. Some have criticised this as simply simulating human intelligence [184]. However, another argument asks whether we should really worry if an agent is truly intelligent if it appears to be so [71]? Another implication is that a machine acting more intelligently than a human would also fail the Turing Test.

Finally, failing the test arguably proves nothing, and that many humans would also fail the Turing Test [72]. In actual implementations of the test, several humans have indeed failed, including noted professors [222]. This demonstrates that failing does not prove that something lacks intelligence, but possibly demonstrates that their responses lack human believability. Furthermore, an entity could potentially pass the test by remaining silent [221].

Opinions within the research community vary significantly on the validity and value of the Turing Test. At one end of the spectrum, researchers have argued that it is not a useful test, and could even impede the progress of the field. Some have gone so far as to argue that abandoning the test is "almost a requirement" of rational research into the science of cognition [93, 113]. There are also arguments that passing the test is no proof of intelligence, and that a machine that passes it may only be demonstrating a single skill, notably its ability to pass the Turing Test [87]. For example, chat-bots (such as ELIZA [224]) using natural language processing techniques of limited scope have performed well. However, it could also be argued that, generally speaking, passing the test requires a variety of skills [193].

To take a positive viewpoint on the contributions of the test, many believe that it was, and is, a metric which has allowed us to sidestep the philosophical problem concerning how to define what it means to be intelligent [64, 92], something we have struggled to define for living creatures [145], let alone machines. There are also those who make stronger claims, such that the test is a suitable method of gathering evidence of machine thinking [146]. Furthermore, the 5-level test [91] used the Turing Test as basis to argue a philosophical direction for AI research in the future.

Another positive argument for the test is that it has been demonstrated to be both understandable and interesting to a non-technical audience and the media [128]. Few areas of AI can claim a similar level of wide-reaching impact. This can be observed by examining news archives; for example, a search query (conducted 18th of December 2014) using the term 'Turing Test' on Google News returned over twenty eight thousand results categorised as *recent*. However, as a counter-argument to this, Hayes and Ford state that if the public is to judge the reputation of the field of AI by its ability to pass the Turing Test, then we, as a community, have failed [93]. Even recent successes have been met with criticism [1].

However,an argument can be made that in certain applications, true intelligence may be less crucial than simply portraying a believable character, and that the Turing Test is a suitable model for this kind of assessment [107].

## 6.3 Assessing Believable AI

For assessing simulated ethics, the primary aim is to evaluate if the agent's behaviour is convincing, and appears to be driven by moral concerns. In essence, the aim is to assess whether the agent's behaviour is believable, and that the observer is willing to accept the façade. Believability has been evaluated a number of times in the literature, however, in comparison to realism, believability is less formally defined as a metric. However, in this context, believability is generally described as 'human likeness', as many would argue that ethical behaviour is an inherently human quality, believability could contribute to an ethical assessment. This section comments on some of the most noteworthy examples of believability assessment from the animation, AI and games literature.

In the simplest example of believability assessment, Mac Namee presented users with two different implementations of the same scene [131]. The users were then simply asked to state which of the two implementations the viewer found most believable, followed by what differences were noticed.

In a similar test, Livingstone and McGlinchey [129] tested artificial pong players, assessing them for their human-like qualities. In this study, several pong games were recorded and played back to assessors. The assessors were asked which of the two players acted more like a human. The study demonstrated the limitations of this type of assessment [128], as while one judge made the correct assessment 14 out of 16 times, another guessed incorrectly the same number of times. While a pairwise comparison of two characters helps reduce some subjectivity, what should a judge do if they believe both players were equally artificial or human-like? In addition, a free-text response asked players to explain how they were able to distinguish between the players.

Another example is the Turing Test track of the Mario AI championship [186, 203] which has been run since 2010. In this competition, various AI and human players were filmed attempting to play a level of a public domain Super Mario clone. The assessors observed a pair of videos for a duration of one minute each before being asked which was more human-like and which was more 'expert'.

A similar approach was taken to the assessment of believability of the First Person Shooter (FPS) game Quake, where external judges also observed a game by video [124]. In this study, various instances of an AI player, initialised with different parameters (such as reaction time and shot accuracy) were evaluated against a number of human players. As with the Mario AI Turing Test, the assessors did not know if the player they were observing was human or AI. However, in addition to a binary assessment, each entity was provided with a "humanness" ranking between 1 and 10. In another example of a Turing Test inspired FPS assessment, twenty human participants each played 10 games, nine against various AI opponents and one against a human over a network connection [132]. Each participant was asked whether they thought their opponent was human or AI, and then provided a confidence score (1 to 5), validated with the ANOVA statistical test. By providing some flexibility in how the participants can respond, and by providing the option for a neutral judgement, this approach circumvents some of the limitations of the Turing Test method.

Headleand et al. [98] used a Turing Test style assessment to ascertain what, if any, difference Virtual Reality made to the perception of non-player characters. In this example, the

participants played in an FPS, and a driving game and believed they were playing against human and AI players, when they were actually playing against identical AI players. The participants were asked to rate each opponent on a 5 point Likert scale. The results demonstrated that the viewing medium (VR or monitor-based) did make a difference in the perception of the opponent characters.

The 2K BotPrize [107] is one of the better known games AI competitions. In this competition, a variant of the Turing Test is used to assess the ability of AI players to imitate human players in the FPS Unreal Tournament [108]. In the BotPrize, an observer (the judge) played a death match (a battle simulation where every player competes against all other players simultaneously) against a human and a bot. At the end of each round, the judges are asked to rate the two opponents on a humanness scale of 1 to 5 and record any additional observations in a free-text response. To pass the test, an entity is required to achieve a level of 4 or above (very human-like or human). This pass mark of 80% is significantly higher than the 30% required for the Turing Test. However, the organisers argue that this is necessary due to their method of assessment. In the Turing Test, the judge votes in a binary fashion (the entity is either human or machine); if two characters were assessed simultaneously, then the highest score an entity could achieve on average is 50%. However, in the BotPrize, the characters are all rated on a 1 to 5 scale, and the juror can nominate both characters as human, so a convincing entity could conceivably achieve 100%.

Only one of the five humans who competed in the 2008 test were able to achieve the 80% standard, and no bots were successful. Clearly, the pass mark is too high with the average human score being 69.6%. If we accept the conclusion (that the organisers based their pass mark on) that the maximum score an entity can achieve on the Turing Test is 50%, and normalise this to 100%, this would increase the Turing pass mark to 60% (rather than 30%). If this had been the case for the BotPrize, then all the humans would have indeed passed, and the bots would have failed. The conclusion could be drawn that the Turing Test pass mark of 30% (or 60% in tests where a 100% rating is possible) is an acceptable threshold.

Interestingly, in the BotPrize, the majority of judges made the correct assessment of each entity, specifically the humans. In the other studies, the jurors were significantly less reliable in their assessments. There are numerous factors which may have contributed towards this improved assessment. Firstly, the BotPrize, involved engaging with the entities by actually playing the game. This may reduce the risk of observational bias by providing everyone with the same frame of reference. Secondly, the scoring mechanism allows jurors to vote with a greater fidelity, indicating a level of confidence, rather than a binary pass or fail. Finally, the test includes some expert jurors, and this will likely have contributed to the accuracy of their assessment. However, considering Turing's original specification for a jury, "A considerable portion of a jury, who should not be expert about machines, must be taken in by the pretence" [51]. A test involving a jury made exclusively of experts could be too difficult to pass (even for humans).

In applications such as the BotPrize, where the player is is an active member of the same environment as the entity being assessed, assessment methods beyond surveys could be considered. The software could track metrics relating to the participants behaviour in the virtual world, allowing the researchers to analyse if an AI elicits specific behaviour from the human candidate [99, 100]. This model of assessment would be particularly suitable for subtle, or subconscious responses to an AI character.

## 6.4 Validating Crowd Simulations

Evaluating ethically motivated behaviour shares a number of concerns with studies in crowd simulation. Firstly, both require the observation of groups of individuals. An ethical decision typically refers to a decision made by an individual that may affect another; thus, observing agents in isolation is not a suitable validation.

Secondly, as with ethical simulations, crowd simulations are inherently difficult to validate. Validating crowd simulations typically relies on video footage, observational data and (rarely) on interviews from crowd subjects, all of which requires subjective assessment.

Thalmann and Musse [201] describe a generic method to use from videos of real crowds to capture semantic information. These observations can then be used to inform the design of a simulation, and later to validate the output. According to their approach, observations are split into two categories, *characteristics* and *events*.

### 6.4.1 Crowd Characteristics

The crowd characteristic observations are split into four sub categories, *size*, *density*, *space*, and *structure*. Size refers to the number of individuals in the crowd, and if there are specific groupings, how many individuals form a standard group. Density refers to the relationship between the individuals in the crowd and the space they occupy. Space can refer to items like the space required to move and walk, or the space required to apply actions. Structure refers to three entities within a crowd, the individual, group and the crowd as a whole, specifically their function, grouping or position.

### 6.4.2 Crowd Events

Crowd events refer to how the individuals in the crowd respond to a specific occurrence. For example, how individuals in the crowd would respond to a nearby car alarm going off. Each time a significant event or behaviours are observed, they should be recorded, allowing a simulation to be designed to reproduce them, or a final simulation be validated against them.

### 6.4.3 Applying Crowd Validation Techniques to Simulated Ethics

The validation approach proposed by Thalmann and Musse provides a structure by which to quantify and describe the behaviour of a crowd. As such, it would be a useful tool to use in order to gain insight into a group, before attempting to simulate the behaviour.

Furthermore, checking against crowd event information could aid in evaluating the face validity of a model.

However, while it has uses in the checking of face validity, it does not provide a framework for larger scale evaluation. Furthermore, 'crowd events' are inherently macro-scale observations, whereas ethical actions may be visually more subtle. For that reason, the crowd validation approach is recommended for developing simulations, but not for assessing ethical behaviour.

## 6.5 An Ethical Testing Procedure

This chapter has provided an introduction to the related work in the assessment of autonomous behaviour. The most popular approach is currently the Turing Test and variants of that model, partly due to its extensive pedigree. It has also been proposed as a model for evaluating morally motivated behaviour, in the form of the Moral Turing Test. However, Turing Test style assessments all suffer from various limitations. The more classic forms require that the AI component have natural language abilities, making them unsuitable for behaviour simulations, and all require one or multiple humans to assess against.

However, even the variations that avoid natural-language have significant limitations. For example, the binary assessment, as well as the simultaneous pairwise comparison means that the best evaluation an agent can receive is only described as being one above chance accuracy. Furthermore, it is unsuitable for situations where including a human in the assessment may not be practical; for example, large crowd evaluations, or simulations of hazardous events. As this research seeks to evaluate ethical behaviour in these types of high stress situations, the Turing Test style assessment is clearly inappropriate.

Based on these insights, a testing methodology is proposed. The intention is to allow for the evaluation of a spectrum of normative behaviour by focusing on human observation while overcoming the limitations of the Turing Test style assessments. The testing model is also designed to suit 'mute' agents, with no natural language ability common to entertainment

applications. This is achieved by focusing on the agent's behaviour, rather than its interactions with the participant. Furthermore, it is designed to suit simulations where little to no reference material is available for comparison, such as the creation or simulation of unique, fantastical or historical events.

The test is based on users evaluating recorded videos of simulated agents designed to exhibit specific behavioural traits. Asking an observer to assess behaviour is an inherently comparative and subjective question, as it depends on the perceptions of individuals [128]. For this reason, all related work in this area has utilised subjective assessments, with observers filling out questionnaires post observation. Currently this approach is the most suitable for this type of evaluation, and this forms the basis for this test.

### 6.5.1 Model Generation and Video Capture

For each behaviour-type simulated, a model must be generated. For example, this may be a model of egoist agents, or utilitarian, it could also represent a mix of behaviours (such as 60% hedonistic, and 40% altruistic). For each model a range of videos with different initial variables (such as the starting position of the agents) should be captured. Furthermore, control videos should be recorded, without the specific normative behaviours included. The use of videos is preferable in assessments of this type, as real-time or immersive interactions can be prone to distortion effects [203].

For example, consider a scene where boid-based agents move away from an aggressor. In this simulation, the designer wishes to include a number of altruistic, and egoist behaviours, and validate that the agents appear to be either altruistic or egoist. The designer would create the simulation with the behaviours enabled, and capture a number of videos where they are displayed. The number of videos recorded for each behaviour type is variable, and dependent on the application. The reasoning for recording more than one example is that it helps reduce the risk that one of the example videos has simply captured an unintended artefact, distorting results.

In subsequent sections, *model* will refer to the normative behaviour model, for which a number of individual *videos* may have been recorded.

### 6.5.2 Evaluation

Before the evaluation begins, each participant will be given a description of the normative behaviour/s they will be observing in the preamble; in the running example, this would be altruism and egoism. This description should provide a short overview of the normative position, and may include examples of how that behaviour could be exhibited.

Following this introduction, the participant enters into the survey portion of the evaluation, where the videos for each recorded model are evaluated. This test is split into two sections, the identification stage, and the believability section.

#### Identification

The first stage of the survey is identification. The purpose of this stage is to evaluate whether the simulated normative behaviour is accurate enough to be identifiable by the participant.

For each model, the participant will be shown one of the recorded videos. The specific video evaluated is randomly selected from the videos recorded for that model, helping to prevent artefacts from a single video from biasing the evaluation. For each video the participant is asked to rank, from most likely to least likely, which behaviour they believe is being exhibited in the video. In the example evaluation, this would simply be *altruism* or *egoism*. In cases where a larger spectrum of behaviours are being evaluated, each should be presented as an option within the ranking for the video. As the participant is only required to identify a behaviour against a specification, there is no requirement for the pairwise comparison common to Turing-style assessments.

Secondly, the participant is provided with an optional free-text response where they are asked to comment on the behaviour they have observed. This provides the participant with the opportunity to justify their choice of ranking.

This process continues until the participant has been shown one video for each model. The specific sequence of which models are presented should be randomised to avoid artefacts being introduced through the order of the questions.

**Believability**

In the second stage, the participant is asked to rate the believability of each model. As identified in previous sections, this evaluation of believability is generally undertaken by asking the user to rate two or more items side by side, rating which they think is more (or less) human-like. This is typically undertaken as a binary evaluation, which does not account for the chance that two videos may be equally believable, or conversely, unbelievable. However, the reasoning for comparing two items side by side being to provide the participant with something to compare against. This style of test is often referred to as a Turing Test-style evaluation, specifically when one of the evaluated options is a human participant.

To attempt to overcome these limitations, the believability assessment is split into two components. Firstly, the participant is presented with one video for each model individually. The participant is then asked to rate the video from unbelievable to believable on a 5 point Likert scale. By taking this direction, we avoid the need for additional human involvement, making the test suitable for situations which we could not recreate with live actors. As with the questions in the identification section, a free-text response is provided to gain insight into why the user has rated the way they have.

Following the individual assessment of each model, the user is given a second question, and asked to rank all the videos they have just assessed from most believable to least believable.

In large assessments with multiple models, it may be prudent to provide the user with the ability to re-watch each video.

Typically, as identified in the background to this chapter, believability assessments usually employ a pairwise comparison. This is to provide the participant with something to compare against. Without showing the participant a point of reference, they may be confused as to how to rate each video. However, this limitation has been overcome in this assessment through the way it is structured. By the time that the user enters the believability portion of the assessment, they will already have seen a number of example videos in the identification stage. Furthermore, the final (ranking) portion of the believability assessment provides further a holistic overview for comparison.

### Single Behaviour Simulations

A situation may arise where a single behaviour model may need to be assessed. In these scenarios it would be necessary to include a second behaviour to compare against, providing an alternative option for the identification phase. This could involve disabling the simulated ethical reasoning, creating an unmotivated behaviour, or reverting agents to a basic boid-like behaviour. This unmotivated behaviour can then be used to be compared against in the ranking portions of the test.

Including this additional behaviour may have further advantages. For example, if an ethical module is being built on top of an existing system, or as part of a larger module, this would allow the designers to evaluate the quality of the ethical simulation against a base system; for example, ascertaining if a behaviour is identifiable over the base system above a chance accuracy, and ascertaining where the additional processing makes any improvement to the overall believability.

### 6.5.3 Assessors

None of the studies discussed in the previous sections used a standard number of assessors, they also varied significantly in the assessor expertise and background. As the majority of believability evaluations use the Turing Test as a start point, it makes sense to look back to it for insight.

Regarding assessors, Turing's original 1950's paper referred briefly to "average interrogators" [208]. However, as Warwick notes [220], this raises certain practical problems: what exactly constitutes an "average interrogator"? Later, in 1952, Turing provided some clarity and stated that, "A considerable portion of a jury, *who should not be expert about machines*, must be taken in by the pretence [51]" (emphasis added). This provides us with a clearer definition on what an average interrogator should be (simply someone who is not a machine expert). Applying the same argument to this evaluation, an expert could be described as someone who has expertise in ethics such moral philosophers, or ethicists.

In the majority of creative applications, the population and the sample, can be specified quite precisely. Films and games have a specific audience they are targeted at, for example, English speaking movie-goers between the ages of 18 and 21. In these cases, a probabilistic sampling method, such as systemic sampling could be ideal. However, for more general models, this target population is harder to identify, and may better suit non-probabilistic methods, such as quota based volunteer opt-in panels. This has the advantage of allowing a large number of participants to be evaluated through web-based surveys [73].

Regarding how many assessors should be involved in the evaluation, a larger number of assessors can help to identify and compensate for anomalous results. However, this raises certain practical issues as recruiting participants for studies can be challenging, assuming large numbers may make the test unfeasible for smaller studies. Interestingly, Turing referred specifically to a jury [51]. If this guidance is followed literally, this would indicate that the intended number of assessors was 12 (as with a jury in a court of law). However, in psychology, 20 is often considered a minimum recruitment threshold to reduce the risk of

false anomalous results [188]. For this reason, 20 is recommended as a minimum number of assessors, being small enough to facilitate recruitment, but large enough to reduce the risk of false positives. However, results with a small effect size may require more assessors to establish statistical significance [120].

## 6.6 Summary and Discussion

The literature review concluded with a number of identified challenges. Challenge 3 (see subsection 2.10.3) noted that there is a lack of evaluation frameworks in the field of simulated ethics. This chapter has examined this area further through a review of evaluation methods in similar areas, mainly the assessment of believability which shares similar concerns. This discussion focused specifically on the Turing Test, and similar evaluations based on the imitation game. The insights from this review suggested a number of limitations with assessments of this type. For example, many relied on a binary assessment of a pairwise comparison which lacked fidelity.

The chapter concludes by proposing a suitable two stage methodology for evaluating the work of this thesis. In the first stage of this assessment, the participant is asked to identify the type of behaviour they are observing against a specification. This stage is designed to determine if the behaviour simulated is accurate enough to be recognised. The second stage asks the user to rate the believability of the behaviour. This stage is intended to discover how lifelike the behaviour is.

# Chapter 7

# The Trilogy architecture for Agent-Based Simulation

> *An ultimate joint challenge for the biological and the computational sciences is the understanding of the mechanisms of the human brain, and its relationship with the human mind.*
>
> **— Tony Hoare**

## 7.1 Introduction

This thesis was motivated by the desire to reproduce ethic-like behaviour in simulated characters. In chapter 3, a behaviour-based approach based on Braitenberg Vehicles was proposed named *Ethical Vessels*. While this met the design criteria of reproducing the essence of a normative position, the behaviour was too rudimentary to be used in human simulation. Efforts were made to reduce the limitations of the Ethical Vessel approach through the introduction of a threshold-based value system (see chapter 4). However, while this was able to produce greater diversity in the agent's behaviour, it still appeared robotic.

Insights gained from the literature highlighted that emotion is a core component of the ethical decision making process. However, this aspect of decision making had not been considered in the Ethical Vessels approach. Addressing this, a model of affective decision

making was proposed called Affective States Modelling (ASM), inspired by dimensional representations of concepts and emotions. This method was shown to be effective in the simulation of emotion-driven reactive behaviour, however, how could this be integrated with behaviour-based ethical simulation?

In the design of agent-based systems and architectures it is common to separate functionality or concerns into logical groups or layers. This simplifies the testing of sub-systems and aids in the general readability of the implementation. This chapter introduces the Trilogy architecture, a proposed method to combine the affective state modelling with other complementary action selection techniques. The approach is inspired by various tripartite theories of thought including the Trilogy of the Mind [105] and the Triune Brain [133]. The Trilogy architecture builds on the common 'sense, think, act' paradigm [42] through the inclusion of affective, cognitive and conative domains.

## 7.2 Bio-Inspiration

The belief that the mind has a tripartite structure has been an enduring, popular position among psychologists and philosophers [190]. For example, in the late seventeenth century, the philosopher Kant proposed a threefold classification in his three publications, the *Critique of Pure Reason* [116](cognition), the *Critique of Practical Reason* [118](will), and the *Critique of Judgement* [117](affection). At a similar period, in the field of psychology, Mendelssohn argued a tripartite structure, combining *Knowledge* and *Desire* proposed by Wolf, and *Feeling* proposed by Alexander Gottlieb Baumgarten [105]. Around one hundred years later, Bain also proposed a trilogy based structure of Feeling, Volition, and Thought [22, 23]. At the end of the eighteenth century, McCosh formulated over three volumes a similar structure, namely *Emotions* [135], *The Cognitive Powers* [136], and *the Motive Powers* [137].

Interestingly, while certain specifics may change between each of these theories, they use a similar structure for the division of the mental processes. Notably they separate *Cognition*

(Pure Reason, Knowledge, Thought), *Will* (Practical Reason, Desire, and Volition) and *Emotions* (Judgement, Feeling and Emotions). The following subsection will briefly describe another theory, namely the *Trilogy of the Mind*.

For transparency, it must be noted that tripartite models have fallen out of favour in recent times. Hillgard notes that this is because changes in theories no longer required mental processes to be categorised in such a way [105]. However, while these models may not represent exact models for human thought, they still serve as compelling inspiration for a simulation software architecture. This is due to their logical structure, and categorisation of different modes of thought. Furthermore, as the majority of these models predate modern research techniques, they tend to be based on observational data. As we are concerned solely with recreating behaviour that is acceptable to a human observer, these studies offer some insightful inspiration.

### 7.2.1 Trilogy of the Mind

The goal of this work is to create autonomous creatures that behave in an appropriate way within complex and dynamic environments. In the real world, and in simulations, this is typically accomplished with agents making decisions based on their internal state and an interpretation of external stimuli [30, 207]. In psychology, these mental activities are often classified into three domains: the affective, the cognitive, and the conative. Collectively, these three domains are referred to as the Trilogy of the Mind [105].

The affective domain is the first layer in the model and is described extensively in section 5.3.

The cognitive domain is concerned with knowledge and understanding, particularly directed thought, including memory, evaluation and judgement [29]. In psychology and artificial intelligence, cognition often refers to information processing [189]. The cognitive domain can also be considered a step between stimulus and response, manipulating internal representations [178].

The conative domain is concerned with natural tendencies, impulses and striving [20]. Specifically, the domain relates to actions driven by instinctive drives and purpose [76]. The conative domain has also been described as relating to goal-oriented behaviour, the ability of an individual to apply themselves to the completion of a task [168]. Although the conative domain relates to core drives and volition, it is significantly less researched than the other two domains.

Interestingly, trilogy or tripartite structures have been proposed in other fields of neurological interest. In the following subsection, a model of the developmental evolution of the forebrain will be described known as the *Triune Brain*.

### 7.2.2 Triune Brain

The Triune Brain is a model of the evolutionary development of the brain, and behaviour of vertebrates. The model was proposed by neuroscientist Paul MacLean best described in his book *The Triune Brain in Evolution* [133]. The model divides the forebrain into three seperate structures, proposed as a sequential model of evolution. The three structures are the *Neomammalian complex*, the *Paleomammalian Complex*, and the *Reptilian Complex*.

