University of Bath

PHD

AI Governance Through a Transparency Lens

Theodorou, Andreas

Award date:
2019

Link to publication

Alternative formats
If you require this document in an alternative format, please contact:
openaccess@bath.ac.uk

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal ?

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 18. Sep. 2021

Citation for published version:
Theodorou, A 2019, 'AI Governance Through a Transparency Lens', University of Bath.

Publication date:
2019

Document Version
Publisher's PDF, also known as Version of record

Link to publication

University of Bath

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 29. Jul. 2019

AI Governance Through a Transparency Lens

submitted by

Andreas Theodorou

for the degree of Doctor of Philosophy

of the

University of Bath

Department of Computer Science

March 2019

COPYRIGHT

Attention is drawn to the fact that copyright of this thesis rests with the author. A copy of this thesis has been supplied on condition that anyone who consults it is understood to recognise that its copyright rests with the author and that they must not copy it or use material from it except as permitted by law or with the consent of the author.

This thesis may be made available for consultation
within the University Library and may be
photocopied or lent to other libraries for the purposes
of consultation with effect from ..... (date)

Signed on behalf of the Faculty of Science .....

Abstract

When we interact with any object, we inevitably construct mental models to assess our relationship with the object. These determine our perceived utility of the object, our expectations of its performance, and how much trust we assign to it. Yet, the emerging behaviour of intelligent systems can often be difficult to understand by their developers, let alone by end users. Even worse, some intelligent system developers have often been using anthropomorphic and other audiovisual cues to deliberately deceive the users of their creations.

This deception alongside with pop-science narratives for the creation of an 'all-powerful' AI system result in a moral confusion regarding the moral status of our intelligent artefacts. Their ability to exhibit agency or even perform 'super-human' tasks leads to many believing that they are worthy of being granted moral agency, a status given only to humans so far, or moral patiency. In this dissertation, I provide normative and descriptive arguments against granting any moral status to intelligent systems.

As intelligent systems become increasingly integral parts of our societies, the need for affordable easy-to-use tools to provide transparency, the ability to request —at any point of time or over a specific period— an accurate interpretation of the agent's status, grows. This dissertation provides the knowledge to build such tools. Two example tools, ABOD3 and ABOD3-AR, are presented here. Both of them are able to provide real-time visualisation of transparency-related information for the action-selection mechanisms of intelligent systems.

User studies presented in this document demonstrate naive and experts end users can use ABOD3 and ABOD3-AR to calibrate their mental models. In the three human-robot interaction studies presented, participants with access to real-time transparency information had not only a reduced perception of the robots as anthropomorphic, but also adjusted their expectations and trust to the system after ABOD3 provided them with an understanding of Artificial Intelligence (AI) by removing the 'scary' mystery

around why "is it behaving like that". In addition, indicative results presented here demonstrate the advantages of implementing transparency for AI developers. Students undertaking an AI module were able to understand the AI paradigms taught and the behaviour of their agents better by using ABOD3.

Furthermore, in a post-incident transparency study performed with the use of Virtual Reality technology, participants took the role of a passenger in an autonomous vehicle (AV) which makes a moral choice: crash into one of two human-looking non-playable characters (NPC). Participants were exposed to one of three conditions; a human driver, an opaque AV without any post-incident information, and a transparent AV that reported back the characteristics of NPC that influenced its decision-making process, e.g. its demographic background. When the characteristics were revealed to the participants after the incident, the autonomous vehicle was perceived as significantly more mechanical and utilitarian. Interestingly, our results also indicate that we find it harder to forgive machinelike intelligent systems compared to humans or even more anthropomorphic agents. Most importantly, the study demonstrates a need for caution when incorporating supposedly normative data, gathered through the use of text-based crowd-sourced preferences in moral-dilemmas studies, into moral frameworks used in technology.

Based on the concerns that motivate this work and the results presented, I emphasise the need for policy that ensures distribution of responsibility, attribution of accountability, and inclusion of transparency as a fundamental design consideration for intelligent systems. Hence, the research outlined in this document aims to contribute to —and has successfully contributed to—the creation of policy; both soft governance, e.g. standards, and hard governance, i.e. legislation.

Finally, future multi-disciplinary work is suggested to further investigate the effects of transparency on both naive and expert users. The proposed work is an extended investigation of how robot behaviour and appearance affect their utility and our overall perception of them.

Acknowledgements

Firstly, I would like to thank my supervisor, Joanna J. Bryson, who constantly pushed me to become the researcher I am today. Her valuable knowledge into Artificial Intelligence allowed me conduct the research in this document. Moreover, her insights the world of academia and research have assisted me with conducting my own supervision, disseminating my research, and get engaged in policy discussions.

My thanks also go to my amazing collaborators. First, to Robert H. Wortham who developed the R5 and run two of the studies presented here and whose friendship definitely help me 'survive' this PhD. Alexandros Rotsidis who trusted my ideas and and developed the ABOD3-AR software. Holly Wilson who is the first student I ever supervise and did an amazing job at carrying out my proposed research project. Alin Coman, Mark Riedl, and Kristin Siu for their feedback on the Sustainability Game and enabling me to spent some time at the Georgia Institute of Technology.

I will like to thanks my examiners, Marina De Vos and Sabine Heart, for the time they spent at reading this hefty document. Their feedback helped me form the final version of this document.

Furthermore, I need to thanks various people at the Department of Computer Science at the University of Bath for all their support, chats, buying me an espresso machine, and helping me expand my knowledge. This includes, in alphabetical order, Alan Hayes, Alessio Santamaria, Anamaria Ciucanu, Christina Keating, Cillian Dudley, David Sherratt, Eamonn O'Neill, Fabio Nemetz, Guy McCusker, Hashim Khalid Yaqub, James (Jim) Laird, Jo Hyde, Joanna Tarko, Julian Padget, Michael Wright, Özgür Şimşek, Rachid Hourizi, Siriphan Wichaidit, Tom S. F. Haines, and Zack Lyons. Outside the department, I will like to thanks Christopher (Chris) Harrison from the Doctoral College, who picked up this copy from the printing services and helped me sort out all the paperwork! My thanks extend to academics beyond Bath, who provided advice or helped me enhance my AI knowledge. This includes Antonis Kakas, Alan Winfield,

Ilse Verdiesen, Frank Dignum, Jahna Otterbacher, Loizos Michael, Scott Hawley, and my—at time of writing—boss Virginia Dignum.

I also need to acknowledge my friends at Bath, who provided support and understanding throughout the PhD. In alphabetical order: Andreas Michael, Daniela De Angeli, John Benardis, and Mojca Sonjak. In addition, I will like to thanks my friends in the UK and Cyprus for all of their support; this includes Afroditi Chari, Andreas Foiniotis, Andreas Alvanis, Andreas Antoniadis, Chris Green, Christos Piskopos, George Flourentzos, Nasia Michaelidou, Stefani Nikolaou, Stella Kazamia, and Xenia Menelaou.

Moreover, I will like to say a huge *Thank You* to my mother, Maria Ioannou, who supported me throughout my life and enabled me to pursue my academic dreams. In addition, I thank my father, Christos Theodorou, my sister, Marianna Theodorou, and my grandparents; Panikos Ioannou, Georgia Ioannou, and Androulla Theodorou. Finally, my thanks extend to my Godmother, Yiannoula Menelaou, and her family, Adamos, Georgia, and Marios, and my uncle Doros Theodorou and his family, Andri, Andreas, and Ioanna.

Last but definitely not least, I will like to thank a special person who supports me in life—and now in research too—my partner, Andrea Aler Tubella. Her support was paramount during the last years of my PhD.

Contents

Chapter 1

Introduction

> "As you set out for Ithaka, hope your road is a long one, full of adventure, full of discovery."
>
> Constantinos P. Cavafy,
> *Ithaka*

1.1 Thesis

Transparency is a key consideration for the ethical design and use of Artificial Intelligence, and has recently become a topic of considerable public interest and debate. We frequently use philosophical, mathematical, and biologically inspired techniques for building artificial, interactive, intelligent agents. Yet despite these well-motivated inspirations, the resulting intelligence is often developed as a black box, communicating no understanding of how the underlying real-time decision making functions. This compromises both the safety of such systems and fair attribution of moral responsibility and legal accountability when incidents occur.

This dissertation provides the knowledge and software tools to make artificially intelligent agents more transparent, allowing a direct understanding of the action-selection system of such a system. The use of transparency, as demonstrated in this document, helps not only with the debugging of intelligent agents, but also with the public's understanding of Artificial Intelligence (AI) by removing the 'scary' mystery around "why is it behaving like that". In the research described in this document I investigate and compare the perception we have of intelligent systems, such as robots and autonomous vehicles, when they are treated as black boxes compared to when we make their action-

selection systems transparent. Finally, I make normative and descriptive arguments for the moral status of intelligent systems and contribute to regulatory policy regarding such systems.

In the rest of this chapter, I first discuss the motivation behind this research and then the structure of this dissertation. I conclude the chapter by outlining the engineering, scientific, and overall societal contributions made by this research.

1.2 Motivation

In this Section, I first discuss the need for policy—both in terms of standards and legislation—for AI and how this research aims to inform policymakers, hence, contribute to regulations and therefore society. Next, I explore the safety and societal concerns that motivated this research in designing and building transparent-to-inspection intelligent systems. I discuss the safety concerns that could arise from the black-box treatment of intelligent systems and how transparency can mitigate them by allowing real-time calibration of trust. Furthermore, I discuss that as incidents—either due to developers' errors or malicious use—will inevitably happen and how transparency can at least help us attribute responsibility and accountability to the right legal person.

1.2.1 AI Governance

Artificial Intelligence technologies are already present in our societies in many forms: through web search and indexing, email spam detecting systems, loan calculators, and even single-player video games. All of these are intelligent systems that billions of people interact with daily. They automate repeating tasks, provide entertainment, or transform data into recommendations that we can choose to act upon. By extending ourselves through our artefacts, we significantly increase our own pool of available behaviours and enhance existing ones.

This technology has the potential to greatly improve our autonomy and wellbeing, but to be able to interact with it effectively and safely, we need to be able to trust it. For example, automatic elevators were fully developed as early as in 1900. Yet, most people at the time were too uncomfortable to ride in them, citing safety concerns and support for elevator operators. It took a strike in 1945 that left New York paralised and a huge industry-led PR push in the early 50s to change people's minds. At the same time, the American National Standards Institute updated the *Safety Code for Elevators first issuance*, later to be known as standard *A17.1*, to establish minimum-safety requirements (ASA 17.1 -1955, n.d.).

As the automatic elevators examples shows, building public trust in robotics and artificial intelligence at large requires a multi-faceted approach; it is both a societal challenge and a technical one. Accidents, misuse, disuse, and malicious use are all bound to happen. The real problem is establishing the social and political will for assigning and maintaining *accountability* and *responsibility* for the manufacturers and users of artefacts. Yet, there is no unanimity between researchers on the need for effective regulations addressing the design and usage of intelligent systems by holding their designers and users potentially liable. Instead, there is a disagreement in literature regarding the moral status of intelligent agents (Bryson and Kime, 1998; Gunkel, 2012) and, therefore, regarding policy.

There is a belief that the capacity to express human-like behaviour is in any way indicative of commonality of phenomenological experience between machines and humans. Due to our frequent lack of understanding and attribution of human-like characteristics —both physical and emotional— to intelligent systems, some believe that human punishments such as fines, prison, and the other tools of human law could be extended to our intelligent artefacts. At the same time, they may believe that our artefacts require *welfare* and need to be protected as other sentient beings are. This could result in the elevation of intelligent systems into being *moral subjects*; part of *our moral spectrum*.

A common argument in favour of granting them a moral status is their actual appearance or their human-like —or sometimes even 'superhuman'— capabilities (Coeckelbergh, 2009; Himma, 2009). Such a move could lead to the attribution of *legal personhood* to them. If we declare artefacts to be legal persons, those artefacts could be used like shell companies to evade justice, potentially leading to further societal disruption, which could corrupt economies and power structures (Bryson, Diamantis and Grant, 2017), leaving ordinary citizens disempowered with less protection from powerful institutions (Elish, 2016).

This *moral confusion* about the moral status of robots is the direct result of the association of a wide range of psychological, ethical, or even religion-related phenomena with the 'briefcase' words: morality, intelligence, cognition, and consciousness. Hence, in the next chapter, I aim to make it clear with straight-forward definitions that *we* have complete control over the design of both our intelligent systems and of the policy that governs them. Once their manufacture nature is clear, either through educational means like this dissertation or proactive approaches, like the implementation *transparency* as proposed by the *EPSRC Principles of Robotics* (Boden et al., 2011), then we can focus in the implementation of governance mechanisms. AI policy includes ethical guidelines, standards, and legislation (discussed in chapter 7) that provide good-design practices

and minimum-performance standards, while they also promote the societal-beneficial development and usage of this technology.

Ultimately, building responsible AI requires a focus in these three core principles: responsibility, accountability, and *transparency* (Dignum, 2017). While the first two, as discussed in chapters 2 and 7, can be solved through socio-legal means and intervention by policymakers, the last requires novel engineering solutions. Hence, I focus the majority of this dissertation at the principle of transparency, but also inevitably make recommendations for the other two by investigating and recommending good-design practices for designing and developing intelligent systems to help establish the necessary political will at AI governance. Once such practices are clear, then with education and good hiring, ordinary legal enforcement of liability standards should be sufficient to maintain human control, while transparency, as discussed next, can help us ensure real-time safety and long-term accountability.

1.2.2 Transparency in AI

1.2.3 Misuse, Disuse, and Trust

The black-box nature of intelligent systems, even in relatively simple cases such as context-aware applications, makes interaction limited and often uninformative for the end user (Stumpf et al., 2010). Limiting interactions may negatively affect the system's performance or even jeopardize the functionality of the system.

Disuse refers to failures that occur when people reject the capabilities of a system, whereas misuse refers to the failures that occur when people inadvertently assign trust that exceeds the system capabilities (Lee and See, 2004). Consider for example an autonomous robotic system built for providing health-care support to the elderly, who may be afraid of it, or simply distrust it, and in the end refuse to use it leading to disusing the system. In such a scenario human well-being could be compromised, as patients may not get their prescribed medical treatment in time, unless a human overseeing the system detects the lack of interaction (or is contacted by the robot) and intervenes. Conversely, if the human user places too much trust in a robot, it could lead to misuse and ultimately to over-reliance on the system (Parasuraman and Riley, 1997). If the robot malfunctions and its patients are unaware of its failure to function, the patients may continue using the robot, risking their health.

To avoid such situations, proper calibration of trust between the human users and/or operators and their autonomous systems is important. Calibration of trust refers to the correspondence between a person's trust in the system and the system's capabilities

(Lee and Moray, 1994). It occurs when the end-user has a mental model of the system and relies on the system within the system's capabilities and is aware of its limitations. If we are to consider transparency as a mechanism that exposes the decision-making of a system, then it can help users adjust their expectations and forecast certain actions from the system. A real-time implementation can help users to calibrate their trust to the machine (Lyons, 2013, and references therein). Therefore, transparency is first and foremost a *safety consideration*.

The relationship between transparency, trust, and utility is a complex one (Wortham and Theodorou, 2017). By exposing the inner 'smoke and mirrors' of our agents, we risk making them look less interesting. Moreover, the wide range of application domains for AI and of the different stakeholders interacting with intelligent systems should not be underestimated. Therefore, what is effectively transparent varies by who the observer is, and what their goals and obligations are. There is however a need for design guidelines on how to implement transparent systems, alongside with a 'bare minimum' standardised implementation (Bryson and Winfield, 2017; Theodorou, Wortham and Bryson, 2017). I discuss these considerations, alongside with how we should define *transparency*, in chapter 3. In chapter 4 I present the implementations of tools developed in line with these principles and considerations. Finally, I examine the effects of transparency in terms of calibrating expectations and trust in the context of unplanned naive robot encounter in chapter 5.

1.2.4 Malfunctioning and Malicious Usage

Accidents due to malfunctioning or incidents due to malicious use are bound to happen. Over the period of 2000-2013, no less than 44 deaths and 1391 injuries were reported as the results of accidents involving medical robots at US-based hospitals (Alemzadeh et al., 2016). In October 2007, an Oerlikon GDF-005 semi-autonomous anti-aircraft gun malfunctioned, entered automatic mode, broke through the traversal-restriction safety mechanisms and began firing, striking the other guns along the firing line. It shot 250 high-explosive rounds at nearby friendly soldiers, resulting in 9 deaths 14 injured. Moreover, there have been 4 accidents with self-driving cars (Josh and Timmons, 2016; Lubben, 2018; Green, 2018) and likewise incidents have been recorded in industry settings. Bystanders, people who do not directly interact with the robot, can also be effected. For example, the pedestrian killed by a self-driving car in April 2018 was crossing the road, when the car misidentified and hit her (Lubben, 2018). At the time of writing, an autonomous robot at an Amazon Warehouse punctured a bear repellent, putting 24 workers in the hospital (Jolly, 2018). These incidents are classified as acci-

dents; they were caused by flaws in the design, bugs in the development of system, or misuse.

Furthermore, like all information technology systems, intelligent systems are —and will— be prone to cyberattacks and other malicious use. We should always try to secure our systems using physical and software cybersecurity techniques. However, we should also consider the new challenges at securing *autonomous* systems compared to more 'traditional' computer systems. Data gathering capabilities embedded into children's toys elucidate how both adults and children may be lulled into a false sense of security. Research reveals some children do not realise when their 'smart toy' was recording (McReynolds et al., 2017). Self-driving cars and other robots are 'moving' data collection machines, where bystanders get themselves and/or property recorded.

It has become increasingly important that AI algorithms be robust against external, ill-natured manipulation. While the biases that pervade our culture will be unintentionally uploaded into our models (Caliskan, Bryson and Narayanan, 2017), there is also the likelihood of deliberately introducing biases into our systems and, even worse, to force them to act on those biases (Bryson, 2017). An example of this is the chatbot *Tay*, which tweeted pro-Nazi messages as a group of people exploited the lack of appropriate filters to feed it with data (Vincent, 2016). In addition, we should not discount the possibility of the developers themselves introducing stereotypes or biases, unbeknownst to their users.

Simulators, testing procedures, and other safety measures can only cover *what the developers thought of*. However, the world is much larger and more complex than any simulator; as Brooks (1991a) says, the world is its own best model. Accidents will happen regardless of how many hours our systems spend on a simulator. Moreover, regardless of any implementation of transparency, equipment can still malfunction and warnings be ignored. We can not 'fool proof' our systems sufficiently to stop all incidents, but rather we can limit them by implementing the best security measures possible —especially for safety-critical systems.

Not all intelligent systems are designed with a legal, societal-beneficial purpose. For example, a *botnet*, a term that combines *robot* with *networks*, is a collection of connected computers, each running one or more bots, which coordinate to perform some task (Hoque, Bhattacharyya and Kalita, 2015). Malicious botnets are used to launch spam and distributed denial-of-service (DDoS) attacks and to engineer theft of confidential information, click fraud, cybersabotage, and cyberwarfare at unprecedented scales.

AI technologies have been used to psychologically profile and manipulate the voting

population at an unparalleled scale. Already, evidence shows that such manipulation altered the outcomes of the UK's EU membership referendum (Howard and Kollanyi, 2016; Bastos and Mercea, 2017), the US presidential election (Howard, Woolley and Calo, 2018), and attempted to disrupt French Elections (Ferrara, 2017). In all three instances, bots used by populist movements disseminated information and engaged in interacting with other users of social media. The manipulation of the public aimed at entrapping voters into echo chambers, in an effort to invest exclusively, through their votes, into their in-group identity. This polarisation may result to people withdrawing from the more profitable, but riskier, out-group transactions, and both aggregate and per capita output necessarily fall (Stewart, McCarty and Bryson, 2018). Behavioural economics research demonstrates that explicit knowledge of the benefits of cooperation in the form of public goods investments does not universally promote that investment, even when doing so is beneficial to the individual and group (Sylwester, Herrmann and Bryson, 2013; Herrmann, Thöni and Gächter, 2008; Binmore and Shaked, 2010). We now risk causing not only a long-term damage to our economies, but also to the democratic institutions of our societies.

These examples of malicious use of AI technologies further raise the need for AI-related legislation, which promotes transparency, responsibility, and accountability. Users interacting with intelligent systems should know when they do so. It is also equally important, at least in political settings, that users should be made aware of the political donors who funded the system.

Where incidents occur, they must be addressed, in some cases redressed, and in all cases used to reduce future mishaps. This is nice in principle, but probably impossible to implement without having an understanding of what lead to the error. We need to work towards implementing *transparency* in general; whether that is in the development process or in the design itself. This dissertation focuses on the later; transparency in the action-selection systems. Such implementations of transparency can, as discussed above and shown in this dissertation in chapter 5, help avoid incidents as users calibrate their expectations. Yet, an appropriate implementation of transparency of the decision-making system is not only beneficial for post-deployment end users, but also for experts. Transparency can help designers and developers design and debug their systems (Wortham, Theodorou and Bryson, 2017b). Furthermore, accident investigators can make better-educated judgements on what happened (Winfield and Jirotka, 2017). Therefore, the end goal of transparency should not necessarily be complete comprehension. Instead, the goal of transparency is providing sufficient information to ensure at least human accountability (Bryson and Theodorou, 2019). We took this into

consideration in the design guidelines presented in chapter 3 and technology shown in chapter 4.

1.3 Dissertation Structure

In this section, I outline the content of each of the following chapters of this dissertation. Moreover, I reference any related publications published, under review, or in preparation containing any of the positions, results, or other contributions presented.

1.3.1 Chapter 2: Morality and Intelligence

This chapter provides the definitions used throughout this dissertation. Most importantly, it emphasises that this dissertation deals exclusively with the design principles that the human agents involved in the design, development, usage, and governance of AI should consider. It aims, by trying to tackle our moral confusion concerning intelligent agents, at sending a message that the implementation of transparency is a design and implementation choice *we can make*, similar to how the usage of a certain technology X over another technology Y is our own choice.

Here, I present an evaluation of the requirements for *moral agency* and *moral patiency*. I examine human morality through a presentation of a high-level ontology of the human action-selection system. Then, drawing parallels between natural and artificial intelligence, I discuss the limitations and bottlenecks of intelligence, demonstrating how an 'all-powerful' Artificial General Intelligence would not only entail omniscience, but also be impossible. I demonstrate throughout this chapter how culture determines the moral status of all entities, as morality and law are human-made 'fictions' that help us guide our actions. This means that our moral spectrum can be altered to include machines. However, there are both descriptive and normative arguments for why a such move is not only avoidable, but also should be avoided

Associated Papers

Theodorou A., Under Review. Why Artificial Intelligence is a Matter of Design.

1.3.2 Chapter 3: Designing Transparent Machines

In this chapter, after considering and expanding upon other prominent definitions found in the literature, I provide a robust definition of transparency as a mechanism to expose the decision making of a robot. This chapter concludes by discussing design decisions

developers may face when implementing transparent systems. Work presented in this chapter is taken into consideration throughout the rest of this document.

The United Kingdom's Principles of Robotics advises the implementation of transparency in robotic systems, yet, it does not specify what transparency really is. This chapter introduces the reader to the importance of having transparent inspection of intelligent agents by examining how we construct mental models for AI. Moreover, by considering and expanding upon other prominent definitions found in the literature, it provides a robust definition of transparency as a mechanism to expose the decision-making of a robot. It also investigates case-specific transparency implementations for embodied agents. The chapter concludes by addressing potential design decisions developers need to consider when designing and developing transparent systems.

Associated Papers

Theodorou, A., Wortham, R.H. and Bryson, J.J., 2017. Designing and implementing transparency for real time inspection of autonomous robots. *Connection Science*, 29(3), pp.230 241.

1.3.3 Chapter 4: Building Human-Centric Transparent AI

In this chapter I demonstrate how by considering the design principles from the previous chapter we can actually develop tools that provide real-time understanding of the decision-making mechanisms of intelligent agents. Two of such tools are the main engineering contributions of this dissertation, *ABOD3* and *ABOD3-AR*, which are used in work presented in later chapters.

The applications presented in this chapter provide real-time visualisation and debugging of an agent's goals and priorities. Both ABOD3 and ABOD3-AR can be used as by both agent developers and end users to gain a better mental model of the internal state and decision making processes taking place within an agent. The former can use such information to tune and debug their creations. End users benefit from better models, as described the previous chapter. Both of these claims are tested in four user studies presented in the next chapter.