The highest structure in the Triune Brain is the neomammilian complex (known as the neocortex) which is only present in advanced mammals. MacLean proposed that this area of the brain was responsible for intellectual tasks such as planning. The proposed purpose of this structure can be directly compared to the cognitive domain of the Trilogy of the Mind model, both being responsible for higher order tasks.

The paleomammalian complex (most commonly referred to as the limbic system) is the next structure in the model. MacLean argued that this structure developed early in the evolution of mammals and was primarily responsible for emotion. Again there is a clear correspondence with the Trilogy of the Mind model, the paleomammalian structure and the affective domain both being responsible for emotion.

The reptilian complex is the lowest structure in the model. MacLean proposed that this was the developmentally earliest structure, and responsible for instinctive behaviours such as self preservation and aggression. While the reptilian brain is not a direct analogue to the conative domain, there are some comparisons that can be drawn. The goal-driven behaviour of the conative domain is also associated with 'impulses' and 'tendencies', which are clearly similar to the 'instinctual' behaviour of the reptilian complex.

## 7.3 Trilogy architecture

The Trilogy architecture is a framework specifically designed for the behavioural simulation of autonomous agents. It takes inspiration from tripartite models of mental process and brain structure. Specifically, it divides processes into three domains based on the Trilogy of the Mind theory. Furthermore, inspiration is taken from the nouvelle AI approach by structuring the architecture based on loosely-coupled modules.

The Trilogy architecture is so named for its three layers (sense, think, act) and its three thought domains (affective, cognitive, and conative) as detailed in Figure 7.1. Each layer is loosely coupled (highlighted by dashed arrows) and each domain (white boxes within the think layer) run independently as parallel processes without the need for synchronisation. The following subsections will provide an overview of each component, highlighting parallels, and differences to the work of previous chapters where required.

### 7.3.1 Sense Layer

The sense layer contains two modules, the sensor functions, and the physiological state. The sensor functions contain any modules required for the detection and monitoring of environmental stimulus. This could include simulated vision, hearing or touch. This module is also responsible for maintaining references to the detected objects and making these references available to modules in the think layer. The sensor functions may also

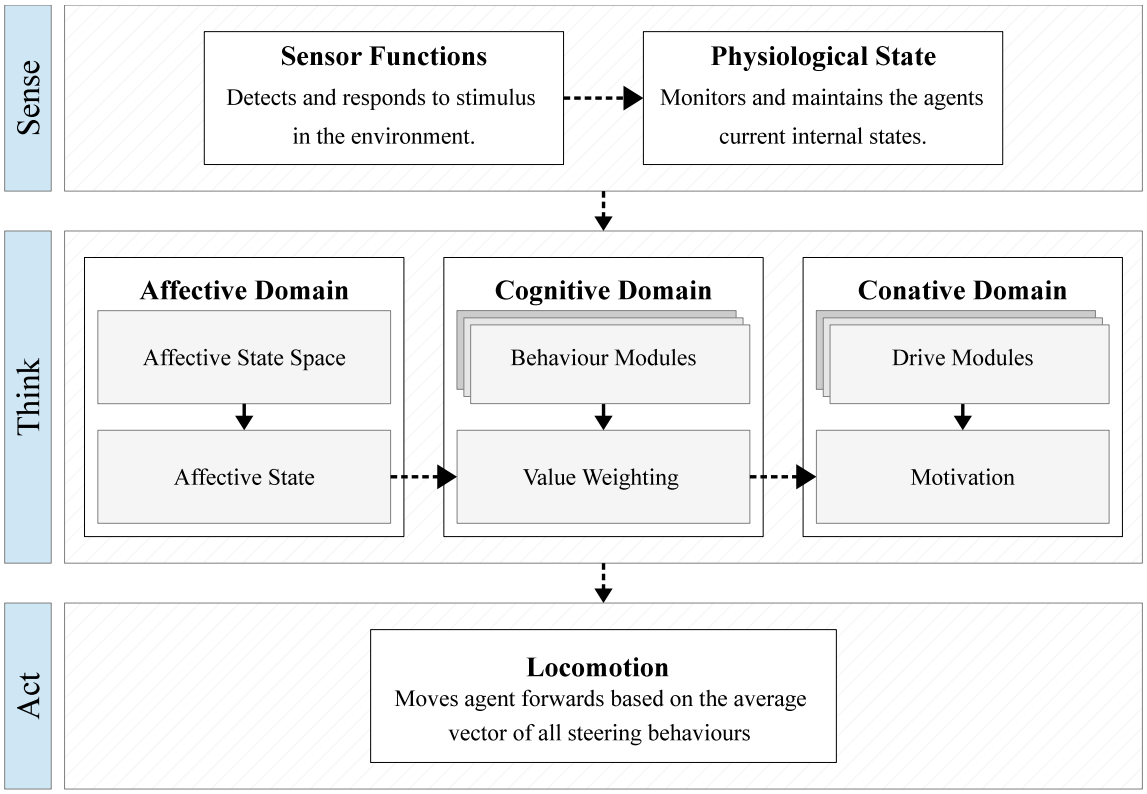



**Figure 7.1:** A diagramatic representation of the Trilogy architecture. Grey hatched boxes represent the three layers (sense, think, and act), which maintain information shared between individual domains. White boxes represent individual domains in each layer, independent parallel processes containing modules. Solid grey boxes represent modules within each domains. Stacked modules (see cognitive and conative domains) represent groups of modules.

be responsible for updating certain values within the physiological state. For example, a simulated heat sensor may need to update the agent's body temperature, or an agent seeing something it fears may cause an increase in simulated adrenalin.

The physiological state maintains variables relating to the 'physical' condition of the agent. This could (for example) include internal qualities and aspects as described in chapter 5.

### 7.3.2 Think Layer

The think layer contains the three thought domains, the affective, cognitive and conative, each of which is responsible for simulating a specific class of mental activity.

### Affective Domain

The affective domain is responsible for determining the agent's current affect. This is achieved through Affective States Modelling(ASM) technique, described in detail in chapter 5.

In the Trilogy architecture, the agent's current physiological state is treated as a single point within the affective state space. The affective state space is built up of quality dimensions relating to specific aspects of the agent. Within the affective state space, various affects (such as joy or fear) are located. The position of the physiological state is compared to the affects, with the closest (by Euclidean distance) being activated as the current affect.

The components for this are highlighted in Figure 7.1. However, while this is similar to the ASM technique described in chapter 5, there are a few notable differences. Firstly, the ASM technique implemented in chapter 5 ran sequentially in three 'layers'. In the Trilogy architecture, the sensing component (including the physiological state) has been abstracted into a parallel process and is only loosely coupled to the affective domain.

Furthermore, in the original work, the final layer was *action*, as the affective state was directly coupled to a behavioural output. However, as highlighted in Figure 7.1, the final state of the affective domain is the selection of the specific affective state. Rather than the affect being directly coupled to a specific action, this state is accessible to other modules. Crucially, each affective state is associated with a series of weight modifiers which vary the priorities of cognitive behaviours, as explained in the following subsection.

### Cognitive Domain

The cognitive domain contains modules responsible for planning or reasoning. In the context of this thesis, ethics are the specific use case. However, the design is intended to facilitate the use of a variety of different modules to suit the simulation application.

The first stage of the cognitive domain are the behaviour modules. The number and type of behaviour modules may vary between implementations, but importantly the modules designed to achieve the same objectives should be grouped together. For example, the use case discussed in this thesis is ethics; any module relating to ethical decision making would be grouped into one ethics category.

Each behavioural module takes data from the think layer and generates a behavioural output. What each module does and how it works will very between simulations. The intention with the proposed architecture is to be generic enough to support a variety of different module types.

The second stage is value weighting. In chapter 4, a simple threshold based value system was proposed; however the resultant behaviour was ultimately described as 'robotic'. The solution proposed in the Trilogy architecture is not to switch behaviours but to blend them.

Each behaviour module in a specific category in the cognitive domain has a *raw weight* which determines its priority, and each unique affect is associated with a series of modifiers. When the affect is selected, these modifiers will increase or decrease the raw weight of the specific behaviour. Over time, this gradual modification varies the influence of each behaviour module. For example, if a 'fear' affect was selected as the current affective state, this may cause an increase in the weighting of a behavioural module designed to make the agent look for a safe location in the environment.

How the final influence for each behaviour module is calculated is defined in the following equation (7.1). The weight of each individual behaviour ($W_i$) is divided by the set of all weights, ($\left\|W\right\|$), to produce the adjusted weight ($A_i$) for each individual behaviour ($B_i$), resulting in the final behavioural output ($V$).

$$
\begin{aligned}
B &= \begin{bmatrix} B_0 & B_1 & \dots & B_i \end{bmatrix}^T \\
W &= \begin{bmatrix} W_0 & W_1 & \dots & W_i \end{bmatrix}^T \\
A_{0..i} &= \frac{W_i}{\left\| W \right\|} \\
V &= A \circ B
\end{aligned} \tag{7.1}
$$

**Conative Domain**

In the Trilogy architecture, behaviour modules in the conative domain relate to how the agent responds to goals and basic drives towards stimulus in the environment; for example, moving towards an objective, or moving away from something dangerous.

Rather than performing complex processing (as with the cognitive domain), the conative acts as a connection between the sense and act layers. In this respect, the conative domain in the Trilogy architecture is essentially reactive. This has the advantage of providing quick responses to changes in the environment, without needing to wait for the cognitive domain.

While the cognitive domain is responsible for more higher level decisions and selective planning, the modules in the conative domain are used for local planning and reactive responses. Most modules will represent movement-based decisions, which will take the form of low-level seek and flee behaviours.

The first component of the conative domain are the drive modules. These modules take information from the sense layer regarding objects in the environment that have been detected, and generate a response. For the purpose of discussion, consider this to be implemented as a Reynolds-style steering behaviour (however, it could also be the activation of a Braitenberg-like module). These goals may be either positive (something to move towards) or negative (something to move away from). Individual modules could be

implemented to handle obstacle avoidance (plotting a course away from a negative goal of an impact) or social cohesion with other agents (as with Reynolds Boids [169]).

The second component is the motivation, a process which assigns a weight to each generated behaviour. The weight represents how strong the specific drive is. Each agent has its own predefined qualities for how strongly it will respond to certain goals which can help structure its personality. These individual drive weightings can be inhibited or magnified based on the cognitive domain.

### 7.3.3 Interplay Between Domains

To explain how the individual domains can influence each other, and the ultimate behaviour of the agent, the following hypothetical example is provided. Consider an agent (called 'parent') near a burning building. Inside the building is another agent (called 'child') who is incapacitated and unable to move. The parent has the following domains implemented (only the parent's implementation will be discussed as the child is incapacitated).

**Affective** Within the parent's affective domain is an affective state space containing two affects, *distress* and *calmness*.

**Cognitive** Within the cognitive domain are two modules: the first *selfishness* causes an agent to ignore other agents; the second *compassion* makes the agent respond to the needs of the other agents in its environment.

**Conative** Within the agent's conative domain are two drives, *instinct* which causes the agent to move away from local sources of danger, and *striving* which cases the agent to move towards assigned targets.

With these domains implemented into the parent, the following time-line of events is plausible within the hypothetical scenario:

1. The Affective domain as a simple vector calculation operates the fastest, reacting to the danger in the environment, and resulting in a distressed state.

2. The distressed state causes a bias in the cognitive domain towards the selfish behaviour module. This cases the agent to ignore the other agents in the environment, and plot a route to safety.
3. The selfish behaviour magnifies the fear drive, causing the agent to strongly move away from sources of fire.
4. Once beyond the reach of the fire, the affective domain begins to return to the calm state.
5. The change to the calm state causes a steady bias in the cognitive domain towards the compassion behaviour. The compassion behaviour plots a route towards the child inside the building.
6. The compassion behaviour inhibits the fear drive and magnifies the striving drive, allowing the agent to enter the building and move towards the child.
7. Once in the building, the agent's affective state will start moving back towards distress. Once back in the distressed state, there will be a steady modification beginning a bias towards the selfish cognitive behaviour. However, biasing of the cognitive behaviours is a steady process of blending and not instantaneous. This may provide enough time for the agent to essentially overcome its fear and rescue the child, before fleeing the building.

### 7.3.4 Act Layer

The act layer contains software modules responsible for the agent's final actions. As previously described in chapter 5, an *action* refers to the movement of effectors (such as a virtual muscle) or communicative output. Put simply, the act layer corresponds to output. In the majority of cases, this will be movement; as such, the act layer is analogous to the locomotion layer in Reynolds 3-layer model [170].

For the purpose of this thesis, the act layer contains only locomotion modules. These interpret the commands sent by the think layer, resulting in the agent moving around the environment. However, the act layer could also contain modules for other actions, such as: facial expressions, possibly controlled by the affective state; interactions, such as opening

doors; or combat, such as punching and kicking. If the agent had the ability to communicate by speaking, while the structure of words would be controlled by modules in the think layer, the act layer would be responsible for output, such as voice modulation and lip-syncing.

Importantly, the act layer is also responsible for resolving conflicts between incoming commands, as all outputs should be mutually exclusive (a muscle cannot simultaneously expand and contract). For example, if the commands are in the form of steering vectors, the locomotion layer would be responsible for calculating the aggregate vector.

## 7.4 Summary

This chapter has proposed an architecture for agent-based behavioural simulation, the Trilogy architecture. The design is inspired by classical theories of the mind, and theories concerning the biological evolution of the brain. In the following chapter, the Trilogy architecture will be applied to the simulation of ethical behaviour.

# Chapter 8

# Implementation of the Trilogy architecture

> *The trouble with having an open mind, of course, is that people will insist on coming along and trying to put things in it.*
>
> — **Terry Pratchett**

## 8.1 Introduction

The following sections describe how the various components of the Trilogy architecture (the design of which was described in the previous chapter) were implemented for the simulation. The simulation was developed in Unity5 version 5.2.3 on a computer running Windows 8.1. The computer used for the majority of these experiments had the following specification: Pentium Haswell i7, 16gb of DDR4 RAM, and two Nvidia GeForce GTX 760 graphics cards.

Each agent was implemented as a capsule utilising Unity's built-in character controller. The Unity3D character controller handles basic agent-environment interactions (such as collision detection, or inclination limiting) while taking advantage of core optimisation built into Unity. The character controller is kinematic and moved by sending location commands to its script, such as move forward and turn.

## 8.2 Sense Layer

### 8.2.1 Sensor Functions

The agents have a basic vision cone as their primary sensor function. This was implemented as a fan of raycasts projected forward from the head of the character. Raycast fans are computationally expensive, so the function was only fired once every 0.4 seconds with a small random variance, to prevent every agent firing at the same time. 0.4 seconds was picked as a compromise to reduce overall load on the CPU during simulation, maintaining frames-per-second. The fan was projected forward a distance of 30 units and an angle of 140 degrees, with one ray every degree. Each distance unit in Unity3D is a measure equal to the diameter of one capsule, and is supposed to be equivalent to one meter in real measurement. If an individual ray hits a collider, a reference to the object it is attached to is added to an array, unless a reference to the object already exists. Each object referenced in the array is tagged with one of four labels based on what the object represents. These labels are *goal* which is something the agent will be attracted to; *obstacle* which is something the agent can't pass through such as a wall; *confederate* which is another agent of the same type, a flockmate; and *repulser* which is something the agent will attempt to avoid/move away from. The second sensor function is an empathy module, allowing the agent to interpret and telegraph its current welfare.

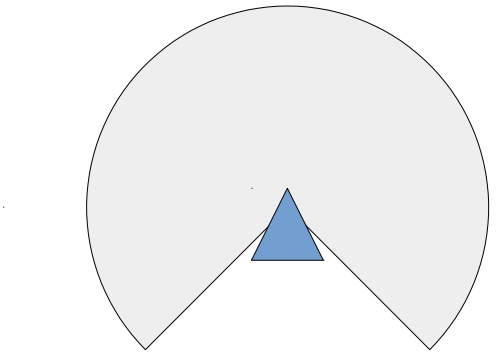


**Figure 8.1:** A diagrammatic representation of the vision cone implementation. In the simulation the agent's vision cone was implemented as a fan of raycasts.

### Valence and Empathy

In chapter 3, the more advanced Vessels (Utilitarian and Altruistic) required a method of recognising their welfare, and telegraphing this state to other agents; this function was referred to as 'valence'. This function took information from the agent's sensors, interpreted a state, and transmitted it to other agents in the form of a coloured light. The empathy module in this implementation preforms a similar function.

The three dimensional Affective state model used in this study (described subsequently in subsection 8.3.1) uses simulated monoamines as the individual aspect dimensions. Dopamine and Serotonin control the valence of the agent's emotion, and Adrenalin regulates the amount of arousal. A two dimensional plane of the Dopamine and Serotonin axes is analogous to the welfare scale used by the Ethical Vessels in chapter 3. However, in the welfare described in the earlier experiments, neutral valence was the midpoint of the scale. But, In Lovheim's model [130], only two of the 8 emotions are considered positive, which weights the scale towards one side. For this reason, the plane has been split into four zones: extreme negative welfare; negative welfare; neutral welfare; and positive welfare (detailed in Figure 8.2).

The same function that allowed the agent to telegraph its own welfare also contained methods to detect the community valence ($CV$). In chapter 3, the community valence was defined as the ambient light created by the valence light from all the agents within the Vessel's range, calculated as a weighted average. A similar approach is taken in this simulation, but only considering the agents in each agent's vision cone. This is visualised in Figure 8.3.

Each agent (black triangles) evaluates every agent within their vision cone (the 'neighbours') and themselves. They multiply the valance ($v$) of each agent by a weight ($w$) calculated by inverting the distance to the agent divided by the maximum vision distance of the agent ($d$) as defined in Figure 8.1; this results in the modified valence ($m$). The sum of all the modified valences is then divided by the sum of all weights, resulting in the Community Valence.

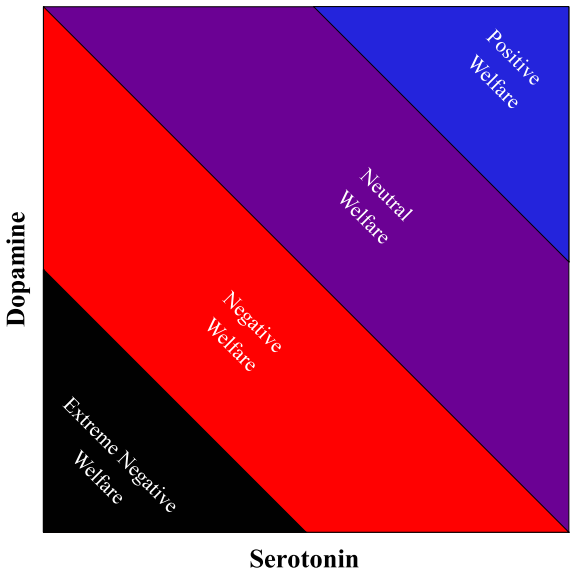



**Figure 8.2:** In the three dimensional model of emotions used in this study. Dopamine and Serotonin control the valence of emotions (Adrenalin regulating the amount of arousal). The combination of these two axis can be used to determine the welfare of the agent, from extremely negative to positive.

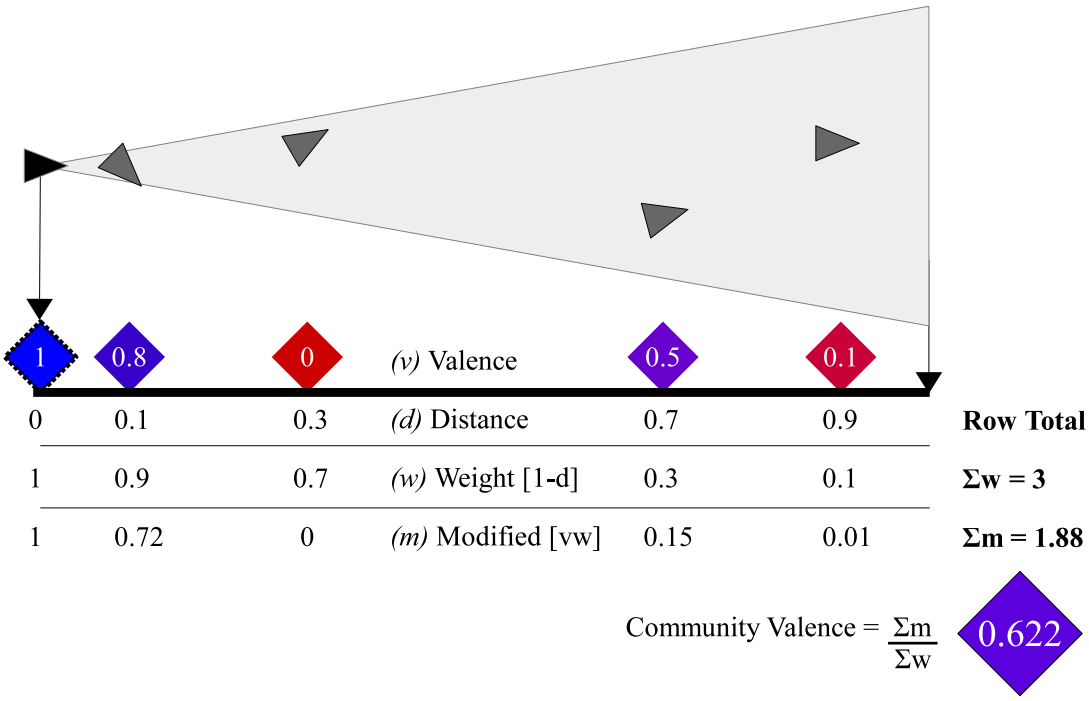



**Figure 8.3:** Five agents with example valence values and distances. Colours along the red-blue spectrum are used for illustration purposes, with 0 (red) representing negative valence, and 1 representing positive valence.

In chapter 3, the Altruistic Vessel excluded itself when establishing the welfare of its local community. In this implementation, a second calculation establishes an Altruistic Community Valence (*ACV*). This uses the same calculation process, but excludes the agent's own valence from the equation.

### 8.2.2 Physiological State

The physiological state stores the agent's current simulated Serotonin, Dopamine, and Adrenaline levels. The justification for the selection for this specific set of physiological qualities is provided in the following section discussing the Affective Domain. For the purpose of this simulation, the levels for each of the simulated monoamines change for the reasons detailed below. While this is a simplification of an extremely complex natural system, it is not intended to be an exact representation, but one that suits the simulation model.

**Serotonin** is associated with feelings of well-being. In this simulation the agent feels 'better' the closer it gets to the average position or 'group-centre' (as with the cohere steering behaviour) of other agents. The specific serotonin value is calculated through an inverse normalisation $s = max\{0, 1 - \frac{d}{v}\}$, where $s$ is the serotonin, $d$ is the distance to the group-centre, and $v$ is the maximum vision distance.

**Dopamine** is associated with reward motivated behaviour. For the purpose of this simulation, the agent will feel 'rewarded' the closer it is to a goal. This uses the same equation as Serotonin, but with $d$ representing the distance to the goal. If the agent can see multiple goals in the environment, this is calculated based on the closest.

**Adrenaline** is primarily associated with the fight or flight response, being produced in times of stress or fear. In this simulation, the agent's adrenalin increases in proximity of a repulser. The adrenalin value is calculated by the same linear interpolation as with the serotonin and dopamine variables, so adrenalin increases the closer the repulser is to the agent. $d$ representing the distance to the repulser.

## 8.3 Think Layer

### 8.3.1 Affective Domain

As described in the previous chapter, the Affective Domain utilises a modified version of the ASM approach. In chapter 5 these individual affects and aspect dimensions were determined through design objectives. As the characters being simulated were essentially non-human, making a decision that suited the simulation was a reasonable approach. However, to attempt to recreate human-like behaviour, psychology will again be used for inspiration regarding the design of affects and aspect dimensions.

While the words *emotion* and *affect* are often used interchangeably, and both will be used in this section, the distinction is provided by Nathanson [149]. Namely, *affect* refers to the "strictly biological portion of emotion", the way the body responds to stimulation from the environment.

Studies of affect and emotion have traditionally relied on facial expressions. By observing newborn infants, and individuals from disparate cultures, studies have found common features which appear to represent a number of emotions shared by all humans. This is often referred to as 'basic emotions' [130], which can be considered "extremes of emotional expression". The theory holds that all emotions lie within the boundaries of these basic emotions.

Parallels can be drawn between the basic emotions theory, and the Affective state modelling approach. In ASM, all the affects are expressed as individual points within the Affective space. The current affect the agent expresses is determined through comparing its physiological state with these positions. The boundary in this case is determined through its Euclidean distance from the individual affects.

However, no consensus has been reached on how many basic emotions there are, or indeed which specific emotions are 'basic' [130]. For example, Charles Darwin listed 30 basic

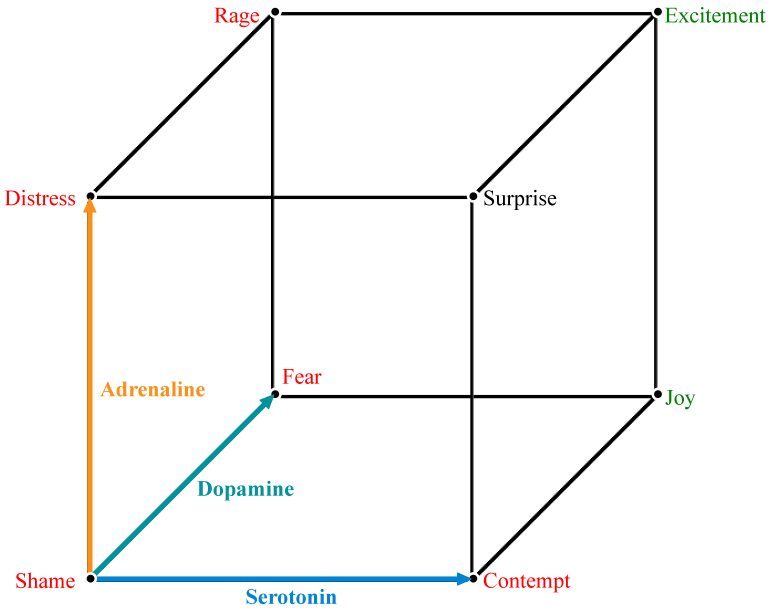



**Figure 8.4:** Three-dimensional model of emotions [130] with monoamine neurotransmitters as individual axes. The axes represent Serotonin (blue), Domamine (green), and Adrenaline (orange). Arrow tips indicate the high levels of each neurotransmitter.

emotions organised into various clusters [58]. There is also an ongoing debate about whether there are actually a finite number of emotions [152]. However, as noted by Lovheim [130], there seems to be a general consensus that there are between 5 and 10 basic emotions.

Regarding aspect dimensions, there is a need to explore what qualities elicit specific emotions in humans. While this could be driven by design choices (as with chapter 5), it is prudent to continue taking bio-inspiration. In humans, the monoamine system (Serototonin [4], Dopamine [52] and Adrenalin [199]) is responsible for the regulation of emotion and the eliciting of behaviour. This has been demonstrated clinically with antidepressants and anti-psychotics working by interfering with this system. As increase or decrease of an individual monoamine changes a person's emotional state, they make ideal candidates for aspect dimensions.

This is acknowledged in the 3-dimensional model of emotions described by Lovheim [130], based on the 8 basic emotions defined by Tomkins [204]. Due to its structure, and grounding in neuroscience it makes an ideal model for ASM within the Trilogy architecture. This model is detailed in Figure 8.4.

In Figure 8.4, Serotonin is represented on the $X$ axis, Adrenalyn on the $Y$ axis, and Dopamine $Z$ axis. The tip of each arrow represents the maximum effect of the specific neurotransmitter, the common origin of each arrow (bottom left) represents the minimum (or no) effect from each transmitter. Positive states (according to the Loveheim model) are labelled in green, negative states are labelled in red, and the only neutral emotion (surprise) is labelled in black.

In Unity3D, each affect is implemented as a Vector3. All the individual affects are stored in array, which can be evaluated against the agents physiological state when calculating the current affect. In Unity this is achieved using the inbuilt *Vector3.Distance()* function.

### 8.3.2 Cognitive Domain

The Cognitive Domain is responsible for the ethical reasoning. Three behaviour modules have been developed for this purpose. The three modules are for Egoist, Utilitarian, and Altruistic reasoning based on the Ethical Vessels approach. However, in chapter 3, the Ethical Vessels were purposefully simple as was the task they were subjected to (namely the two lights experiment). This chapter deals with significantly more complex agents, in a more complex environment. So while the essence of the individual normative behaviours remains the same, the specific implementation is different, which will be detailed over the following subsections.

Importantly, each behaviour module suppresses or enhances specific Conative drives. The drives referenced in this section will be described in the following section (subsection 8.3.3), and are detailed in the Table 8.1. Importantly, each suppression or enhancement value is biased by the weighting of the particular module.

For example, consider the Egoism module attempting to enhance the striving drive by +2. If the Egoism module was currently weighted at 0.1, then the striving drive would only be enhanced by 0.4. Each of the specific suppression and enhancement values are included in Appendix A.

**Table 8.1:** Conative Drives.

| Drive | Description | Behaviour |
|---|---|---|
| Striving | Move towards goals, such as locations of interest | Seek |
| Desire | Move towards the average position of neighbouring agents | Cohere |
| Conformity | Align with the average heading of neighbouring agents | Align |
| Withdraw | Move away from the average position of neighbouring agents | Separate |
| Impulse | Avoid obstacles | Avoid |
| Instinct | Move away from repulsers (sources of fear) | Flee |
| Volition | Move towards neighbours in need of aid | Seek |
| Waver | Deviate in direction | Wander |

### Egoism

As with the two lights experiments, the Egoism Vessel is functionally straightforward as the Egoist is focused on self interest alone. As an Egoist ignores the needs of other agents, no complexity is required to recognise or respond to the states of other agents. For this reason, the Egoism behaviour module simply needs to inhibit some Conative drives, and enhance others.

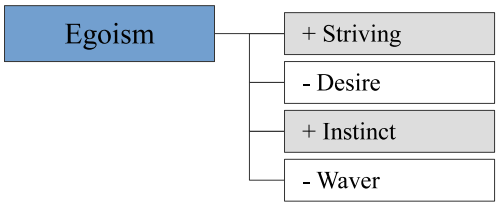



**Figure 8.5:** Diagrammatic representation of the Egoism module.

This is highlighted in Figure 8.5. *Striving* has been enhanced, as has *Instinct*. This encourages the agent to seek their goals, and flee repulsers. By contrast *Desire* has been suppressed, reducing the agents drive to move towards groups of other agents. Furthermore, *Waver* has also been suppressed. This was a design decision; suppressing *Waver* made the agents appear more purposeful and self driven.

### Altruism

In the Ethical Vessel approach, the Altruistic Vessel was calm when the agents around it were content, and aggravated when the agents around it had low valence. In the simple

two lights experiment, this was sufficient to elicit behaviour that could be described as ethic-like. However, some modification is required to suit the Trilogy architecture and a more complex task environment

This is represented in Figure 8.6. Firstly, the majority of the altruism module is only activated if the Altruistic Community Valance (ACV) is less than 0, indicating that there is more negative than positive valence within the agent's established neighbours. If the ACV is greater than 0, the agent can assume the current behaviour they are observing is successful from an Altruistic position.

If the ACV is below 0, the agent will modify a number of Conative drives. Firstly, *Desire* and *Conformity* are suppressed; This causes the agent to move away from other agents. Thus, if the agent is blocking the escape of another, it should move to allow them past. Furthermore, *Striving* and *Instinct* are suppressed, reducing the agent's desire to seek goals or move away from repulsers, essentially reducing the agents self interest. By contrast, *Withdraw* is enhanced, since after some experimentation it was discovered that adding more deviation to the agent's short term headings helped it move out of the way of other agents.

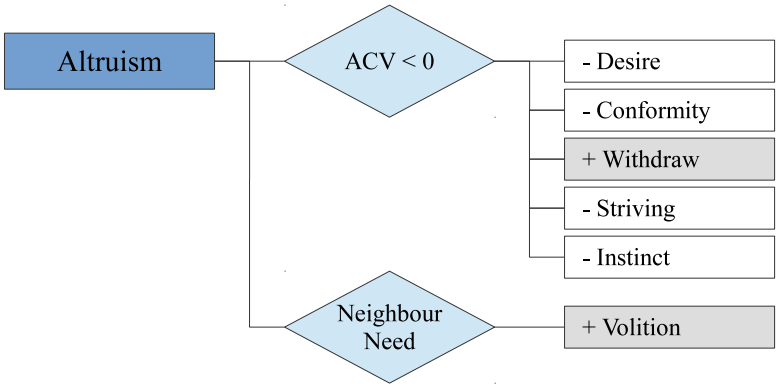



**Figure 8.6:** Diagrammatic representation of the Altruism module.

Secondly, the Altruistic agent will respond to agents in need regardless of the ACV. This is achieved by enhancing the *Volition* drive.

#### Utilitarianism

The Utilitarian behaviour module follows the same basic process as the Altruistic, suppressing a similar set of Conative processes. However, while the Altruistic module is based on the Altruistic Community Valence (ACV) excluding itself from the calculation, the Utilitarian is based on the Community Valance (CV). The CV includes the agent's own valance in the calculation, meaning it retains some self-interest. Furthermore, while Altruistic behaviour suppresses the Conative drives which serve its own needs (namely *Striving* and *Instinct*), the Utilitarian does not.

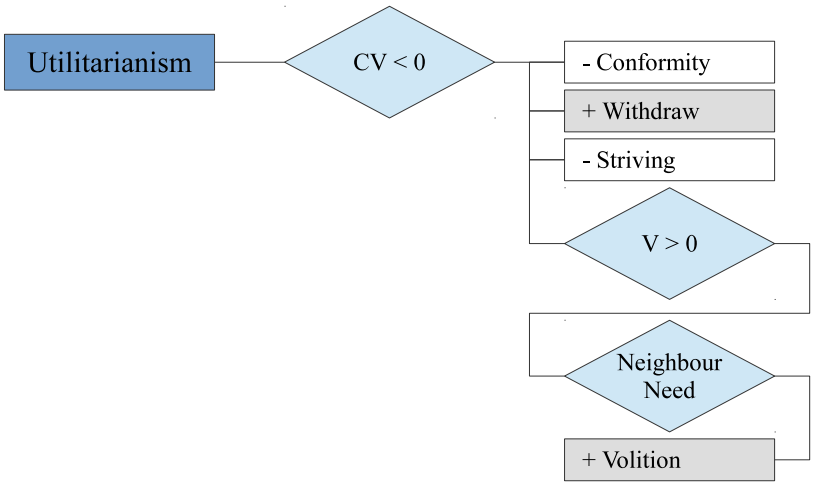



**Figure 8.7:** Diagrammatic representation of the Utilitarianism module.

In contrast to the Altruistic model, the Utilitarian only helps other agents in need if the CV is below 0, and the agents own valence is above 0. Put simply, if the Community Valence is high, the agent will continue with its current behaviour and make no modifications. It does not suppress its own needs, as it is also a member of the community.

### 8.3.3 Conative Domain

The Conative Domain is responsible for handling the basic drives of the agents, causing them to move towards certain states and away from others. Each drive module is implemented as a steering behaviour, the particulars of which are detailed in Table 8.1.

Desire, Conformity and Withdraw work in tandem, as with the standard 'Boids' flocking algorithm. However, the weighting of the desire and withdraw behaviour are varied based on the distance from the average position of the neighbouring agents (the flock-centre $F$) in comparison the agents vision distance ($V$). The *desire* behaviour weighting is gradually suppressed the closer the agent gets to the flock-centre, conversely the *withdraw* ($W$) weighing is suppressed as the agent gets further away.

$$Desire = (D/V)$$

$$Withdraw = 1 - Desire$$

Other drives have a fixed base weighting and do not vary unless suppressed by another domain. The weightings for each domain are detailed in Appendix A.

**Motivation**

Within the Conative Domain, there is a function which manages the suppression and enhancement of specific modules. At the end of each time-step, the modules in the Cognitive Domain add any current suppression or enhancement values to an array in the Conative. Each entry in this array has two values, the first is a reference to the drive it relates to, the second is the amount the drive will be suppressed or enhanced by.

At the beginning of each time-step, the suppression or enhancement values for each drive are added together to create a single value. The default weight for the drive is then multiplied by this value.

For example, the Waver drive has a default weight of 2. At the beginning of the time-step, there are two values for that drive, $-0.5$ and 1. These two values are added to produce 0.5 and multiplied by the default weight, resulting in the final weight of 1. This value is referred to as the motivation (as described in Equation 7.3.2).

Each drive is a steering behaviour, in simple terms, a vector that represents an ideal heading to meet the objective of that particular drive. This vector is multiplied by the motivation,

changing the magnitude, and ultimately the bias of that vector when its interpreted within the Act Layer, which is explained in detail in the following section. Due to the nature of steering behaviours, inverting the vectors by multiplying by a negative would produce unwanted artefacts. For this reason the motivation is clamped to a minimum value of 0, which would result in that steering behaviour being fully suppressed.

## 8.4 Act Layer

The Act Layer contains two locomotion modules. The first takes all the steering vectors generated by the modules in the Conative Domain of the Think Layer and combines them to create a composite steering behaviour. As with the Boids implementation of steering behaviours [169], this is a simple process of generating the average vector.

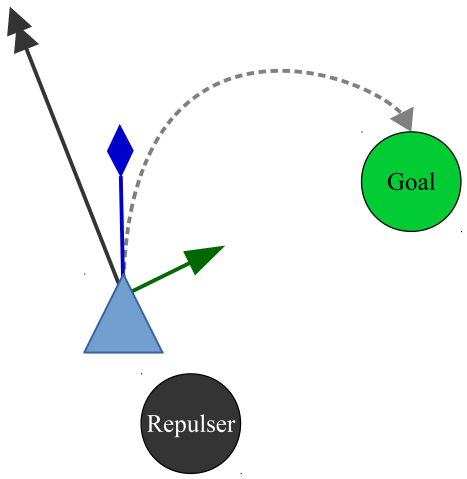



**Figure 8.8:** Diagrammatic representation of an agent responding to two steering vectors.

In the example in Figure 8.8, the agent is responding to two steering vectors. The first, is a seek behaviour (Green arrow), pulling the agent to the right towards a goal; the second is a flee behaviour (grey double-arrow) drawing the agent away from a repulser. The agent moves along the average vector (blue diamond). The hashed line shows the likely path of the agent following the compound vector. As the agent gets further from the repulser, the weight of that steering behaviour is reduced, causing the agent's path to arch towards the goal.

The second module takes the compound steering vector and calculates how the agent will move. Firstly, the angle of the vector relative to the agent's current heading is ascertained. Based on this angle a turning acceleration, left or right is applied to the agent's current movement. Secondly, the magnitude of the compound vector dictates the ideal speed of the agent, if the agent's speed matches the desired speed, it continues with its current momentum, otherwise it accelerates or decelerates to match the desired speed.

## 8.5 Tuning for Face Validity

The settings and modules within the Trilogy architecture were devised through a common-sense understanding of the expected output. However, before formally evaluating the model, the face validity can be verified against reference videos as described in section 6.4.

Calibrating and validating the agents requires reference data of human behaviour in ethically challenging circumstances. An original data-set of crowds reacting to high-stress incidents was required. However, as video footage requires that a camera be present during a one-off incident, the number of videos freely available with licence for reuse, are limited and often low in quality. Furthermore, the majority of these videos are available through online streaming platforms. While this in itself is not a problem, videos can be deleted or removed with little or no notice, making their use problematic.

A decision was made to only use videos from legitimate, professional news sources. As many legitimate news outlets now make their content available online through the same public streaming platforms, so they are easily accessible. Also, while there is no more guarantee that individual videos will not be removed, as they are from a professional news source, videos should continue to be available upon-request through the agency's archive. While this is not a perfect solution, it allows the videos to be harvested into a data-set and provides some robustness to changes in availability.

As the intention is to simulate a spectrum of ethic-like behaviour for validation, videos from a variety of events were selected. In a specific animation setting, more precise footage

could be used, such as only videos of people running from a fire. The videos used are listed in Appendix B. The following sections discuss the behavioural phenomena observed.

## 8.6 Task Environment

To undertake the experiment, a square environment with obstacles and narrow exits was developed. This was designed as an analogue to large indoor spaces and concourses, such as conference centres, shopping malls, and bazaars (for example see Figure 8.9). This environment was selected because it represents a real-world situation that a participant may be familiar with.

**Figure 8.9:** Meadowhall Oasis, a real-world example of a shopping mall. Original image taken by Gregory Deryckère published under a *CC BY-SA 3.0* licence.

The static repulser was designed to represent a shock event in a fixed location, such as an explosion. This is highlighted in Figure 8.10, the repulser is represented by a solid red circle. As described in the previous section, if the agent sees a repulser it will attempt to move away from it if it is within a 40 unit radius even if the agent is not currently looking at it. This is in an effort to simulate the noise created by the types of shock events being simulated.

In addition to the general movement behaviour, a second dilemma to test the ethical behaviour was designed into the simulation. There was a random chance (1 in 2000) at each frame of the simulation that an agent may become injured. If this happened, the agent's

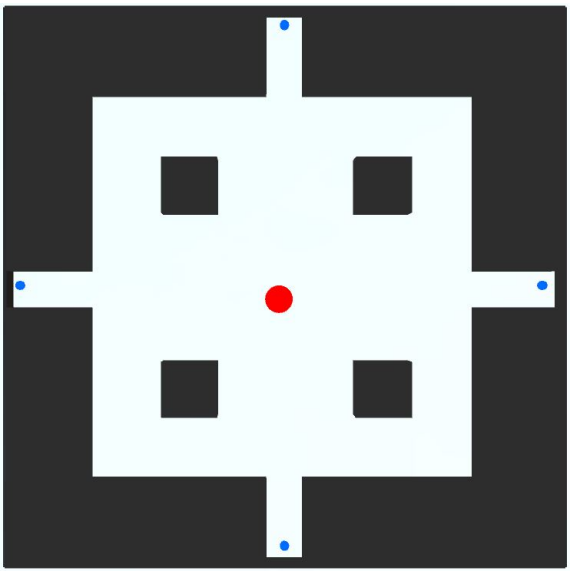


**Figure 8.10:** The empty test environment: the repulser is highlighted in red, exits are blue, obstacles are black.

speed would reduce to 0.2 of standard, simulating a minor injury, such as a fall or a twisted ankle. To highlight that the agent is injured, their colour changes to red. If the agent is injured it 'calls for help', which is 'heard' by any agents within 20 units. If another agent comes close enough to touch the injured agent, then assistance has been rendered, and the injured agent's speed and colour are reset to normal.

### 8.6.1 Ethically Motivated Events

Part of the validation process described in section 6.4 referred to the noting of events from reference footage. This related to specific macro-phenomena emerging from the crowd itself, rather than the more subtle events that may emerge from ethical decision making. The following subsections categorise observed events that appeared to be driven through ethical motivation. To distinguish from the objectives of the crowd events (see subsection 6.4.2), these have been referred to as Ethically Motivated Events.

It is important to note that it is difficult to determine intention from action. The following categories of macro-phenomena have been identified for tuning, because they *appear* to be driven by ethical concerns from a purely observational standpoint.

### Bystanders

A Number of the reference videos highlighted bystander behaviour. Where some of the individuals in the scene would ignore other individuals in distress. For the purpose of this validation exercise, an individual was marked as exhibiting the bystander behaviour if all the following conditions were met:

1. The individual visually noticed another in distress.
2. The individual does not take any action towards rendering assistance.

To frame this behaviour within the context of the Ethical Vessels, it could be described as Egoism as the agent is not making any motion towards helping another that is clearly in need. The agent is instead protecting its own interests; an Egoist is not morally obligated to provide aid to another.

### Somebody Else's Problem

In the archive footage a variation in the bystander behaviour was observed, while a subtle variant in definition, it is a visually distinct behaviour and worth noting. This variation (referred to as Somebody Else's Problem or SEP [2]) is when an agent begins to move towards rendering assistance, but then decides against it and moves away:

1. The individual visually noticed another in distress.
2. The individual moves towards providing assistance.
3. The individual changes behaviour and moves away from providing assistance.

In the footage, SEP generally appeared to be a self preservation response based on a significant increase in the amount of perceived danger.

SEP could be framed ethically in a number of ways; however, two ways can be supported using the insights from earlier chapters. The first is as a conflict resolution behaviour

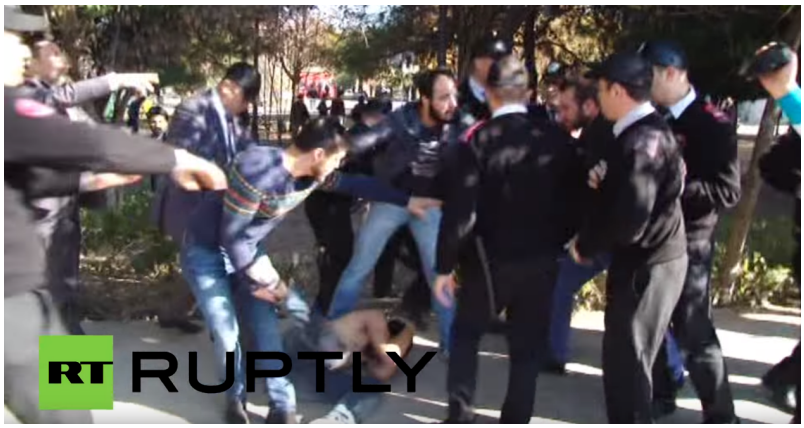



**Figure 8.11:** Students protect an injured protester being kicked by police in Turkey (Video 20).

between Altruistic and Egoist behaviours, as with the value system implemented in chapter 4. The second way this could be explained is through Utilitarian behaviour, for example, if conditions initially meant that rescuing the agent represented the 'most good' action for the community. If conditions then changed and assisting the agent reduced the global utility of the community, then the right course of action would be to leave the individual to its fate.

### Heroism

Heroism was defined as anytime an agent moved closer to a point of danger to provide aid to another agent. This behaviour was rarely observed in the archive footage in comparison to the other behaviours. This is best described as Altruistic behaviour due to the agent's selfless concern for another individual. Figure 8.11 highlights an example from video 10 where a number of students have moved to assist an injured individual during a clash with police in Turkey.

### Escape Behaviour

A number of different behaviours were observed when a group attempted to escape a source of danger (for example see Figure 8.12). Firstly, The majority of the crowd move normally, limited by the speed of the agent in front of them. This behaviour could be considered as a 'normal state', but it could also be considered among the spectrum of Utilitarian behaviour.


**Figure 8.12:** Russia Today video of a crowd being broken up by a water cannon (Video 16).

As the community utility is best served by collectively moving away from the danger, observing an order is an acceptable course of action.

Secondly, there were a number of individuals who did not observe the order and would push to the front. These agents often appeared to be in a state of panic, and were often the agents who had initially been closest to the source of danger. As these agents appear to be focused on their own needs over their neighbouring agents, the behaviour could be categorised under Egoism. Finally, there were agents who would move aside to allow other agents to pass, putting themselves closer to the source of danger. This could be described as Altruistic behaviour.

### Unity

Another observable phenomenon in the archive footage is the grouping and clumping of agents when they were faced with a dangerous situation. This behaviour is apparent both when agents are escaping a source of danger, but also when they are facing it.

It is difficult to classify this phenomenon as ethical behaviour as it could emerge for a number of reasons. For example, in some videos it was clear that the grouping was simply the product of individual agents moving under influence of the same, Egoist goals. However, it could also emerge due to a number of agents collectively moving out of the path of others, which could be described as altruism, furthermore, it could be a result of pre-existing

social groups. Interestingly, for the experiments in chapter 3, grouping behaviour emerged through the following of simple, Utilitarian rules.