In addition to ABOD3 and ABOD3-AR, a new action-selection system *UNity-POSH* (UN-POSH), is introduced. The UN-POSH Planner is a new lightweight reactive planner, based on an established behaviour based robotics methodology and its reactive planner representation —the POSH (Parallel-rooted, Ordered Slip-stack Hierarchical) planner implementation. UN-POSH is specifically designed to be used in modern video

games by exploiting and facilitating a number of game-specific properties, such as synchronisation between the action-selection system and the animation controller of the agent. It can provide a feed transparency-related information, which can be interpreted by ABOD3 to visualise plan execution. UN-POSH is used in the BOD UNity Game (BUNG) presented in the same chapter and the Sustainability Game presented in chapter B.

Associated Papers

Rotsidis A., Theodorou A., and Wortham R.H., 2019. Robots That Make Sense: Transparent Intelligence Through Augmented Reality. *1st International Workshop on Intelligent User Interfaces for Algorithmic Transparency in Emerging Technologies*, Los Angeles, CA USA.

Bryson, J.J. and Theodorou A., 2019. How Society Can Maintain Human-Centric Artificial Intelligence. In Toivonen-Noro M. I, Saari E. eds.*Human-centered digitalization and services.*

Theodorou A., 2017. ABOD3: A Graphical Visualization and Real-Time Debugging Tool for BOD Agents. *CEUR Workshop Proceedings*, 1855, 60-61.

1.3.4 Chapter 5: Improving Mental Models of AI

Autonomous robots can be difficult to design and understand. If designers have difficulty decoding the behaviour of their own agents, then naive users of such systems are unlikely to decipher a system's behaviour simply through observation. In this chapter, we demonstrate that providing even a simple abstracted real-time visualisation of an agent's action-selection system, by using the software introduced in chapter 4, we can radically improve the transparency of machine cognition. The four studies included in this chapter demonstrate the need for transparency by testing the claims made in chapter 3 about real-time transparency.

First, I present results from two previously-published studies where ABOD3 was used; an online experiment using a video recording of a robot and one from direct observation of a robot. Next, findings from an ABOD3-AR study, contacted as part of an art exhibition, are also presented and discussed. Finally, indicative results from survey to measure the effectiveness of ABOD3 as a debugging tool are presented.

Associated Papers

Rotsidis A., Theodorou A., Bryson, J.J., and Wortham R.H., In Prep. Understanding Robot Behaviour through Augmented Reality.

Theodorou A. and Bryson J.J., In Prep. Transparency for Killer Teams.

Wortham, R.H., Theodorou, A. and Bryson, J.J., 2017. Improving robot transparency: Real-time visualisation of robot AI substantially improves understanding in naive observers.*26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, Vol. 2017-January, pp.1424-1431.

1.3.5 Chapter 6: Keep Straight and Carry on

This chapter investigates the use of transparency in a post-incident situation to understand our moral intuitions. The work of this chapter relates to the moral confusion discussed in this chapter and further in chapter 2. Moreover, it acts as a continuation of the studies in chapter 5, by looking into transparency in a scenario where the user is directly effected by the actions of the system.

We used a moral dilemma, a version of the trolley problem built in a Virtual Reality simulator, to compare how we perceive moral judgements and actions taken by humans to ones taken by intelligent systems. Participants took the role of a passenger in an autonomous vehicle (AV) which makes a moral choice: crash into one of two human-looking non-playable characters (NPC). Experimental subjects were exposed to one of three conditions; a human driver, an opaque AV without any post-incident information, and a transparent AV that reported back the characteristics of NPC that influenced its decision-making process, e.g. its demographic background. Human drivers were perceived to be significantly more morally culpable and human-like than self-driving cars. Yet, when the characteristics were revealed to the participants after the incident, the autonomous vehicle was perceived as significantly more mechanical, intentional, and utilitarian. Participants found it harder to forgive the actions taken by a 'mechanical' agent. Most importantly, in contrast to high-profile results in similar studies, participants expressed distress at decisions based on attributes, such as social value. Hence, a need for caution when incorporating supposedly normative data, gathered through the use of text-based crowd-sourced preferences in moral-dilemmas studies, into moral frameworks used in technology.

Associated Papers

Wilson H., Bryson J.J., and Theodorou A. Under Review. Slam the Breaks! Perceptions of Moral Dilemmas in a Virtual Reality Car Simulation.

1.3.6 Chapter 7: Transparency and the Control of AI

This chapter contains a high-level overview of how different AI governance initiatives—and how they *should*—interact with each other. It discusses steps forward for policy-makers to establish a comprehensive AI policy.

1.3.7 Chapter 8: Conclusions

This chapter provides a synopsis of the work presented in this document and makes suggestions for further work. It discusses possible next steps and open research directions related to tools, user studies, and general recommendations made throughout this document. Finally, it reiterates the purpose of the research presented in this dissertation and summarises the main conclusions drawn throughout this dissertation.

1.4 Research Contribution

The main contributions of my research can be summarised as follows:

1. Engineering Contributions & Software Delivered:

   (a) Design recommendations: In chapter 3, I present design considerations developers for AI developers about how-to develop transparent-to-inspection systems.

   (b) ABOD3: Described in chapter 4, it is a novel reactive plan editor and real-time debugger I was the sole developer of. ABOD3 has to conduct the HRI experiments in chapter 5. Finally, ABOD3's code has been used in the development of ABOD3-AR.

   (c) ABOD3-AR: An Android application version of ABOD3, which I provided the idea and helped in the design of, which has been used in an HRI experiment described in chapter 5.

   (d) The Sustainability Game: A novel gamified agent-based model used to communicate implicit knowledge in behaviour economics to its users. It is described in and used by an experiment in chapter 4 and appendix B.

(e) BOD-Unity Game (BUNG): A serious game, described in chapter 5, designed to be used in teaching AI development to final-year undergraduate and master-level students.

(f) UN-POSH: A lightweight reactive planner, first described in chapter 4, made to be used for Unity games; it has been used to develop The Sustainability Game and BUNG and now taught as part of University of Bath's final-year undergraduate and masters-level AI module, Intelligent Cognitive Control Systems.

(g) Moral Dilemma VR: A Virtual Reality simulation, which I provided the idea and design of, of a moral dilemma involving a self-driving car, described in and used by experiments in chapter 6.

2. Human-Robot and Human-Computer Interaction Contributions:

(a) Real-time transparency studies with end users: chapter 5 demonstrates through three distinct experiments how the use of the visualisation software presented in chapter 4, ABOD3 and ABOD3-AR, can help naive users improve the mental models for mechanical-looking agents. I helped in the design of ABOD3-AR study for which I also performed the results analysis. I have also rewrote all the Discussion and interpretation of some results from the first two ABOD3 studies.

(b) Real-time transparency study with developers: After including BUNG and ABOD3 as part of a final-year AI module, indicative results have been gathered and presented in chapter 5. I carried the data collection and analysis for the BUNG/ABOD3 study presented in the same chapter.

(c) Post-incident transparency study: A study, presented in chapter 6, was conducted to further understand the impact of transparency on human moral intuitions. While I did not gather the data, I provided the experimental design and data analysis instructions. Furthermore, the Discussion section is my own.

3. Policy Contributions:

(a) In chapter 2, I argue why intelligent systems should only be treated as artefacts, without any moral status.

(b) Definition of Transparency: I provide a definition for the keyword 'Transparency' when used within the context of Artificial Intelligence. This defini-

tion is used and cited by the upcoming *IEEE Standards Association P7001 Transparency in AI* standard.

(c) While conducting this research, I made direct and indirect contributions to AI-related policy through participation and dissemination of my research output in policy initiatives, meetings, and conferences. chapter 7 contains a summary of my policy recommendations. Appendix A contains a list of policy initiatives and events I participated in.

I use *I* in sections that are the results of my own individual research, reflecting my own insights and intentions. However, multi-disciplinary research requires collaboration with people of various expertise. Therefore, *we* stands for an insight or a result which was produced in a collaboration with somebody else. Moreover, *we* maybe used to refer to *me and the reader*, our *our society*, or even *humanity* at large. Since this voice is not the default one, it is important to clarify my contributions and those of my co-authors, who through their work and our discussions, helped me deliver and shape my final output.

Firstly, this research would not have been conducted without my PhD supervisor, Joanna J. Bryson. She originally envisioned ABOD3's debugging functionality and the Sustainability Game. Attending her class, *Intelligence Cognitive Control Systems*, and some of her numerous talks, helped me formulate the ideas presented in chapter 2. Moreover, she provided guidance and feedback throughout this research programme — including proofreading this dissertation. The *I* mentioned in the above paragraph has been extensively 'altered' over the past years with Bryson's insights and contributions.

Secondly, I would like to acknowledge the valuable contribution made by my colleague Robert H. Wortham to this research. Wortham developed the Instinct Planner and built the R5 robot used throughout the experiments presented in chapter 5. Moreover, Wortham performed the data gathering and analysis for two of the four experiments presented in that chapter. He was the lead author in those two experiments, hence, any data presented are not considered contributions of this dissertation. Wortham, together with Bryson and myself, engaged in lengthy discussions about real-time transparency in intelligent systems.

Furthermore, the application ABODE-AR was developed by Alexandros Rotsidis. I provided the original project idea to Rotisidis. In terms of programming, Wortham helped him understand Instinct, while I provided to Rotisidis the original ABOD3 code and explanations of the debugging methodology. Furthemore, I arranged for him a venue for data gathering, helped him with the experimental design, and explained to him how to analyse his results presented in chapter 5. Using the results he gathered, I

performed the analysis and provided the discussion that appears in chapter 5 and will appear to the associated paper.

Another important person I feel obliged to acknowledge is Holly Wilson, whose work is featured in chapter 6. Wilson performed her research as part of an MSc Research Project under my supervision. I provided the original idea for and guidance throughout the project. The guidance offered included advising her on tools to use, feedback on a self-driving car simulation environment, how to develop the intelligent agents used in the project, and how-to present transparency information. Finally, I helped her design the experiments and make sense of the results. She did all the coding, ran the experiments, and did the final analysis of the data.

Chapter 2

Morality and Intelligence

> "So far, about morals, I know only that what is moral is what you feel good after and what is immoral is what you feel bad after."
>
> Ernest Hemingway, *Death in the Afternoon*

> "Morality is simply the attitude we adopt towards people we personally dislike."
>
> Oscar Wilde, *An Ideal Husband*

2.1 Introduction

The words *ethics* and *morals* are often regarded as interchangeable, when they are not. The word ethics derives from the Greek word $\eta\acute{\theta}o\varsigma$ (ethos), meaning *moral character*. However, *moral* comes from the Latin word *mos*, meaning *custom* or *manner*. Ethics is the *system* of moral values, providing the framework for examining these values. When a society agrees to follow an ethical system, it produces *policy* to ensure the enforcement of those values.

Since the days of the Ancient Greek philosophers, morality has been recognised as unique to humans. We are deceived by our own genes into believing that there is a pro-social objective morality that binds us all (Ruse and Wilson, 1986). We can still maintain that morality is rooted in our unique action-selection system; like all animals, we are not born as blank slates without goals, biases, and means to trigger behaviours. However, our uniqueness is in our ability to produce cultural tools to increase our list of possible behaviours and dramatically enhance the performance of our existing ones. Central to our culture, as we will discuss further, is our language. It allows us to not

only signal intentions, but also facilitate social hierarchies and cooperation (Bryson, 2007).

Now, we are at an unprecedented point in human history, where we create artefacts that can perform action selection based on the decision-making systems found in nature. The same artefacts exploit our language and influence our culture. As demonstrated in this dissertation, we have a dualistic understanding of such artefacts, which leads to the creation of inaccurate mental models and the consequential attribution of anthropomorphic elements to them. This anthropomorphism leads to many believing in an Asimovian future, where all-knowing human-like machines will be part of our societies and be able to follow and be owed legal rights and obligations (Carsten Stahl, 2004; Gunkel, 2012). This over-identification with intelligent systems creates a *moral confusion* about the moral status of these human-made objects into moral subjects (Bryson and Kime, 2011). If we want to produce effective governance mechanisms to regulate the development, deployment, and use of intelligent systems, then first we must agree upon their moral status and capabilities. Otherwise, we run into the risk of allowing fictitious omniscience to distract us from current—and near future—issues.

In this chapter, first I provide clarification of terms, such as *Artificial Intelligence* (AI), and then define the various 'labels' we place on entities to denote their moral status and subsequently any rights and obligations they have or are owed. I discuss what *moral agents* and *moral patients* are and the requirements for an entity to gain either of the two moral statuses. Next, I provide a high-level overview of human morality, by examining our own action-selection mechanism from an evolutionary perspective. I explain how our ability for cultural accumulation, which language both exemplifies and further enables, is the key behind human uniqueness in the animal kingdom. I exploit this discussion to draw parallels to Artificial Intelligence and how we prescribe agency to our robots. Further, I debunk the 'myth' of the possibility of creating an 'all-powerful' Shelleyan[1] machine that will turn against its creators and instead discuss *what is possible*. Finally, I make descriptive and normative arguments against the elevation of intelligent systems into moral subjects.

2.2 Terminology

Some of the greatest challenges of appropriately regulating AI are social rather than technical. Especially as we cannot even agree on a definition of the term, even though there are perfectly well established definitions of both *artificial* and *intelligence*. The

[1] From Mary Shelly author of the classic Frankestein book.

primary problem is that we as humans identify ourselves as intelligent, which certainly is one of our characteristics, but that does not imply that *intelligent* means exclusively 'human-like'. Hence, to ensure a common vocabulary with the reader, widely-used terms are defined for at least the scope of this chapter.

Agency refers to the ability of an entity to effect change; through the usage of sensors, an agent can perceive its environment and internal state and through the use of actuators to change them. Throwing rocks off a cliff in order to satisfy hunger is not optimal—or rational—but it is a sign of agency as the rocks move and the environment changes. We shouldn't confuse agency with *intelligence*. The later implies the selection of the *right action* at the right time; e.g. eating hallumi to satisfy hunger. Thus, *intelligent agents* are a special category of entities characterised by their ability to select rationally and as optimally as possible their actions. If the agent can perform real-time search between all its available actions to select how to react to a situation, then following from Bryson and Winfield (2017) it may also be termed *cognitive*. Hence, *artificial intelligence* refers to the human-made action-selection systems that exhibit signs of intelligence.

Physical agents, such as humans and robots, inhabit, populate, and alter the material world. Within this chapter, the term *robot* denotes the combination of both the software and hardware of an autonomous robotic system (Bryson, 2010b) and always implies as such agency; I do not discuss less intelligent robotics here. Virtual agents can be either an independent piece of software, running autonomously, or part of a larger system. Such agents, as is case of simulations and video games, they may inhabit dynamic, virtual worlds, while other virtual agents may be a simple agent running at the background performing tasks various tasks without any user interaction.

Agents that act individually, instead of as a dependent of a Multi-Agent System (MAS), are called *complete* agents, while agents that have multiple—and often conflicting—goals and mutually-exclusive methods of fulfilling those goals are *complex* agent (Bryson, 2001). These different goals may also be of different priorities. For example, we, humans, prioritise satisfying hunger (and ultimately survival) over entertainment. If we are both very hungry and bored, we will first try to satisfy the former by finding food and eating it.

Agents, such as humans, animals, plants, and micro-organisms, are *natural agents*; their existence is the consequence of biological reproduction. Their traits, abilities, and limitations are due to *biological evolution*. While all natural agents may behave intelligently, not all of them are cognitive. Manufactured agents are *artificial* agents. They are designed and produced to serve a particular purpose, they are in other words,

intentionally made. Their functionalities are set by their creators and their limitations are either set due to resources constraints or again, intentionally. In this chapter, the terms *agent* and *intelligent agent* are used interchangeably to refer specifically to artificial agents. Unless explicitly stated, these agents can be either *virtual* or *physical*.

2.3 Our Morality Spectrum

A special case of agency is *moral agency*. A moral agent is an entity that can be held accountable for its action, that is, it can be held *morally responsible*. Over the last three millennial, there have been arguments on what constitutes an action to carry a *moral value*. Various schools of ethics have been debating on what is the 'correct way' to judge an action. For example, for a utilitarian an action must maximise utility, while for a consequentialist the ends may justify the means.

While there are various schools of normative ethics, most refer back to Aristotelian conditions of moral responsibility and in turn moral agency. Hence, in this section, I discuss the Aristotelian requirements for moral agency and compare them with the ones set by other philosophers. Further, I discuss another category of moral subjects: *moral patients*. I conclude the section by discussing the relationship between morality and law.

2.3.1 From Aristotle to Himma

Aristotle in his well-know *Nicomachean Ethics* linked *autonomy* with moral responsibility. He distinguishes actions of whose origins were 'inside' a person from those whose origins were from the 'outside'. In his model of autonomy, what defines an action as autonomous is seen as its point of origin; it must have an 'immaculate conception'. Actions performed without external influences, i.e. pressure and compulsion, are to be considered voluntary and the only ones worthy of a moral value.

Ultimately, the Aristotelian version of autonomy depends on how we define external influences. Our actions, as I will discuss in detail in the next section, can often be in response to an environmental change and external stimuli but still considered 'our own.' Aristotle took this into account in his discussion on the problems that arise from defining ideas such as 'compulsion' and by estimating the degree of severity of pressure that could make an action not voluntary. Aristotle considers not only an agent's autonomy as an essential requirement, but also for the actor to have knowledge of the action.

Aristotle, building upon Plato's work, is the Western founder of *Virtue Ethics*, one of the major frameworks of normative ethics. For a virtue ethicist, a moral action should be performed only if it will help its actor to gain a *virtue*, a positive trait of character. Only agents which can accept a distinctive range of considerations as reasons for an action that lead to the acquisition of a virtue can possess it. In layman's terms, the agent should not just perform the morally-correct action, but also be able to understand why this is the right action and what the wrong one could be.

Similar to Aristotle, Kant reduced the definite conception of morality to the idea of freedom. Kant argues that full autonomy implies that an agent is the lawmaker of its own rules. Kantian Ethicists refer to such rules as *duties*. A free agent acts with no foreign forces dictating its actions and obligations. Yet, a moral agent can still act contrary to its duty—it can act immorally. In order to assess whether an action is morally permissible or not, an agent must consider if it is consistently universal; all moral agents, given the same context, would act in the same manner.

Kant argues that the *moral law* stems from reason alone, acting as a moral compass for all rational beings. Yet, rational beings can still act irrationally. Their actions might be influenced by basic animalistic needs for survival, the temptations of pleasure, and other emotions. Thus, for Kant the moral law acts as a constraint to natural desires. A moral agent is acting on 'good will' only if its actions are wholly determined by their moral demands. Kant's moral framework is *Deontological*; it is founded on the idea that doing what is right is nothing else other than doing one's duty. These duties are self-determined and exclusive for rational beings, as an agent needs to be able to rationalise and understand moral concepts; what is right and what is wrong. The ability to reason before selecting an action is essential for a Kantian moral agent.

While Aristotle, Kant, and other philosophers *disagree on what constitutes an action as morally right*, *they overlap on which are the principal requirements an agent needs to fulfil to be granted moral agency.* Himma (2009) summarises and generalises these requirements as the following two conditions:

1. The agent has a *capacity to freely choose its actions*; they are not forced by an external agent, e.g. held hostage or under the influence of a demon. This is widely understood as having autonomy or self-control.

2. The agent has the *ability to apply moral concepts and principles*; it has an understanding of what is *right* and *wrong*.

The second requirement is what the different schools of ethics disagree upon; *how to*

judge the moral worth of an action. Albeit the disagreement, there is still an agreement on having *the ability to do so*. This is why children—and even adults with severe cognitive disabilities, such as schizophrenia—are not considered moral agents. We consider the cognitive abilities to be under development or insufficient and the agent unable to make clear rational choices. Societies across the world have different age requirements for adulthood (usually post secondary education) and thus for moral agency.

Himma (2009) argues that there is a precondition implied in these two conditions; the agent has the capacity for *consciousness or at least self-consciousness*. As a result, he discards doings such as 'breathing' and 'waking up' from being considered as (moral) actions. Himma argues that the lack of an intention—of a mental state or desire—behind the doings is what discounts them from being considered as actions. Yet, when we breath the air around us changes. Plants, in the process of photosynthesis, consume carbon dioxide and produce oxygen. These are signs of agency. Implicit awareness is not a requirement for a *doing* to be consider as an *action* or even for an action to have a moral value. I clarify this further, in the next section, where I provide an ontological view of our action-selection system and how our culture and language remains a *sine qua non* for our concept of morality.

2.3.2 Moral Patiency

It is important to make distinction between a moral agent and a *moral patient*. Moral patients are entities, which are owed at least one right or obligation by moral agents. On this definition, all moral agents are also moral patients, but not all moral patients are moral agents. Unlike moral agency, there are no widely-accepted requirements on which entities we can attribute moral patiency. Humans who suffer from a cognitive disability are not expelled from our moral spectrum, but granted moral patiency. Similarly, animals are not considered moral agents, but are widely accepted as moral patients.

Animals are a good example of how moral patiency is spread over a wide area of our moral spectrum. Different animals are treated differently (Franco and Olsson, 2015). Our legal system reflects how we find it harder to kill cats, dogs, and non-human primates than 'game animals', e.g. pheasants, rodents, and rabbits. For example, an EU legislation regulating animal welfare explicitly says: "animals such as dogs and cats should be allowed to be re-homed in families as there is a high level of public concern as to the fate of such animals" (Commission, 2010). In Western countries, we currently consider it moral to raise in captivity specific breeds of animals and to 'domesticate' others. While an extended discussion of animals' mortality is far beyond the scope of this chapter, the message here is how the utility of an animal, alongside with the

culture of the moral agent interacting with it, determine its exact position in the moral spectrum of a society.

In the Greek Orthodox world, icons (paintings of Saints and other deities) and the Evangelion are to be treated with respect. In Islamic countries, a similar respect is provided to copies of the Qur'an. Likewise, flags are to be treated with specific rules—they must never touch the ground and uniformed personnel, e.g. military, have to formally salute them. These objects, without agency, have been granted moral patiency due to the entities they symbolise. An Orthodox believer will not kiss an icon to honour the wood and paint pigments that it is made of. Instead, she will perform this 'ritual' to request the blessing of the deity drawn on the wooden plank. A non-orthodox person may admire such an icon as a piece of art, but will not follow the same social norms as an Orthodox believer.

In short, artefacts do not receive any obligations, but rather indirectly 'benefit' from the duties and protection we show towards the entities they represent. Therefore, objects receive protection and participate in 'rituals' based on their utility and moral values of the agents' interacting with them.

2.3.3 Morality and Law

Each society has different moral values they want to enforce. For example, there is a significant global variation in humans' willingness to and treatment of those who engage in cooperative behaviour. In low-GDP countries, especially where there is evidence of corruption and lack of rule of law, individuals punish those who behave prosocially (Sylwester, Herrmann and Bryson, 2013). When a society decides to protect and promote specific moral values, it does so through its legal framework.

Moral agents' rights and obligations towards a society are enforced through its *legal framework*, by assigning a *legal personhood* to them (Smith, 1928). Legal persons have been granted rights, such as owning property and conducting financial transactions, but they are also bound by the laws of the societies in which they operate. Such laws include taxation to ensure contributions to public goods. If a legal person performs an action forbidden or did not perform one mandated by the law, they can also be held *legally liable* for these actions or inactions. Legal liability can be enforced through the placement of sanctions, such as momentary fines or time spent in prison, when there is a breach of the law.

The law is itself an artefact. It is made by humans, a tool aimed to enable human agents protect themselves and their societies. Yet, when the term 'person' comes to

legal concepts, it is not necessarily confined to and synonymous with humans. Any entity *or even specific objects* can be granted a legal personhood. Lawyers call this a 'legal fiction'. Bodiless entities such as corporations, which are not physical objects, are considered as legal persons (Johnson, 2012; Solaiman, 2017).

The law treats organisations, corporations, and even countries as legal persons. Such legal entities have rights, obligations, and can be held legally liable. Still, the law recognises that an organisation can't act on its own. Each action performed by it is the result of the collective actions and decisions made by the individual agents that are affiliated with it. Individuals, such as shareholders and employees, are the residual claimants of corporate assets and the ones benefited by profits in the form of dividends, capital gains, or residual payments. Similarly, if a corporate acts unlawfully, depending on the country and the severity of the act, individuals may also be held responsibly in addition to the company. Even if individuals are not held directly responsible, any sanctions against a company will result in damages and an indirect punishment of its key stakeholders.