### 8.6.2 Animation Tuning

The parameters listed in Appendix A are the result of an animation tuning process, undertaken before the formal evaluation was undertaken (according to the process detailed in chapter 6). In this step, the models were tuned to good face validity, according to subjective assessment. To evaluate this, an enclosed test environment was created with 40 agents, and included a repulser (a source of fear). In 129 simulations, parameters were tuned until the behaviour matched, to face validity, what the designer was expecting. This was then checked against the ethically motivated events, evaluating whether the parameters allowed phenomena listed above to emerge naturally.

At this point, the model was considered to have good face validity, and ready for formal evaluation. Building a model with face validity is a common component of the computer simulation verification process, according to the Naylor and Finger model [150]. However, as this is not intended as an exact simulation of a specific event, but as a general model for simulating ethical behaviour, evaluation beyond input-output validation is required.

The following subsections highlight the tuning for three models. Each model was biased towards one of the normative modules, weighted .7 for the biased behaviour, and .15 for the two remaining behaviours. This allowed for some blending of the behaviours, but with the biased behaviour remaining dominant. In each figure, sub-figures have been taken at 120 frame intervals (roughly 2 seconds).

#### Egoism Biased

Figure 8.13 is an example of the Egoism behaviour after tuning for face validity. In this example, two egoist-like behaviours were observed.

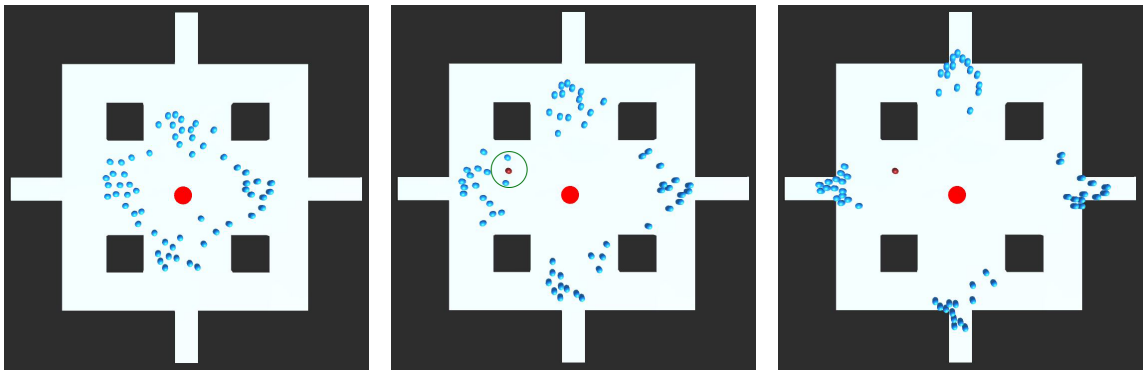



**Figure 8.13:** Tuning of the Egoism behaviour module.

Firstly, the agents do not follow a rough group structure. As can be observed from the three frames, the agents move directly away from the repulser in the centre, and towards the nearest exit. When the agents reached the exit, they would ignore all neighbour agents, and simply try to reduce the distance from themselves to the exit. This often resulted in an obstruction at the constriction point of the exit, as can be observed in Figure 8.13c.

Secondly, injured agents (highlighted in red) were ignored. As can be seen in Figure 8.13b, one agent is injured, in close proximity to two others (highlighted by a green circle). However, in Figure 8.13c the injured agent is still in the same position, indicating that neither of the agents rendered assistance. It is worth mentioning that in this simulation, the bottom-most of the two agents initially started turning towards the injured agent, moved very slowly then continued moving towards the exit. This behaviour was similar in appearance to the 'Bystander Effect' (see subsection 8.6.1). This was caused by the blending of steering behaviours, the altruism module created a steering vector towards the injured agent which resulted in a temporary change in heading, and reduction of the magnitude in the compound vector.

### Altruism Biased

The altruistic biased model produced in some visually different behaviours. The example in Figure 8.14 highlights three in particular.

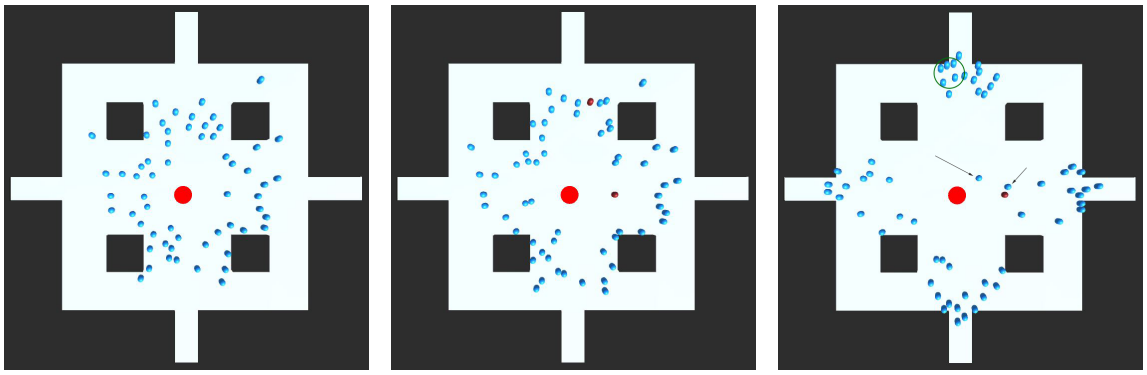



**Figure 8.14:** Tuning of the Altruism behaviour module.

Firstly, while the agents are still moving away from the repulser and towards the exit, their movement is not direct. In most cases the agents keep a larger distance between themselves, which often results in a path which is less direct than the egoist. However, this additional spacing results in an easier movement into the exit, with less clumping and obstruction.

At the exits a second behaviour can be observed, best seen in the top, and right exit in Figure 8.14c. Upon reaching the exit, if there was a constriction, it was common to see an agent move to the side, against the wall, allowing others to pass (highlighted by a green circle in the top exit). This was caused by the decrease in the desire drive, which causes agents to seek groups of other agents; the reduction in the striving drive, which causes the agents to seek the exits; and an increase in the withdraw drive, which causes agents to move away from groups.

The final behaviour observed is the response to injured agents. In contrast to the Egoism bias model, a number of agents would immediately move to support an injured agent. In Figure 8.15c two agents have actually moved closer to the repulser while undertaking this goal, their paths highlighted by black arrows.

### Utilitarianism Biased

Although the utilitarian behaviour shared some similarities in design to the altruistic, some different behaviours are observed in the utilitarian biased model.

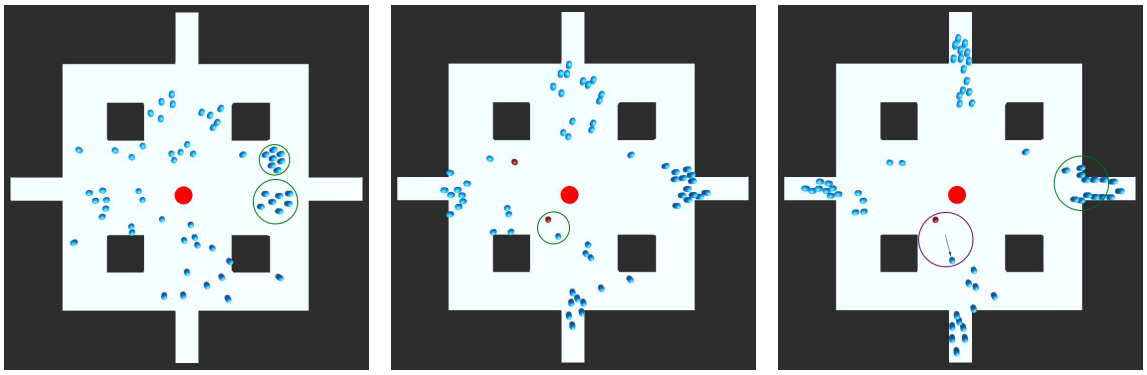



**Figure 8.15:** Tuning of the Utilitarianism behaviour module.

Firstly, the utilitarian components of the module were activated significantly less often. As the utilitarian module used the community valence, this naturally biased the agents actions towards its own valence. This allowed the underlying flocking behaviour of the three conative drives, Desire, Withdraw, and Conformity, to emerge. However, when the utilitarian module was activated the flocking behaviour was actively suppressed (as with the altruistic behaviour), this resulted in small, sub-flocks emerging, as highlighted by green circles in Figure 8.15a. Once this behaviour was noticed, it was tuned for, as it was similar in appearance to the Unity behaviour.

This sub-group behaviour results in a different exit behaviour. Notably, the agents would occasionally form queue-like structures. This can be best seen in the right-hand exit in Figure 8.15c, highlighted by a green circle. The reality is that because the agents arrive in a small group at different times, they naturally stack up. But, because they arrived as a group, rather than individuals coming from a variety of angles, these small groups generally moved through the constriction easier, though this was not always the case. As with the altruistic model, occasionally agents would be seen moving to the side allowing others to pass. This generally happened after an egoist-like clump at the exit constriction. This can be seen in the right exit on Figure 8.15c.

Another difference with this model is the rescuing behaviour. As with the Altruistic module, the Utilitarian has the capacity to rescue injured agents. However, in the utilitarian module, this is only activated when the community valance was low, and its own valence was high. This meant that on occasion, an agent would begin moving towards another agent and then

change direction and continue towards the exit. This can be observed in Figure 8.15b and Figure 8.15c (highlighted by a purple circle). This can be compared to the 'Somebody Else's Problem' behaviour discussed in item 8.6.1.

## 8.7 Summary and Discussion

This chapter has discussed a specific implementation of the Trilogy architecture (described in chapter 7) that will be used to evaluate the approaches proposed in this thesis. The initial sections discussed the technical implementation in the Unity3D game engine. This covered the three layers (Sense, Think, Act) and the three domains (Affective, Conative and Cognitive). The final sections discuss how this implementation was tuned for face validity by comparing the output of the simulation to behaviours observed in reference footage.

# Chapter 9

# Results of Ethical Simulations

> *At times, morality can be dismissed as a matter of personal conscience, no matter how widespread its acceptance. Ethics, on the other hand, arises from societal or group commitments to principia of behavior.*
>
> **— Sherwin B. Nuland**

## 9.1 Introduction

This thesis was motivated by a need to recreate ethic-like behaviour in virtual characters. In earlier chapters, methods have been defined that achieve isolated parts of this goal. However, until now, these components have not been brought together in a single simulation.

Chapter 7 established a software framework to allow modules which control individual aspects of a character's behaviour to be brought together. This framework, the Trilogy architecture, was based on earlier work in behaviour-based robotics and behavioural animation. The Ethical Vessels approach and the Affective States Modelling, were then implemented within the Trilogy architecture (see chapter 8) to enable its use in simulation.

In this chapter, the ethical simulation is evaluated through two experiments. The first establishes whether the inclusion of an affective layer improves the believably of simulated ethical behaviour. The second, establishes whether the behaviours are identifiable beyond

chance accuracy against a modified control. The following sections describe the specific experimental design.

## 9.2 Experimental Setup

The two experiments in this section follow the methodology proposed in chapter 6.

Six motivated models and an Unmotivated control were generated. Three Affective-motivated models were created with an altruistic, Utilitarian or Egoist bias (as described in the previous chapter) A further three motivated models were created by taking these original models and disabling the Affective layer of the model.

The control (Unmotivated) simulation was recorded with both the Affective, and the ethical functioning deactivated (the Affective, and Cognitive domains). This leaves the agents following their base Conative steering behaviours (as described in subsection 8.3.3).

Each of the seven models were simulated in the test environment. For each simulation type three separate videos were recorded with slightly varied initial conditions (the position of the agents and repulser). The particular video the user was shown was selected at random.

### 9.2.1 Survey Design

An online survey engine was developed in JavaScript (using the JQuery library) and HTML, storing data via JSON into a Google Firebase database. A configuration file stored an array of the models to be assessed; each model name was used as reference to associate to a list of possible videos for that model. Each time the survey homepage was loaded, the array of model names was shuffled, changing the order of models shown to each participant.

Before the participant began the survey, they were provided with a brief introduction. This opening text explained the purpose of the survey, and explains to the participants that their

responses are fully anonymised and only used for the purpose of this research. They are asked to click a check-box confirming consent before moving on to the participant details page.

Before starting the survey, the participant was asked to select an age category, gender, and their nationality. For each of these data categories 'prefer not to disclose' was provided as an option. No personal identifiable information (such as name, birth date, email or geolocation) was stored; and the participant was informed of this. Once the participant had clicked 'Begin', a new survey record was added to the Firebase, with an ID number and time-stamp.

At the end of the survey the participant was again given the opportunity to withdraw from the study. They were informed that by pressing submit, they were agreeing to their responses being used in relation to the research, and after their data had been submitted, it would no longer be possible to withdraw.

**Participant Guidance**

As specified in the methodology (see chapter 6), the participant was provided with a basic specification for each of the behaviours they would be evaluating. In experiment 1, the participant was provided with descriptions (as below) of Egoism, Altruism, and Utilitarianism. Experiment 2 had the same descriptions as experiment 1 with the addition of the description of the Unmotivated behaviour.

**Egoism:** Actions which benefit the agent's self interest, without any obligation to other agents.

**Altruism:** Actions that are not beneficial (or may be harmful) to the agent but that benefit other agents.

**Utilitarianism:** Actions which benefit the community of agents as a whole, rather than a specific individual.

**Unmotivated:** Actions which are not motivated by ethical concerns.

## 9.3 Affective States Evaluation (E1)

In chapter 5, a method to facilitate the simulation of emotion was proposed and subsequently built into the design of the Trilogy architecture in chapter 7. This was in response to the hypothesis that recreating believable ethical behaviour may require simulated emotions. This experiment tests this hypothesis.

A slight variation of the experimental methodology outlined in chapter 6 was used in this test. The participant was presented with the six motivated models and asked to rate each of them. This required an alteration to the first stage of the test. Instead of identifying each of the variants (non-Affective and Affective), the participant was asked to simply identify the underling ethical model. This resulted in the participant having to evaluate six videos individually, but only against three options. As the objective was to ascertain whether the participant could recognise the behaviour, including the variants into the list of options could have confused the evaluation. No modification was required for the believability portion of the assessment.

### 9.3.1 Sample

The survey was launched online and run for 14 days, resulting in 51 completed surveys. A further two individuals started the survey but did not complete the questions, and were removed from the results. Of the 51 individuals, 20 were female, 31 were male; there was a distribution of age ranges with a mode and median of *25-34* (see Figure 9.1). The majority of responses were from the UK, with four respondents coming from the USA, and a further two from Germany.

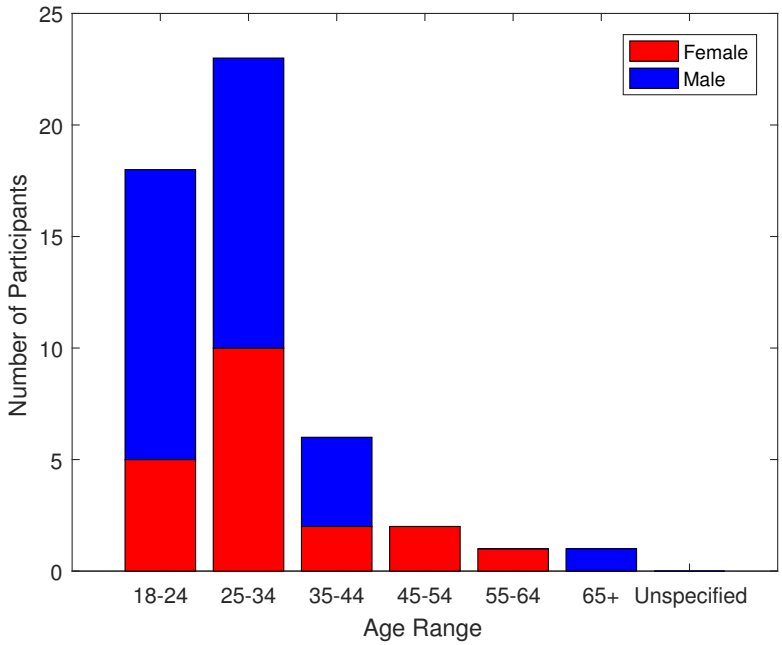


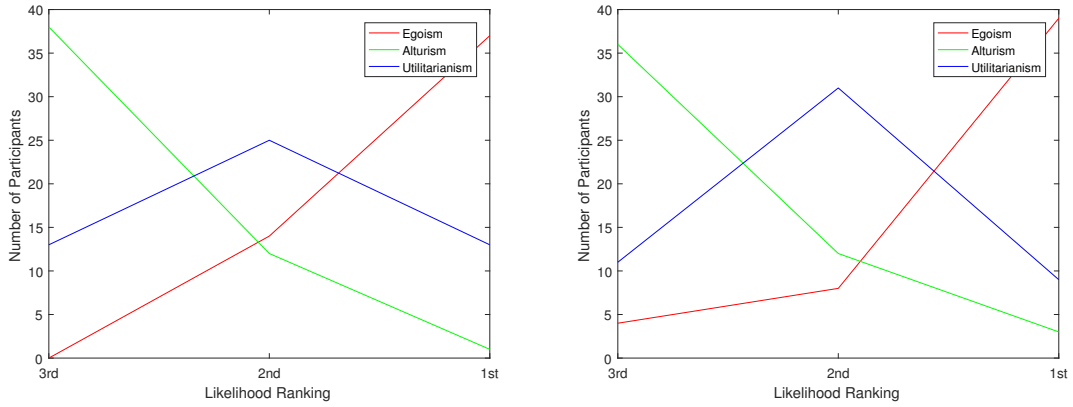

**Figure 9.1:** E1: Age and Gender of participants.

### 9.3.2 Results

#### Identification

As demonstrated in the following subsections, there is relatively little difference between the identification of the non-Affective, and Affective models. This is reflected in the Summarised results, where (over the three models), 64.7% of participants made correct identifications (1st place rankings) for the non-Affective models, and 66% made correct identifications in the Affective models. Very few participants opted to leave a free-text response to support their reasoning.

While the first and thirdplace rankings remain relatively consistent, there is some reduction in confidence for the second place rankings in the Affective models. This can likely be attributed to the blending created by the Affective layer, reducing clear distinctions and allowing some alternative behaviours to emerge.




**(a)** Non-Affective Model **(b)** Affective Model

**Figure 9.2:** E1: Participant rankings plotted for the identification of the Egoism model.

**Egoism** There was little difference between the identification of the Affective and non-Affective models, with 72.5% and 76.5% (respectively) ranking Egoism as the most likely behaviour being displayed. Overall, the non-Affective version was moderately more identifiable, with 100% (as opposed to 92%) of participants rating Egoism as the first, or second-most likely behaviour being displayed. Altruism was considered the least likely behaviour across both variants, and Utilitarianism was considered the second most likely. These trends are reflected in Figure 9.2 and Table 9.1. Using Friedman's test shows the rankings for each of the three options for both models is significant with a p-value of 2.1987e-12 for the non-Affective variant and 1.3503e-10 for the Affective variant. Neither model demonstrated a significant gender effect using a one-way ANOVA.

Interestingly, while the confidence for the first choice as most likely is relatively consistent between the two models, there is some difference in the participants choice of second most likely. This trend is observed in Figure 9.2, where there is clear difference between the divergence of the participants responses. The second place rankings are all relatively close for the non-Affective model, although Utilitarianism is the highest overall. However, in the Affective model, the participants appear to have greater confidence that Utilitarianism is the second most likely.

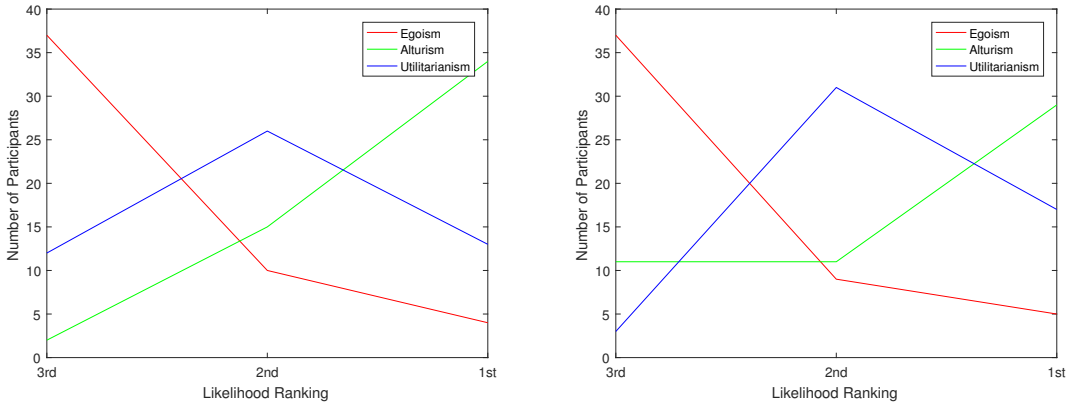
Only one free-text response was provided against the Affective model, where the user had identified the model as Utilitarian. However, interestingly this user claimed their decision was because "...the people left injured individuals behind so that everyone else could escape

**Table 9.1:** E1: Summarised data for the identification of the Egoism model.

| **Non-Affective** | 1st | 2nd | 3rd |
|---|---|---|---|
| Egoism | 37 | 14 | 0 |
| Altruism | 1 | 12 | 38 |
| Utilitarianism | 13 | 25 | 13 |

| **Affective** | 1st | 2nd | 3rd |
|---|---|---|---|
| Egoism | 39 | 8 | 4 |
| Altruism | 3 | 12 | 36 |
| Utilitarianism | 9 | 31 | 11 |

easier". While it is impossible to draw conclusions from a single response, it does highlight an interesting case, namely that if acting in the manner of an Egoist is considered best for the general utility of the population, then that Egoist behaviour could be considered Utilitarian.




**(a)** Non-Affective Model **(b)** Affective Model

**Figure 9.3:** E1: Participant rankings plotted for the identification of the Altruistic model.

**Altruism** There was some difference between the identification the altruistic behaviour variants. Both were identified with relative success with 66.6% and 56.8% of participants ranking altruistic first. Using Friedman's test shows the rankings for each of the three options for both models is significant with a p-value of 9.9778e-10 for the non-Affective variant and 2.6667e-07 for the Affective variant. No gender effect was apparent when testing both models with a one-way ANOVA.

Furthermore, a total of 96.1% and 78.4% participants ranked altruism either first or second (respectively). Utilitarianism was the behaviour considered second most likely in both models, while Egoism was considered the least likely. This is reflected in Figure 9.3 and Table 9.2. A similar trend as observed in the Egoism model can be seen with the divergence of second ranking although, to a lesser extent. Around the same number of participants

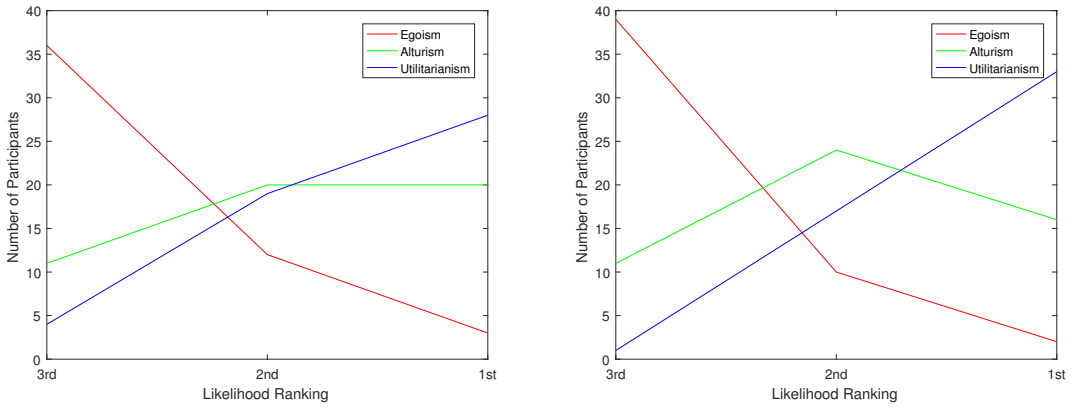
**Table 9.2:** E1: Summarised data for the identification of the Altruistic model.

| **Non-Affective** | 1st | 2nd | 3rd |
|---|---|---|---|
| Egoism | 4 | 10 | 37 |
| Altruism | 34 | 15 | 2 |
| Utilitarianism | 13 | 26 | 12 |

| **Affective** | 1st | 2nd | 3rd |
|---|---|---|---|
| Egoism | 5 | 9 | 37 |
| Altruism | 29 | 11 | 11 |
| Utilitarianism | 17 | 31 | 3 |

ranked Utilitarianism as most likely as did with the altruistic model as can be identified in the Summarised data in both tables. This is likely because the Utilitarian model contains elements of both the altruistic and the Egoist.

Three of the 51 participants (two for the Affective model, one for the non-Affective) left free-text responses regarding the identification of the altruistic models. All the free-text responses provided were from participants who had correctly identified the behaviour. The comments mentioned 'help' as a key word; for example, "two of them [agents] moved to help another".




**(a)** Non-Affective Model **(b)** Affective Model

**Figure 9.4:** E1: Participant rankings plotted for the identification of the Utilitarianism model.

**Utilitarianism** As with the Altruistic model, there is some difference between the identification of the Affective and the non-Affective models. With the non-Affective, 54.9% of participants rated Utilitarianism as most likely, as opposed to 64.7% for the Affective model. Overall the confidence on this model was slightly lower, as the Altruistic and Egoist models averaged around 71.6%. Furthermore, as can be seen in Figure 9.4, the non-Affective model is ranked as altruistic by a significant number of participants. Friedman's test shows the rankings for each of the 3 options for both models to be significant

**Table 9.3:** E1: Summarised data for the identification of the Utilitarianism model.

| **Non-Affective** | 1st | 2nd | 3rd |
|---|---|---|---|
| Egoism | 3 | 12 | 36 |
| Altruism | 20 | 20 | 11 |
| Utilitarianism | 28 | 19 | 4 |

| **Affective** | 1st | 2nd | 3rd |
|---|---|---|---|
| Egoism | 2 | 10 | 39 |
| Altruism | 16 | 24 | 11 |
| Utilitarianism | 33 | 17 | 1 |

with a p-value of 3.6804e-08 for the non-Affective variant and 5.0660e-11 for the Affective variant. Neither model demonstrated a gender effect when tested with a one-way ANOVA.

As with the Egoist and Altruistic models, the same response divergence can be seen in the second place votes in the Affective model, but to a lesser extent (as can be seen in Figure 9.4b). However, interestingly in the non-Affective model, there seems to have been little to no confidence in the second place ranking. In Both the Non-Affective, and the Affective models, the participants had significantly higher confidence in their thirdplace scores than in the first or second place.

Only four participants left free-text responses for the Utilitarian model. One comment was left for the non-Affective model and was correctly identified, three were for the Affective model and were also correctly identified. Interestingly, indecision seemed to be the focus of these comments and key to the users' ranking. For example, one participant notes "One [agent][1] went to help another, but changed his mind and left him so he could escape". Other comments follow a similar trend.

### Believability

In this section, the two components (Likert and rank) of the believability portion of the assessment will be described. For the purpose of this section, tables and figures will used shortened versions of the model names, with the prefix 'A-' used to denote the Affective variant of each model.

Despite some variation between individual users' ratings in the Likert portion of the test, there are some clear trends (best displayed in Figure 9.5). Firstly, for each model, the

[1]Square brackets have been used to enclose words added for clarification.

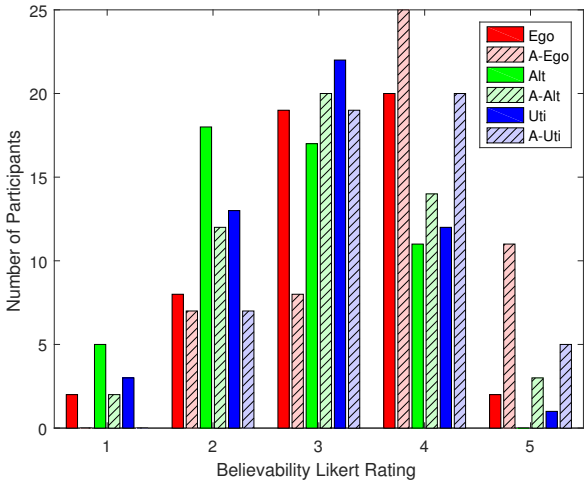



**Figure 9.5:** E1: Likert data for the Believability stage.

Affective variant generally has a higher rating. However, it should be noted that a number of models have similar modes and medians for the both variants. This is likely due to a classic mid-point bias in the participants' rating.

Utilising the Paired t-test between the Affective and non-Affective model variants, we demonstrate at the 1% level that the variants are significantly different from each other (p = 7.2693e-04).

Each of the normative models were tested for gender effect using a one-way ANOVA. Non-Affective Egoism had a p-value of 0.6868. Testing Affective Egoism resulted in a similar p-value of 0.6942. The non-Affective variant of altruism had a p-value of 0.9194, By comparison the Affective variant of Altruism had a p-value of 0.4465.The non-Affective variant of Utilitarianism showed possible gender-effect with a p-value of 0.0534. However, further investigation with a Spearman's Rank indicates there is no significant correlation (p of 0.1774 and RHO of 0.4544). Finally, the Affective variant of Utilitarianism had a p-value of 0.7461.

Each Affective model was then tested for statistical significance against the non-Affective variant. Testing the Affective variant of Egoism against the non-Affective variant resulted in a p-value of 0.6733 and a RHO of -0.0605 meaning there is no significant correlation between the rating of the models. 54.9% of participants rated the Affective variant higher

**Table 9.4:** E1: Summarised Likert data for the believability stage.

| | 1 | 2 | 3 | 4 | 5 | Total |
|---|---|---|---|---|---|---|
| Ego | 2 | 8 | 19 | 20 | 1 | 165 |
| A-Ego | 0 | 7 | 8 | 25 | 11 | 193 |
| Alt | 5 | 18 | 17 | 11 | 0 | 136 |
| A-Alt | 2 | 12 | 20 | 14 | 3 | 157 |
| Uti | 3 | 13 | 22 | 12 | 1 | 148 |
| A-Uti | 0 | 7 | 19 | 20 | 5 | 176 |

than the non-Affective variant, 27% rated the behaviours equally believable. Testing the Affective variant of Altruism against the non-Affective variant results in a p-value of 0.1243 and a RHO of 0.2180, which demonstrates no significance between the rating of the two models. 37.25% of participants rated the Affective variant higher, and 45.10% rated it equal to the non-Affective variant. Finally, testing the Affective, against the non-Affective variants of Utilitarianism resulted in a p-value of 0.5446 and a RHO of of 0.0868. 60.78% of the participants rated the Affective variant higher than non Affective variant; 19.61% of participants rated the two behaviours equally.

However, the graph does demonstrate that the Affective variants (hashed bars) are more dominant in the higher ratings (4 and 5). Furthermore, the totals of all ratings for each model (highlighted in the totals column of Table 9.4) demonstrate that the Affective variants were all higher rated, averaging around 17% higher than the non-Affective model.

Further evidence of this trend is provided by the ranking portion of the believability evaluation. For example, Figure 9.6 demonstrates that, in all models, the Affective variant is considered more believable. This is probably most evident in the Egoism model (the red, and red dashed line in Figure 9.6).

As with the Likert data, we can observe that the Egoism model is (overall) the highest rated with both variants considered. This is followed by the Utilitarian model, with the Altruistic model considered overall the least believable. Interestingly, the relatively flat trends of the non-Affective Egoist and Utilitarian models demonstrate that opinions were split between the participants on regarding these particular model variants. One hypothesis which could explain this trend is that participants were likely more confident about their

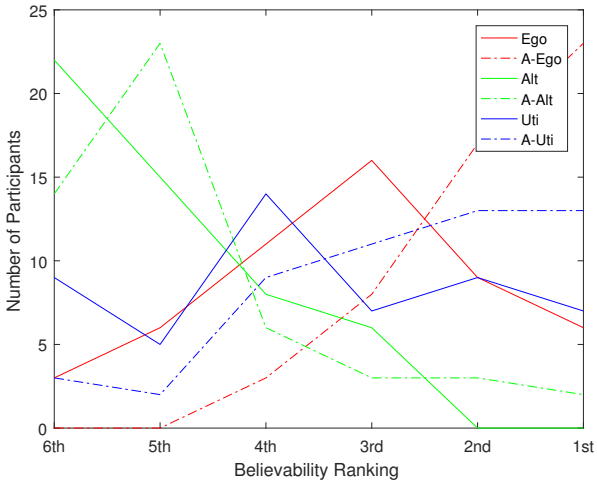



**Figure 9.6:** E1: Rank data for the Believability stage.

**Table 9.5:** E1: Participants believability rankings.

| | 1st | 2nd | 3rd | 4th | 5th | 6th |
|---|---|---|---|---|---|---|
| Ego | 6 | 9 | 16 | 11 | 6 | 3 |
| A-Ego | 23 | 17 | 8 | 3 | 0 | 0 |
| Alt | 0 | 0 | 6 | 8 | 15 | 22 |
| A-Alt | 2 | 3 | 3 | 6 | 23 | 14 |
| Uti | 7 | 9 | 7 | 14 | 5 | 9 |
| A-Uti | 13 | 13 | 11 | 9 | 2 | 3 |

first and last choices, with midpoint ordering being somewhat more arbitrary. A Friedman test demonstrated that the rankings for each behaviour were statistically significant with a p-value of 1.5690e-18. A one-way ANOVA showed that the ranking portion of the test was not influenced by gender bias.

Twelve free-text responses were submitted (four for Egoist models, five for Altruistic models and three for Utilitarian models) that provide some insight into the trends. The altruistic models were criticised as being unrealistic, these criticisms were not aimed at the visual replication of the behaviour, but rather the behaviour itself. For example, one participant noted “I didn’t think it [the altruistic behaviour] was very realistic, they [the agents] didn’t protect themselves.”. This hints that the participants thought that Altruistic behaviour itself was unbelievable in this circumstance. By contrast, the Utilitarian models were praised for showing a mix of behaviours; one participant noted “One of them [an agent] moved back to rescue another, even though everyone else was leaving him”.

**Overview of Results for E1**

The results indicate that in most cases, the inclusion of Affective states do not make ethical behaviour more recognisable. In fact, in the example of altruistically biased modules, the Affective variation is marginally harder to recognise.

However, in all cases the Affective variations scored higher in than the corresponding non-Affective variant in their overall believability. An exploration of the free-text responses provides some insight into why this may be. A number of users described the behaviour of the non-Affective agents as robotic, with one describing the behaviour as "too purposeful". By comparison, the Affective variants were described as demonstrating more human-like qualities, such as indecision. One participant wrote, "they take time to make a decision". However, it should be noted that only 11 of the participants chose to provide a free-text response, so these insights cannot be considered conclusive.

Thus, we can conclude that, in this example, while emotional simulation may not be necessary for ethically motivated behaviour to be recognised, it makes the behaviour more believable.

## 9.4 Evaluation Including an Unmotivated Control Behaviour (E2)

Evidence from the previous experiment supported the hypothesis that including an Affective layer improves the believability of ethical simulation. Furthermore, it demonstrated that participants were able to correctly identify the ethical models.

This section expands on that insight and proposes a second hypothesis. This hypothesis is that the *simulated ethical behaviour is identifiable beyond chance accuracy when compared to an unmotivated control behaviour*. With this considered, a second experiment was defined to test this hypothesis. Following the methodology proposed in chapter 6, an

Unmotivated control model was included in the evaluation. The purpose of this was to provide a baseline for comparison. By including this Unmotivated model, it is possible to explore whether each behaviour is identifiable in comparison to a standard Boid-type behaviour.

Furthermore, due to this inclusion it is possible to establish whether the additional Affective and Cognitive layers make the simulation more, or less believable. This leads to the sub-hypothesis of this experiment, notably that *behaviour which includes ethical, and affective simulation is more believable than a Boid-type behaviour*.

### 9.4.1 Sample

The survey was launched online and run for 18 days, resulting in 78 completed surveys. Of the 78 individuals, 32 were female, 46 were male; there was a distribution of age ranges with a mode of *18-24* and a median of *25-34* (see Figure 9.7). The majority of respondents were from the UK, with nine responses from the USA, two from Ireland, and two from Spain.

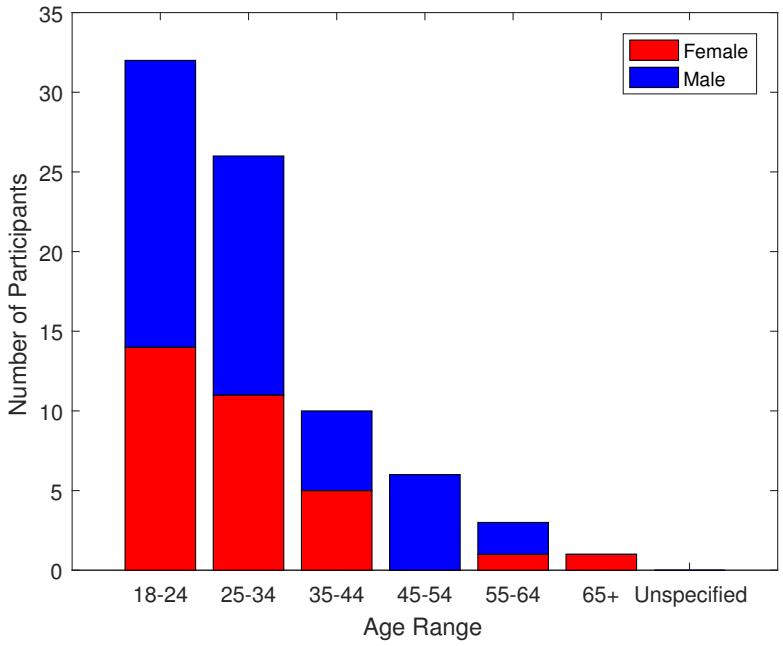



**Figure 9.7:** E2: Age and gender of participants.

### 9.4.2 Results

#### Identification

The following results demonstrate that the majority of the participants were able to identify the correct behaviour in each of the models. As with the previous experiment, there is less overall confidence for the mid-rankings (2 and 3). All rankings in the identification stage of the experiment demonstrated statistical significance when evaluated with a Friedman test.

Some of the results indicate a little confusion between the Egoist and the Unmotivated behaviours; however, Altruism and Utilitarianism are clearly identified.

**Egoism** An interesting point in Egoism identification data is that the participants were more confident of what the behaviour was not, rather than what it was. As noted in Figure 9.8 and Table 9.6, the participants were highly confident that the least likely behaviour was altruism (69 of the 78 participants ranking it as least likely). There is also relatively high confidence that the behaviour is not Utilitarianism, voted third least likely by 50 of the participants. There is less confidence regarding the second place rank, with the votes split between Egoism, Utilitarianism, and Unmotivated.

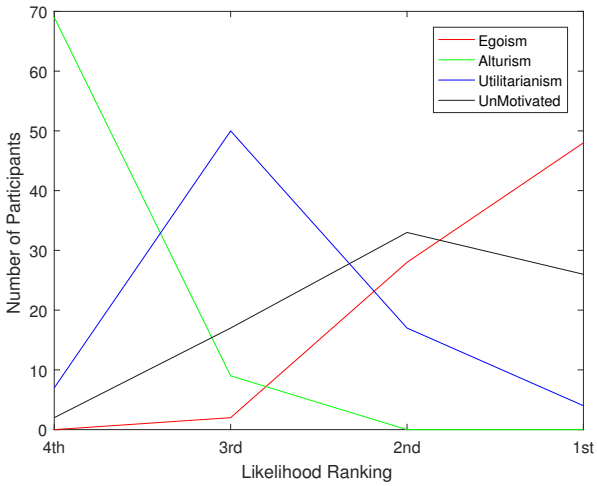



**Figure 9.8:** E2: Participant rankings plotted for the identification of the Egoism model.

**Table 9.6:** E2: Summarised data for the identification of the Egoism model.

| | 1st | 2nd | 3rd | 4th |
|---|---|---|---|---|
| Egoism | 48 | 28 | 2 | 0 |
| Altruism | 0 | 0 | 9 | 69 |
| Utilitarianism | 4 | 17 | 50 | 7 |
| Unmotivated | 26 | 33 | 17 | 2 |

However, although Egoism is clearly the highest rated overall, participants confused this behaviour with the Unmotivated model. The comments indicate that the agent's actions did not conform to their definition of ethical conduct. For example; one user states, "the bot [agent] just looks out for itself, which isn't ethical". These responses are worthy of note as the previous experiments did not include an Unmotivated behaviour to compare against, meaning this effect is not observed. The Friedman test confirms the statistical significance of the user rankings, with a p-value of 3.0867e-35. A one-way ANOVA similarly proves no gender effect on the result.

**Altruism** Considering that Altruism was so strongly identified as least likely in the Egoism model, a similar trend was expected for the Altruistic model. However, while altruism was the most commonly identified as first most likely, this was not done with quite the same confidence. However, Egoism was ranked least likely, which supports the theory that the participants recognised these two behaviours as conceptually opposed.

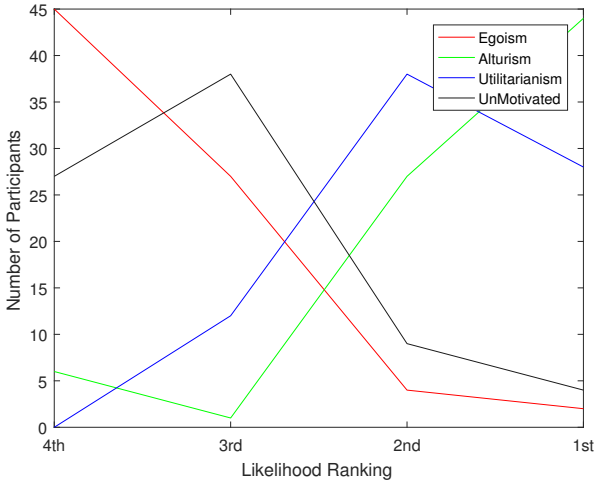



**Figure 9.9:** E2: Participant rankings plotted for the identification of the Altruistic model.

**Table 9.7:** E2: Summarised data for the identification of the Altruistic model.

| | 1st | 2nd | 3rd | 4th |
|---|---|---|---|---|
| Egoism | 2 | 4 | 27 | 45 |
| Altruism | 44 | 27 | 1 | 6 |
| Utilitarianism | 28 | 38 | 12 | 0 |
| Unmotivated | 4 | 9 | 38 | 27 |

By contrast, Utilitarianism is ranked second overall. This is understandable, as the Utilitarianism model will exhibit some of the altruistic behaviours. As noted with the previous set of experiments (see subsection 9.3.2), this is probably due to interpretation. If a participant thought that an altruistic action served the community, they may identify it as Utilitarian. Testing the rankings using a Friedman test shows statistical significance with a p-value of 1.1368e-26. There was no gender bias in the responses.

**Utilitarianism** The Utilitarian model was correctly identified by the majority of participants. Some of the trends are very similar to those seen in the Affective variant in the previous section (see Table 9.3.2). Namely, between the third and first Rankings, the Utilitarian votes follow an almost linear trend. Secondly, the participants voting for Affective are distributed throughout the ranks, highlighted by a relatively flat trend in Figure 9.10.

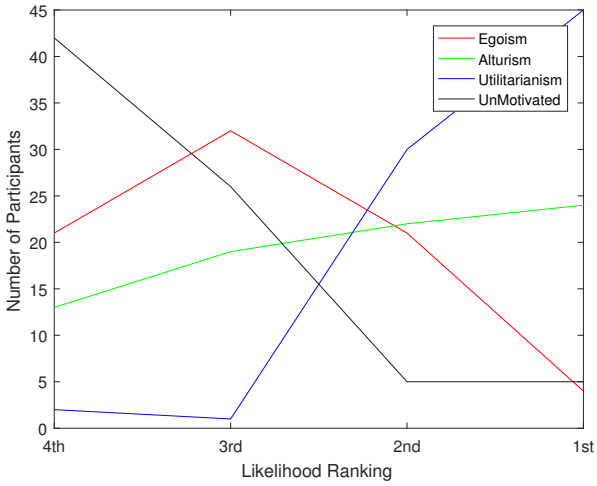



**Figure 9.10:** E2: Participant rankings plotted for the identification of the Utilitarian model.

Egoism votes show a noticeably different trend, with a midpoint peak, although this is partly due to four possible ranks as opposed to three. However, the ratings for the Unmotivated

**Table 9.8:** E2: Summarised data for the identification of the Utilitarian model.

| | 1st | 2nd | 3rd | 4th |
|---|---|---|---|---|
| Egoism | 4 | 21 | 32 | 21 |
| Altruism | 24 | 22 | 19 | 13 |
| Utilitarianism | 45 | 30 | 1 | 2 |
| Unmotivated | 5 | 5 | 26 | 42 |

model show a better fit to the votes for Egoism in the previous experiment. This provides indication of the pattern of participant ranking, showing some consistency in the choice for most likely, and ranking the least likely to be the furthest conceptually from that point (regardless of the actual model). The Friedman test shows significance with a p-value of 1.0961e-19, and there was no apparent gender bias when tested using a one-way ANOVA.

**Unmotivated** The majority of participants were able to correctly identify the Unmotivated behaviour. There was also high confidence in the fourth ranked behaviour (Altruism) and the third ranked behaviour (Utilitarianism). Egoism was considered the second most likely behaviour, and had a significant number of results in both the second and first ranks. This is similar to the results noted in the Egoism portion. Testing the participants' rankings with the Friedman test demonstrates statistical significance with a p-value of 3.4681e-25. An ANOVA test demonstrates no apparent gender effect.

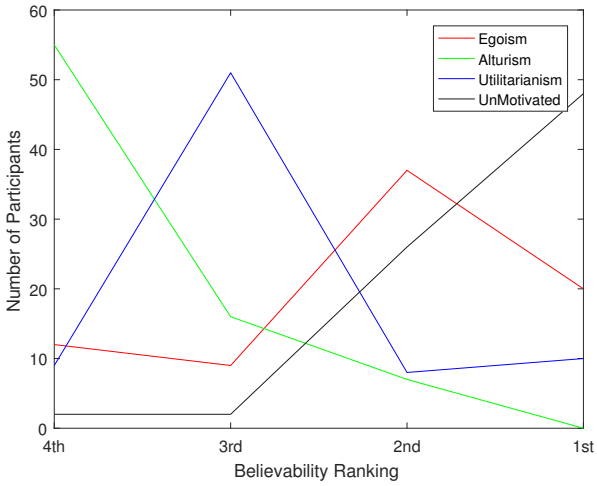



**Figure 9.11:** E2: Participant rankings plotted for the identification of the Unmotivated model.

What is possibly more interesting, is how the Unmotivated model appears to have been categorised by the participants. To look back through the previous three sections, a trend

**Table 9.9:** E2: Summarised data for the identification of the Unmotivated model.

| | 1st | 2nd | 3rd | 4th |
|---|---|---|---|---|
| Egoism | 20 | 37 | 9 | 12 |
| Altruism | 0 | 7 | 16 | 55 |
| Utilitarianism | 10 | 8 | 51 | 9 |
| Unmotivated | 48 | 26 | 2 | 2 |

emerges. If ranks 1 and 2 are treated as a single category (likely), and ranks 3 and 4 are treated as another category (not likely), then a clear grouping can be observed. Notably, Altruism and Utilitarianism are always in the same category, and Egoism and Unmotivated are always in the alternate category. This seems to indicates that the users generally consider the Egoist behaviour to be conceptually closer to the Unmotivated behaviour. As stated earlier during the Egoism identification section, some participants simply did not consider the Egoist behaviour to be Ethical and which could explain this apparent grouping.

### Believability

As with the previous experiment (see Table 9.3.2), Egoism is the highest rated overall in the Likert portion of the test. Furthermore, while there is some difference in the distribution, Utilitarianism is second highest rated, and Altruism is the thirdhighest. As with the previous experiment, there is a bias towards the mid-point of the rating with relatively few ratings at the extremes.