It is not unheard of to provide legal personhoods to non-cognitive objects, such as rare religious idols and even rivers (Solaiman, 2017). There is a lack of uniformity across legal systems—a potential by-product of moral values variation between societies—to dictate which entities can (or should) be considered legal persons. Thus, the extent of what is 'allowed' to be granted by a legal personhood depends upon a given jurisdiction of each country's independent legal system. This also brings an interesting case of misalignment between moral philosophy and the law: for thousands of years the concept of moral agency has been universally accepted as only applicable to humans. We created law to enforce it and now we are developing legal fiction to allow non-natural agents to be granted rights and obligations towards our societies. If we are giving legal personhood to non-human entities, should we give a 'higher' moral status and perhaps even moral responsibility to artificial agents?

In the next section, I will discuss morality from an evolutionary perspective and why we are the only species considered as moral agents. I exploit this discussion in the section following after to answer the above question, by arguing why artificial agents are neither eligible nor should ever be.

2.4 Natural Intelligence

We breathe air to generate energy and stay alive, drink coffee while writing our dissertations to stay awake, and so on. Throughout the day we perform doings without what

Kant calls *reasoning*. 'Deciding' to use our legs to walk to the coffee machine is chosen from a pool of *possible acts*. Emotions, past experiences, and even 'gut feelings' often trigger what we consider *spontaneous* behaviours.

We are complex agents with multiple goals and an even larger pool of possible actions to select from. There could have been numerous combinations of sensory inputs and actuator output, but natural selection trims down this search space. Contrary to Skinner (1935), Gallistel et al. (1991) demonstrates that pigeons can learn to peck for food and flap their wings to escape shock, but not to flap their wings for food or to peck to avoid shock (Bryson, 2001).

Biological evolution provides the architecture to scaffold viable working systems that withstand natural selection. This section discusses this scaffolding from a high-level ontological view, by exploring the relationship between *conscious* and *unconscious* actions. It elaborates on the costs and bottlenecks associated with real-time search. It concludes by arguing why humans are the only moral agents.

2.4.1 Kinds of Minds

Dennett (1996) suggests a high-level ontology of four 'different minds' that make up our own and to some extent other biological creatures'. We are products of biological evolution, making us what Dennett calls 'Darwinian creatures'. When we are born our brains are not a *tabula rasa*, without any drives, behaviours, unable to use our sensors and actuators. We know how to breath, how to 'eat' food, and how to cry. We have inherited hardwired goals and behaviours, that are executed—often in parallel to other actions—to ensure our survival and eventually reproduction.

The Skinnerian Mind allows *ABC learning* (Associationism, Behaviourism, and Connectionism) by testing actions in the external environment. This reinforcement-learning system allows us to make associations and generalisations through trial and error. When we first received building toys, in our early childhood, we probably spent hours and hours learning what to do with them. Now, as adults, we are able to build objects with the same blocks in a matter of minutes. That joy we felt on our first successful construction helped us learn how to properly use these blocks. This positive reinforcement happened so early in our lives, that it makes us assume that this ability was always there. Such prior-learned skills are considered as common sense.

Common-sense knowledge is the collection of behaviours that we can learn, most at such an early enough age that we can't recall their origins (Minsky, 2006). Our Skinnerian mind helps us build this collection, allowing us to create plans, sequences of actions

needed to achieve a specific goal, reducing the need to re-formulate such plans, when we encounter similar scenarios.

Popperian creatures run an internal environment, a simulation, after importing sensory input to preview and select amongst possible actions. Unlike Skinnerian creatures, the hypothesis testing is done internally, allowing hypothesises to die without being executed. Our Popperian Mind is what helps us deliberate options and select sequences of actions to form behaviours. It is one of our *conscious* minds; it allows an online modelling of the expected outcomes across a range of candidate behaviours.

Finally, unique to humans, is the Gregorian mind. Our inner environment is informed by the designed parts of the outer environment. We import cultural tools to facilitate this self-reflection. These imported tools expand our pool of available behaviours, enhance our decision making, and give us an unprecedented power over our environment (Wheeler, 2010).

These 'different minds' allow us to compute which is the right action. We can loosely group Dennet's Kinds of minds into two groups: an implicit minds group, consisting of the Darwinian mind, and an explicit mind group consisting of the Popperian and Gregorian minds. The Skinnerian belongs to both groups; model training is done explicitly, but action repetition can be done implicitly.

2.4.2 Conciousness and Action Selection

Our brain represents about 2% of our body weight, yet accounts for over 20% of our daily consumption of calories (Raichle and Gusnard, 2002). Not every agent can accommodate the energy and time requirements of System 2. Consciousness is, by necessity, adaptive in nature. Unless a new or surprising situation arises, a conscious creature will keep using its fast inexpensive System 1 to perform automatic habitual actions (Isoda and Hikosaka, 2007).

By consciousness, I use the definition placed by Bryson (2012) to exclusively refer to the mechanism that generates awareness of the moment and episodic memory, enabling the learning of new behaviours and explicit decision making. I do not consider any of the other psychological, ethical, or even religion-related phenomena that consciousness has been associated with over the years (Dennett, 2001).

Any cognitive processing can delay taking action, as consciousness introduces a lag and noise into action selection (Norman and Shallice, 1986; Cooper, Shallice and Farringdon, 1995; Schneider and Chein, 2003). This is shown by Libet (1985), who ran a study where

participants had to flex their hand at the wrist while noting the position of a revolving spot. Albeit a simple task, when the experimental subjects were making conscious decisions, there was a delay between 350 and 400 msec behind the onset readiness potentials. This delay may reflect an allocation of time to real-time search for a better solution. However, if the agent acts too slow, another agent may take advantage of a situation before it (Bryson, 2009, 2010a).

If consciousness is so ineffective, why and how do conscious creatures survive natural selection? Intelligent species can only survive natural selection due to the actions they perform (West, Griffin and Gardner, 2007). At the same time natural selection tunes these actions over generations to maximise their effectiveness (Dawkins, 1996). At the cost of performance speed, we can perform real-time search (cognition) to solve problems and take advantage of opportunities that change more rapidly than other ways of performing action selection (Bryson, 2012). Consciousness allows individuals to flexibly adjust their behaviour in previously-unseen dynamic environments (Wortham and Bryson, 2016).

2.4.3 The Power of Language

While biological evolution plays a key role in the capacities of all species, including *Homo sapiens*, it is our cultural evolution—particularly in its extent—that sets us apart. We have an unsurpassed ability to participate in a collective cognition, allowing us to domesticate other animals or even build intelligent artefacts, to extend our physical and cognitive capabilities beyond our originated evolutionary scaffold.

Our unsurpassed competence for cultural accumulation is enabled by our—both written and spoken—language (Bryson, 2007). The ability to communicate is not unique to us; many animals have vocal and other means of signalling each other. However, language is unique to humans. We can communicate not only information about the surrounding world, like animals do, but also gossip. Gossip leads to information exchange about others and eventually the development of social hierarchies and organisation (McAndrew and Milenkovic, 2002). Even if that information comes with a high degree of uncertainty, cooperation can be sustained (Mitchell et al., 2016).

Language also facilitates the creation of fictional constructs, such as religion and names for our tribes. No monkey will call 'the spirits of the forest' to protect it. These constructs provide us a system of guiding beliefs and symbols of in-group identities Ysseldyk, Matheson and Anisman (2010). Migration between groups results in spatial structuring of cultural skill accumulation, as specialised skills and tools become avail-

able to both the migrants and their receiving group (Powell, Shennan and Thomas, 2009). Due to the acquisition and ascription of such memes, alongside with gossip, we are able to cooperate and organise ourselves in far larger and more complex societies than any other animal. Cultural accumulation at scale is the incontrovertible aspect of human uniqueness compared to other biological creatures. This accumulation significantly enhances and extends our consciousness and general cognition capabilities Wheeler (2010).

2.4.4 The Problem of Dithering

Explicit decision making at all times has both a cost and is subject to *dithering*: switching from one goal to the other so rapidly that little or no progress is made in achieving any goals (Humphrys, 1996; Rohlfshagen and Bryson, 2010). Nonetheless, an agent *should be able to* devote sufficient resources to at least achieve survival-related goals. An appropriate balance between the responsiveness and persistence at pursuing the currently selected goal is needed to avert dithering. In order to achieve this balance, we have both built-in individual and external regulation mechanisms to modulate how the agent invests its resources in trying to complete different goals.

In nature, drives and emotions may be seen as a chemically-based latching system, evolved to provide persistence and coherence to the otherwise electrically based action selection, provided by the central nervous system (Bryson and Tanguy, 2010). The hormone and endocrine systems, which underlie drives and emotions, are evolved to provide smooth regulation of behaviours in all animals. They determine the current focus of attention by managing the trade off of losing resources, such as energy and time, when switching between the different goals.

However, such regulation mechanisms focus primarily on helping the agent achieve some of its *own goals*. If multiple complete, complex agents work together, they form a community with both common and individual objectives (unlike a multi-agent system, where there is a common objective for all agents). A common objective can be the acquisition or development and maintenance of *public goods*, resources shared within a single social group. For a society to remain sustainable its agents must spend a sufficient amount of resources at both their own private goals and the common public ones. When group members take advantage of a good shared among the community members, without contributing towards it, they are *free riders*. Communities do not passively accept the free riding of others. Instead, when they have the opportunity to punish free riders, they do so.

Punishment takes the form of an actor paying a cost to reduce a fitness value of the punished target. However, for this to be consider as punishment, the punished agent needs to experience an unpleasant mental state (Himma, 2009). We ordinarily aim to avoid receiving punishment, and in so doing allow the law to influence our decision making. The claim is not that the law 'limits our free will' and that actions taken under the threat of sanctions have no moral value. Taxes are meant to ensure sufficient contributions to public goods and the wellbeing of the community at large. We still have the option to break the law. When people break the law, e.g. tax evaders, they have an understanding that they will be sanctioned if caught and then they make a conscious choice to damage their societies (or at least their governments; the societal damage may be collateral) by acting anti-socially. The law, like all cultural tools, is imported by our Gregorian mind. It helps us focus our decision-making process, by adding biases that influence our action selection to avoid pursuing behaviours that may inflict damage to our societies.

2.4.5 Morality for Humanity

The legal system is a reflection of the moral values a society wants to protect. As a society, we have complete control over it. Creating a new law is not different from software development. Bugs (loopholes) and flaws (exclusion clauses) are bound to happen during the implementation, which is why we try to find them and fix them during internal testing—when policy-makers debate before ratifying a new law. Arguably, in the process we may introduce further bugs and flaws. Once the system (law) is made available, some users may discover new bugs. In Common Law, judges do not just interpret the law, but often fix it through their rulings. As a society grows and evolves, we may update, retract, or replace completely a piece of legislation by voting new lawmakers to represent our views and interests. In short, the law and its foundation —morality—are artefacts.

There is no universal objective morality. We may all have the same kinds of minds, but our Skimmerian mind received different training, our Gregorian imported different tools, and so on. This does not mean that we cannot evaluate the optimality of an action, but that our evaluations will be different if an action holds any moral gravity. At the end, morality is not only subject to cultural variation, but it also provides group-specific moral norms for group members to follow and enhance their in-group identity (Ellemers, Pagliaro and Barreto, 2013). Like language, it influences culture and is being influenced by it.

Animals, which are biological agents like us, are not considered able to satisfy the

requirements for moral agents (Gruen, 2017). If morality is a matter of benevolent inclinations, accepting as good that which is agreeable or useful to ourselves for others, then some animals can be moral agents. For example, a lioness providing her hunt to the whole pride is an act of altruism. Yet, due to human exceptionalism and speciesism, we discard such actions as 'mere instinct' (Johnson, 1983; Gruen, 2017). I argue—as an exceptionalist myself—that as morality is part of culture, therefore, access to the unique-to-humans Gregorian mind is needed to understand moral gravity of an action. This does not exclude animals from performing acts of moral value, but exonerates them from being held morally responsible.

2.5 Artificial Intelligence

Artificial Intelligence is not a newly evolved life form (Brundage and Bryson, 2016), but a field of research aimed at producing products which provide some utility to us. We have the ability to decide their shape, action-selection system, and so on. Sometimes even the act of developing an agent is by itself the purpose of act, e.g. for educational purposes. The same can be said with any other tool, from a scythe made to help us mow to nuclear weapons that provide us a competitive advantage over others. In short, we are not morally or otherwise obliged to develop any artificial agents, but we do so to improve our own performance.

Miller (2015) argues that their human-made nature is the exact reason why there is *no question on what moral status intelligent artefacts deserve*; the only question is what *we may want to assign them*. Yet, not everyone considers AI as *just* an artefact; purposely designed and developed to work as extensions of our own agency. Instead, there have been advocates of granting a moral status to (some) intelligent artefacts (Carsten Stahl, 2004; Gunkel, 2012). Others argue that an elevation of artefacts' moral status does not apply to current specialist systems, but only to future powerful and potentially conscious system (Himma, 2009; Coeckelbergh, 2009).

In this section, I discuss how the bottlenecks of real-time search in Natural Intelligence exist also in Artificial Intelligence. I argue that the development of an 'all-powerful' system is unlikely and even practically unnecessary. Furthermore, I discuss why even a hypothetical system does not fulfil the requirements for moral agency set earlier in this chapter. Finally, I discuss through normative and descriptive arguments that making AI moral agents or patients is not only an intentional and avoidable decision, but also an undesirable one.

2.5.1 The Omniscience of AGI

In section 2.4 I discussed how real-time search costs time and energy, which is why biological agents limit or even avoid it. These two requirements impose limits not only to natural intelligence, but also to Artificial Intelligence (Bryson, 2012, 2018). No agent is able to compute all possible solutions to all problems. The recent progress is largely due to a combination of substantial corporate investments in AI research, improvements in the design of computer hardware, and the availability of larger datasets. Yet, 'there is no such thing as a free lunch'; regardless of how energy efficient and whichever fabrication process we use, our processors still require *some* time to run an algorithm, consume energy, and take space—not only for the processor itself, but also for its cooling system, as heat is a by-product of energy consumption.

Moravec (1988) argues that: 'It is comparatively easy to make computers exhibit adult level performance on intelligence tests or playing checkers, and difficult or impossible to give them the skills of a one-year-old when it comes to perception and mobility'. We already have systems that can outperform humans in 'hard' problems, such as complex games like go and DOTA2 (OpenAI, 2018). While we are making steps at object identification and classification (He et al., 2015), due to the combinatorial complexity of explicit action selection, we are nowhere near *Artificial General intelligence* (AGI). Even if we can build and power a high-performing computing centre at the size of Sweden to run an 'all-powerful' AGI, what will be the actual gain? There is no practical benefit to use the same agent someone develops to play DOTA2 against to also drive our cars, manage our calendars, and do our taxes.

This does not stop AGI advocates from claiming that machine learning techniques, such as reinforcement learning, can lead to 'human-like' AI—some go as far as claiming that such a discovery can happen 'accidentally' (Bostrom, 2014)! In reinforcement learning, we specify a reward function giving us an imperfect control over what the agent should consider as good behaviour and what not to. Our designs can include fail-safe mechanisms and other means to focus learning and action selection in general. In Bryson and Theodorou (2019) we present such a design methodology and ontology to maximise our control and hence safety. There can be no 'accidental' development of a machine that will 'decide to turn us all into paperclips'. However, bad code may still lead to accidents costing human lives. Artificial agents are coded and in the process we, intelligent-systems developers, prescribe our own moral agency to them.

Consider a pet, such as a dog, trained to respond to specific commands. If you shout *sit* and it sits, you reward this behaviour by either petting it or handing a treat to

it. At each re-occurrence of the *sit* command followed by a treat, the action-reward association is enforced. On the sound of *sit* a dog can also be trained to attack humans, roll on the floor, and so on. The behaviours that can be associated to *sit* are only limited by the physical and cognitive constraints of the dog. A dog *does not understand the context behind our words*; it lacks a Gregorian mind. It will never understand it—let alone communicate back using—our language. This does not reduce a dog's ability to react to the specific sounds that make up words. Similarly, the words *sit*, *eat*, *Hitler*, and so on have no real context to an intelligent system. If there is a sensory trigger associated with the detection of that sound or input of text, an action will take place.

An example of this is the chatbot *Tay*, which tweeted pro-Nazi messages as a group of people exploited the lack of appropriate filters to feed it with malicious data (Vincent, 2016). The bot did not actively support Nazism or far-right political views. Instead, as its creators failed to develop the necessary filters. It then acted in the way it was programmed, to maximise interaction by using—what it perceived due to the attack as—commonly-used words and popular phrases. This also demonstrates that there is no need for intelligent agents to understand our culture to exploit it. In fact, search tools, such as Google, already exploit our language by looking at the distribution of keywords in documents. Dating apps do not need to understand relationships or romance to suggest potential matches.

Any training data used to train our models, inevitably contains our biases (Caliskan, Bryson and Narayanan, 2017). This not only weakens the argument for granting moral agency to them, but makes the argument for transparency stronger. Transparency can ensure sufficient auditing of the system, its debugging, and overall help attribute accountability and responsibility (Theodorou, Wortham and Bryson, 2017; Bryson and Winfield, 2017)—this argument is revisited in the next chapter.

2.5.2 Extending of Our Agency

Let us consider the game *catch* for a second. Its roots go back to when dogs were trained to be used by hunters to bring their shot-down prey. Dogs are faster and better at tracking than humans are, thus, their use in hunting. We use pets as an extension of ourselves to improve our abilities, expand our pool of available actions, and increase our performance. Dogs are autonomous intelligent agents, but through reinforcement training we domesticate them.

A bird eats worms to survive, flaps its wings to fly, and procreates for its genes to carry over through its offspring. These are behaviours refined by natural selection.

Its Darwinian Brain drives their execution. There is an unconscious intention behind them; the short-term survival of the individual and long-term survival of species. I argue that an artificial agent does not have this intention; such agents are not bound by natural selection with a need to perform any action. Their deployment and actions *should* serve—ultimately—their owners.

Artificial agents are developed to be used as enhancement of our agency—like a domesticated dog is. During the development and training of artificial agents, we ascribe agency to them. Intelligent agents, as their name suggests, perform action selection; *they decide amongst a human-defined list of possible actions* which one to follow. Artificial agents get sensors/actuators based on a number of criteria, e.g. the utility of the agent, costs and availability of the sensor/actuator, and so on. We have the same control over deciding whether to add an ultrasonic sensor as we have on which bathtub to install in our new bathroom. In both cases, one or more human stakeholders decide to increase the cost and add a certain utility.

In the end, it is our design choices that define their pool of available actions. We prescribe any 'freedom' of choice they have and even our own biases to them. Moreover, like animals, there is no understanding of the moral gravity of its actions. An artificial agent *is intelligent and can act autonomously*, but its actions are context agnostic, as its agency is prescribed. Still, as morality is a human-defined concept, the moral law can be updated to accommodate artificial agents. Next, I will discuss why we should not do that.

2.5.3 Incidents Happen

Even the best trained dog may still ignore its owner and instead start barking at a passing car. Our control over all domestic animals is precarious. When they ignore us, we shout at them, we pull them by the collars, and so on. Our response to their misbehaviour does not aim to punish the animal by attributing responsibility to them. Instead, it aims at teaching them that their actions were wrong—as always, based on our own understanding of what is right and wrong. We do not hold the dog responsible, but rather we try to educate it in order to *fix the bugs in its training*. Dogs, like all animals, are not subject to moral governance and hence not morally accountable for their actions (Himma, 2009; Bryson, 2018). Any moral responsibility is passed to their owner and trainer.

From a functional perspective, the attribution of responsibility is as effective as the punishment that it carries. For an act to be considered as *punishing*, there must be a

successful introduction of an unpleasant mental state and a reduction of fitness value. Himma (2009) writes on punishment "You cannot punish someone who loves marshmallows, as conceptual matter, by giving them marshmallows; if it doesn't hurt, it is not punishment, as a matter of definition—and hurt is something only a conscious being can experience". Our legal systems map different sanctions with different behaviours we, as a society, consider worthy of punishing. We only live a finite amount of time, any time spent in prison reduces our ability to acquire wealth or competitive advantages. Potentially, it leads to long-term disadvantages. Our back-up mechanism, written language, takes a significant amount of time and effort. It is imperfect; we can't just read and 'restore' someone's memories. Two biological agents can never be exactly the same —even monozygotic twins raised in a shared environment (Freund et al., 2013). There is an element of uniqueness that affects both ourselves directly when we loose time under punishment, but also others who benefit from our actions.

Autonomous systems perform various actions on our behalf. Defects, operational mistakes, and even malicious interference with their operation can result in enormous harm to humans, animals, and property. When such incidents happen any errors (e.g. bugs that allowed the malicious hack to take place) need to be addressed—and often to be re-addressed. We need to not only deter occurrences, but also distribute responsibility. In such the correct implementation of post-incident transparency can help with the accurate discovery of a fault and attribution of responsibility to the right parties (Winfield and Jirotka, 2017).

The idea of punishing an intelligent agent that caused an incident is similar to punishing a gun instead of the shooter in a murder. For the agent the 'trigger' is pulled indirectly by the system's stakeholders upon its design, development, and sequential deployment. However, both the agent and the gun are artefacts; they are items developed and owned.

Providing robots with a moral agency requires us to not only re-examine the concept of morality, but also its legal implications. A precedent in favour of granting moral agency non-human entities is the legal personhood granted to organisations. However, there are two distinct differences between granting personhood to a company and to a robot: 1. the effects of punishing a company, and 2. the motivation and benefits from such an act.

In a company, there is a collective responsibility for the actions taken. Any sanctions placed on a company affect directly and indirectly all personnel. If a regulator fines a company, its stakeholders lose money and individual employees may be found also responsible. In heavily regulated industries, companies are required to have compliance

and legal officers. These officers are require to audit and consult other executives. They discourage any wrongdoing, as they have to stand to answer for any violations of the law found in their organisations. Over the years, we have been enforcing these and other mechanisms to discourage—and deal with—any wrongdoings for the protection of our societies.

If we were to held a robot responsible, similarly to how could a company, how can we punish it? A robot can be made to *simulate pain* (Kuehn and Haddadin, 2017). Similar to the language processing, the agent does not have a context, but rather simulates a response as if it is in pain. AI affords the capacity to fake similarity (Wortham and Theodorou, 2017). Emotions simulation can be useful as latching mechanism to avoid dithering (Tanguy, Willis and Bryson, 2003) or even facilitate communication with end users (Collins, Prescott and Mitchinson, 2015). I am not arguing that robots should be made to suffer, but that we have control over the ability to make them simulate such emotions. Like with all other attributes we describe on them, simulating pain is a deliberate design decision. Likewise, the pain-simulation algorithm can be turned off at any moment with the robot remaining operational. Furthermore, putting a machine in jail is inefficient at best. Why do we want to keep a machine 'switched on' to experience time spent in a confined space, wasting electricity and taking space? A machine can be repaired and its memories restored prior to entering jail. At the same time, its owners—who are the main benefactors of its actions—can make replicas of the machine and receive the same benefits. The impracticalities of punishing machines do not stop here. If an infringement is found a company can hire lawyers to challenge the prosecutors and mitigate damages. A robot granted moral agency would either hold assets itself or be given legal representation by the state. This raises another question: will robots—which are property themselves—be granted the right to own assets to pay their lawyers? Such a right, alongside with the general notion of granting them rights, could lead to societal disruption and to an eventually increase of (human) inequality (Bryson, Diamantis and Grant, 2017; Solaiman, 2017).

Corporate personhood allows individuals to take higher risks and participate in larger ventures (Johnson, 2012). It increases the stability, size, and complexity of corporations. Our legal fiction aims at contributing to our socio-economic growth. Granting legal personhood to companies is not a decision taken overnight. It took years to establish how to treat corporations as separate persons and up to this day, we still update our corporate law to better organise and protect our societies. Allowing machines to be granted personhood has arguably no possible socio-economics benefits. On the contrary, it may seriously disrupt our ability to govern, as well as our economy (Bryson, 2018).

2.5.4 Patiency not Agency

So far, I have focused the discussion on robots as moral agents. However, our spectrum of morality is far wider. Entities that are not eligible for moral agency may be given moral patiency. Gunkel (2012) states "The question of moral patiency is, to put it rather schematically, whether and to what extent robots, machines, non-human animals, extraterrestrials, and so on might constitute an other to which or to whom one would have appropriate moral duties and responsibilities."

Coeckelbergh (2010) argues that even we regard robots as 'just' property, we still have *indirect* obligation towards them. He bases his argument on the fact that robots, like all objects, have a value. Coeckelbergh argues that out of respect of their owners, we have indirect obligations towards robots. Moreover, due to their value, robot owners and users will protect robots—even with violence—from others. I agree with Coeckelbergh that robots have the value that we place on them. Their value is not only the costs associated with their development, marketing, and selling, but also of the utility they provide.

However, I disagree with Coeckelbergh's argument that we have any obligation towards the objects. We do not protect items because we feel obliged to do that. Instead, we protect items—and all public goods at large—because we aim to avoid a reduction of our fitness value and the loss of any competitive advantage we had from the ownership of those items. No one enters an unpleasant mental state when she drops her phone due to feeling bad for 'making the phone suffer' but rather to its repairs cost or property loss. This is why the law specifies how any compensation for property damage goes to the owner and not to the property. The owner does not have to repair or replace with a replica the damaged object. Even artefacts such as religious icons, flags, and so on, as discussed in section 2.3.2, are not directly owed any obligations. Any signs of respect, salute, and so on are due to their status as in-group identifiers—in other words, such signs are directed to the group.

Ultimately, morality is an artefact, which varies 'in shape' across different social groups. Like all cultural tools, it is the product of our social evolution. It is used by our Gregorian mind to regulate our decision making. Assigning a moral status to any entity is an intentional action. In this chapter I have shown that moral patiency is not a binary *all of them or none of them* choice, but rather a spectrum. I retain my earlier stated position that robots—like all artefacts—are purposely developed and can only be considered as property. In this sense, they are no different than a computer, a calculator, or even a hammer. Agents are granted an indirect protection, as they

contribute to the fitness value of their owners or even of a society.

2.6 Conclusions

Morality is itself an artefact, a fictitious concept developed by us, varied across societies. Like all other social constructs, we developed it to help our own action-selection system and ensure cooperation within a society. Like all fiction, it provides a common identity. As it is stated in the introduction of this chapter, there are serious concerns regarding the anthropomorphism and misunderstanding of the mechanical context-agnostic nature of robots, which leads to a confusion about the moral status of the robots (Bryson and Kime, 2011; Coeckelbergh, 2010; Gunkel, 2012).

In this chapter, I presented an evaluation of the generalised requirements for moral agency, by tracing them back into Aristotelian and Kantian ethics. Then, I discussed another type of moral entity: moral patients, demonstrating how not only the utility of an entity, but also the cultural background of the moral agents interacting with it determines its moral status. I provided a high-level ontology of the human action-selection system, arguing that morality and law are human-made 'fiction' to help us guide our actions. They are the consequence of our cultural evolution that is enabled and is enhanced by the source of our uniqueness, language. I discussed the limitations and bottlenecks of natural intelligence; dithering and the costs associated with cognition. Then, I explained that AI is actually subject to the same limitations, thus, the idea of AGI or omniscience is impractical. Instead, we should purposely limit the application domain of our systems to ensure their performance, similar to how cognition—and consciousness to an extent—is adaptive in nature.

I argued that even an Asimovian robot does not fit the requirements for moral agency. However, I acknowledged the fact that morality, as it is a human-made construct, can be altered to include machines. Hence, I made further descriptive and normative arguments why such a move is not only avoidable, but also disruptive to our societies. Finally, I discussed moral patiency and demonstrated why *objects are not moral patients*. Instead, in rare cases, such as when they are used as embodied symbols of our fiction, the object 'benefits' from the obligations and rights attributed toward the fiction it represents. It is important to specify that I do not claim that an artificial agent cannot take actions of moral worth. In fact, chapter 6 shows a user study with an agent making life-ending moral choices. Rather I claim, as established in this chapter, that all actions performed by an intelligent system are executed as an extension to their developers and/or owners moral agency.

Ultimately, this chapter aimed to cut through the 'smoke and mirrors' surrounding AI and communicate to its readers the manufacture nature of intelligence systems to motivate the creation of governance mechanisms. AI governance can include not only legislation related to accountability and responsibility, but also standards and guidelines to establish good-design practices, quality assurance procedures, and performance metrics. While I revisit the subject of AI-related policy at a high level in chapter 7, in the following four chapters I focus exclusively in one good practice; the principle of *transparency*. Transparency is defined the next chapter as a mechanism to expose the decision-making system of an agent. It can help us ensure the long-term accountability by providing an audit trail on what contributed to a decision. Moreover, as investigated in chapter 5, transparency can be used by end users to calibrate their mental models and, therefore, improve the safe use of the system. Finally, similar to the aim of this chapter, transparency aims to make the machine nature of all artificial systems explicit.

Chapter 3

Designing Transparent Intelligents

> "A lack of transparency results in distrust and a deep sense of insecurity."
>
> Dalai Lamma

3.1 Introduction

In order to navigate and interact with the world we inevitably construct mental models to understand and predict behaviour, utility, and attribute trust to both other agents and objects (Craik, 1943; Collins and Gentner, 1987; Johnson-Laird, 2010). If these models are incorrect or inadequate, we run the risk of having non-realistic expectations and sequentially placing too much or too little trust in an agent or object.

The black-box nature of intelligent systems, even in relatively simple cases such as context-aware applications, makes interaction limited and often uninformative for the end user (Stumpf et al., 2010). Moreover, limiting interactions may negatively affect the system's performance or even jeopardize the functionality of the system. Consider for example an autonomous system built for providing health-care support to the elderly, who may be afraid of it or simply distrust it, and in the end they may refuse to use it.

In such a scenario human well-being could be compromised, as patients may not get their prescribed medical treatment in time, unless a human overseeing the system detects the lack of interaction (or is contacted by the robot) and intervenes. Conversely, if the human user places too much trust in a robot, it could lead to misuse, over-reliance, and ultimately disuse of the system (Parasuraman and Riley, 1997). In the previous example of a health-care robot, if the robot malfunctions and its patients are unaware of its failure to function, the patients may continue using the robot, risking their health.

To avoid such situations, proper calibration of trust between the human users and / or operators and their robots is critically important, if not essential, in high-risk scenarios, such as the usage of robots in the military or for medical purposes (Groom and Nass, 2007). Calibrating trust occurs when the end-user has a mental model of the system and relies on the system within the system's capabilities and is aware of its limitation (Dzindolet et al., 2003).

We[1] believe that enforcement of transparency, which is defined in this chapter as a mechanism that exposes the decision-making of a system, is not only beneficial for end users, but also for intelligent agents' developers. Real-time debugging of a robot's decision making—its action selection mechanism—could help developers to fix bugs, prevent issues, and explain potential variance in a robot's performance. Despite these possible benefits of transparency in intelligent systems, there are inconsistencies between the definitions of transparency and no clear criteria for a robot to be considered a transparent system.

In this chapter, first we define what mental models are and then discuss how we continuously calibrate them, as we perceive—or receive—new information. Then, we discuss why we do not understand AI and create inaccurate mental models with roots in folk science fiction. As a result of this, we create unrealistic anthropomorphic mental models for AI. In the next section, we elaborate on the dangers of having such inadequate mental models. We propose a revised definition of transparency: a design principle aimed at helping us calibrate our mental models. Finally, we discuss the design decisions a developer needs to consider when designing transparent robotic systems.

3.2 Understanding AI

While navigating the world, we inevitably interact with other agents and non-cognitive objects. Each interaction is grounded by a system of models that we have for the entity (Johnson-Laird, 1983). We use this system of models to attribute trust and expectations to guide our interactions with the world that lies outside us. When it comes to objects, like robots, we have control over their appearance and other elements, we are able to tune their designs to inspire trust, likeability, and other attributes depending on the occasion.

Consequently, understanding how people perceive robots is essential to formulating

[1]This chapter contains text and ideas previously published in: Theodorou, A., Wortham, R.H. and Bryson, J.J., 2017. Designing and implementing transparency for real time inspection of autonomous robots. *Connection Science*, 29(3), pp.230 241.

good design practices and regulations for such artefacts. Most people interact with intelligent systems, such as search engines, daily. Yet, many consider AI to be futuristic and unknown(Weiss et al., 2011). Human-robot interaction combines software and mechanical engineering with behaviour theory from fields of communication, organisational behaviour, and human-computer interaction to investigate how users perceive robots and why so.

In this section, first we define *mental models*, a term that is used throughout this document. Further, we discuss how our mental models are not static, but rather can—and must—be calibrated, when we receive new information. Next, we elaborate on some of the factors that affect our models for AI. Finally, we examine the dangers of having inadequate models.

3.2.1 Mental Models

Mental models have been investigated in a wide variety of phenomena and different cognitive processes have been attributed to them. Craik (1943) first proposed that people reason, in general, by carrying out thought experiments on internal mental models. Gentner and Gentner (1983) and later Collins and Gentner (1987) expand upon this proposition and demonstrate that people use analogies in their cognitive processes. They define mental models as inferential frameworks used to generate hypotheseses on what will happen in real-world contexts. Rouse and Morris (1986) expand upon this proposition and argue that their purpose is to generate descriptions and explanations of an entity's purpose, functionality, and state.

Johnson-Laird (1983) defines mental models as the knowledge structures constructed from sensory input, imagination, or the comprehension of discourse. He argues that we use them to provide semantic information to reason with. In later work, Johnson-Laird (2010) demonstrates that we create a mental model for each distinct prediction we have in a situation. We use these predictions to perform action selection. This claim is similar to what Dennett (1996) calls a *Popperian mind*; one of 'minds', which runs an internal simulation to preview and select the best appropriate action—a further discussion of Dennett's high-level ontology of minds is found in chapter 2. Here, we hypothesise that our 'other minds' construct and update our mental models, which our Popperian mind exploits to influence *action selection*.

While *mental models* –like *consciousness*—is a 'briefcase term', there is consensus that mental models are typically analogous representations used for reasoning. In the context of this dissertation, the term *mental model* refers to the *cognitive structures and*

operations that we create in order to assign narratives to the world, its objects, our fiction, and other agents. Our mental models are *based on our beliefs and not necessarily on facts.*

Shared Mental Models

Cannon-Bowers, Salas and Converse (1993) demonstrate how we create *shared mental models* (SMMs) when interacting with others, especially in team settings. SMMs are a special type of mental models for when team members have elements of their individual mental models in common. They provide information on team members' past and current state, and predicted actions; such information influences our decision-making process, as we adjust our behaviour based on our predictions (Cooke et al., 2003). Thus, SMMs host crucial information for us to adjust our expectations of others, as we predict their performance and utility. Such predictions are essential for us to anticipate a reaction to our actions, whether that is approval or disapproval. In turn, understanding the intentions of others is a fundamental building block of social behaviour.

Calibrating Mental Models

While there is an agreement on the existence of mental models, there is a disagreement on their exact location in our brains. While some (Johnson-Laird, 1983; Wilson and Rutherford, 1989) claim they are part of the working memory, others consider them part of the long-term memory (Craik, 1943; Bainbridge, 1992). More recent findings by Nersessian (2002) argue that mental models exist as knowledge structures in long-term memory, but are used and updated by the mental models formed in working memory. Thus, as we gain new experiences or continue interacting with others, we keep calibrating our mental models.

The calibration of our long-term models is underpinned by the *Predictive Coding* theory (Elias, 1955; Glimcher, 2011). We make predictions of what may occur based on our existing mental models. Feedback received and reasoning aids helps recalibrate our mental models (Geffner, 1992). For example, one may form an expectation that a humanoid robot has sensors where its 'eyes' are. However, upon closer inspection or reading its technical manual, we update the mental model of the robot. The argument that we update our mental models is also supported by the studies conducted as part of this dissertation, where we demonstrate update our mental models, when exposed to additional information participants adjust their perception of embodied intelligent agents.

Our mental models contain the contextual information we have gathered for an entity, whenever that is another human, an animal, or an object. If these models are inadequate by having only partial information or—even worse– incorrect due to misinformation and/or deception, we run into the risks of assigning too much or too little trust to an entity. This assignment of trust happens based on our perceived utility and performance. Jones et al. (2011) demonstrates how the accuracy of mental models affects the probability of our predictions and sequential ability to reason over them. This is essential for optimal engagement with our external environment.

3.2.2 Creating Mental Models for AI

Our ability to create shared mental models has long served us in understanding others and communicating context. Our shared phylogenetic history and cognitive mechanisms, such as motor matching mechanisms, evolved schemata, and empathy for pain from the social cognition domain, allow us to interpret—often in anthropomorphic means—the behaviour of animals (Urquiza-Haas and Kotrschal, 2015).

On the other hand, as discussed in the previous chapter, artificial agents have afforded the capacity to simulate emotions, 'fake' an understanding of our culture, and be designed in any shape or form we desire. We lack adequate priors to allow us to build adequate mental models of these (Wortham and Theodorou, 2017). The *Social Representation Theory* (SRT) by Moscovici (1981) might explain why people have already been forming insufficient and incorrect mental models of AI. According to SRT, our construction of the representation of a phenomenon is collective and results from common cognition. For example, as discussed below influenced by the media.

Media Representation

Höijer (2011) argues that our mental models of intelligent agents are often sculpted by the representation of AI in the media and contemporary science fiction, and are then reinforced by interpersonal communication. Television series, like *Westworld* and *Battlestar Gallactica*, and films, such as *The Terminator* and *Her*, are examples of the media representation of AI. Intelligent agents appear as 'all-powerful' human-like machines, which 'rebel' against us with the intention to enslave, subjugate, or destroy humanity. This repeated narrative fuels our mistrust of AI and autonomous robotics. In other science fiction media, robots appear as the 'good guys'; able to experience love, pain, and other emotions. For example, Commander Data in *Star Trek* often saves—and is saved by—other members of the starship Enterprise. We argue that such narratives are also not beneficial; they fuel our expectations of what an intelligent agent should

be able to do, elevating the artefact to a moral agent, and demonstrate human-level emotional attachment to a man-made object *as acceptable*.

Liang and Lee (2017) shows that media exposure to science fiction alters the perception of intelligent agents regardless of an individual's demographic background. There is research into Hollywoodian representation of AI on elderly peoples' mental models (Sundar, Waddell and Jung, 2016). For example, their anxiety towards real-life robots, and their perceived usefulness, is influenced by the amount of films they have watched with a robot as the leading character, and how human-like these robots are.

Embodied Agents

The physical appearance of a robot influences the mental model we construct about it, which in turn underpins our interactions (Fong, T., Nourbakhsh, I., & Dautenhahn, 2003). For example, a human-like 'face' can make a system likeable and engaging to its users (Koda and Maes, 1996). A study conducted by Goetz, Kiesler and Powers (2003) further support Koda and Maes' findings. In said study, participants systematically preferred to use a particular robot over others, when its design was anthropomorphic with a matched sociability required in those jobs. Otterbacher and Talias (2017) demonstrate how gender-based stereotyping has been observed to extend to human-robot interaction. The physical appearance of a robot, i.e. with male or female characteristics, triggers uncanny reactions in its 'other-gender' observer.

Kiesler and Goetz (2002) demonstrates the exposure of computer parts, such as control boards and wires, alters the perception users have of a robot. Participants had less positive perception of the robot's reliability but had a more positive perception of its power, if mechanical and computer parts were visible. The amount of human-like characteristic the robot has, changes how much anthropomorphising we attribute to it (Koda and Maes, 1996; Kiesler et al., 2008b). Wortham (2018) shows that even trivial changes, such as a small colourful animal-like cover, could alter the perception—and sequentially the mental model– someone has of a robot.

Non-visual cues, such as audio played or words 'spoken' also affect mental models. Lee et al. (2005) demonstrate how the language used by a robot influences the perception of a robot's knowledge, as if it is a person. Their study shows that participants are more likely to associate a robot with knowledge of tourist landmarks in Hong Kong compared to the ones in US, if the said robot is speaking Chinese instead of English, and vice versa when a robot is speaking in English. In both conditions participants anthropomorphised the robot. They did not consider that the robot can have knowledge of or real-time

access to information for both locations.

Deception in Games

In games, we want to achieve a particular behaviour to suit the design goals, by using simple sets of controls, with understandable and predictable effects. Complexity is not only computationally intensive, whereas games AI needs to run in real-time on limited resources, but also can make it more difficult to attain the desired game experience. Often, in games, it is satisfactory if we create an illusion of advanced, complex, and autonomous intelligence. Players, as seen in examples below, create mental models of the agents in the games they are playing and they often attribute far more advanced complex behaviours to them.

The cult-classic game *Pac-Man* is one of the first games with intelligent agents. The agents, in the form of 'Ghosts', have two states: one normal state, when the player is collecting pips, and another one when the player is under the influence of a power-up. In the former state, each of the four ghosts moves in a straight line until it reaches a junction. Once at a junction the ghost selects to either follow the direction of the player or to take a random route. Each of the four ghosts has a different likelihood of doing one or the other. In the second state, when the player is in pursuit of the ghosts, then they simple turn 180 from their current position and move in a similar way, only if at a junction, they decide between moving on the opposite direction of the player or a taking a random route (Millington and Funge, 2009). This simple approach proved to be effective. The AI, to this day, confuses naive observers into believing that far more elaborate decision-making system is in place. Players often report that the ghosts are able to anticipate their movements and act accordingly.

A game praised by AI researchers, press, and players is *FEAR*. FEAR is mainly known to AI researchers for its Goal-Oriented Action Planning; the press, and players remember the game for its coordination between the player's enemies (Orkin, 2006). Enemy agents in FEAR, who are introduced to the player in squads of 4 or 5, are having simple dialogues between themselves. If the player is firing towards the enemies, one of the agents may ask another "What's your status?" and the corresponding enemy will reply back "I am hit." or "I am alright.", reinforcing the illusion that the agents are working together as a human-like squad. If the player successfully kills a number of enemy characters, one of the remaining squad members would shout "I need reinforcements!" As in all shooters, it is likely that as players progress through the level, they will see more enemies. Having recently heard "I need reinforcements!", the player may conclude that the new enemies are the reinforcements coming to help the now dead, previously

encountered squad. In reality the new group of enemies has nothing to do with the previous squad. Yet, this confused both the press and players alike (Orkin, 2015). Using simple audio cues as a means for agents to interact with each other, even if it did not necessarily affect their actions, FEAR managed to promote the perception of a complex intelligence to its players. FEAR serves as an example of a game where simple, easy to implement actions can create the illusion a far more complex behaviour.

3.2.3 Issues

We have seen that unpacking sparency and trust is complex, but can be partly understood by looking at how humans come to understand and subsequently trust one another, and how they overcome evolutionary fears in order to trust other agents, through implicit non-verbal communication. Unacceptable levels of anxiety, fear and mistrust may result in an emotional and cognitive response to reject robots.

Privacy Concerns

We tend to assume functions of an agent's 'eyes' and that it can only sense within the our own spectrum. Yet, the real location and capabilities of its audiovisual sensors can be different. We can take the SoftBank Robotics' social robot, Pepper, which is a humanoid robot [2] as an example. While Pepper does have cameras for surveying the environment, these are not placed where people would assume—what appears to be its 'eyes'. Instead, these cameras are placed in the forehead with microphones on the top of the head. Therefore, as Schafer and Edwards (2017) asks: why give mammalian-looking designs to our robots, when their sensors and actuators are not comparable with ours?

McReynolds et al. (2017) shows how owners of 'smart toys', usually designed and sold for children use, do not realise how their toys actively gather data through audiovisual sensors, e.g. microphones. Even when the robot has no data gathering capacity, its morphology can still be deceptive and, therefore, lead to issues regarding privacy. Kiesler et al. (2008a) conducted a study where individuals interacting with a human-like robot were more likely to choose a healthier snack-bar rather than a candy-bar and report less socially undesirable information than those interacting with a robot more machine-like in appearance. They viewed the prior as being significantly more dominant, trustworthy and sociable. This suggests the presence of a human-like robot may make one feel observed.

The aforementioned illusion of observation is in direct contrast to many smart home

[2] `https://www.ald.softbankrobotics.com/en/robots/pepper`

devices. Intelligent agents acting as 'personal assistants', such as Alexa, are designed to save us time and energy as they handle tasks for us, help us stay connected, and adapt to our personal preferences. Meanwhile, their sensors may harvest video, vocal, and personal preference data. However, unlike humanoid agents, the morphology of such devices do not make there sensors explicit. There are no ears or eyes as clues to the users of their surveillance. Consequently, users may let their guards down. Significant others, data collectors, and hackers can take advantage of this to survey individuals, or to intercept and hijack devices (Batalla, Vasilakos and Gajewski, 2017). Even when data capturing functions are explicit, users may reveal personal information, when they wrongly believe the device is off. Alexa's microphone is always on—recording is initiated with a wake word. Yet unknown to the user, close approximations to the wake word can trigger recording.

Social Engineering

The malicious use of AI technologies resulted in behaviour change at an unparalleled scale, through the dissemination and even in some instances the generation of disinformation; such as propaganda messages and biased media articles. Already, evidence shows that such manipulation altered the outcomes of the UK's EU membership referendum (Howard and Kollanyi, 2016; Bastos and Mercea, 2017), the US presidential election (Howard, Woolley and Calo, 2018), and attempted to disrupt French Elections (Ferrara, 2017). In all three instances, bots used by populist movements disseminated information and engaged in interacting with other users of social media. The aim was manipulation of the public by entrapping them into echo chambers.

It is becoming increasingly important not only to identify and remove disinformation, but also when an interaction—at least in a virtual environment—is with an artefact. While a lengthy conversation could potentially reveal the machine nature of the bot, that takes time and does not significantly reduce the damage already done.

3.3 Defining Transparency

Despite the importance assigned to *transparency* back in 2011 by the *EPSRC Principles of Robotics* (Boden et al., 2011), research into making systems transparent, until the start of the present research project in 2016, was still in its infancy with few publications focused on the need of transparent systems and even fewer have attempted to address this need. In this section, we first provide our own definition on the keyword *transparency*, which has now influenced the definition placed on transparency by the

P7001 Standard on Transparency (P7001, n.d.). Next, we provide commentary on the advantages of adhering to our definition. Finally, we conclude by providing a survey of other definitions placed on the keyword transparency throughout the literature that contributed to our definition.

It should also be noted that the term *transparency* in the distributed systems and human-computer interaction literature, implies that the system has become 'invisible' to the user. Any changes on back-end components, such as the deployment of a new feature, should not be noticeable by users or interfere with other components. In the context of autonomous systems, at least in this document, we will not be using this definition.

3.3.1 Our Definition: Exposing the Decision-making Mechanism

We propose (in Theodorou, Wortham and Bryson, 2017) that *to consider an agent transparent to inspection, its user should have the ability to request accurate interpretations of the agent's status*; i.e. its capabilities, goals, current progress in relation to its goals, its sensory inputs, its reliability, as well as reports of any unexpected events. The information provided by or for the agent should be presented in a human-understandable format. Our definition implies that transparency might better be thought of as a more of *a general characteristic of intelligence*. Our definition, informed by the literature in this section, goes significantly beyond (though by no means deprecates) the requirement of providing access to adequate documentation.

A fully transparent system may imply a mechanism integral to its intelligence for providing information concerning its operation at any specific moment or over a specific period. We can consider two distinct implementations of the transparency mechanism: one for real-time transparency, providing information as the status of the agent changes, and one for post-incident transparent, which deals with information related to a past decision. These implementations are not mutually exclusive; i.e. an intelligent system can provide both. Next, we visit each one of them, discussing their potential uses.

Real-time Transparency

A transparent agent, with an inspectable decision-making mechanism, could also be debugged in a similar manner to the way in which traditional, non-intelligent software is commonly debugged. The developer would be able to see which actions the agent is selecting, why this is happening, and how it moves from one action to the other. This is similar to the way in which popular Integrated Development Environments (IDEs)

provide options to follow different streams of code with debug points. Note that the necessary requirement for human understandability requires tradeoffs in detail, as real-time decision-making events may easily occur far faster than humans can discriminate between stimuli (Pöppel, 1994).

The ideal for the games industry would be if skilled game designers and writers could directly adjust or even create the characters they design (Brom et al., 2006a). Game designers, who may lack technical expertise, require simple interfaces and methodologies to create agents (Orkin, 2006).Yet, even with such software, the designers may have trouble understanding the emergent behaviour of their agents. If the decision-making mechanism reports the execution and status real time, as we have done in the game BOD-UNity Game presented in chapter 5, it allows developers to implicitly capture the reasoning process within the agent that gives rise to its behaviour. This should improve debugging and allow the usage of highly autonomous agents, without the fear of them going "off script". Similarly, in other applications, such as robots, interaction designers could tune the behaviours to maximise both user engagement and the utility of the agent.

In the following chapter, I present two applications we developed, ABOD3 and its mobile-centric version ABOD3-AR, to facilitate real-time transparency for both developers and end users, through visualisation. Both tools have user-customisable interfaces, allowing them to be deployed for both expert and naive users transparency. A non-text base solution was proposed by Wortham and Rogers (2017), who argue in favour of robot vocalisation as an alternative methodology. In their approach, the robot generates audible sentences. A filtering mechanism is used to output only high-level behaviours and avoid overloading its user with information.