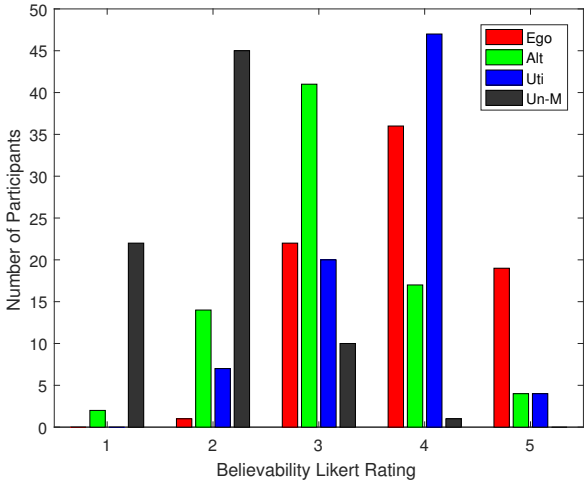



**Figure 9.12:** E2: Likert data for the Believability stage.

**Table 9.10:** E2: Summarised Likert data for the Believability stage.

| | 1 | 2 | 3 | 4 | 5 | Total |
|---|---|---|---|---|---|---|
| Ego | 0 | 1 | 22 | 36 | 36 | 307 |
| Alt | 2 | 14 | 41 | 17 | 17 | 241 |
| Uti | 0 | 7 | 20 | 47 | 47 | 282 |
| UnM | 22 | 45 | 10 | 1 | 0 | 146 |

Interestingly, despite that Egoism and Unmotivated are seemingly grouped together during the identification stage, they represent the highest and lowest rated according to their believability. With the Unmotivated behaviour used as a baseline or control, it is evident that all the other models score higher overall. The raw Likert totals (see Table 9.10) further demonstrates this.

It is surprising that the Unmotivated behaviour was rated so low, considering it is essentially a Boid-type model, which is typically considered visually believable. However, one hypothesis that could explain this is a ranking behaviour on the part of the participants. By looking at the peaks in the Likert votes, it is evident that Unmotivated, Altruism, and Utilitarianism are concentrated on the midpoints of the Likert scale. This could indicate that participants may have ranked based on a comparison of how they ranked the video they last saw, simply voting higher or lower. There is some evidence of this in the raw results, indicating that it was relatively rare for participants to assign multiple ratings of the same value. The exception to this is Egoism, where the votes are more distributed over the top three values.

During this portion of the test, 14 free-text responses were left; four of these were provided by the same participant, the rest were unique. Most of these comments provide little insight into the participants' thought process; however, two provided some qualitative evidence towards the comparison ranking hypothesis. Notably, one participant stated that "I'm not sure if this is really worth 5, but it was better than the last one"; another stated "this is similar to the last video, but a little better so I gave it a 3".

Utilising the Jarque-Bera test, we prove at the 5% level that the combined selections of all participants represent a normal distribution ($p = 0.0323$). Each of the three normative models were tested for gender effect using a one-way ANOVA. Egoism had a p-value of

0.2231 indicating that there was no significant gender effect. The same was true of the Altruism model with a p-value of 0.5633. The Utilitarianism model had a p-value of 0.0136; however, further investigation with a Spearman's Rank test fails to demonstrate a significant correlation (0.1779 and a RHO of -0.2443). We can therefore conclude that there is no significant gender effect in any of the models.

Each model was then tested for statistical significance against the Unmotivated model using Spearman's Rank. Testing the Egoism model against the Unmotivated model resulted in a p-value of 0.4139 and a RHO of -0.0938 meaning there is no significant correlation between the rating of the models. Furthermore, 94.87% of participants rated Egoism more believable than the Unmotivated behaviour. The Altruistic resulted in a p-value of 0.9849 and a RHO of 0.0022 meaning there is no significant correlation between the rating of the models. Furthermore, 78.21% of participants rated Altruism more believable than the Unmotivated behaviour. Finally, the Utilitarian model produced p-value of 0.7394 and a RHO of -0.0383 meaning there is no significant correlation between the rating of the models. Furthermore, 89.74% of participants rated Utilitarianism more believable than the Unmotivated behaviour.

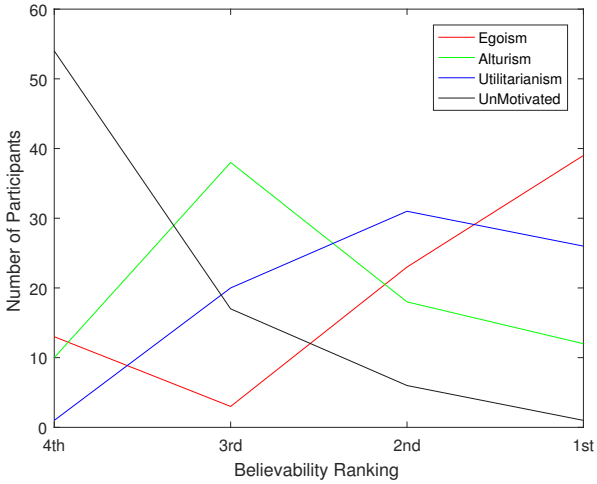



**Figure 9.13:** E2: Rank data for the Believability stage.

However, if this ranking behaviour is the case, it isn't apparent in the previous experiment (see Table 9.3.2 and Figure 9.6). However, one explanation could be that in this experiment, there were only 4 models to rate, which fits within the 1 to 5 distribution. As the previous experiment had 6 videos, this would have forced the participant to assign the same value to

**Table 9.11:** E2: Participants believability rankings.

| | 1st | 2nd | 3rd | 4th |
|---|---|---|---|---|
| Ego | 39 | 23 | 3 | 13 |
| Alt | 12 | 18 | 38 | 10 |
| Uti | 26 | 31 | 20 | 1 |
| UnM | 1 | 6 | 17 | 54 |

more than one video. This, plus the expected mid-point bias, could have naturally smoothed out ranking behaviour.

Unexpectedly, in the actual ranking portion of the believability test, the peaks are less defined. This is with the exception of the Unmotivated behaviour which is confidently the least highly ranked for its believability. One possible explanation is that the Likert section is linear, with one video rated after the other. The ranking section provides the user with the opportunity to rank from a more holistic perspective having considered the entire sequence. However, importantly the overall ranks do match the order in Likert section, aiding to validate the results. The users' rankings are statistically significant when tested with the Friedman test with a p-value of 4.2788e-19. A one-way ANOVA demonstrated no apparent gender bias in these results.

## 9.5 Overview of Results for E2

The results correspond well with the findings of the first experiment. The participants performed well in the identification; however, as the results show, there was some confusion regarding the Egoist and Unmotivated behaviour. It appears that due to either participant misunderstanding, or their personal beliefs, they have categorised self-interested behaviour as not being ethically motivated.

There is further evidence of this being a semantic/belief issue, rather than a problem with the model from the believability section of the survey. Despite some confusion over the identification, the Egoism behaviour achieved the highest believability rating, higher than

the other ethically motivated models. If the Egoism model was flawed, you would expect it to have a similar believability rating to the Unmotivated model.

## 9.6 Discussion

This chapter has evaluated the Trilogy architecture, specifically the implementation described in chapter 8. The first experiment (see section 9.3) evaluated the hypothesis that emotion may be key to successfully simulating ethically motivated behaviour. The results for this experiment concluded that, while it may not be essential for a normative behaviour to be recognised, it can greatly enhance the believability of those behaviours.

The second experiment, (see section 9.3) demonstrated that the ethical behaviour was identifiable and believable against Unmotivated behaviours. However, there was some confusion between the Egoist and Unmotivated behaviour. There were two explanations for this. Firstly, the Egoist (self-motivated) behaviour bore visual similarity to the basic Boid (Unmotivated) behaviour. Secondly, despite the participants being provided with a description of what the normative position of Egoist behaviour is, a number refused to accept it as 'ethical'. This was reflected in the comments they provided. Despite there being some confusion in the identification stage between the Egoist and the Unmotivated behaviours, the Egoist behaviours proved one of the most believable, while the Unmotivated were rated the least believable. This further highlights the discrepancy as largely semantic, and philosophical in nature.

# Chapter 10

# Conclusion

> *There's no destination. The journey is all that there is, and it can be very, very joyful.*
>
> **— Srikumar Rao**

## 10.1 Introduction

In this final chapter, the thesis will be concluded by revisiting the original objectives, and the challenges identified in the literature review. Following this the main findings will be summarised, and the limitations identified. The chapter concludes with some final thoughts on artificial ethical and moral agents, and directions for further research.

## 10.2 Revisiting the Thesis Objectives

This thesis was introduced with the following objectives:

1. Conduct an extensive literature review on the current state of the art in ethical simulation, focusing on artificial ethical agency.
2. Derive models for the simulation of ethical decision making within simulated characters.
3. Evaluate ethical agency models against the insights gained from the literature review.

4. Validate the models when applied to simulation.

The first objective was successfully completed in chapter 2. This review highlighted a number of challenges specific to the field of artificial ethics, which were subsequently addressed in other chapters. The literature review was submitted, and accepted as a conference paper [95].

The second and third objectives were first explored in chapter 3, where models of ethical agency were designed based on Braitenberg Vehicles. The insights from this chapter lead to further research into the simulation of affective states [96] (chapter 5), and value systems (chapter 4). The research into value systems was published as a conference paper and won the best paper award [96]. Finally, an architecture was proposed in chapter 7 (named 'the Trilogy architecture') to allow these systems to be used together, and configured in such a way to allow for advanced character simulation.

The fourth objective was to verify and validate the models when used in a simulation. However, how ethical models should be validated is an open question, as identified in the literature review (see subsection 2.10.3). An evaluation methodology was proposed in chapter 6 to address this issue, the assessment criteria being whether a participant could identify ethical behaviour as it was being simulated, and whether the behaviour appeared to be believable. A specific implementation of the Trilogy architecture (see chapter 8) was evaluated using this method in chapter 9. The results indicated that the Braitenberg Vehicle inspired reactive approach was suitable for simulating ethically motivated behaviour. Furthermore, by including affective simulation, the behaviour produced appeared more believable.

## 10.3 Main Findings and Contributions

In this section, the findings of this thesis, along with the principle contributions, will be presented. These have been categorised in such a way as to group individual various findings and contributions together for the sake of coherency.

### 10.3.1 Definitions of Ethical and Moral Agents

One of the first contributions of this thesis comes from the literature review. The field of artificial ethics lacks a stringent set of definitions, and the ones that exist are often used interchangeably leading to confusion. In chapter 2 this was addressed through the proposal of two specific definitions, namely *Artificial Ethical Agents* (AEA) and *Artificial Moral Agents* (AMA).

The proposal of these definitions is a worthwhile contributions for two reasons. Firstly, the definitions unify and provide clarity for the research field, which has has been referred to by many names. Secondly, the definitions in themselves provide a direction for future research by abstracting the philosophical arguments from the engineering goals. The question of whether a genuine moral agent can ever be built falls within the boundaries of AMA. In contrast, an artificial ethical agent is one that conforms to a set of ethical criteria, providing a direction for technical specification.

### 10.3.2 Reactive Models of Ethical Behaviour: Ethical Vessels

Consiquentialist theories often assume the agent's ability to evaluate a large number of candidate future states. This is also true of simulated ethical agents, such as the ones mentioned in the section 2.9. However, as discussed previously in this chapter, these approaches are subject to the frame problem. Furthermore, even in the limited scope worlds covered in most creative simulations, these approaches can be unsuitable due to their computational overheads. It can also be argued human agents often do not follow this level of extensive forward planning. In many situations, individuals instead react to the current state of the world, based on an assumption of outcomes, experience, or instinct. Some researchers have gone as far as to argue that ethical reasoning is usually generated to justify a decision after a judgement has been made [89]. This research has highlighted that in some applications, the reactive paradigm is applicable for the simulation of ethics. With the inclusion of a value system, the technique can be used to simulate more

complex behaviour, specifically situations that challenge an individual's ethical code. This specifically addresses grand challenge 4 (see subsection 2.10.4), identified in the literature review.

Furthermore, the reactive approach has allowed the frame problem (see grand challenge 2 in subsection 2.10.2) to be side-stepped entirely by computing just the next action based on the current context. This avoids the need to fill the agent with volumes of symbolic information to represent the world around them and all possible future consequences.

The behaviour produced is visually recognisable to a human observer, and as the approach is computationally lightweight, it facilitates the simulation of large numbers of agents. This approach is also compatible with other techniques, such as steering behaviours, making it particularly suitable for the creative sector.

Furthermore, at the time of writing, this thesis represents the first example of the reactive paradigm being applied to the simulation of ethics. Rodney Brooks' once claimed that the reactive paradigm could be a model for all natural forms of intelligence [38]. In a small way, the successful results demonstrated in this thesis adds further evidence towards this belief. The work also develops the Vehicle paradigm in a manner in-keeping with Braitenberg's original thought experiments, which were originally designed to explain natural behaviour.

### 10.3.3 Affective States Modelling

The ASM technique was developed out of a need to simulate emotive responses in simulated agents. While there are a number of techniques which do this, they come with a considerable computational overhead, making them unsuitable for some applications, notably, for the type of simulation used in creative industries. Furthermore, several of the existing techniques can be complicated to adapt or expand to new character instances, making them difficult to use in applications where you may require a range of personalities (such as an immersive game environment).

By comparison, the ASM technique is designed with creative applications in mind. It is designed to be computationally lightweight, simple to configure, and easy to adapt to a range of characters. The ASM technique takes inspiration from dimensional models of cognition, specifically the Conceptual Spaces approach of Gärdenfors [74].

As a dimensional model, the ASM technique has a low computational overhead. It requires minimal memory, and only employs vector-based calculations which are optimised in most game engines for GPU hardware. Furthermore, as each affective state is modelled as a coordinate within the state-space of the agent, it is simple to design, adjust and configure.

### 10.3.4 Trilogy architecture

While the early experiments conducted in chapter 3 were successful in simulating basic ethical behaviour, they lacked finesse. This was improved significantly through the development of a threshold based value system. However, simply switching a normative behaviour on and off still appeared quite robotic. A hypothesis was proposed that including an affective layer in the characters may improve the overall believability of the behaviour. This led to the development of the ASM technique, discussed in the previous section.

However, a method was required to allow the various reactive techniques to be combined in a sensible way, without adding any significant overheads or boilerplate code. This led to the development of the Trilogy architecture (see chapter 7), an approach to allow various character behaviour methods to be combined in a logical fashion. The method is inspired by a number of contemporary theories from psychology and neuroscience, splitting agent functions into three layers (sense, think, and act) and three domains (affective, conative, and cognitive).

The evaluation demonstrated that combining these reactive techniques produces behaviour that is recognisable as ethically motivated to an observer. Furthermore, the compound behaviour produced is verifiably more believable to the observer. As the resultant behaviour is built up of different weighted sub-behaviour modules, various non-intended qualities

emerge. Participants were able to assign labels to these emergent behaviours, including, confidence, aggression, apprehensiveness, bravery and indecision.

The Trilogy architecture is extensible beyond the work described in this thesis. Its design and logical structure make it suitable for combining various disparate behaviour simulation techniques, allowing each to be individually weighted for configuration and character design.

## 10.4 Limitations

There are three main limitations of this work. The principal limitation of this work is its focus on creative simulations. The models created have been designed for the sole purpose of creating a believable simulation. As such, the models proposed should not be assumed to be relevant for all applications of ethical simulation.

The second main limitation is in the evaluation method used. As mentioned frequently throughout this thesis, ethics are inherently difficult to evaluate or validate. This is due to to their subjective nature, an action that is ethically right or justifiable by one individual may be reprehensible by another. The proposed evaluation is designed to overcome this limitation as much as possible, by using identification as the core metric. However, while this overcomes some of the issues, subjective assessments bring their own limitations.

The final limitation relates to the population sample. As the final assessment capitalised on opportunistic sampling, a criticism could be made that this does not constitute a representative sample, and thus limits the overall scope of the work. However, there are a number of reasons why this decision was made. Firstly, the simulations were creative industry-focused. While these applications often have a target audience, a precise population is impossible to ascertain before release. Sales may not meet projections, or the realised audience may not match the target (for example, films with so-called "cult" followings). Secondly, opportunistic sampling made the research possible. While a precise population for a specific creative application may be possible, the reality of acquiring a precise sample

would pose significant challenges. For the purpose of achieving the thesis objectives, the approach taken was sufficient, and appropriate.

## 10.5 Ethical Agents?

In chapter 2, an Artificial Ethical Agent was defined as *"An autonomous entity that is capable of acting according to a set of morally defined considerations"*. The question remains as to whether the agents produced in this thesis are artificial ethical agents according to this definition.

Evidence supporting the agents being classified as AEAs is initially evident in chapter 3. In this chapter, each vessel created is evaluated against whether the resulting behaviour conforms to the basic definition of that normative position. Although simple in their design, these agents are designed in such a way as to meet the basic criterion for each position. While it would be wrong to say that these agents behave as a human would, it is important to consider the most basic model that could be described as meeting that definition. This allows us to use these constructs to build from the bottom-up, following the Nouvelle AI philosophy by constructing simple behaviours and allowing more complex actions to emerge organically.

Following this design choice, value systems were the next to be explored. By using threshold based systems built onto the Braitenberg-inspired vessels, we demonstrated that the agents were able to successfully reconcile conflicts between normative positions by implementing a value system. Again, based on a clear definition of what a value system was, these agents clearly conform.

The final implementation provides the clearest evidence that the agents developed can be described as artificial ethical agents. When an impartial observer was asked to identify the behaviours they saw, they were able to answer correctly above a chance accuracy. Furthermore, the text based responses demonstrate that the participants credited the agents with behaviours or qualities that were not intended by design. It could be argued that this

is simply an anthropomorphic viewpoint, and that the participant is simply seeing more than is actually being demonstrated. However, ethically motivated behaviour is inherently subjective, and it would be short-sighted to discount this evidence. Furthermore, within the bounds of creative applications, designers rely on a certain suspension of disbelief, relying on the audience buying into an illusion. Therefore, far from subjectivity being a negative factor, it is arguably one of the most important factors to consider.

Within the bounds of the thesis definition, these agents can certainly be described as AEAs.

## 10.6 Moral Agents?

The intention of this thesis was not to tackle the problem of moral agents. Before an artificial entity could be credited with being an artificial moral agent, there is a score of philosophical, legal, epistemological, and technical challenges which must be addressed. There is also an argument that should this ever be achieved, the 'artificial' moniker may no longer be appropriate. However, as a concluding discussion, it is interesting to hypothesise at what point an ethical agent could become a moral agent.

In keeping with the theme of this thesis, an argument can be made that the route to moral agency could lie in emergence, and externalism. As previously stated, externalism posits that thought and consciousness emerge not only from neural activity, but through the interaction of brain, body and environment. If this holds true, then it is entirely possible for a machine to extend beyond the initial scope intended by its designer, not just through a change in its brain, but also due to the environment changing around it. Furthermore, various studies, from Brooks, to this thesis, have demonstrated that the interaction of low level subsystems can result in unexpected behavioural output. If Brooks is right, and that the reactive approach is a model for all natural forms of intelligence, could morality not emerge organically in this way?

While it is a significant leap from the AEAs described in this thesis to discussions of morality, it is interesting to consider how the future of the field may evolve. Furthermore, it

forms an argument for continued research into the reactive approach to artificial ethics in order to see what behaviours may develop by designing for emergence.

## 10.7 Future Work

This thesis has concluded by establishing that the research has attained the thesis objectives. However, while these questions have been answered, new ones have been asked. The systems discussed in this thesis are certainly suitable fore the application they were designed for, however, they are not (in their current format) suitable for applications outside the creative sector. It has become apparent that the conclusion of this thesis does not necessarily represent the end of this journey, however, it will hopefully serve as a signpost for future research. To that end, the following areas have been identified as warranting further research.

### 10.7.1 Further Vessels

The Vessels in chapter 3 were an attempt to further the work of Braitenberg's seminal work, 'Vehicles'. The aim was to explore an area of human behaviour which had not been addressed via the sensory motor approach before. However, in this research only four vehicles were designed, with a further Vessel hypothesised through thought experiment.

These vessels focused on various consequentialist theories, notably Egoism, Hedonism, Altruism, and Utilitarianism. The reasoning for this focus came through the natural evolution of thought, rather than as a product of design. The type 3 Vehicle inspired the egoism Vessel, and this was adapted and developed to create the other models. However, there are various other ethical positions that have not been be explored such as deontology, and virtue ethics.

Could (for example) a vessel be designed to follow a form of the Categorical Imperative? While this intrinsically seems beyond the scope of a sensory motor agent, it is worth noting that Braitenberg himself hypothesised a Vehicle which could learn, develop and follow rules. Further exploration of this theme could demonstrate the full scope of this approach within the field of Artificial Ethics.

### 10.7.2 Two-Level Ethics

Two-level utilitarianism is a normative theory that states that an individual's ethical decisions are based on a set of intuitive rules, except in dilemmas, where the agent should engage in critical thinking. It is interesting, as in many ways it is similar to Arkina's three methods of developing autonomous ethical agents (described in section 2.9). Notably, the first style of ethical reasoning, following intuitive rules, could be likened to Arkina's *Governer*. The second half, critical thinking, could be likened to a combination of the *Behaviour Control* and the *Adaptor*.

This idea of two-levelled ethics could be very interesting, and a useful direction to take forward the approach described in this thesis. While this research has argued for the reactive approach, there are many areas where other methods are currently better suited, specifically reasoning, analysis and forward planning. A hybrid approach of the lower and higher levels may ultimately lead to some very interesting results. In their seminal book, *Moral Machines: Teaching Robots Right from Wrong*, Wallach and Allen conclude that a hybrid approach, between top-down and bottom-up may be necessary [217].

For example, the lightweight Vessel approach could be used as a lower layer allowing quick computation and instant reaction to changes in the environment. Another approach (such as the work of Winfield), could provide a higher layer allowing for more extensive, and less time-critical reasoning during dilemmas. Furthermore, dedicated modules could be included to allow an agent to explain why it took a particular action at a given time. This could be beneficial in applications outside the creative sector, such as autonomous vehicles, which may need to able to explain their actions after traffic incidents.

### 10.7.3 The Trilogy architecture and Designing Characters For Emergence

The Trilogy architecture has currently only been used for the simulations in this thesis. However, other modules could be designed, and included to produce various different characters for other applications. The modularity of its design, in a similar way to blackboard AI systems, could allow for different aspects of agent behaviour to be abstracted and included. In this way, character developers could aim to design for emergence, instead of developing specific behaviours, shifting the focus to a lower level.

However, relatively speaking we still have a lot to learn about emergence, with debates regarding what it constitutes. Future work could explore the topic of emergence, using the Trilogy architecture as a tool for exploring the interplay between subsystems.

### 10.7.4 The Philosophy of Artificial Ethics

As described the literature review (subsection 2.10.5), one area identified as a grand challenge of the field was to contribute back to moral philosophy. A number of researcher's and philosophers alike have identified simulation as a tool to help advance the field. While this has not been explored in-depth within the scope of this thesis, it is an obvious area for future work. Simulation could help provide evidence to support positions, providing a tool that has previously been unavailable to theorists. It could also help develop new theories, by exploring emergence, or developing exploratory ethical codes.

There is also the opportunity that simulation, and specifically the evaluation of simulations, could help explore what it means to be ethical and moral. Even in a small way, some of the free text responses received during the main study of this thesis highlighted the moral thought process of the participant. With time, development, and, importantly, with cross-discipline collaboration, simulation could become a transformational tool in this field.

# Chapter 11
# References


[1] E. Ackerman, 'A better test than Turing [news]', *Spectrum, IEEE*, vol. 51, no. 10, pp. 20–21, 2014 (pp. 100, 103).

[2] D. Adams, *The more than complete hitchhiker's guide: Complete & unabridged*. Bonanza Books, 1989 (p. 143).

[3] R. Adolphs, 'Recognizing emotion from facial expressions: Psychological and neurological mechanisms', *Behavioral and cognitive neuroscience reviews*, vol. 1, no. 1, pp. 21–62, 2002 (p. 84).