Post-incident Transparency

Other than real-time transparency, there is a need for a *post-incident* transparency. Incident investigators, persons or organisations tasked with discovering the root cause of an incident, establish who is responsible, for bug fixing, insurance-claim purposes, or in a court of law. Such investigators gather and analyse evidences from multiple sources, e.g. witnesses, CCTV, interviewing stakeholders, etc.

In aviation, *Flight Data Recorders*, or as commonly referred to *black boxes*, have been installed in planes to record data and assist investigators—similar boxes can be found in other mission-critical equipment. Winfield and Jirotka (2017) propose that intelligent agents should be equipped with a similar 'black box' to record sensor and relevant inter-

nal status data. Access to such information as possible, can help incident investigators to distribute responsibilities and even accountability.

I agree with the integration of such recording boxes to at least agents that operate in heavily regulated industries, e.g. medicine, finance, or directly effect public safety, e.g. self-driving cars. However, alongside with the development and integration of such boxes in intelligent agents, their developers should work towards the development of relevant tools to help investigators understand the data collected. Otherwise, raw data, such as lines in a log file, albeit useful, are an inefficient methodology to debug an agent. Our real-time debugger, ABOD3, allows non-real-time debugging based on logged performance. Finally, there are a number of security and privacy concerns, including data access and handling, which I will discuss in the next section.

3.3.2 Other Definitions

Different ways of understanding transparency can be found in the literature and high-level ethical guidelines produced by nations, research funding bodies, and other organisations. Unlike our definition, the majority of the work presented here considers implementations of transparency that can provide information exclusively for either real-time decisionmaking or for past decisions. Here we review related work that motivated our definition and our research at large.

The EPSRC's Principles of Robotics includes the keyword *transparency* in principle four. Its definition is implied by contrast: "Robots... should not be designed in a deceptive way to exploit vulnerable users; instead their machine nature should be transparent." The EPSRC definition of transparency emphasizes keeping the end-user aware of the manufactured, mechanical, and thus artificial nature of the robot. However, the phrasing used allows us to consider even indirect information, such as online technical documentation, as a sufficient methodology to provide transparency (Bryson, 2012). Such a solution places at least part of the burden of responsibility with the user, which implies that not all users will find the robot transparent. A user would have to find, read, and understand the documentation or other information provided by the manufacturer, which might be opaque for some user groups.

One of the earliest publications to define *transparency* did so in terms of communicating information to the end user, regarding the system's tendency for errors within a given context of data (Dzindolet et al., 2003). While the Dzindolet *et al.* interpretation covers only part of what we think would be desirable in a definition of transparency, the study presents interesting findings concerning the importance of transparent systems. The

study shows that providing extra feedback to users regarding system failures, can help participants place their trust in the system. The users knew that the system was not completely reliable, but they were able to calibrate their trust to the autonomous system in the experiment, as they became aware of when they could rely on it and when not to.

Military usage of robotic systems is becoming increasingly widespread, especially in the form of Unmanned Aerial Vehicles (UAVs). Transparency in combat systems is essential for accountability. Consider the situation where an artificial agent identifies a civilian building as a terrorist hideout and decides to take actions against it. Who is responsible? The robot for being unreliable? Or the user, who placed their trust in the system's sensors and decision-making mechanism? While the Principles are intended to ensure that responsibility falls to humans or their organisations, given that the damage done is irreversible accountability needs to be about more than the apportionment of blame. Where errors occur, they must be addressed, in some cases redressed, and in all cases used to reduce future mishaps. Wang, Jamieson and Hollands (2009) recommend that robots working autonomously to detect and neutralize targets have transparent behaviours, in the sense that their users, who oversee the system, are alerted to contextual factors, e.g. weather conditions, that affect the system's reliability. The overseers should have constant access to measurements of the system's reliability in its current situation and use such metrics to calibrate their trust towards the system.

Transparency is also often linked to *traceability*; the ability to request a record of information (e.g. inputs, outputs, considerations, etc) related to a decision (Bryson and Winfield, 2017; IEEE, 2016). Traceability is particularly important for verification and validation (Fisher, Dennis and Webster, 2013), but also for post-incident transparency that can be used to assist incident investigators (Winfield and Jirotka, 2017). Studies by Kim and Hinds (2006) and Stumpf et al. (2010) focus on providing feedback to users regarding unexpected behaviour of an intelligent agent after a decision was made. In these studies, the user is alerted only when the artefact considers its own behaviour to be abnormal. Kim and Hinds (2006) demonstrates that when increasing autonomy, the importance of transparency is also increased, as control over the environment shifts from the user to the robot. These results are in line with a study conducted by Kahn et al. (2012). These two studies demonstrate that humans are more likely to blame a robot for failures than other manufactured artefacts or even human co-workers.

However, in Kim and Hinds (2006) implementation, the robot alerts the user only when it detects that it behaves in an unexpected way. This solution might be seen as an attempt to 'fix' one black box by adding another, since there is no guarantee that an

agent would recognise its own misbehaviour. In fact, a monitoring system will need to be checking for patterns of abnormal behaviour by the 'primary decision-making system'. Even if the behaviour is detected in time, there is no guarantee that the user can take control in time. In practice it is often easier to recognise than to diagnose (let alone prevent) misbehaviours (Gat, 1992). For example, most contemporary systems that construct models of their environment can recognise an unexpected context—and even express a measure of its unlikelihood—without necessarily knowing what caused the failure of its models to predict its sensor readings. While ideally transparency could be used to enforce persistent real-time guarantees, in practice the implausible capacity to create such a perfect system might render communication to human users unnecessary. Nevertheless, a system of cognizant error detection does afford one concept of AI transparency: providing at least some ability to detect when something has or might go wrong with a system.

Miller (2014) equates transparency to predictability; the possibility to anticipate imminent actions by the autonomous system based on previous experience and current interaction. Miller argues that by providing information related to each decision may lead to information overload, making the system unusable. Instead, a transparent system should be able to provide sufficient information to improve comprehension of the system's actions and, therefore, increase its predictability. We agree with the concern regarding overloading a user with unnecessary low-level information. However, as different stakeholders have different needs, e.g. developers and incident investigators require access to low-level information, we argue that any definition of transparency should take into consideration the existence of multiple stakeholders with different objectives and needs.

Vitale et al. (2018) considered a robot transparent when it was able to communicate to users the privacy policies for data processing and storage. In a study run, the availability of this high-level information did not lead to lead to significant effects on users' privacy, but significantly improved the user experience as users. Albeit an interesting experiment, similar to the EPSRC's Principles of Robotic, the burden of responsibility is shifted to the user to find and read the said policies. Our critique does not by no means deprecate the importance of having all the relevant documentation accessible by users, as not all information may be communicable through more interactive lower-level approaches.

Participants interacting with a robot in a user study to investigate effects of transparency and communication modality on user trust in a human-robot interaction scenario, reported higher trust levels in the constant level of information on why a par-

ticular task is being conducted by a robot Sanders et al. (2014). Other transparency definitions, describe transparency as the communication of information regarding the machine's abilities (Mercado et al., 2016) and capabilities (Wohleber et al., 2017). Roundtree, Goodrich and Adams (2019) provide a meta-definition on their work by describing transparency as the principle of providing information that is easy to use to promote comprehension of shared awareness, intent, roles, interactions, performance, future plans, and reasoning processes. Hellström and Bensch (2018) argues that transparency is similar to *understandability*; the ability to think about a robot and then use concepts to deal adequately with that robot.

Finally, in data-driven systems transparency is often referred to as *explainable AI*, which in turn is related to the concept of *interpretability* (Biran and Cotton, 2017; Anjomshoae et al., 2019). Choo and Liu (2018) defined the interpretability of a deep learning model as identifying features in input layer which are responsible for the prediction result at the output layer. Doshi-Velez and Kim (2017) considered interpretability of machine learning models and proposed a taxonomy of three approaches: application-grounded, which judges explanations based on how much they assist humans in performing a real task; human-grounded, which judges explanations based on human preference or ability to reason about a model from the explanation; and functionally-grounded, which judges explanations without human input, based on some formal proxy for interpretability. Overall, there is an agreement in the literature that interpretability implies understanding through introspection or explanation (Biran and Cotton, 2017). We revisit the discussion about explainable AI and transparency later in this chapter.

3.3.3 Hardware-level transparency

We should not design our robots with the purpose of making them 'likeable' in all situations, but only when this deliberate deception provides contextual information of their functionality or increases the robots' utility. However, at no time should we hide the location of its sensors; for example hide advanced camera sensors by placing unnecessary mammalian-esque 'eyes' on our robots to deceive their users. The locations and capabilities of their sensors should be visible to provide a bare minimum physical layer of transparency.

Schafer and Edwards (2017) argue that we should take cues from current CCTV-related laws and practices, which require signs on CCTV-monitored spaces. Similarly, 'Robot in operation with AV recording' signs should become mandatory for shops, restaurants, and places where robots with data-capturing capabilities are used. Even if this a functional solution, it only solves part of the problem. Driverless cars and robot-delivery

vehicles are 'mobile CCTV' units, akin to the ones employed by law enforcement agencies. Similar to police cars fitted with camera equipment, I propose that there should be a legally-enforced minimum requirement is a set of stickers, each indicating the different type of data the agent can capture. Robots should, through the use of LEDs or otherwise, indicate when they actually record data. Such a system will be like the LEDs used by our computers to alert us of hard drive activity.

Yet, a careful design and usage of signs, stickers, and LEDs can only provide a small degree of 'passive' transparency—there is still a lack of informed consent. We might be aware that a car passing next to us is filming or recording audio, but we get no choice as to whether we are on these recordings (Bloom et al., 2017). There is a larger discussion on how-to secure data gathered, enforce data access control, and even provide procedures in place for bystanders to 'remove their consent' and have their faces (and other identifications) blurred out from any data saved.

3.4 Design Considerations

To date, prominent research in the field of designing transparent systems focuses in presenting transparency only within the context of human-robot collaboration. Thus, it focuses on designing transparent systems able to build trust between the human participants and the robot (Lyons, 2013). It is as important to build trust as it is to enable stakeholders to know when *not* to trust a system. Developers should strive to develop intelligent agents that can efficiently communicate the necessary information to the human user and sequentially allow her to develop a better mental model of the system and its behaviour. In this section, we discuss the various decisions developers may face while designing a transparent system.

3.4.1 Usability

In order to enforce transparency, additional displays or other methods of communication to the end-user must be carefully designed, as they will be integrating potentially complex information. Agent developers need to consider both the actual relevance and level of abstraction of the information they are exposing and how they will present this information.

Relevance of information

Different users may react differently to the information exposed by the robot. Tullio et al. (2009) demonstrate that end-users without a technical background neither un-

derstand nor retain information from technical inputs such as sensors. In contrast, an agent's developer needs access to such information during both development and testing of the robot to effectively calibrate sensors and to fix any issues found. However, within the same study, they demonstrate that users are able to understand at least basic machine-learning concepts, regardless of a non-technical educational and work-history background.

Tullio et al. (2009) establishes a good starting point at understanding what information maybe relevant to the user to help them understand intelligent systems. Nevertheless, further work is needed in other application areas to establish both domain-specific and user-specific trends regarding what information should be considered of importance.

Abstraction of information

Developers of transparent systems need to question not only *what*, but also *how much* information they expose to the user by establishing a level of complexity with which users may interact with the transparency-related information. This is particularly important in multi-robot systems.

Multi-robot systems allow the usage of multiple, usually small robots, where a goal is shared among various robots, each with its own sensory input, reliability and progress towards performing its assigned task for the overall system to complete. Recent developments of nature inspired swarm intelligence allow the usage of large quantities of tiny robots working together in such a multi-robot system (Tan and Zheng, 2013). The military is already considering the development of swarms of autonomous tiny robotic soldiers. Implementing transparency in a such system is no trivial task. The developer must make rational choices about when low or high level information is required to be exposed. By exposing all information at all times, for all types of users, the system may become unusable as the user will be overloaded with information.

We believe that different users will require different levels of information abstraction to avoid information overload. Higher levels of abstraction could concentrate on presenting only an overview of the system. Instead of having the progress of a system towards a goal, by showing the current actions the system is taking in relation to achieving said goal, it could simply present a completion bar. Moreover, in a multi-robot system, lower level information could also include the goal, sensor, goal-process, and overall behaviour of individual agents in a detailed manner. Conversely, a high-level overview could display all robots as one entity, stating averages from each machine. Intelligent agents built with a modular cognitive architecture, such as the Behaviour Oriented Design

(BOD) that I will discuss in chapter 4, could present only high level plan elements if an overview of the system is needed. In the case of an agent designed with BOD, users may prefer to see and become informed about the states of Drives or Competences but not individual Actions. Other users may want to see only parts of the plan in detail and other parts as a high level overview.

A good implementation of transparency should provide the user with the options described above, providing individuals or potential user-groups with both flexible and preset configurations in order to cater for a wide range of potential users' needs. We hypothesize that the level of abstraction an individual needs is dependent on a number of factors:

1. User: We have already discussed the way in which different users tend to react differently to information regarding the current state of a robot. Similarly, we can expect that various users will respond in a similar manner to the various levels of abstraction based on their usage of the system. End-users, especially non-specialists, will prefer a high-level overview of the information available, while we expect developers to expect access to lower level of information.

2. Type of robotic system: As discussed in our examples above, a multi-robot system is most likely to require a higher level of abstraction, to avoid infobesity for the end user. A system with a single agent would require much less abstraction, as less data are displayed to its user.

3. Purpose of the robotic system: The intended purpose of the system should be taken into account when designing a transparent agent. For example, a military robot is much more likely to be used with a professional user in or on the loop and due to its high-risk operation, there is much greater need to display and capture as much information about the agent's behaviour as possible. On the other hand, a robotic receptionist or personal assistant is more likely to be used by non-technical users, who may prefer a simplified overview of the robot's behaviour.

Presentation of information

Developers needs to consider how to present to the user any of the additional information regarding the behaviour of the agent they will expose. Autonomous robotic systems may make many different decisions per second. If the agent is using a reactive plan, such as a POSH plan (Bryson, Caulfield and Drugowitsch, 2005b), the agent may make thousands of call per minute to the different plan elements. Such an amount of information is hard to handle with systems providing only audio output. Still, with sufficient abstraction

and filtering of information, audio is a feasible mean to provide transparency to naive users (Wortham and Rogers, 2017).

Visualizing the information, i.e. by providing a graphical representation of the agent's plan where the different plan elements blink as they are called, should make the system self-explanatory and easy to follow by both experts and less-technical users. Finally, a graph visualization as a means to provide transparency-related information has additional benefits in debugging the application. The developer should be able to focus on a specific element and determine why it has been activated by following a trace of the different plan elements called and viewing the sensory input that triggered them.

3.4.2 Utility of the system

So far in this chapter we have expanded upon the importance of transparency and the design choices regarding the implementation of it. However, we believe the developer also needs to consider whether implementing transparency may actually damage the utility of a system. We argued in Wortham and Theodorou (2017) that in certain applications the utility of an agent may increase with the degree to which it is trusted. Increasing transparency may reduce its utility. This might, for example, have a negative effect for a companion or health-care robot designed to assist children. In such cases, the system is designed without regards for the EPSRC Principles of Robotics, since it is trying to actively exploit the users feelings to increase its utility and performance on its set task.

If we are able to understand the workings of the intelligence, does it inherently appear to become less intelligent and less interesting to interact with? If we consider video games, an application domain where AI is frequently used, transparency is at variance with deception. In games we actively aim to deceive the user, by presenting our agents as far more intelligent than they often are. As discussed earlier, games use audiovisual cues, conceal the decision making mechanisms of their agents, and even use an element of randomness the agents' decision making to deceive the user to increase the illusion of complex AI. Furthermore, in games where immersion and storytelling are fundamental elements of their experiences, we try to present our agents as believable characters, which inhabit their virtual words and the player can interact with. Game developers often script events and actions performed by the agents, reducing the autonomy of the agents they develop, to make sure they will fit their intended role within the game.

Exposing the decision making mechanism of the agents to the player, could not only make the AI look less intelligent or uninteresting to players, but also reveal tricks em-

ployed by the developers to create believable, complex virtual characters. Finally, there is a gameplay consideration that if the player is competing against an agent, it should not know what the agent is planning to do next. Instead, in most games, an important part of the gameplay is the player trying to predict and counter an enemy agent's actions. Should we ignore transparency in video games altogether? Perhaps players who like spoilers or desire to better understand and train in playing a game competitively, could benefit from transparency. Imagine playing against AI and loosing the match. If you could watch a replay of the match, where the enemy agent's decision making mechanism is understandable, you will be able to perform better at the next match. This adds an opportunity for players to understand a game's mechanics and improve their performance, potentially, making the game a more fun experience. Similarly, agents who team-up with the player may use prompts to alert the player of their actions and environmental perception, improving their cooperation and eventually the win rate of the player.

In embodied agents, transparency may have a negative effect for a companion or healthcare robot designed to assist children. In such cases, the system is designed to actively exploit the users' feelings to increase its utility and performance on its set task. In some situations robot transparency may therefore be at odds with utility, and more generally it may be orthogonal rather than beneficial to the successful use of the robot. An example of this type of design decision which affects the system is the physical transparency of the system. The physical appearance of an agent may increase its usability (Fischer, 2011), but also it may conflict with transparency by hiding its mechanical nature. Back in our companionship robot example, a humanoid or animal-like robot may be preferred over an agent where its mechanisms and internals are exposed, revealing its manufactured nature (Goetz, Kiesler and Powers, 2003). Developers should be aware of this trade–off as they design and develop robots, but also aim to achieve the minimum passive-transparency practices established in the previous section.

3.4.3 Security and Privacy

It has become increasingly important that AI algorithms to be robust against external, malicious manipulation (Brundage et al., 2018). For example, a machine vision system in an autonomous weapon can be hacked to target friendly targets instead of hostile ones. An even more likely scenario is the hacking for an autonomous car, potentially leaking private information or even turning it into a weapon. In line with well-established computer security practices; 'security through obscurity is no security', transparency may improve the overall security of a system. Transparency can help us

trace such incidents, even as they occur, as we can have a clear, real-time understanding of the goals and actions of the agent.

However, to implement transparency sensitive data captured by the sensors and regarding the internal state of the robot need to be made retrievable, thus, traceable. Such data are prone to be targets of third-party unauthorised hackers and may even be misused by corporations and governments for user profiling, raising privacy concerns. Developers of robotics systems should cater to address such concerns by not only securing any data collected, but also by providing the users of their systems with a clear overview on which data are collected, how the data are used, and how long its kept.

While it is beyond the scope of this dissertation to argue and propose methods to develop secure systems, in our view, Artificial Intelligence researchers and developers should start thinking not only about improving the performance of their solutions, but also of their security.

3.4.4 Explainable vs Transparent AI

An aspect of transparency is *explainability*. A system is considered to be explainable only if it is possible to discover why it behaves in a certain way. For example, if a robot is asked "Why did you stop?", an explainable system can produce an answer a human-like language explanation "Obstacles detected!", thus, explainability involves being able to describe causality behind a system's actions, at a high level of abstraction.

Transparency also includes the capacity to understand a system without seeking an explanation. For example, the hardware-level transparency discussed in the previous section, a system that displays its present priorities, technical manuals, and open-source code.

3.5 Conclusion

In this chapter, we have reviewed the concept of transparency, both as used in the EPSRC Principles of Robotics, and as used elsewhere in the AI literature prior to placing our own definition; having the ability to request accurate integrations of an agent's status at any point of time. We have determined that the Principle requires the accessibility of an agent's ordinary decision-making, not only in situations of accountability, collaboration, or cognizant error-detection. Artificial intelligence is defined by the fact it is authored, and as such needs never be the kind of mystery evolution provides us.

We believe the implementation and usage of intelligent systems which are fundamentally

transparent can help not only with debugging AI, but also with its public understanding, hopefully removing the potentially-frightening mystery around "why that robot behaves like that". Transparency should also allow a better understanding of an agent's emergent behaviour. Thus, we redefined transparency as an always-available mechanism able to report a system's behaviour, reliability, senses, and goals. Such information should help us understand an autonomous system's behaviour. Further, we suggested the need for a minimum-level transparency at the hardware level.

Futhermore, we discussed the design decisions a developer needs to consider when designing transparent robotic systems. These requirements include not only the application domain of the system, but also the stakeholder that will be using the transparency information—but not necessarily the system. Once these are identified, the developer should consider *what*, *how much*, and *how to present* information.

In the next chapter I present tools, ABOD3 and its mobile-centric version ABOD3-AR, to facilitate real-time transparency for both developers and end users, through visualisation. Both of these applications have been developed in line with the discussion on the design considerations presented in this chapter. In the rest of this dissertation, I investigate how transparency alters our mental models for intelligent systems by providing us with an understanding of their decision-making system.

Chapter 4

Building Human-Centric Transparent AI

> "Programming today is a race between software engineers striving to build bigger and better idiot-proof programs, and the Universe trying to produce bigger and better idiots. So far, the Universe is winning."
>
> Rich Cook, *The Wizardry Compiled*

4.1 Introduction

A myth of AI is that systems should become as intelligent as humans and therefore not require any more training than a human. In reality, very few will want to put as much energy into training an AI system as is required to raise a child, or even to train an intern, apprentice, or graduate student. In Chapter 2 we discussed why constraining learning or planning allows an intelligent agent to *operate more efficiently* by limiting its downtime due to search and dithering due to having multiple conflicting goals. In the previous Chapter, we presented suggested design principles and considerations for intelligent systems, namely, we discussed the importance of transparency. Programming is generally a far more direct, efficient, and accurate way to communicate what is known and knowable about generating appropriate behaviour. However, debugging a complex, modular, real-time system requires more insight than ordinary programming. Further, we may well want to allow non-programmers, e.g. user-experience designers, to set priorities and choose between capacities for their agents once reliable behaviour libraries have been defined (Orkin, 2015). For example, the reactive planning approaches

described in this Chapter offer a sensible means of transparency for either of these two applications: expert debugging or ordinary user understanding. At the highest level, AI safety may be also achieved by maintaining ordinary levels of human accountability through legislation, something I discuss further in Chapter 7, which is necessary even with the use of the 'right' technologies.

First, we [1] describe Behaviour Oriented Design (BOD), an approach to systems engineering real-time AI. BOD, as a cognitive architecture, provides both a development methodology and an ontology for developing intelligent agents. In addition, it provides specifications for action-selection systems. We discuss three such systems: POSH, Instinct, and finally UN-POSH, one of the contributions of this research programme. Next, we present ABOD3, a thick-client application designed with a user-customisable interface and extensibility. ABOD3 implements a novel real-time visualisation methodology, which can be used for both end-user transparency and by developers to debug BOD-compliant plans. We conclude the chapter by presenting ABOD3-AR, an Android Augmented-Reality (AR) application version of ABOD3, developed exclusively for debugging robots using Instinct. ABOD3-AR is designed with an emphasis on resources optimisation to run in embedded hardware.

Finally, we emphasise that this Chapter does not discount the uses of other AI technologies, such as formal methods or machine learning, as means of developing complete complex agents. Our purpose here is to present technologies, developed as part of this research or by the wider research group[2], to facilitate the development of AI systems while maintaining control and the ability to audit them. However, their basic design principles ay well be generalised to other systems. In fact, it provides a discussion on how learning systems can be intergrated and be used as part of UN-POSH in order to achieve high-level transparency.

4.2 Prior Work: Behaviour Oriented Design

It has long been established that the easiest way to tackle very large engineering projects is to decompose the problem wherever possible into subprojects, or *modules* (Bryson, 2000a). A method for designing modular decomposition for a system is to assess what

[1] This Chapter contains text and research previously published in: (1) Bryson, J.J. and Theodorou A., 2019. How Society Can Maintain Human-Centric Artificial Intelligence. In Toivonen-Noro M. I, Saari E. eds.*Human-centered digitalization and services.* (2) Rotsidis A., Theodorou A., and Wortham R.H., 2019. Augmented Reality: Making sense of robots through real-time transparency display. *1st International Workshop on Intelligent User Interfaces for Algorithmic Transparency in Emerging Technologies*, Los Angeles, CA USA.

[2] BOD was developed by Joanna J. Bryson and Instinct by Robert H. Wortham

the system needs to know, and for each aspect of that knowledge, the best way to maintain that knowledge, as well as to exploit it. Here, we describe one approach to systems engineering real-time AI, Behaviour Oriented Design (BOD).