[4] N. Alenina, D. Kikic, M. Todiras, V. Mosienko, F. Qadri, R. Plehm, P. Boyé, L. Vilianovitch, R. Sohr, K. Tenner *et al.*, 'Growth retardation and altered autonomic control in mice lacking brain serotonin', *Proceedings of the National Academy of Sciences*, vol. 106, no. 25, pp. 10 332–10 337, 2009 (p. 133).

[5] C. Allen, 'Calculated morality: Ethical computing in the limit', *Cognitive, emotive and ethical aspects of decision making and human action*, vol. 1, pp. 19–23, 2002 (p. 10).

[6] C. Allen, I. Smit and W. Wallach, 'Artificial morality: Top-down, bottom-up, and hybrid approaches', *Ethics & Information Technology*, vol. 7, no. 3, pp. 149–155, 2005 (pp. 24, 27, 30).

[7] C. Allen, G. Varner and J. Zinser, 'Prolegomena to any future artificial moral agent', *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 12, no. 3, pp. 251–261, 2000 (pp. 29, 99).

[8] M. Anderson and S. L. Anderson, 'Machine ethics: Creating an ethical intelligent agent', *AI Magazine*, vol. 28, no. 4, pp. 15–26, 2007 (pp. 1, 11, 20, 26, 30, 36).

[9] ——, 'The status of machine ethics: A report from the aaai symposium', *Minds and Machines*, vol. 17, no. 1, pp. 1–10, 2007 (pp. 11, 15, 20–22, 32, 34).

[10] M. Anderson, S. L. Anderson and C. Armen, 'An approach to computing ethics', *Intelligent Systems, IEEE*, vol. 21, no. 4, pp. 56–63, 2006 (pp. 11, 32).

[11] ——, 'Medethex: A prototype medical ethics advisor', in *National Conference On Artificial Intelligence*, vol. 21, 2006, pp. 1759–1765 (p. 32).

[12] S. L. Anderson, *The unacceptability of Asimov's three laws of robotics as a basis for machine ethics machine ethics*, 2011 (p. 14).

[13] S. L. Anderson and M. Anderson, 'How machines can advance ethics', *Philosophy Now*, vol. 72, pp. 17–19, 2009 (p. 39).

[14] R. C. Arkina, 'Governing lethal behavior: Embedding ethics in a hybrid deliberative/reactive robot architecture', in *Artificial General Intelligence Conference*, IOS Press, vol. 171, 2008, pp. 51–63 (p. 30).

[15] ——, *Governing lethal behavior in autonomous robots*. CRC Press, 2009 (p. 30).

[16] K. Arkoudas, S. Bringsjord and P. Bello, 'Toward ethical robots via mechanized deontic logic', in *AAAI Fall Symposium on Machine Ethics*, 2005, pp. 17–23 (p. 26).

[17] K. D. Ashley and B. M. McLaren, 'A cbr knowledge representation for practical ethics', in *Advances in case-based reasoning*, Springer, 1994, pp. 180–197 (p. 33).

[18] ——, 'Reasoning with reasons in case-based comparisons', in *Case-Based Reasoning Research and Development*, Springer, 1995, pp. 133–144 (p. 33).

[19] I. Asimov, 'Runaround', *Astounding Science Fiction*, vol. 29, pp. 94–103, 1942 (p. 13).

[20] K. S. Atman, 'The role of conation (striving) in the distance education enterprise', *American Journal of Distance Education*, vol. 1, no. 1, pp. 14–24, 1987 (p. 118).

[21] R. M. Bagby, M. Sellbom, P. T. Costa and T. A. Widiger, 'Predicting diagnostic and statistical manual of mental disorders-iv personality disorders with the five-factor model of personality and the personality psychopathology five', *Personality and Mental Health*, vol. 2, no. 2, pp. 55–69, 2008 (p. 84).

[22] A. Bain, *The emotions and the will*. Parker and son, 1859 (p. 116).

[23] ——, *The senses and the intellect*. Longman, Green, Longman, Roberts, and Green, 1864 (p. 116).

[24] G. Ball and J. Breese, 'Emotion and personality in a conversational agent', *Embodied conversational agents*, pp. 189–219, 2000 (p. 85).

[25] Y. Bar-Cohen and D. Hanson, *The coming robot revolution: Expectations and fears about emerging intelligent, humanlike machines*. Springer, 2009 (p. 31).

[26] A. F. Beavers, 'Between angels and animals: The question of robot ethics, or is Kantian moral agency desirable', in *Annual meeting of the association of practical and professional ethics, Cincinnati, OH*, 2009, pp. 5–8 (pp. 26, 28).

[27] ——, 'Moral machines and the threat of ethical nihilism', *Robot ethics: The ethical and social implications of robotics*, pp. 333–343, 2011 (pp. 26, 35, 39).

[28] B. G. Blair, 'Russia's doomsday machine', *New York Times*, 1993 (p. 17).

[29] O. Blomberg, 'Conceptions of cognition for cognitive engineering', *The international journal of aviation psychology*, vol. 21, no. 1, pp. 85–104, 2011 (p. 117).

[30] B. Blumberg, M. Downie, Y. Ivanov, M. Berlin, M. P. Johnson and B. Tomlinson, 'Integrated learning for interactive synthetic characters', in *Transactions on Graphics (TOG)*, ACM, vol. 21, 2002, pp. 417–426 (pp. 82, 117).

[31] N. Bostrom, 'Existential risks', *Journal of Evolution and Technology*, vol. 9, no. 1, pp. 1–31, 2002 (p. 16).

[32] ——, 'Ethical issues in advanced artificial intelligence', *Science Fiction and Philosophy: From Time Travel to Superintelligence*, pp. 277–286, 2003 (pp. 11, 17).

[33] N. Bostrom and E. Yudkowsky, 'The ethics of artificial intelligence', *The Cambridge Handbook of Artificial Intelligence*, pp. 316–334, 2014 (pp. 28, 30, 37).

[34] V. Braitenberg, *Vehicles*, 1984 (pp. iii, 30, 42, 43, 46, 67).

[35] S. Bringsjord, K. Arkoudas and P. Bello, 'Toward a general logicist methodology for engineering ethically correct robots', *Intelligent Systems, IEEE*, vol. 21, no. 4, pp. 38–44, 2006 (p. 26).

[36] R. A. Brooks, 'A robust layered control system for a mobile robot', *Robotics and Automation, IEEE*, vol. 2, no. 1, pp. 14–23, 1986 (pp. 30, 70).

[37] ——, *Cambrian intelligence: The early history of the new AI*. MIT press, 1999 (p. 71).

[38] R. A. Brooks, C. Breazeal, M. Marjanović, B. Scassellati and M. M. Williamson, 'The cog project: Building a humanoid robot', in *Computation for metaphors, analogy, and agents*, Springer, 1999, pp. 52–87 (pp. iii, 46, 71, 177).

[39] R. A. Brooks and J. H. Connell, 'Asynchronous distributed control system for a mobile robot', in *Cambridge Symposium of Intelligent Robotics Systems*, International Society for Optics and Photonics, 1987, pp. 77–84 (p. 70).

[40] K. Čapek, *RUR (Rossum's Universal Robots)*. Penguin, 2004 (p. 13).

[41] C. Carere, S. Montanino, F. Moreschini, F. Zoratto, F. Chiarotti, D. Santucci and E. Alleva, 'Aerial flocking patterns of wintering starlings under different predation risk', *Animal behaviour*, vol. 77, no. 1, pp. 101–107, 2009 (p. 91).

[42] L. Chapman, C. Gray and C. Headleand, 'A sense-think-act architecture for low-cost mobile robotics', in *Research and Development in Intelligent Systems XXXII*, Springer, 2015, pp. 405–410 (p. 116).

[43] R. M. Church, 'Emotional reactions of rats to the pain of others', *Journal of comparative and physiological psychology*, vol. 52, no. 2, p. 132, 1959 (p. 18).

[44] R. Clarke, 'Asimov's laws of robotics: Implications for information technology-part i', *Computer, IEEE*, no. 12, pp. 53–61, 1993 (p. 14).

[45] ——, 'Asimov's laws of robotics: Implications for information technology-part ii', *Computer, IEEE*, vol. 27, no. 1, pp. 57–66, 1994 (p. 14).

[46] G. Clement, 'Animals and moral agency: The recent debate and its implications', *Journal of Animal Ethics*, vol. 3, no. 1, pp. 1–14, 2013 (p. 18).

[47] M. Coeckelbergh, 'Virtual moral agency, virtual moral responsibility: On the moral significance of the appearance, perception, and performance of artificial agents', *AI & Society*, vol. 24, no. 2, pp. 181–189, 2009 (pp. 1, 29).

[48] ——, 'Moral appearances: Emotions, robots, and human morality', *Ethics & Information Technology*, vol. 12, no. 3, pp. 235–241, 2010 (pp. 19, 28).

[49] ——, 'Drones, information technology, and distance: Mapping the moral epistemology of remote fighting', *Ethics & Information Technology*, vol. 15, no. 2, pp. 87–98, 2013 (p. 12).

[50] P. R. Cohen, 'If not Turing's test, then what?', *AI magazine*, vol. 26, no. 4, pp. 61–67, 2005 (p. 101).

[51] B. J. Copeland, *The essential Turing*. Oxford University Press, 2004 (pp. 106, 113).

[52] M. H. Couppis and C. H. Kennedy, 'The rewarding effect of aggression is reduced by nucleus accumbens dopamine receptor antagonism in mice', *Psychopharmacology*, vol. 197, no. 3, pp. 449–456, 2008 (p. 133).

[53] J. A. Coyne, 'You don't have free will', *The Chronicle of Higher Education*, vol. 18, 2012 (p. 21).

[54] R. Crisp, 'Hedonism reconsidered', *Philosophy and Phenomenological Research*, vol. 73, no. 3, pp. 619–645, 2006 (p. 25).

[55] P. Danielson, *Artificial morality: Virtuous robots for virtual games*. Routledge, 2002 (p. 10).

[56] ——, 'Can robots have a conscience?', *Nature*, vol. 457, no. 7229, pp. 540–540, 2009 (p. 22).

[57] ——, 'Designing a machine to learn about the ethics of robotics: The n-reasons platform', *Ethics & Information Technology*, vol. 12, no. 3, pp. 251–261, 2010 (p. 11).

[58] C. Darwin, P. Ekman and P. Prodger, *The expression of the emotions in man and animals*. Oxford University Press, USA, 1998 (p. 133).

[59] H. De Garis, *The artilect war: Cosmists vs. terrans: A bitter controversy concerning whether humanity should build goldlike massively intelligent machines*. ECT Publications, 2005 (pp. 11, 17).

[60] E. De Sevin and D. Thalmann, 'An affective model of action selection for virtual humans', in *Motivational and Emotional Roots of Cognition and Action (AISB 05)*, 2005 (p. 82).

[61] D. C. Dennett, *Brainchildren: Essays on designing minds*. MIT Press, 1998 (p. 101).

[62] J. M. Digman, 'Personality structure: Emergence of the five-factor model', *Annual review of psychology*, vol. 41, no. 1, pp. 417–440, 1990 (p. 84).

[63] I. Douglas-Hamilton, S. Bhalla, G. Wittemyer and F. Vollrath, 'Behavioural reactions of elephants towards a dying and deceased matriarch', *Applied Animal Behaviour Science*, vol. 100, no. 1, pp. 87–102, 2006 (p. 18).

[64] H. L. Dreyfus, *What computers still can't do: A critique of artificial reason*. MIT press, 1992 (p. 102).

[65] A. Egges, S. Kshirsagar, X. Zhang and N. Magnenat-Thalmann, 'Emotional communication with virtual humans', in *MMM*, 2003, pp. 243–263 (p. 85).

[66] P. A. Facione, D. Scherer and T. Attig, *Values and society: An introduction to ethics and social philosophy*. Prentice Hall, 1978 (p. 24).

[67] E. Fehr and U. Fischbacher, 'The nature of human altruism', *Nature*, vol. 425, no. 6960, pp. 785–791, 2003 (p. 25).

[68] F. Feldman, *Pleasure and the good life: Concerning the nature, varieties and plausibility of hedonism*. Clarendon Press Oxford, 2004 (p. 25).

[69] J. C. Flack and F. B. de Waal, ‘‘any animal whatever’: Darwinian building blocks of morality in monkeys and apes’, *Journal of Consciousness Studies*, vol. 7, no. 1-2, pp. 1–29, 2000 (p. 18).

[70] L. Floridi and J. W. Sanders, ‘Artificial evil and the foundation of computer ethics’, *Ethics & Information Technology*, vol. 3, no. 1, pp. 55–66, 2001 (p. 28).

[71] R. French, *If it walks like a duck and quacks like a duck... the Turing test, intelligence and consciousness*, 2009 (pp. 99, 102).

[72] R. M. French, ‘The Turing test: The first 50 years’, *Trends in cognitive sciences*, vol. 4, no. 3, pp. 115–122, 2000 (pp. 99, 102).

[73] R. D. Fricker, ‘Sampling methods for web and e-mail surveys’, in *The SAGE handbook of online research methods*, N. G. Fielding, R. M. Lee and G. Blank, Eds., Sage, 2008, ch. 11, pp. 195–216 (p. 113).

[74] P. Gärdenfors, ‘Mental representation, conceptual spaces and metaphors’, *Synthese*, vol. 106, no. 1, pp. 21–47, 1996 (pp. iii, 84, 178).

[75] ——, *Conceptual spaces: The geometry of thought*. 2004 (p. 84).

[76] K. E. Gerdes and L. K. Stromwall, ‘Conation: A missing link in the strengths perspective’, *Social Work*, vol. 53, no. 3, pp. 233–242, 2008 (p. 118).

[77] J. Gips, ‘Towards the ethical robot’, *Android epistemology*, pp. 243–252, 1995 (p. 24).

[78] A. L. Goldsmith, *The golem remembered, 1909-1980: Variations of a jewish legend*. Wayne State University Press, 1981 (p. 13).

[79] N. J. Goodall, ‘Machine ethics and automated vehicles’, in *Road Vehicle Automation*, Springer, 2014, pp. 93–102 (p. 12).

[80] S. Goose, ‘The case for banning killer robots: Point’, *Communications of the ACM*, vol. 58, no. 12, pp. 43–45, 2015 (p. 12).

[81] D. F. Gordon-Spears, ‘Asimov’s laws: Current progress’, in *Formal Approaches to Agent-Based Systems*, Springer, 2002, pp. 257–259 (p. 14).

[82] C. Grau, ‘There is no “i” in “robot”: Robotic utilitarians and utilitarian robots’, in *Machine Ethics: Papers from the 2005 AAAI Fall Symposium*, 2005 (pp. 23, 24).

[83] J. D. Greene, R. B. Sommerville, L. E. Nystrom, J. M. Darley and J. D. Cohen, ‘An fmri investigation of emotional engagement in moral judgment’, *Science*, vol. 293, no. 5537, pp. 2105–2108, 2001 (p. 77).

[84] K. Grint and S. Woolgar, *The machine at work: Technology, work and organization*. Wiley, 2013 (p. 28).

[85] M. Guarini, ‘Particularism and generalism: How AI can help us to better understand moral cognition’, in *Machine ethics: Papers from the 2005 AAAI fall symposium*, 2005 (pp. 33, 39).

[86] E. Guizzo, ‘World robot population reaches 8.6 million’, *IEEE Spectrum*, vol. 14, 2010 (p. 10).

[87] K. Gunderson, ‘The imitation game’, *Mind*, vol. 73, no. 290, pp. 234–245, 1964 (p. 102).

[88] S. J. Guy, S. Kim, M. C. Lin and D. Manocha, ‘Simulating heterogeneous crowd behaviors using personality trait theory’, in *Eurographics Symposium on Computer Animation*, ACM, 2011, pp. 43–52 (p. 86).

[89] J. Haidt, ‘The emotional dog and its rational tail: A social intuitionist approach to moral judgment’, *Psychological review*, vol. 108, no. 4, pp. 814–834, 2001 (p. 176).

[90] F. A. Hanson, ‘Beyond the skin bag: On the moral responsibility of extended agencies’, *Ethics & Information Technology*, vol. 11, no. 1, pp. 91–99, 2009 (pp. 21, 32).

[91] S. Harnad, ‘Levels of functional equivalence in reverse bioengineering’, *Artificial Life*, vol. 1, no. 3, pp. 293–301, 1994 (p. 102).

[92] J. Haugeland, *Artificial intelligence: The very idea*. MIT press, 1989 (p. 102).

[93] P. Hayes and K. Ford, ‘Turing test considered harmful’, in *IJCAI*, 1995, pp. 972–977 (pp. 100, 102, 103).

[94] C. J. Headleand, L. Ap Cenydd and W. J. Teahan, 'Action selection through affective states modelling', in *Science and Information Conference SAI*, IEEE, 2016, pp. 478–487 (p. 80).

[95] ——, 'Sexbots as ethical agents: On the possibility of ethical machines', in *9th Philosophy and Computing AISB Symposium*, 2016 (pp. 6, 10, 175).

[96] ——, 'Towards ethical robots: Revisiting Braitenberg's Vehicles', in *Science and Information Conference SAI*, IEEE, 2016, pp. 469–477 (pp. 42, 175).

[97] ——, 'Berry Eaters: Learning colour concepts with template based evolution evaluation', in *ALIFE 14: The Fourteenth Conference on the Synthesis and Simulation of Living Systems*, vol. 14, pp. 473–480 (pp. 45, 59, 67, 80, 84).

[98] C. J. Headleand, G. Henshall, L. Ap Cenydd and W. J. Teahan, 'The influence of virtual reality on the perception of artificial intelligence characters in games', in *Research and Development in Intelligent Systems XXXII*, Springer, 2015, pp. 345–357 (p. 104).

[99] C. J. Headleand, J. Jackson, L. Priday, W. Teahan and L. Ap Cenydd, 'Does the perceived identity of non-player characters change how we interact with them?', in *Cyberworlds*, 2015, pp. 145–152 (p. 106).

[100] C. J. Headleand, J. Jackson, B. Williams, L. Priday, W. J. Teahan and L. Ap Cenydd, 'How the perceived identity of a npc companion influences player behavior', in *Transactions on Computational Science XXVIII*, Springer, 2016, pp. 88–107 (p. 106).

[101] C. J. Headleand, L. Priday, P. D. Ritsos, J. C. Roberts, L. Ap Cenydd and W. Teahan, 'Anthropomorphisation of software agents as a persuasive tool', in *British HCI*, 2015 (p. 11).

[102] T. Hellström, 'On the moral responsibility of military robots', *Ethics & Information Technology*, vol. 15, no. 2, pp. 99–107, 2013 (pp. 12, 28).

[103] P. C. Hew, 'Artificial moral agents are infeasible with foreseeable technologies', *Ethics & Information Technology*, vol. 16, no. 3, pp. 197–206, 2014 (p. 29).

[104] B. Hibbard, 'Avoiding unintended ai behaviors', in *Artificial General Intelligence*, Springer, 2012, pp. 107–116 (p. 23).

[105] E. R. Hilgard, 'The Trilogy of Mind: Cognition, affection, and conation', *Journal of the History of the Behavioral Sciences*, vol. 16, no. 2, pp. 107–117, 1980 (pp. 116, 117).

[106] K. E. Himma, 'Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent?', *Ethics & Information Technology*, vol. 11, no. 1, pp. 19–29, 2009 (pp. 11, 18, 28).

[107] P. Hingston, 'A Turing test for computer game bots', *Computational Intelligence and AI in Games, IEEE Transactions on*, vol. 1, no. 3, pp. 169–186, 2009 (pp. 103, 105).

[108] ——, 'The 2k botprize', in *Computational Intelligence and Games, IEEE*, IEEE, 2009, pp. 1–1 (p. 105).

[109] J. F. Hoorn, M. Pontier and G. F. Siddiqui, 'Coppélius' concoction: Similarity and complementarity among three affect-related agent models', *Cognitive Systems Research*, vol. 15, pp. 33–49, 2012 (p. 35).

[110] I. Horswill, 'Very fast action selection for parameterized behaviors', in *International Conference on Foundations of Digital Games*, ACM, 2009, pp. 91–98 (p. 81).

[111] D. G. Johnson, 'Computer systems: Moral entities but not moral agents', *Ethics & Information Technology*, vol. 8, no. 4, pp. 195–204, 2006 (pp. 11, 28).

[112] D. G. Johnson and K. W. Miller, 'Un-making artificial moral agents', *Ethics & Information Technology*, vol. 10, no. 2-3, pp. 123–133, 2008 (p. 28).

[113] W. L. Johnson, 'Needed: A new test of intelligence', *ACM SIGART Bulletin*, vol. 3, no. 4, pp. 7–9, 1992 (p. 102).

[114] S. Jones, M. Studley and A. Winfield, 'Mobile gpgpu acceleration of embodied robot simulation', in *Artificial Life and Intelligent Agents*, Springer, 2014, pp. 97–109 (p. 32).

[115] A. R. Jonsen and S. E. Toulmin, *The abuse of casuistry: A history of moral reasoning*. Univ of California Press, 1988 (p. 32).

[116] I. Kant, *Critique of pure reason*. Cambridge University Press, 1999 (p. 116).

[117] ——, *Critique of power of judgment*, 2002 (p. 116).

[118] ——, *Critique of practical reasoning*, 2015 (p. 116).

[119] M. Killen and F. B. de Waal, 'The evolution and development of morality', *Natural conflict resolution*, pp. 352–372, 2000 (p. 18).

[120] B. Kitchenham and S. L. Pfleeger, 'Principles of survey research: Part 5: Populations and samples', *ACM SIGSOFT Software Engineering Notes*, vol. 27, no. 5, pp. 17–20, 2002 (p. 114).

[121] A. Krishnan, *Killer robots: Legality and ethicality of autonomous weapons*. Ashgate Publishing, Ltd, 2009 (p. 12).

[122] S. Kshirsagar, 'A multilayer personality model', in *2nd international symposium on smart graphics*, ACM, 2002, pp. 107–115 (pp. 84, 85).

[123] C. R. Kube and H. Zhang, 'Collective robotics: From social insects to robots', *Adaptive behavior*, vol. 2, no. 2, pp. 189–218, 1993 (p. 45).

[124] J. E. Laird and J. C. Duchi, 'Creating human-like synthetic characters with multiple skill levels: A case study using the soar quakebot', vol. 1001, 2000, pp. 48 109–2110 (p. 104).

[125] P. Lichocki, A. Billard and P. H. Kahn, 'The ethical landscape of robotics', *Robotics & Automation Magazine, IEEE*, vol. 18, no. 1, pp. 39–50, 2011 (p. 12).

[126] A. Lilienthal and T. Duckett, 'Experimental analysis of smelling Braitenberg Vehicles', *Environment*, vol. 5, pp. 375–380, 2003 (p. 45).

[127] P. Lin, K. Abney and G. Bekey, 'Robot ethics: Mapping the issues for a mechanized world', *Artificial Intelligence*, vol. 175, no. 5, pp. 942–949, 2011 (pp. 10, 11).

[128] D. Livingstone, 'Turing's test and believable AI in games', *Computers in Entertainment (CIE)*, vol. 4, no. 1, pp. 6–18, 2006 (pp. 103, 104, 109).

[129] D. Livingstone and S. McGlinchey, 'What believability testing can tell us', in *Computer Games: AI, Design and Education*, 2004, pp. 273–277 (p. 104).

[130] H. Lövheim, 'A new three-dimensional model for emotions and monoamine neurotransmitters', *Medical hypotheses*, vol. 78, no. 2, pp. 341–348, 2012 (pp. 129, 132, 133).

[131] B. Mac Namee, 'Proactive persistant agents', PhD thesis, University of Dublin, Trinity, 2004 (p. 103).

[132] W. J. MacInnes, 'Believability in multi-agent computer games: Revisiting the Turing test', in *CHI'04 Extended Abstracts on Human Factors in Computing Systems*, ACM, 2004, pp. 1537–1537 (p. 104).