BOD provides an ontology of required knowledge and a convenient representation for expressing timely actions as the basis for modular decomposition for intelligent systems (Bryson, 2001, 2003). It takes inspiration both from the well-established programming paradigm of object-oriented design (ODD) and its associated agile design, extreme programming (Gaudl, Davies and Bryson, 2013a), and an older but still very well-known AI systems-engineering strategy, called *Behaviour-Base AI* (BBAI) (Brooks, 1991b). Behaviour-based design led to the first AI systems capable of moving at animal-like speeds, and many of its innovations are still extremely influential. Its primary contribution was to emphasise design —specifically, modular design.

In this Section, we first provide a brief overview of BBAI, explaining how its similarities and differences from BOD. Then, we present two extant BOD-compliant action-selection systems, POSH and Instinct.

4.2.1 From BBAI to BOD

Prior to the introduction of BBAI, AI developers were trying to model the entire world in a system of logical perfection in order for the agent to select to reach the 'optimal action' (Chapman, 1987). BBAI, by taking cues from philosophy and psychology, aimed at producing *reactive* systems; the agent acts upon changes in the environment below or above a threshold. As Brooks famously claimed: "The world is its own best model" (Brooks, 1991a). Thus, a BBAI developer instead of modelling the environment focuses on:

1. the actions the system is intended to produce, and
2. the minimum, maximally-specialised perception required to generate each action.

BBAI led to the development of the first environment-agnostic systems, capable of moving at animal-like speeds. The Subsumption Architecture by Brooks (1986) emphasises organising pairs of actions and perceptions into modules. Each action is triggered whenever its associated sensor(s) record a value below/above a user-set threshold. If multiple modules could get activated, only the one with the highest pre-selected priority will be executed. Upon each execution, the system goes back into checking perception values to trigger another module —potentially even the same one. Its highly distributed system of inhibition and suppression is also its great weaknesses: developers may not only

find decomposition 'just for simplicity' difficult to achieve, but also coordinating the modules can be proven intractable (Arkin, 1998). While Brooks never managed to fully address this issue, a later revision of the architecture introduced learning, but limited only within sub-modules, as a mean to address the other major criticisms over the lack of memory (Brooks, 1991a).

BOD solves these issues by moving from arbitrating between highly distributed, difficult to conceptualise or design network of dependencies, to hierarchical representations of priorities. The BOD-specified hierarchical constructions express the priorities and goals of their actions and the contexts in which sets of actions may be applicable (Bryson, 2003). Bryson's approach simplifies agents' development, by maintaining a clear and succinct way of representing an agent's action-selection system.

Moreover, it helps AI developers as it provides not only an ontology, an answer on the challenge of 'how to link the different parts together', but also a development methodology; a solution to the 'how do I start building this system'. It includes guidelines for modular decomposition, documentation, refactoring, and code reuse. BOD aims to enforce the good-coding practice 'Don't Repeat Yourself', by splitting the behaviour into multiple modules. Modularisation makes the development of intelligent agents easier and faster. Once a module is written, it can be used by multiple agents —even ones with different goals. Behaviour modules also store their own memories, e.g. sensory experiences, addressing the lack of memory in Brooks' original Subsumption Architecture. Multiple modules grouped together form a *behaviour library*. This 'library' can be hosted on a separate machine, e.g. on the cloud, from the decision-making part —called the *planner*— of the agent. The planner is responsible for exploiting a plan file; stored structures describing the agent's priorities and behaviour. This separation of responsibilities into two major components enforces further code reusability. The same planner, if coded with a generic-written API to connect to a behaviour library, can be deployed in multiple agents, regardless of their goals or even if they are embodied or virtual agents. For example, the Instinct planner, described in Section 4.2.3, has been successfully used in both robots and agent-based modelling (Wortham and Bryson, 2016), while POSH-Sharp has been deployed in a variety of games (Brom et al., 2006b; Gaudl, Davies and Bryson, 2013b).

BOD affords safety and auditing, by exploiting its BBAI-like modular architectures to limit the scope of learning, planning, or any other real-time plasticity to the actions or skills requiring the capacity to accommodate change. Still, even if learning is limited, it doesn't mean that it is removed altogether. In-module memory can be used to keep track the state of the agent, akin to how parameters keep track the state of an ob-

ject in Object-Oriented Programming. For example, Gaudl, Davies and Bryson (2013b) demonstrates a BOD agent playing STARCRAFT. The agent was able to store in memory the status of production, accumulated resources, and other important information. Machine learning can also be used within specialised modules, such as computer vision for accurate object identification and tracking.

This modular architectural design is essential not only for safety, but also simply for computational tractability —learning systems are faster and more likely to succeed if they are conducting their search over relevant possible capacities. In Chapter 2 we discussed how biological agents are the same; evolution has limited organisms' perception and action abilities. To limit the time penalty of real-time search, there are even restrictions in biological agents on which sets of associations, between perceptions and actions, can learn. The relationship between such specialised modules and the higher-level reactive plan is the same as the one between our System 1 and System 2 discussed in Chapter 2. The plan performs the 'heavy lifting' to ensure fast responses, while the specialised modules exploit learning opportunities. Finally, BOD —like Subsumption— allows multiple behaviour modules to work in parallel, if no competition for resources exists.

4.2.2 POSH

POSH planning is an action-selection system introduced by Bryson (2001). It is designed as a reactive planning derivative of BOD to be used in embodied agents. POSH combines faster response times, similar to reactive approaches for BBAI, with goal-directed plans. A POSH plan consists of the following plan elements:

1. Drives Collection (DC): The root node of the plan's hierarchy. It contains a set of Drives and is responsible for giving attention to the highest priority Drive. To allow the agent to shift and focus attention, only one Drive can be active in any given cycle. On each plan cycle, the planner alternates between checking for what is currently the highest level priority that should be active and then progressing work on that priority.

2. Drive (D): Allows for the design and pursuit of a specific behaviour. Each drive maintains its execution state, even when it is not the focus of planner attention. This allows the pseudo-parallelism execution of multiple drives, even within prioritised actions, as well as independently by modules not requiring arbitration. Each drive has its own releaser, one or more Senses, to determine if the drive should be pursued. The Drive execution frequency limits the rate at which the

Drive can be executed.

3. Competence (C): A self-contained basic reactive plan, representing the priorities within a particular plan. Each containing one or more Competence Elements (CE), which also are associated with both a priority relative to the other CEs, and a context which can perceive and report when that element can execute. The highest-priority action that can execute will when the Competence receives attention. They are similar to the Drive Collection, but without any support for concurrency.

4. Action Pattern (AP): Fixed sequences of actions and perceptions used to reduce the design complexity, by determining the execution order in advance. Used to reduce the computational complexity of search within the plan space.

5. Action (A): A 'doing' of the agent, such as the usage of an actuator. Each Action corresponds to a block code in the behaviour library that sets a skill in motion, e.g. turns on a motor.

6. Sense (S): Senses are very much like Actions, they correspond to code in behaviour library. Senses, as their name suggests, provide perception, e.g. sensor readings, or even internal readings, e.g. status of the agents. Senses must return a value which may be used to determine for example whether a Drive or Competence should be released to execute, or even an Action Pattern to be aborted.

POSH makes use of the reactive planning paradigm and only plans locally, which allows for responsive, yet, goal-oriented behaviour, allowing a high degree of autonomy in dynamic environments. Another important feature is the usage of the parallel-rooted hierarchy, which allows for the quasi-parallel pursuit of behaviours and a hierarchical structure to aid the design. Bryson (2000b) argues that the approach of combining a reactive hierarchy not only outperforms fully reactive systems, but also shows how a simplification in the control structure can be achieved using a hierarchical approach.

The enforcement of a modular design and the grouping of all the primitives (Actions and Senses) into a behaviour library are the major strength of POSH. They decouple the plan design from any underlying agent environment-dependent implementation. Once a POSH planner is coded, it can be used in multiple environments and scenarios. Unlike other cognitive architectures, memory and learning are not essential parts of the core system. This reduces the computational resources needed by the agent, thus increasing the overall performance of the system, making it ideal for both games and agent-based modelling, where computational resources are scarce. Once the planner

and the behaviour library are coded, there is little-to-none programming required to create the plan files and tune them.

POSH in Agent-Based Models and Games

BOD agents using a planner such as POSH are asynchronous by design; actions are performed in response to stimuli and not at scheduled intervals. Moreover, actions may block the execution of the next plan cycle, if the agent is performing a lengthy action. Games and simulation environments are stepped; graphics, animations, and agent' decision-making mechanisms need to update at each step.

In a simulation environment aimed to run agent-based models (ABMs) the world updates on a set frequency of 'ticks'. Yet, the POSH action selection is cycle based; as long as primitive action doesn't blocks its calls for any length of time, a POSH planner has a rate of hundreds of cycles per second. The system was originally designed to allow an agent with hierarchical action selection to operate in a fully responsive and reactive manner. Any method calls to behaviours should not block or delay the planner, even if they wait for a protracted action to happen. Instead, where a lengthy action occurs, such as movement, method calls should only initialise or reparameterise the action. The prolonged action is sustained in its behaviour module until its completion or failure. If the external or internal stimuli that prompted the planner to perform an action remains unchanged, then it will instantly perform the same behaviour as in its last cycle. POSH requires to hold in memory the last behaviour performed, keeping track of its state throughout.

Bryson, Caulfield and Drugowitsch (2005a) demonstrate how POSH can be easily adapted to ABMs. Simply, instead of the agent continuously calling the planner to cycle through the plan, the control is passed to the simulation environment of an ABM. The simulation environment, at each step, signals the planner to perform one internal cycle. A new expressed action may not be chosen on every cycle, but the last performed one may recur. The modular design of BOD agents allows an easy integration of a behaviour library with popular ABM environments, such as *NetLogo* and *MASON*, through the usage of APIs.

From a technical point of view, ABMs focus exclusively on the agents. Little computational resources are sacrificed in aspects such as graphics, user interface, or even physics (unless needed for the simulation). This allows hundreds —if not thousands— of agents at once. Video games, like ABMs, provide virtual simulated environments, but focus in the graphical presentation of those virtual worlds —often aiming to provide

photo-realistic graphics. Real-time rendering of graphics and animations is computational expensive and AI developers are usually left with little resources to work with (Millington and Funge, 2009). To counter that, games AI developers often employ tricks and deceptive cues to create an illusion of complexity (Orkin, 2015).

A crucial technical target for games developers is to ensure the rendering of at least 30 frames per second (FPS) to achieve realistic-looking animations. A frame update in a game is equivalent to a 'step' in an ABM. However, unlike ABMs where delays between steps will only prolong the experiment, a delay between frames update might cause distracting stuttering or even nausea. Each in-game animation requires multiple frames —depending on its complexity it could well be hundreds of frames. Hence, each action needs to be synchronised between the action-selection system and the animation controller of the agent.

Prior implementations of POSH in games solved this issue by having a two-way communication between the games engine and the planner; signalling whenever the action was successfully performed or not (Gaudl, Davies and Bryson, 2013a). This implementation requires the state of the planner to be kept and checked at each cycle. If the action has not yet finished, then the cycle might be interrupted. The planner becomes essentially a third-party entity to a game character instead of an intergrated component. This solution works well in games where the planner is running external to the games environment and is connected to it through a memory-manipulation API. Albeit a functional solution, it is not the easiest to implement and a potential delay may be introduced due to the two-way communication.

4.2.3 Instinct

Wortham, Gaudl and Bryson (2016) introduce Instinct as a lightweight alternative to POSH, which incorporates elements from the various variations and modifications of POSH released over the years. The planner was first designed to run on low resources available on the ARDUINO micro-controller system, such as the one used by the R5 robot seen in Figure 4-1. A number of changes have been introduced to increase its execution performance, while maintaining a lower memory footprint to allow the deployment to embedded micro-controller systems.

A major difference between Instinct and POSH is how the two action-selection systems handle Action Patterns. In Instinct, during the execution of an AP, any sensory input will be temporarily ignored. The planner focuses solely at the execution of the AP. Another difference between the two systems is the inclusion of Action Pattern Element

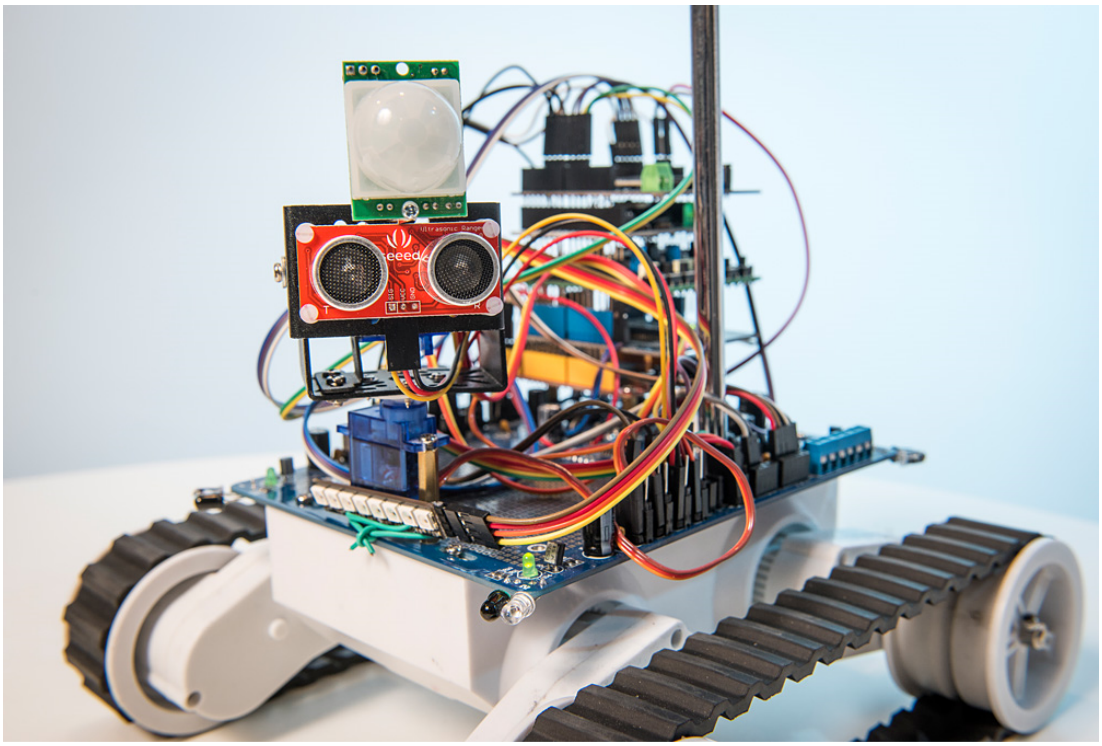

Figure 4-1: The R5 is an ARDUINO-powered low-cost robot developed by Wortham, Gaudl and Bryson (2016). Albeit its limited power, it runs the full version of the Instinct action-selection system.

(APEs). An AP instead of triggering Actions directly, it will trigger APEs. Each AP in an Instinct plan has a set of APEs in a defined order, each responsible to trigger a single Action. APEs ensure the order of Actions in a sequential, developer-defined order.

Instinct incorporates the RAMP model, first developed by Gaudl and Bryson (2014), to allow runtime alteration of drive priority. This, biology-inspired mechanism, allows lower-priority drives to be executed and potentially 'unstick' an agent from being in a loop by performing only high-level actions. A real-life example can be the graduate student writing her dissertation; the task of writing is a fairly high-priority behaviour, but as it gets closer to bed time, the normally lower-priority behaviour 'change to pyjamas' may take control [3].

Another anti-dithering mechanism, called *Flexible Sense Hysteresis* (FHS), has been integrated into the Instinct planner. FHS is based on the flexible-latching mechanism

[3] Author's note: We are aware that a dissertation-writing student would be wearing pyjamas in the first place. No grad students were harmed for this analogy.

first introduced by Rohlfshagen and Bryson (2010). FHS allows noise from the world and sensors to influence drives selection. Finally, Instinct introduces priority group for Competence Elements and two logical operations, i.e. AND and OR, in Competences. Each Competence can group its Competence Elements together based on the priority and specify if in a group all CEs must be executed or whether only one item needs to be for the Competence to be successful.

Overall, Instinct is a low-resources alternative to POSH. It provides significant improvements, compared to prior POSH implementations, in terms of memory and processing management.

4.3 UN-POSH

UNity-POSH (UN-POSH) is a new reactive planner based on Bryson's POSH; it a trimmed-down lightweight version developed to be used exclusively in games. UN-POSH exploits direct access to Unity's animation controller to reduce its processing time, memory footprint, and to allow parallelism. Unlike prior games-centric implementations of POSH, UN-POSH is designed to be run within the game engine as part of an agent instead of as a third-party application with a memory manipulation API, like GameBots (Brom et al., 2006a) and BAWPI (Gaudl and Bryson, 2014; BWAPI: An API for interacting with StarCraft: Broodwar, n.d.).

The UN-POSH planner was first prototyped as part of *The Sustainability Game*, an ecological simulator developed in the modern video games engine Unity. The 'game', discussed in detail in appendix B, is a gamified ABM. It is developed with two-dimensional colourful graphics and other games elements to increase engagement and communicate knowledge to non-expert users. A core requirement of the game, similar to a 'traditional' ABM, is to be able to run hundreds of agents on-screen at once. Unlike ABM-specific environments, such as NetLogo, Unity is a graphics-rich games engine. As previously discussed, game engines are not optimised to run complex intelligent agents. Instead, they dedicate the majority of computational resources available to graphics, physics, and animations rending. Any action-selection systems used must be as lightweight as possible. Later, UN-POSH was imported and polished as part of *Behaviour-oriented design UNity Game (BUNG)*, a shooter game designed to be use by final-year undergraduate and postgraduate students to learn how-to develop BOD agents. In this Section, we[4] talk about the differences—and similarities—between UN-POSH and the

[4]I developed UN-POSH while developing the Sustainability Game. Joanna J. Bryson provided advice and feedback to the project.

more traditional POSH.

4.3.1 The Anatomy of an UN-POSH Agent

The original aim of UN-POSH was to provide a resources-efficient native implementation of POSH in Unity. Its secondary goal is to help students understand Behaviour-Oriented Design and Behaviour-Based AI at large, without having to deal with low-level technical implementation details.

UN-POSH agents are designed to be modular; modularity helps achieve both objectives. It is far easier to implement, test, and long-term maintain smaller modules than a large 'single cut' of code. Splitting the code of a complex complete virtual agent into manageable parts also achieves a level of abstraction. A student learning BOD does not necessarily need to worry about (or at least implement) animations, physics calculations, or even low-level code for sensors and actuators. Instead, the student developer can focus on designing new plans and coding behaviour modules.

A Unity game consists of one or more scene files, and each scene consists of any number of game objects. Every game object is composed from one or more *components*. Modularity is not just a good practice, it is *actively enforced* through the game design and implementation. An agent's components can be categorised into four thematic categories:

1. Physics and Animatronics: All game objects in Unity contain the *Transform* component; a table defining various parameters relating to the object's geometric state (its position, orientation, and size). Agents also have a *Rigidbody* component, which is used by Unity to facilitate physics, collisions detection, and animations. The *Animation Controller* is responsible for manipulating the agent's Rigidbody to play the animation.

2. Internal State: Keeps track of various information regarding the state of the agent, such as: its location, health, stamina. The information stored varies depending on the game.

3. Sensors and Actuators: Consists of all the low-level code that agent uses for input and output, e.g. seeing objects, moving its hand, etc.

4. AI-related components: The system is designed by following the BOD specifications; it consists of the reactive Planner and the Behaviour Library.

These modules can be seen in the architecture diagram presented in Figure 4-2. The di-

agram shows how the UN-POSH Planner is central to the Agent. It is responsible for all the high-level decision making, while allowing low-level specialised modules to perform context-specific decisions. An example of such a module, for navigation, is presented in this section. The Planner interacts with the Behaviour Library; it accesses Senses' values and executes Actions. At each cycle, it reports its execution to a monitoring class responsible for sending a feed of information to a 'Transparency Monitor'.

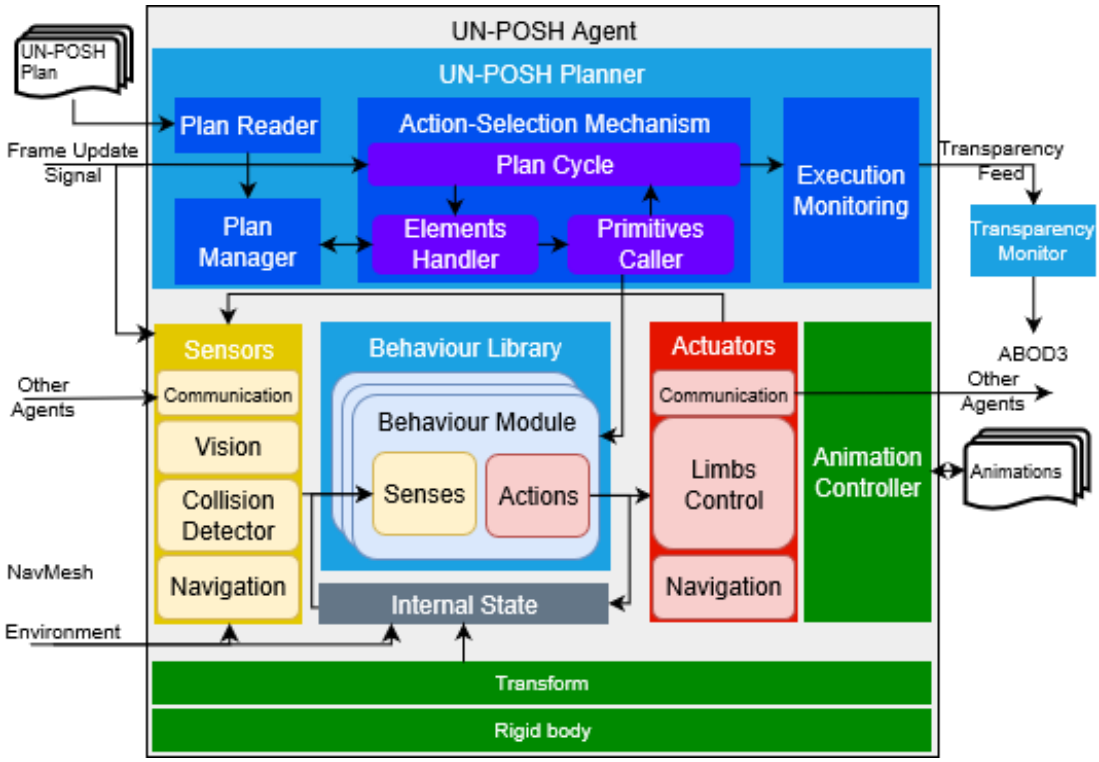

Figure 4-2: The architecture of an UN-POSH agent with sample Sensors and Actuators. The agent consists of the UN-POSH Reactive Planner, which starts a new cycle only when prompted by the game engine. At each cycle, its Callback Monitor module sends a transparency feed to a TCP/IP Client that can connect to ABOD3. The Planner also interacts with the Behaviour Modules in the Behaviour Library. Any Actions triggered alter the environment and/or the Internal State of the agent; such alterations are picked up by the Sensor Model to influence decision making in the next cycle. The Animation Controller monitors changes in the state of the agent and plays relevant animations. The modules are colour coded: Sensors are in yellow, Actuators in red, Internal State is in grey, Physical-body components are in green,, and AI modules are in blue.

Sensors —like the Planner– are updated at the beginning of each frame update and as Actions are executed by the Actuators. Sensors parse updates to Senses in the Behaviour Library and to the Internal State. For example, in BUNG, if an agent is under attack, the attack is perceived by the Sensors, which update the `UnderAttack`

variable in the Internal State of the agent. When an Action is triggered, it will either activate an Actuator or update the Internal State.

Stateless Cycles

A major difference between UN-POSH and POSH/Instinct is that each cycle is independent of the rest. The planner does not keep track of the last executed behaviour. POSH check if the environmental and internal stimuli remains unchanged from the cycle; if so, the planner instantly performs the same behaviour. Instinct requires to keep track of the state, as primitives can return an *in process* status. If an element does so, on the next cycle, the plan hierarchy is traversed again but continues from where it got to in last plan cycle.

UN-POSH stateless approach is inspired by the widely-used *Behaviour Trees*. In Behaviour Trees, the system will traverse down from the root of a tree at every single frame. At each traverse, the system tests each node, from left to right of the tree, to see which is active, rechecking any nodes along the way, until it reaches the currently active node to tick it again. If a change in the environment or in the agent's internal state occurs, it will be detected on the next evaluation and the triggered behaviour will change accordingly. The main disadvantage of this approach is the significant cost of the traversal in larger trees.

The UN-POSH planner at each cycle iterates over the high-level Drives in order of priority. If a Drive is triggered, the planner will go through its children elements until it can no longer trigger a child element. Once a Drive's substructure is traversed, the planner checks and activates any other Drives of the same priority as of the one just executed. However, albeit its lack of memory, it does not have the same computational cost as Behaviour Trees do. UN-POSH by triggering one Drive's subtree at a time remains efficient.

Finally, the UN-POSH planner incorporates the solution suggested by Bryson, Caulfield and Drugowitsch (2005a) for ABMs, with the game engine signalling the planner when to execute a new plan cycle. The new frame will not be rendered until the cycle is complete. The planner, to avoid delaying the frame refresh, does not wait for a behaviour to be resolved. If an Action is initiated, it is instantly considered 'executed' by the planner. Any protracted Actions are sustained by their behaviour modules. There is no waiting for an action to complete itself.

The Role of the Animation Controller

Unity's Animation Controller (AC) is a collection state machines that determines which animations are currently being played to ensure its successful completion and blends between animations seamlessly (Animator Controller, 2018). Otherwise, the activation of different Action could interrupt an animation and sequentially break the game's immersion. The AC works independent to the whole action-selection system. Instead of waiting for direct calls to play an animation, it monitors the state of the character and triggers relevant animations automatically. If for example a character is moving, the animation controller is picking the change of location and plays the 'Movement' animation. This modular approach for total separation of animation handling from action selection is common in the gaming industry; it facilitates developers of different expertise to work in their respective parts independently.

An Action may take multiple frames to complete. Due to the lack of an 'in process' report mechanism and a state memory, as explained above, the same Action may be triggered repetitively. In such cases, as the state of the agent does not change, for example an agent in motion remains in motion, the AC continues playing its current animation. If the same Action is triggered again, it will start playing in a loop. In the eyes of a player/observer, the agent performs the behaviour uninterrupted. There are no visual cues to indicate the agent actually decides at each frame to continue its current behaviour.

Let us consider an example with the Action `MoveToNextNode`, its code is seen in Listing 4.1. `MoveToNextNode` is using a Unity's `NavAgent` to move the agent from node A (its current location) to node B of a pre-generated navigation-mesh path. Depending the distance between the two points, this type of movement may take thousands of frames to complete. At each frame, until the agent reaches its destination, unless a higher priority Drive gets active, the planner will keep triggering the `MoveToNextNode` Action. When so, the Animation Controller will keep the 'Moving' animation in a loop, until at least the character reaches node B. If upon reaching B, if `MoveToNextNode` is initiated again and the agent starts moving towards node C, the AC will start playing the same animation.

```
1 public void MoveToNextNode()
2 {
3 ___if (NavAgent.pathGenerated.Count > 0)
4 ___{
5 ___ ___MoveTowards(NavAgent.pathGenerated[0]);
6 ___}
7 }
```

Listing 4.1: The Action `MoveToNextNode` is part of the Navigation behaviour module. Upon its activation the agent moves from its current location to the next one a pre-generated navigation-mesh path.

UN-POSH allows animations that are not mutually exclusive, i.e. actions using the same limbs, to take place at the same time. This facilitates parallelism. An agent can walk and move its head or shoot a gun. While these actions are triggered sequentially by the planner, they will all be executed at the same frame and appear in parallel to the viewer.

Navigation: A Specialised Behaviour Module

Navigating the world is not an easy task. It involves a combination of real-time sensing and acting. Intelligent agents, e.g. the R5 robot, can 'navigate' the world by avoiding obstacles. Still, unless memory is used, there is no contextual information about the locations an agent is in. In games, we want to achieve a particular behaviour to suit the design goals, by using simple sets of controls, with as understandable and predictable effects as possible. Complexity is not only computationally intensive, when games AI needs to run in real-time on limited resources, but also can make it more difficult to attain the desirable game experience. Often, it is satisfactory if we create an illusion of a highly-advanced intelligence (Millington and Funge, 2009). Thus, it is a common practice for agents to have additional information about the world, e.g. locations of objects. This addition information is used for pathfinding algorithms; algorithms used to find and move an agent, by using the shortest route, between two points.

In the previous section, we talked how BOD agents can have specialised Behaviour Modules, like *Navigation* is for an UN-POSH agent. More specifically, UN-POSH uses Unity's built-in pathfinding solution, the *Navmesh Agent*. A Navigation Mesh (navmesh) is a widely-adopted method to facilitate pathfinding in games (Millington and Funge, 2009). A navmesh is a collection of two-dimensional convex polygons. Adjacent polygons are connected to each other in a graph. It defines which areas of an environment are traversable and at what cost by agents. Due to its graph structure, pathfinding between polygons in the mesh can be done with any graph-search algorithm, such as A* that the Navmesh Agent uses. In UN-POSH, the Navmesh Agent is used through a Wrapper-Decorator class, called *NavmeshController*, to add capabilities, such as memory.

In short, the NavmeshController provides low-level decision making, answering the ques-

tion "how can we get there". Still, the Planner decides not only "if we should there", but also "where is".

Transparency Enhancements

The UN-POSH planner can report its activity as it runs, by means of callback functions, to a monitor class, called *Transparency Monitor*. The Monitor is external to the agent, but is included in both implementations of UN-POSH, and can be used by multiple agents simultaneously. It writes textual data to a TCP/IP stream over network, which can be picked up by a TCP/IP Server to write logs or in case of ABOD3 to facilitate the testing and debugging of the planner.

4.3.2 Drive Elements

UN-POSH introduces a new plan element, *Drive Elements* (DEs) to increase Drives' abstraction level by treating them similarly to how *Selector* nodes are in Behaviour Trees. When a Drive gets activated, instead of calling a Competence or an Action Pattern (or an Action in case of Instinct), it calls one or more Drive Elements.

In the original POSH and Instinct, a Drive only get activated if all of the Sense values assigned to its releaser meet a user-defined threshold. This approach is essentially an 'AND' configuration for all the Senses within a Drive's releaser. In UN-POSH, Drives do not necessarily need to have a releaser as their children elements, DEs, have their own. Each Drive Element can have the same releaser as other DEs in the same Drive. This allows 'OR', 'XOR', and 'NAND' configurations for releases, as DEs trigger the same element.

The inclusion of DEs allows Drives to be responsible for a higher level of behaviour, facilitating a 'partial' context-specific execution of a behaviour. For example, in The Sustainability Game, described below, agents have a Drive `D-Survive`, seen in Figure 4-3. The Drive `D-Survive`, like our own Darwinian Mind, is triggered whenever the agent's 'life' is at stake. The two Drive Elements represent the two difference scenarios that the agent's 'Darwinian Mind' needs to take control, when it is running out of food to satisfy hunger or it is night and the agent needs to find shelter to protect itself from predators. In POSH and Instinct these two Drive Elements would had been two separate Drives. The shift of behaviours one level down to reduce the number of high-level nodes also makes UN-POSH plans behave more like Behaviour Trees.

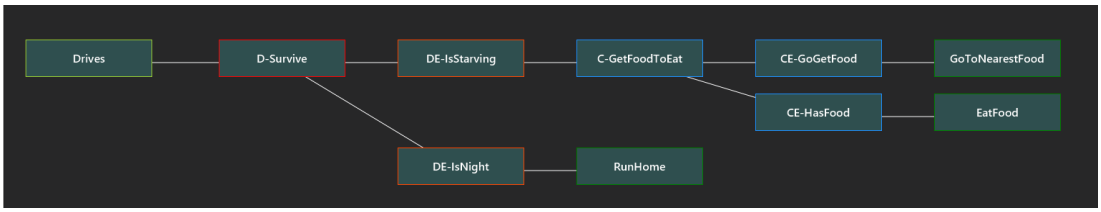

Figure 4-3: A small UN-POSH plan, used in The Sustainability Game, showing Drive `D-Survive` and its subtree.

4.3.3 Use Case: The Sustainability Game

The Sustainability Game is a serious game, developed in the popular game engine *Unity*, designed based on ecological modelling and scientific theory (Theodorou, Bandt-Law and Bryson, 2019). The game has two distinct goals: (1) communicate behavioural economics principles to naïve users and (2) display the measured impact of the player's different investment strategies on the population and individual agents. Both of these goals are tested in a user study found in appendix B.

Gameplay Overview

A society of agents, called *Spiriduși*, populate a fictional two-dimensional world (see Figure 4-4). The agents compose a collective agency; they must invest some resources in their own survival but can also invest in communal goods: bridges and houses. The key gameplay mechanic is that the player selects the percentage of time the agents spend per day on food gathering and consumption, reproduction, building houses for their families, and on benefiting the entire society by building bridges.

The question of where and how much to invest one's resources is complex; there may be multiple viable solutions. Harvested food (apples, grown in two forests) becomes a private good. When an agent eats its stamina level (which normally decreases as time passes) goes up. As a Spiriduș' stamina changes, its colour switches to indicate its status. If a Spiriduș is turns red, signifying a critically low stamina level, then it will stop whatever it is doing and try to find food, regardless of user input. If food is not found within the next moments, it will starve to death. Similarly, the agent will stop its current actions to find shelter from predators at night time.

Technical Details

A technical concern we had was the amount of computational resources needed to run hundreds of agents, each with its own decision-making system in Unity at once. BOD

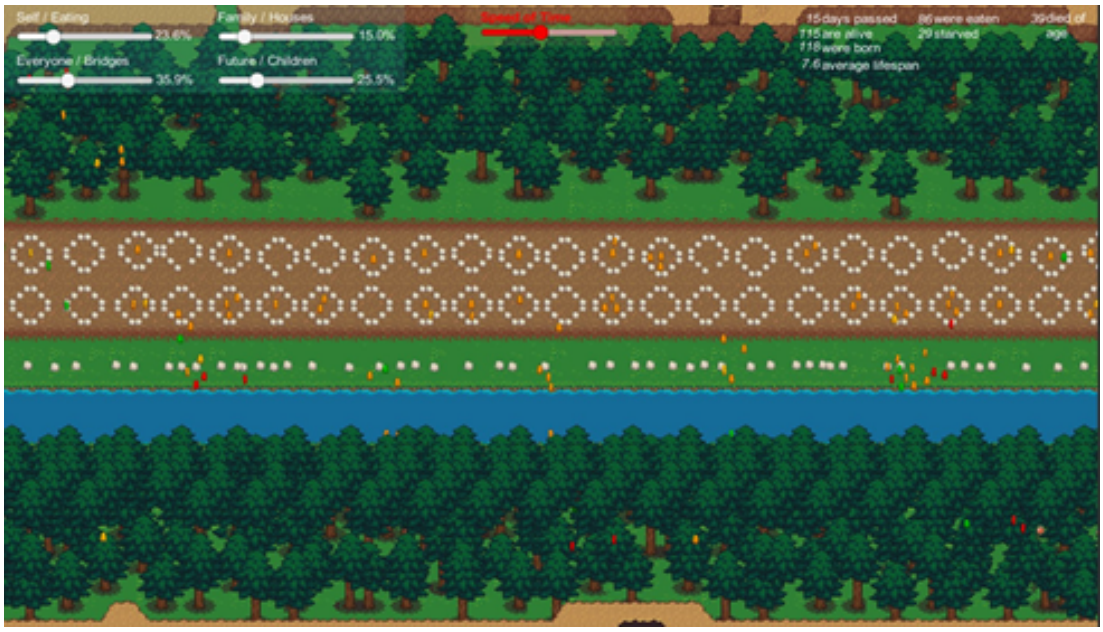

Figure 4-4: Screenshot of game (top-down). Player allocates time spent on given tasks using the sliders in the top left-hand corner. Clockwise: (1) eating; (2) houses; public good (bridge); (4) procreation.

was picked, as it is a lightweight cognitive architecture requiring little computational resources and specifies a modular robust methodology at developing intelligent agents. More specifically, during the development of this game, the first version of the UN-POSH action-selection system was implemented.

UN-POSH facilitated the use of a plan, seen in Figure 4-5, of variable priorities and different levels of abstractions. Each of the four possible behaviours agent could spend its day on (eating, reproduction, building homes, and building bridges) has its own high-level Drive element and relevant subtree consisting of related Competence, Action Patterns, and primitive Actions. All four Drives have the same priority and will only be triggered if the player dedicated sufficient time to the behaviour they facilitate. A fifth Drive, `D-Survive`, was coded to simulate the Darwinian Mind the agents have. The `D-Survive` has the highest priority of all the drives and will be automatically triggered if it is night time or the agent is about to die due to starvation.

Following BOD, each behaviour, e.g. gathering and eating, was coded and tested before work on the next one started. ABOD3, a real-time debugging tool presented next in this chapter, was used for testing. The debugging software allowed real-time visualisation of the agent's decision-making system by presenting a tree-like graph of the plan.

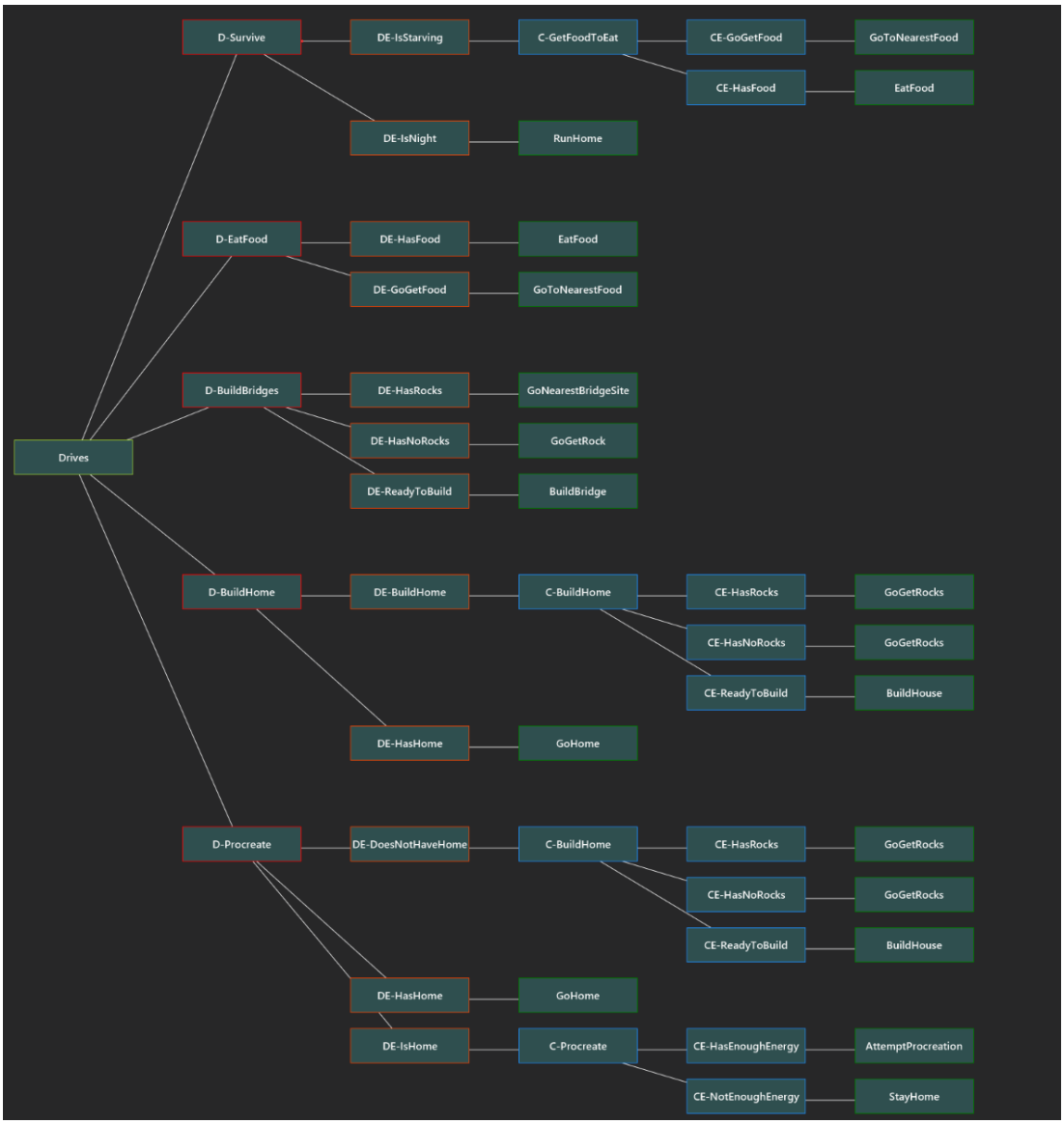

Figure 4-5: The UN-POSH plan used by the agents in The Sustainability Game. The `D-Survive` can be activated based on input from the environment or the internal state of the agent. The rest of the Drives activate only if the user allocated sufficient time to their corresponding behaviours.

This enabled me to see if the correct behaviours were triggered by planne and, hence, categorised bugs to either decision-making bugs or in the underlying behavioural code.

4.3.4 Conclusions and Other Related Work

The UN-POSH planner is a re-engineering of Bryson's original POSH planner designed explicitly for deployment in modern real-time game engines. It combines practices adopted by the games community, more specifically by Behaviour Trees, to help developers understand and work with UN-POSH. By using a simplified lean coding style and efficient use of Unity built-in components engine, it allows the development in a variety of games, from shooters to strategy games, while remaining resources efficient. The transparency capabilities provide the necessary infrastructure to deliver real-time debugging with the ABOD3 software. The importance of transparency was discussed in more detail in the previous Chapter, while the next Chapter contains preliminary results on how UN-POSH with ABOD3 can be used to teach some paradigms of AI thanks to the real-time transparency functionality. UN-POSH has also been used in the development of the Sustainability Game, introduced in this section and described in appendix B.

4.4 ABOD3

In this Section, we[5] present ABOD3, a real-time visualisation system and debugger for BOD-based agents. The system, ABOD3 is based on, but a substantial revision and extension of ABODE (A BOD Environment, originally built by Steve Gray and Simon Jones, Brom et al., 2006a). ABOD3 directly reads and visualises POSH, Instinct, and UN-POSH plans. The biggest extension of ABODE is that ABOD3 enables debugging by providing real-time visualisations of the prioritises of an intelligent system using any POSH-like action-selection mechanism. In addition, it reads log files containing the real-time transparency data emanating from the Instinct Planner, in order to provide a real-time graphical display of plan execution. The ABOD3 is also able to display a video. In this way it is possible to debug either in real time or by using the recorded logs. This provides a new level of transparency for human-like AI.

[5] Joanna J. Bryson had the original idea of ABOD3's debugging functionality. I am the sole developer on the project and came up with the UI and implementation details. Robert H. Wortham provided valuable feedback and beta tested the software.

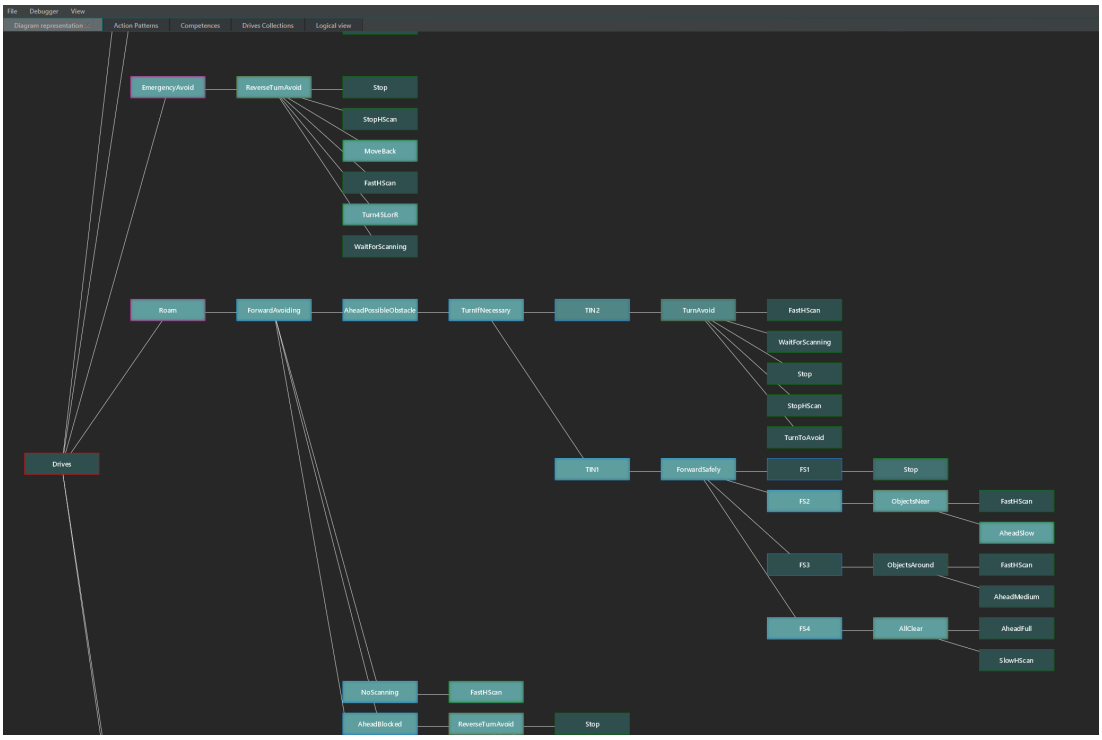

Figure 4-6: The ABOD3 Graphical Transparency Tool displaying an Instinct plan in debugging mode. The highlighted elements are the ones recently called by the planner. The intensity of the glow indicates the number of calls coherent with their recency.

4.4.1 Prototyping

Meetings with major stakeholders (potential users, and developers of prior versions of ABODE) helped me establish basic functional and non-functional requirements before development. Early on it was clear that the new software had to:

1. Run on multiple consumer-oriented operating systems.
2. Provide a customisable User Interface (UI), enabling the application to be deployed by both developers of variable experience and end users.
3. Use a tree-like directed graph to visualise the plan.

The first requirement instantly ruled out a number of programming languages and frameworks, such as C# and the Windows Forms Platform. Java was selected as the development language, as it facilitates a platform-agnostic deployment for the final deliverable. Previous versions of ABODE have been developed in Java, with the AWT and Swing Graphical User Interface (GUI) toolkit. Despite the availability of re-usable code, I decided to rewrite the software from scratch. Taking advantage of starting with

a clean slate, I used the modern JavaFX GUI toolkit.

The editor was developed with an extreme-programming approach; a functional interactive prototype, seen in Figure 4-7, was first developed. The aim of building a prototype was to test the tree-like visualisation of POSH and Instinct plans. After informal feedback gathering, the editor switched to its current high-contrast dark theme and the default tree orientation moved from horizontal to vertical.

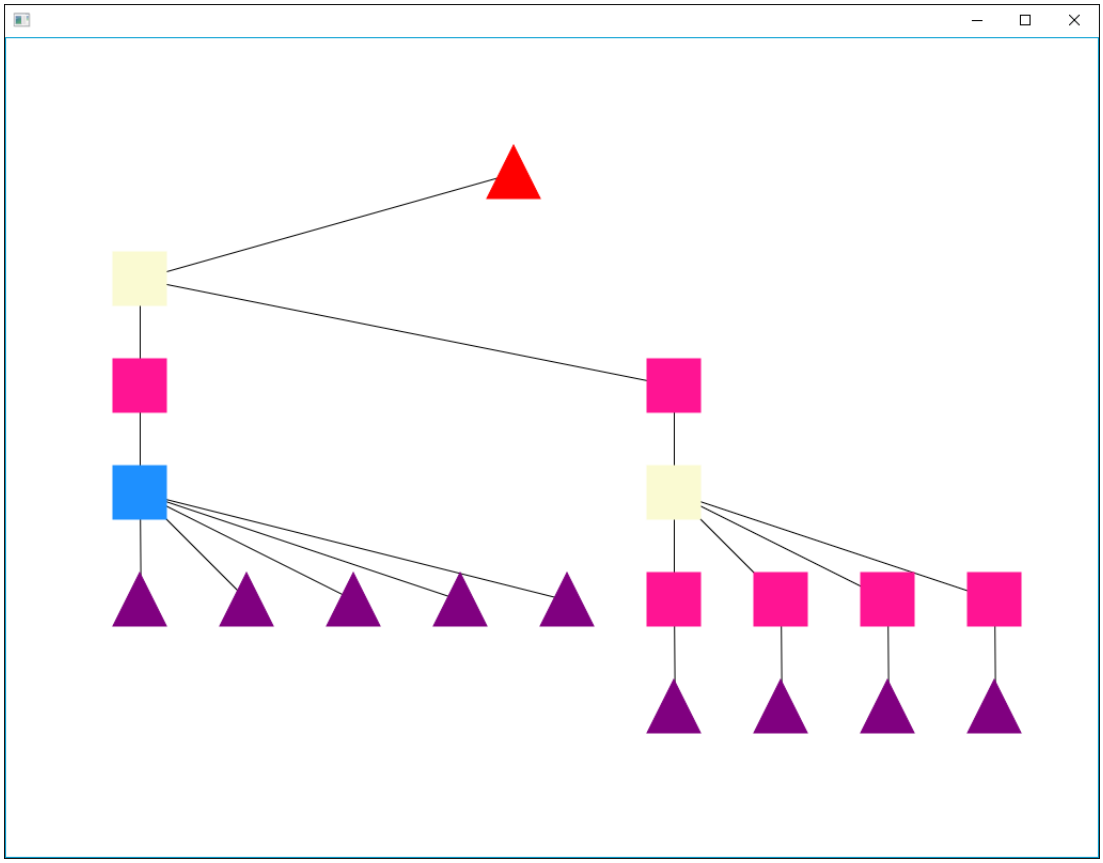

Figure 4-7: Screenshot of the first prototype version of ABOD3. Plan elements are presented in geometrical shapes of various colours. The aim of this interactive prototype was to showcase the tree-like visualisation of POSH and Instinct plans.