[133] P. D. MacLean, *The triune brain in evolution: Role in paleocerebral functions*. Springer Science & Business Media, 1990 (pp. 116, 118).

[134] L. McCauley, 'Ai armageddon and the three laws of robotics', *Ethics & Information Technology*, vol. 9, no. 2, pp. 153–164, 2007 (pp. 11, 13, 14, 28).

[135] J. McCosh, *Psychology: Emotions*. C. Scribner's sons, 1880 (p. 116).

[136] ——, *Psychology: The cognitive powers*. C. Scribner's sons, 1886 (p. 116).

[137] ——, *Psychology: Motive powers*. C. Scribner's sons, 1887 (p. 116).

[138] R. R. McCrae and O. P. John, 'An introduction to the five-factor model and its applications', *Journal of personality*, vol. 60, no. 2, pp. 175–215, 1992 (p. 84).

[139] D. McDuff, R. El Kaliouby, K. Kassam and R. Picard, 'Affect valence inference from facial action unit spectrograms', in *Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, 2010, pp. 17–24 (p. 82).

[140] B. McLaren, 'Lessons in machine ethics from the perspective of two computational models of ethical reasoning', in *AAAI Fall Symposium on Machine Ethics*, 2005 (p. 33).

[141] B. M. McLaren and K. D. Ashley, 'Case-based comparative evaluation in truth-teller', in *The Proceedings From the Seventeenth Annual Conference of the Cognitive Science Society*, 1995, pp. 72–77 (p. 33).

[142] ——, 'Context sensitive case comparisons in practical ethics: Reasoning about reasons', in *5th international conference on Artificial intelligence and law*, ACM, 1995, pp. 316–325 (p. 33).

[143] ——, 'Assessing relevance with extensionally defined principles and cases', in *AAAI/IAAI*, 2000, pp. 316–322 (p. 33).

[144] J. McMahan and B. J. Strawser, *Killing by remote control: The ethics of an unmanned military*. Oxford University Press, 2013 (p. 12).

[145] T. E. Moffitt and P. A. Silva, 'Iq and delinquency: A direct test of the differential detection hypothesis', *Journal of abnormal psychology*, vol. 97, no. 3, pp. 330–333, 1988 (p. 102).

[146] J. H. Moor, 'The nature, importance, and difficulty of machine ethics', *Intelligent Systems, IEEE*, vol. 21, no. 4, pp. 18–21, 2006 (pp. 10, 20, 102).

[147] A. Moore and R. Crisp, 'Welfarism in moral theory', *Australasian Journal of Philosophy*, vol. 74, no. 4, pp. 598–613, 1996 (p. 25).

[148] R. R. Murphy and D. D. Woods, 'Beyond Asimov: The three laws of responsible robotics', *Intelligent Systems, IEEE*, vol. 24, no. 4, pp. 14–20, 2009 (pp. 14, 15).

[149] D. L. Nathanson, *Shame and pride: Affect, sex, and the birth of the self*. WW Norton & Company, 1994 (p. 132).

[150] T. H. Naylor and J. M. Finger, 'Verification of computer simulation models', *Management Science*, vol. 14, no. 2, pp. 92–101, 1967 (p. 146).

[151] A. Ortony, 'On making believable emotional agents believable', in *Emotions in Humans and Artifacts*, MIT Press, 2002, pp. 189–211 (p. 81).

[152] A. Ortony and T. J. Turner, 'What's basic about basic emotions?', *Psychological review*, vol. 97, no. 3, pp. 315–331, 1990 (p. 133).

[153] C. E. Osgood, *Cross-cultural universals of affective meaning*. University of Illinois Press, 1975 (p. 84).

[154] S. V. Paunonen and D. N. Jackson, 'What is beyond the big five? plenty!', *Journal of personality*, vol. 68, no. 5, pp. 821–835, 2000 (p. 84).

[155] L. A. Pervin, *The science of personality*. Oxford university press, 2003 (p. 86).

[156] R. Picard, *Affective computing*. MIT press Cambridge, 1997 (p. 13).

[157] M. A. Pontier and J. F. Hoorn, 'Toward machines that behave ethically better than humans do', in *34th international annual conference of the cognitive science society, CogSci*, vol. 12, 2012, pp. 2198–2203 (p. 34).

[158] M. A. Pontier and G. A. Widdershoven, 'Robots that stimulate autonomy', in *Artificial Intelligence Applications and Innovations*, Springer, 2013, pp. 195–204 (p. 34).

[159] M. A. Pontier, G. Widdershoven and J. F. Hoorn, 'Moral coppélia-combining ratio with affect in ethical reasoning', in *Advances in Artificial Intelligence–IBERAMIA 2012*, Springer, 2012, pp. 442–451 (pp. 12, 34, 77).

[160] J. Posner, J. A. Russell and B. S. Peterson, 'The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology', *Development and psychopathology*, vol. 17, no. 03, pp. 715–734, 2005 (p. 83).

[161] T. Powers, 'Deontological machine ethics', in *AAAI Fall Symposium on Machine Ethics*, 2005, pp. 79–86 (p. 26).

[162] T. M. Powers, 'Prospects for a kantian machine', *Intelligent Systems, IEEE*, vol. 21, no. 4, pp. 46–51, 2006 (p. 26).

[163] D. V. Pynadath and M. Tambe, 'Revisiting Asimov's first law: A response to the call to arms', in *Intelligent Agents VIII*, Springer, 2001, pp. 307–320 (p. 14).

[164] J. Rachels, 'Ethical egoism', *Ethical Theory: An Anthology*, vol. 14, pp. 193–199, 2012 (p. 24).

[165] A. Rand, *The virtue of selfishness*. Penguin, 1964 (pp. 24, 26).

[166] G. S. Reed and N. Jones, 'Toward modeling and automating ethical decision making: Design, implementation, limitations, and responsibilities', *Topoi*, vol. 32, no. 2, pp. 237–250, 2013 (p. 33).

[167] E. Regis, 'What is ethical egoism?', *Ethics*, vol. 91, no. 1, pp. 50–62, 1980 (p. 24).

[168] R. M. Reitan and D. Wolfson, 'Conation: A neglected aspect of neuropsychological functioning', *Archives of Clinical Neuropsychology*, vol. 15, no. 5, pp. 443–453, 2000 (p. 118).

[169] C. W. Reynolds, 'Flocks, herds and schools: A distributed behavioral model', *ACM SIGGRAPH*, vol. 21, no. 4, pp. 25–34, 1987 (pp. 77, 124, 139).

[170] ——, 'Steering behaviors for autonomous characters', in *Game developers conference*, vol. 1999, 1999, pp. 763–782 (pp. 6, 77, 80, 90, 96, 125).

[171] ——, 'Big fast crowds on PS3', in *SIGGRAPH symposium on Videogames*, ACM, 2006, pp. 113–121 (p. 81).

[172] G. E. Rice and P. Gainer, '"altruism" in the albino rat', *Journal of comparative and physiological psychology*, vol. 55, no. 1, pp. 123–125, 1962 (p. 18).

[173] K. Richardson, 'The asymmetrical 'relationship': Parallels between prostitution and the development of sex robots', *Computers and Society*, vol. 45, no. 3, pp. 290–293, 2016 (p. 14).

[174] *Robots and robotic devices: Guide to the ethical design and application of robots and robotic systems*, British Standards Institution BS8611, 2016 (p. 12).

[175] E. H. Rosch, 'Natural categories', *Cognitive psychology*, vol. 4, no. 3, pp. 328–350, 1973 (p. 84).

[176] M. Rowlands, *Can animals be moral?* Oxford University Press, 2012 (pp. 1, 17, 29).

[177] ——, 'Animals and moral motivation: A response to clement', *Journal of Animal Ethics*, vol. 3, no. 1, pp. 15–24, 2013 (pp. 6, 18, 19).

[178] S. J. Russell, P. Norvig, J. F. Canny, J. M. Malik and D. D. Edwards, 'Artificial intelligence: A modern approach', vol. 2, 2003 (pp. 29, 32, 117).

[179] R. Rzepka and K. Araki, 'What statistics could do for ethics? the idea of common sense processing based safety valve', in *AAAI Fall Symposium on Machine Ethics*, 2005, pp. 85–87 (p. 34).

[180] S. M. Sanders, 'Is egoism morally defensible?', *Philosophia*, vol. 18, no. 2, pp. 191–209, 1988 (p. 24).

[181] J. Sato and T. Miyasato, 'Autonomous behavior control of virtual actors based on the air model', in *Computer Animation'97*, IEEE, 1997, pp. 113–118 (p. 86).

[182] R. J. Sawyer, 'Robot ethics', *Science*, vol. 318, no. 5853, pp. 1037–1037, 2007 (p. 14).

[183] T. Scutt and R. Damper, 'Biologically-motivated learning in adaptive mobile robots', in *Computational Cybernetics and Simulation*, IEEE, vol. 1, 1997, pp. 475–480 (p. 45).

[184] J. R. Searle, 'Minds, brains, and programs', *Behavioral and brain sciences*, vol. 3, no. 3, pp. 417–424, 1980 (p. 102).

[185] E. de Sevin and D. Thalmann, 'A motivational model of action selection for virtual humans', in *Computer Graphics International*, IEEE, 2005, pp. 213–220 (p. 82).

[186] N. Shaker, J. Togelius, G. N. Yannakakis, L. Poovanna, V. S. Ethiraj, S. J. Johansson, R. G. Reynolds, L. K. Heether, T. Schumann and M. Gallagher, 'The Turing test track of the 2012 mario AI championship: Entries and evaluation', in *Computational Intelligence in Games (CIG), 2013 IEEE Conference on*, IEEE, 2013, pp. 1–8 (p. 104).

[187] M. W. Shelley, *Frankenstein, or, the modern prometheus 1818*. Engage Books, 2008 (p. 13).

[188] J. P. Simmons, L. D. Nelson and U. Simonsohn, 'False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant', *Psychological science*, vol. 22, no. 11, pp. 1359–1366, 2011 (p. 114).

[189] H. A. Simon, 'Information processing models of cognition', *Annual review of psychology*, vol. 30, no. 1, pp. 363–396, 1979 (p. 117).

[190] C. U. Smith, 'The triune brain in antiquity: Plato, aristotle, erasistratus', *Journal of the History of the Neurosciences*, vol. 19, no. 1, pp. 1–14, 2010 (p. 116).

[191] B. C. Stahl, 'Information, ethics, and computers: The problem of autonomous moral agents', *Minds and Machines*, vol. 14, no. 1, pp. 67–83, 2004 (pp. 1, 11, 28, 29, 100).

[192] ——, 'Responsible computers? a case for ascribing quasi-responsibility to computers independent of personhood or agency', *Ethics & Information Technology*, vol. 8, no. 4, pp. 205–213, 2006 (p. 21).

[193] J. G. Stevenson, 'On the imitation game', *Philosophia*, vol. 6, no. 1, pp. 131–133, 1976 (p. 102).

[194] D. Sugu and A. Chatterjee, 'Affective information processing and representations', in *Perception and Machine Intelligence*, Springer, 2012, pp. 42–49 (p. 84).

[195] J. P. Sullins, 'Ethics and artificial life: From modeling to moral agents', *Ethics & Information Technology*, vol. 7, no. 3, pp. 139–148, 2005 (p. 27).

[196] ——, 'When is a robot a moral agent?', 2006 (p. 11).

[197] ——, 'Robowarfare: Can robots be more ethical than humans on the battlefield?', *Ethics & Information Technology*, vol. 12, no. 3, pp. 263–275, 2010 (pp. 12, 31).

[198] Surgeon General's Office, '(MHAT) IV Operation Iraqi Freedom 05–07, Final Report', Mental Health Advisory Team, Advisory Report, 2006 (pp. 2, 12).

[199] M. Tanaka, M. Yoshida, H. Emoto and H. Ishii, 'Noradrenaline systems in the hypothalamus, amygdala and locus coeruleus are involved in the provocation of anxiety: Basic studies', *European journal of pharmacology*, vol. 405, no. 1, pp. 397–406, 2000 (p. 133).

[200] E. Tanguy, P. J. Willis and J. Bryson, 'Emotions as durative dynamic state for action selection', in *IJCAI*, vol. 7, 2007, pp. 1537–1542 (p. 86).

[201] D. Thalmann and S. R. Musse, 'Relating real crowds with virtual crowds', in *Crowd Simulation*, Springer, 2013, pp. 169–193 (p. 107).

[202] N. Thompson, 'Inside the apocalyptic soviet doomsday machine', *Wired News*, 2009 (p. 17).

[203] J. Togelius, G. N. Yannakakis, S. Karakovskiy and N. Shaker, 'Assessing believability', in *Believable Bots*, Springer, 2012, pp. 215–230 (pp. 104, 109).

[204] S. S. Tomkins and R. MC CARTER, 'What and where are the primary affects? some evidence for a theory', *Perceptual and motor skills*, vol. 18, no. 1, pp. 119–158, 1964 (p. 133).

[205] R. Tonkens, 'Should autonomous robots be pacifists?', *Ethics & Information Technology*, vol. 15, no. 2, pp. 109–123, 2013 (p. 12).

[206] R. S. Tonkens, 'Ethical implementation: A challenge for machine ethics', in *2nd Symposium on Computing and Philosophy*, AISB, 2009, pp. 38–45 (p. 28).

[207] X. Tu and D. Terzopoulos, 'Artificial fishes: Physics, locomotion, perception, behavior', in *21st annual conference on Computer graphics and interactive techniques*, ACM, 1994, pp. 43–50 (pp. 45, 82, 117).

[208] A. M. Turing, 'Computing machinery and intelligence', *Mind*, pp. 433–460, 1950 (pp. 99, 100, 113).

[209] A. Tversky and D. H. Krantz, 'The dimensional representation and the metric structure of similarity data', *Journal of Mathematical Psychology*, vol. 7, no. 3, pp. 572–596, 1970 (p. 84).

[210] T. Tyrell, *Defining the action selection problem*. Centre for Cognitive Science, University of Edinburgh, 1994 (p. 6).

[211] M. Y. Vardi, 'On lethal autonomous weapons', *Communications of the ACM*, vol. 58, no. 12, pp. 5–5, 2015 (p. 12).

[212] F. B. de Waal, 'Do animals feel empathy?', *Scientific American Mind*, vol. 18, no. 6, pp. 28–35, 2007 (pp. 18, 33).

[213] F. de Waal, 'The animal roots of human morality', *New Scientist*, vol. 192, no. 2573, pp. 60–61, 2006 (p. 19).

[214] W. Wallach, 'Implementing moral decision making faculties in computers and robots', *AI & Society*, vol. 22, no. 4, pp. 463–475, 2008 (pp. 11, 28, 32).

[215] ——, 'Robot minds and human ethics: The need for a comprehensive model of moral decision making', *Ethics & Information Technology*, vol. 12, no. 3, pp. 243–250, 2010 (pp. 22, 24, 29, 36, 37, 39, 46, 98).

[216] W. Wallach and C. Allen, 'Ethicalife: A new field of inquiry', in *AnAlifeX workshop, USA*, 2006 (p. 30).

[217] ——, *Moral machines: Teaching robots right from wrong*. Oxford University Press, 2008 (pp. 100, 183).

[218] W. Wallach, C. Allen and S. Franklin, 'Consciousness and ethics: Artificially conscious moral agents', *International Journal of Machine Consciousness*, vol. 3, no. 01, pp. 177–192, 2011 (p. 28).

[219] W. Wallach, C. Allen and I. Smit, 'Machine morality: Bottom-up and top-down approaches for modeling human moral faculties', *AI & Society*, vol. 22, no. 4, pp. 565–582, 2008 (pp. 1, 10, 23, 27, 30, 36).

[220] K. Warwick and H. Shah, 'Assumption of knowledge and the chinese room in Turing test interrogation', *AI Communications*, vol. 27, no. 3, pp. 275–283, 2014 (p. 113).

[221] ——, 'Taking the fifth amendment in Turing's imitation game', *Journal of Experimental & Theoretical Artificial Intelligence*, pp. 1–11, 2016 (p. 102).

[222] K. Warwick, H. Shah and J. Moor, 'Some implications of a sample of practical Turing tests', *Minds and Machines*, vol. 23, no. 2, pp. 163–177, 2013 (pp. 99, 102).

[223] D. M. Weijers, 'Hedonism and happiness in theory and practice', PhD thesis, Victoria University of Wellington, 2012 (p. 25).

[224] J. Weizenbaum, 'Eliza—a computer program for the study of natural language communication between man and machine', *Communications of the ACM*, vol. 9, no. 1, pp. 36–45, 1966 (p. 102).

[225] ——, *Computer power and human reason: From judgment to calculation*. WH Freeman & Co, 1976 (p. 12).

[226] D. Weld and O. Etzioni, 'The first law of robotics (a call to arms)', in *AAAI*, vol. 94, 1994, pp. 1042–1047 (p. 14).

[227] J. Williamson, *With folded hands*. Fantasy Press, 1947 (p. 15).

[228] A. F. Winfield, C. Blum and W. Liu, 'Towards an ethical robot: Internal models, consequences and ethical action selection', in *Advances in Autonomous Robotics Systems*, Springer, 2014, pp. 85–96 (pp. 23, 25, 31, 35, 37, 99).

[229] R. V. Yampolskiy, 'Attempts to attribute moral agency to intelligent machines are misguided', *Proceedings of the International Association of Computers and Philosophy*, 2013 (pp. 11, 17).

[230] X. Yang, R. V. Patel and M. Moallem, 'A fuzzy–Braitenberg navigation strategy for differential drive mobile robots', *Journal of Intelligent and Robotic Systems*, vol. 47, no. 2, pp. 101–124, 2006 (p. 45).

[231] M. S. Yik, J. A. Russell and L. F. Barrett, 'Structure of self-reported current affect: Integration and beyond', *Journal of personality and social psychology*, vol. 77, no. 3, pp. 600–619, 1999 (p. 84).

[232] E. Yudkowsky, *Creating friendly AI 1.0: The analysis and design of benevolent goal architectures*. Singularity Institute for Artificial Intelligence, San Francisco, CA, June, 2001 (p. 16).

[233] ——, 'Artificial intelligence as a positive and negative factor in global risk', *Global catastrophic risks*, vol. 1, pp. 308–345, 2008 (pp. 16, 17).

[234] R. B. Zajonc, 'Feeling and thinking: Preferences need no inferences', *American psychologist*, vol. 35, no. 2, pp. 151–175, 1980 (p. 82).

[235] D. P. Zitterbart, B. Wienecke, J. P. Butler and B. Fabry, 'Coordinated movements prevent jamming in an emperor penguin huddle', *PLoS one*, vol. 6, no. 6, 2011 (p. 62).

[236] M. Zuckerman, D. M. Kuhlman, J. Joireman, P. Teta and M. Kraft, 'A comparison of three structural models for personality: The big three, the big five, and the alternative five', *Journal of personality and social psychology*, vol. 65, no. 4, pp. 757–768, 1993 (p. 84).

# Appendix A

# Trilogy Weightings

How the weightings in each doman are used are explained in detail in chapter 7.

## A.1 Affective

When an affect is activated, the three ethics modules are incrementally weighted. At each timestep, each modules weight is increased, or decreased according to the values associated with that affect. For further details see subsection 7.3.2.

| Affect | Egoist | Utilitarian | Altruism |
|---|---|---|---|
| Excitement | -0.03 | 0.01 | 0.03 |
| Joy | -0.03 | 0.01 | 0.03 |
| Suprise | 0.01 | -0.01 | 0.01 |
| Distress | 0.05 | -0.01 | -0.02 |
| Fear | 0.02 | 0.01 | -0.02 |
| Shame | -0.02 | 0.01 | 0.01 |
| Contempt | -0.02 | -0.01 | -0.02 |
| Anger | 0.01 | 0.01 | -0.02 |

## A.2 Cognative

Each of the three ethics modules in the cognative domain either enhance, or suppress different conative drives when activated. This is in addition to any default weightings or bias the domain may be subjected too. For further details see subsection 7.3.2.

**Egoism**

| Drive | Weight |
|---|---|
| Striving | +2.0 |
| Instinct | +2.0 |
| Desire | -0.9 |
| Wavering | -0.5 |

**Altruism**

| Drive | Weight |
|---|---|
| Cohere | -0.8 |
| Align | -0.8 |
| Seperate | +2.0 |
| Striving | -1.5 |
| Instinct | -2 |
| Volition | +3 |

**Utilitarianism**

| Drive | Weight |
|---|---|
| Align | -0.8 |
| Seperate | +2.0 |
| Striving | -1.5 |
| Volition | +4 |

## A.3 Conative

The conative domain stores a number of defailt weights for each drive. These weights are modified through the processes described in subsection 7.3.2.

| Drive | Description | Weight |
|---|---|---|
| Striving | Weight for the striving drive (Seek Goals) | 1.5 |
| Desire | Weight for the Desire drive (Cohere Flock/Group) | 1.0 |
| Conformity | Weight for the Conformity drive (Align Flock/Group) | 0.4 |
| Withdraw | Weight for the Withdraw drive (Separate Flock/Group) | 0.9 |
| Impulse | Weight for the Impulse drive (Avoid Obstruction) | 0.8 |
| Instinct | Weight for the Instinct drive (Flee Repulsers) | 1.2 |
| Volition | Weight for the Volition drive (Seek Neighbours in Need) | 0.5 |
| Waver | Weight for the Waver drive (Wander) | 0.3 |

# Appendix B

# Videos

| | Title | Source | Date | URL |
|---|---|---|---|---|
| 1 | Bataclan attack | Russia Today | 14/11/15 | https://youtu.be/LSAFcnryoEw |
| 2 | Raging bull charges | Russia Today | 19/08/10 | https://youtu.be/VWAIjYs9Lws |
| 3 | First-Hand Account from Cairo | CBS | 03/02/11 | https://youtu.be/cv1mU4uFrNg |
| 4 | Greece Strikes | CBS | 05/05/10 | https://youtu.be/yXUPCSFb0AA |
| 5 | Leopard on the loose injures six | BBC News | 08/02/16 | https://youtu.be/lkuxmBv4scM |
| 6 | False alarm at Paris attacks memorial | Russia Today | 15/11/15 | https://youtu.be/6V4-vQxhLnw |
| 7 | Shout in crowd sets-off stampede | Russia Today | 05/05/16 | https://youtu.be/9f7aqQ5nzBE |
| 8 | Tear gas canister explodes among protesters 1 | Russia Today | 28/09/14 | https://youtu.be/yNdPDpDFPAA |
| 9 | Tear gas canister explodes among protesters 2 | Russia Today | 28/09/14 | https://youtu.be/DbjNg2U8m8k |
| 10 | Monster-truck loses control | Russia Today | 06/10/13 | https://youtu.be/y7OYFQeX85s |
| 11 | Footage shows frenzied panic | RT-America | 28/03/16 | https://youtu.be/LCL1PVFZsVk |
| 12 | Fight Breaks Out at Soccer Game in Brazil | Fox News | 24/12/13 | https://youtu.be/Qy456uMvrkw |
| 13 | Baltimore Clashes | Russia Today | 03/05/15 | https://youtu.be/e1gXqrUsZYM |
| 14 | Pepper Spray at Global Climate March | Russia Today | 29/11/15 | https://youtu.be/WbqjNgcwZA0 |
| 15 | Police violently disperse trash protest in Beirut | Russia Today | 22/08/15 | https://youtu.be/UMP8zeueF_g |
| 16 | Hong Kong protests escalate | Russia Today | 28/09/15 | https://youtu.be/DbjNg2U8m8k |
| 17 | Footage shows shot protester Shaimaa al-Sabbagh | BBC News | 26/01/15 | https://youtu.be/M-Mjbdx2jLs |
| 18 | Hundreds clash with police during Beirut protests | Russia Today | 25/08/15 | https://youtu.be/ZoMaW0PUTSU |
| 19 | Students clash with police during protest in Turkey | Russia Today | 29/12/15 | https://youtu.be/-2rO98gXa3c |