4.4.2 User Interface

The editor provides a user-customisable user interface (UI) in line with the good practices for transparency introduced in Chapter 3. Plan elements, their subtrees, and debugging-related information can be hidden, to allow different levels of abstraction and present only relevant information to the present development or debugging task. The application, as shown in Figure 4-8, always starts with only a Drives Collection

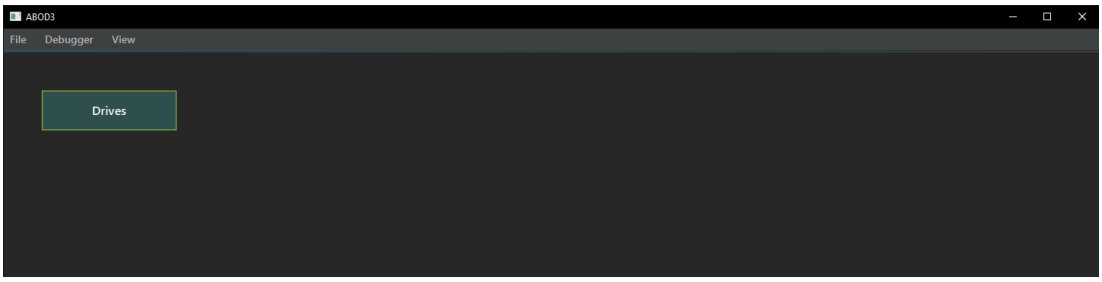

Figure 4-8: The default view of ABOD3 when it starts. Only the rout Drives Collection element is present. The user can either load an existing plan or start creating a new one by adding Drives to the DC.

node visible. *Reader extensions* allow ABOD3 to open BOD plan files in a variety of formats, including but not limited to plaintext for Instinct, LISP for POSH, and XML for UN-POSH. Once a plan is read, it is loaded in memory and then rendered by ABOD3's visualisation engine. The graphical representation of the plan is generated automatically, but the user can override its default layout by moving elements and zooming in/out the camera to suit needs and preferences. Layout preferences can be stored as and restored from a separate file.

As well as visualising plans, ABOD3 also allows editing of them. Right clicking on an element, displays a pop-up menu with possible actions, such as *Remove*, *Browse*, and adding an element. If *Browse* is selected, a pop-up window, seen in Figure 4-10, appears that allows the user to edit the plan element. Developers, through an API, can introduce *Writers extensions* to export the in-memory plan to files compatible with POSH, UN-POSH, or any other system that follows the high-level specifications set by BOD.

The simple UI and customisation allows the editor to be employed not only as a developer's tool, but as demonstrated in the next Chapter also to present transparency related information to the end-user to help them develop more accurate mental models of the agent.

4.4.3 Debugging

ABOD3 is designed to allow not only the development of reactive plans, like its predecessors do, but also the debugging of such plans in real time. Plan elements flash as they are called by the planner and glow based on the number of recent invocations of that element. Plan elements without any recent invocations start dimming down, over a user-defined interval, until they return back to their initial state. This offers abstracted

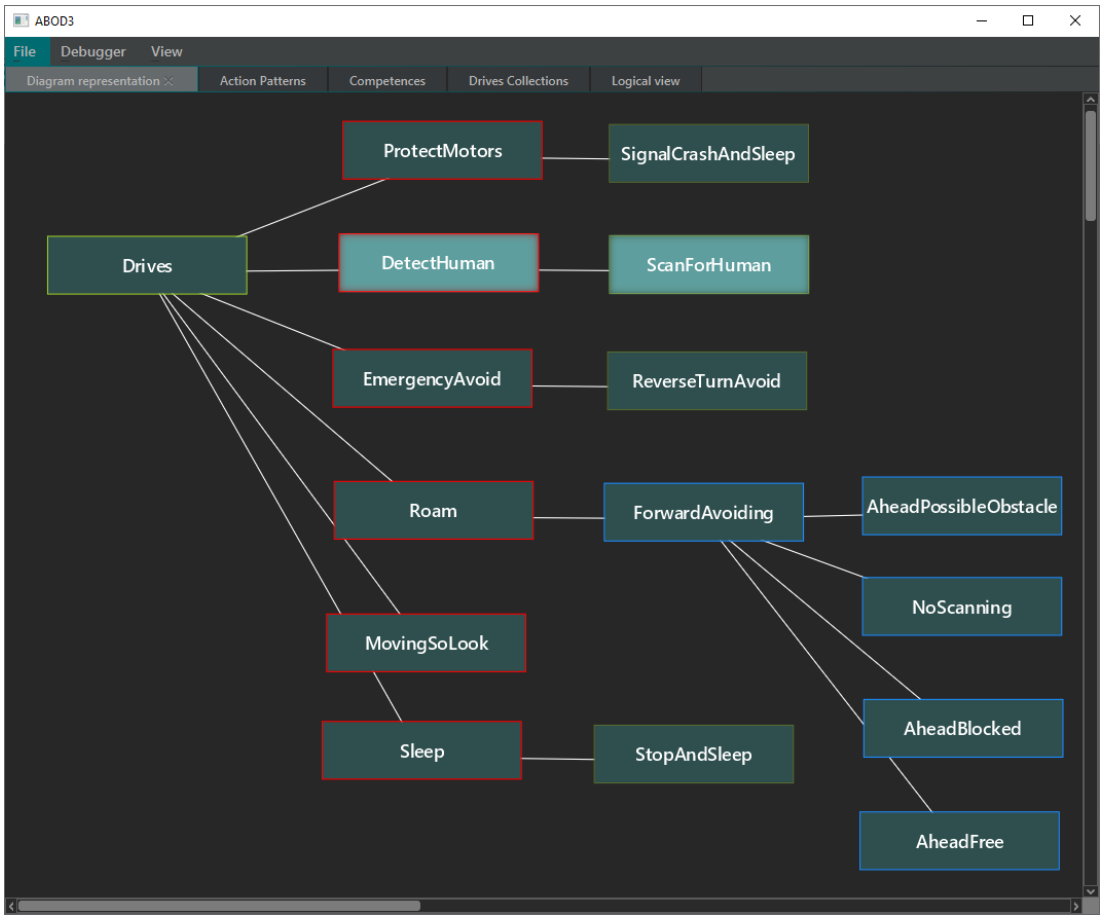

Figure 4-9: The ABOD3 Graphical Transparency Tool displaying the same Instinct plan as the one seen in Figure 4-6. ABOD3 is in debugging mode, but with various subtrees hidden and the camera zoomed-in to the plan. ABOD3, thanks to its user-customisable interface, can be deployed not only for developers but also for end-user transparency display who require a higher level of abstraction.

backtracking of the calls, and the debugging of a common problem in distributed systems: race conditions where two or more subcomponents are constantly triggering then interfering with or even cancelling each other.

During development of the R5 robot, we can report anecdotal experience of the value of offline analysis of textual transparency data, and the use of ABOD3 in its recorded mode. These tools enabled us to quickly diagnose and correct problems with the reactive plan that were unforeseen during initial plan creation. These problems were not so much 'bugs' as unforeseen interactions between the robot's various Drives and Competences, and the interaction of the robot with its environment. As such these unforeseen interactions would have been extremely hard to predict. This reinforces our assertion

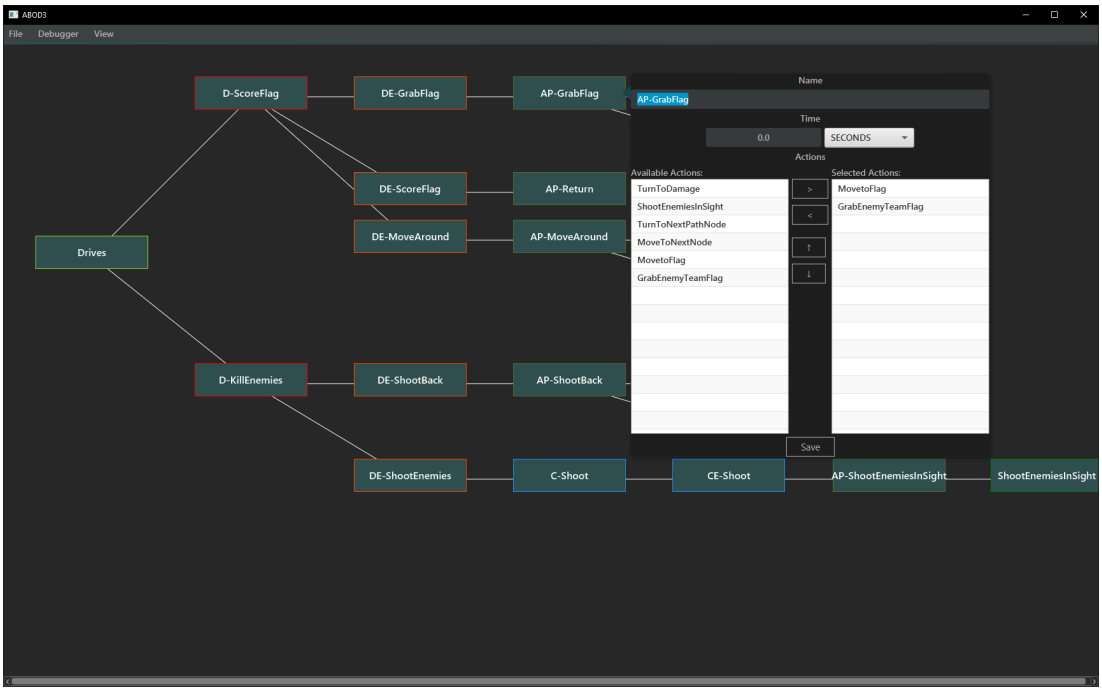

Figure 4-10: Through a simple interface, users can edit POSH and UN-POSH plans. In this example, the user can add new Actions to an AP.

that iterative behaviour oriented design (BOD) is an effective and appropriate method to achieve a robust final design. The BOD development methodology, combined with the R5 Robot hardware and the Instinct Planner has proved to be a very effective combination. The R5 Robot is robust and reliable, proven over weeks of sustained use during both field experiments and demonstrations. The iterative approach of BOD was productive and successful, and the robot designers report increased productivity resulting from use of the Instinct transparency feed and the ABOD3 tool.

ABOD3 can also support integration with videos of the agents in action, allowing for non-real-time debugging based on logged performance. Logging of actions taken and contexts encountered is a substantial aspect of AI accountability and transparency and alongside visual observation, it is often the most typical method of debugging complex intelligent agents. Finally, if ABOD3 is connected via TCP/IP to a remote agent, the server window can be used to send back commands to the agent.

4.4.4 Architecture & Expandability

The editor, as seen in its architecture diagram in Figure 4-11, is implemented in such a way as to provide for expandability and customisation, allowing the accommodation of

a wide variety of applications and potential users. The application is developed with a code-first approach, where each plan element is represented as an object derived from its own domain class and stored in the running memory. The alternative solution is to follow a model-first approach with a dedicated persistent database. The lack of a persistent model increases the ease of deployment and cross-platform compatibility, while reducing the size and complexity of the final deliverable. All data saved in run-time memory, are in an 'action-selection system agnostic state'. An important technical issue considered was the CPU usage. The system relies on multi-threading; thus, special care was taken to ensure thread safety and reduction of CPU load.

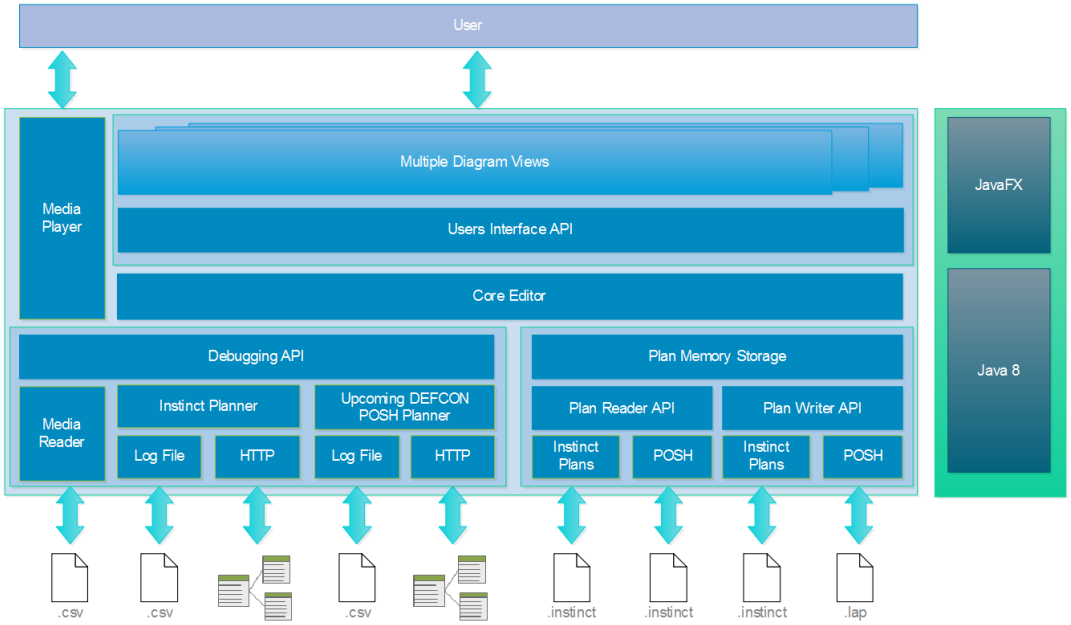

Figure 4-11: System architecture diagram of ABOD3, showing its modular design. All of ABOD3 is written in Java to ensure cross-platform compatibility. APIs allow the support of additional BOD planners for real-time debugging or even multiple file formats for the plans. The editor is intended, through personalisation, to support roboticists, games AI developers, and even end users (Bryson and Theodorou, 2019).

ABODE3 provides an API that allows the editor to connect with planners, presenting debugging information in real time. For example, it can connect to the Instinct planner by using a built-in TCP/IP server. R5 uses a WiFi connection to send transparency data back to the ABOD3.

4.4.5 Conclusions and Other Related Work

In this Section we presented a real-time visualisation tool, ABOD3, which communicates transparency-related information to end users. Our tool uses a user-configurable user interface and debugging capabilities that takes into consideration the design principles set in the previous Chapter. ABOD3 has already become part of our research tools.

It was used for the development of the R5 robot, the embodied agent used in the HRI studies presented in the next Chapter. Moreover, ABOD3 has been used for the development of the two UN-POSH applications: BUNG (see Chapter 5) and the Sustainability Game. When combined with the iterative approach of BOD, AI developers developers report increased productivity and understanding of the emerging behaviour of the artefacts (Wortham, Theodorou and Bryson, 2017b, and in detailed the next Chapter). In the HRI studies discussed in the next Chapter we see another use of ABOD3: end-user transparency. Thanks to its configurable UI, ABOD3 can be deployed to provide real-time transparency to non-expert users by hiding low-level plan elements.

We plan to continue developing ABOD3; not only we plan to port ABOD3 to run in embedded hardware, e.g. SoftBank Robotics Peppers, but also adding features such as "fast-forward" debug functions in pre-recorded log files. Finally, at future releases we will like to expand its use to cover non-POSH-like reactive systems, e.g. Behaviour Trees.

4.5 ABOD3-AR

Chapter 5 demonstrate through use cases how ABOD3 successfully provides real-time transparency information to both intelligent systems end users and developers. Yet, despite its effectiveness, there is a major disadvantage in the ABOD3 solution: a computer is needed, in addition to any on-board the robot, to run the software. A solution is to port ABOD3 to run directly on robots with built-in screens, such as SoftBank Robotic's Pepper. Although a technologically feasible and potentially interesting approach, it also requires that custom-made versions of ABOD3 need to be made for each robotics system. Moreover, this is not a compatible solution for robots without a display.

Nowadays, most people carry a smartphone. Such mobile phones are equipped with powerful multi-core processors, capable of running complex computationally-intensive applications, in a compact package. Modern phones also integrate high-resolution cameras, allowing them to capture and display a feed of the real world. That feed can be

enhanced with the real-time superimposition of computer-generated graphics to provide Augmented Reality (AR) (Azuma, 1997). Unlike Virtual Reality that aims for complete immersion, AR focuses on providing additional information and means of interaction with real-world object, locations, and even other agents.

We have been working on a new program, *ABOD3-AR*, which can run on mobile phones. ABOD3-AR, as its name suggests, uses a phone's camera to provide AR experience by superimposing the ABOD3's tree-like display of Instinct plans over a tracked robot. ABOD3-AR builds on top of the good practices tested by and lessons learned through our extended use of ABOD3. It provides a mobile-friendly interface, that faciliatates transparency to both end users and experts. In this Section, we do not only present the final product, but also look at the technical challenges and design decisions we faced during development.

4.5.1 AR in HRI

Augmented Reality has already been applied in fields such as military training, surgery, and entertainment. An area in which AR has found profound success is in manufacturing. AR applications are being used to simulate, assist, and improve manufacturing processes (Ong, Yuan and Nee, 2008); such processes include robotics-assisted and even fully automated manufacturing (Michalos et al., 2016).

Green et al. (2007) argue that as AR supports natural spatial dialogue, by displaying the visual cues, it can be used to facilitate human-robot collaboration. The use of spatial cues, for both local and remote collaboration, and the ability to visualize the robot relative to the task space (exo-centric view) can help human and robot to reach common ground and maintain situational awareness. Green *et al.* suggest that robots could communicate to human 'collaborators' internal state through graphical overlays. Their hypothesis about the usefulness of AR in HRI is supported by prior studies, such as the one conducted by Maida, Bowen and Pace (2007). Maida *et al.'s* study shows how the use of AR to communicate information related to the operation of a robot results to significant reduction of positioning errors by the human operators and time to task completion.

Further related work by Walker et al. (2018) shows that AR can be successfully used to display additional information which improves objective task efficiency in human-robot interaction. Walker *et al.* solution communicates the robot's motion intent. It superimposes, next to the robot, a line with an arrow pointing towards its planned direction of movement. Similar work to display the path-finding decision-making system

of a robot had been done by Giesler et al. (2004). The maturity of AR solutions over the 14 years between the two publications is evident in the final solutions presented. Giesler et al. requires fiducial markers to be placed and the update on the graphics happen only post-action.

Makris et al. (2016) demonstrate a solution where not only the trajectory of the robot is displayed, but also the 'non-safe zone' around the trajectory for a human observer. In addition, the solution by Markis *et al.* is able to display passive information, e.g. names of components, about a robot. They demonstrated that users of industrial robots, when they had access to their AR solution, have a *'safety feeling'* and acceptance. Both of these solutions aim at providing a lower level of transparency information; showcasing the path-finding algorithm to answer 'where will move on'. Our ABOD3-AR aims at providing a higher level of real-time transparency information to communicate to the user 'if the robot will move'.

A study conducted by Subin, Hameed and Sudheer (2017) demonstrates how users of AR applications aimed at developers that provide transparency-related information require an AR interface that visualizes additional technical content compare to naive users. These results are in-line with the claims made in Chapter 3 on how different users require different levels of abstraction and overall amount of information. Still, as discussed in the next subsection, we took these results into consideration by allowing low-level technical data to be displayed in ABOD3-AR upon user request.

4.5.2 Deployment Platform and Architecture

We selected Android Operating System (OS) [6] as our development platform. Due to the open-source nature of the OS, a number of computer vision and AR libraries already exist. Moreover, no developer's license is required for prototyping or even releasing the final deliverable. Further, Android applications are written in Java, like ABOD3, making it possible to reuse its back-end code. Unlike the original ABOD3, ABDO3-AR is aimed to be used exclusively for embodied agents' transparency. At the time of writing, Instinct is the only action-selection system supported.

Our test configuration, as seen in Figure 4-12, includes the tried-and-tested R5 robot. In the ARDUINO robot, the callbacks write textual data to a TCP/IP stream over a wireless (WiFi) link. A JAVA based Instinct Server receives this information, enriches it by replacing element IDs with element names and filters our low-level information, and sends this information any mobile phones running ABOD3-AR. Clients do not

[6] https://www.android.com/

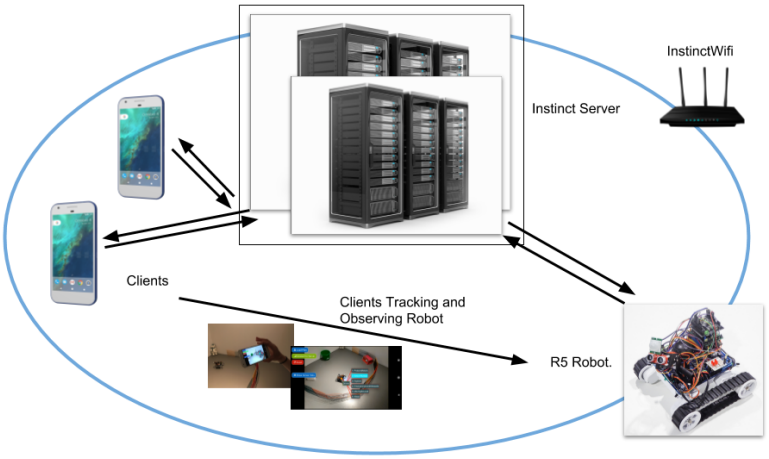

Figure 4-12: R5 uses a WiFi connection to send the transparency feed to the Instinct Server for processing. Smartphones, running ABOD3-AR, can remotely connect to the server and receive the processed information.

necessarily need to be on the same network, but it is recommended to reduce latency. We decided to use this 'middle-man server approach' to allow multiple phones to be connected at the same time.

4.5.3 Robot tracking

Developing an AR application for a mobile phone presents two major technical challenges: (1) managing the limited computational resources available to achieve sufficient tracking and rendering of the superimposed graphics, and (2) to successfully identify and continuously track the object(s) of interest.

Region of Interest

A simple common solution to both challenges is to focus object tracking only within a region of the video feed, referred to as the Region of Interest (ROI), captured by the phone's camera. It is faster and easier to extract features for classification and sequentially tracking within a limited area compare to across the full frame. The user registers an area as the ROI, by expanding a yellow rectangle over the robot, as seen in Figure 4-13. Once selected, the yellow rectangle is replaced by a single pivot located at the centre of the ROI.

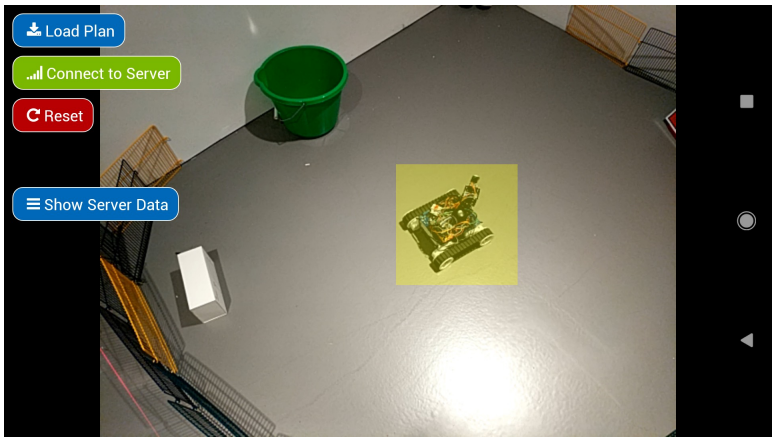

Figure 4-13: Screenshot from ABOD3-AR showing a user-selected Region of Interest marked in a translucent yellow rectangle. Under the rectangle is the R5 robot running the Instinct Planner.

Tracker

Various solutions were considered; from the built-in black-box tracking of *ARCore* [7] to building and using our own tracker. At the end, to speed-up development, we decided to use an existing library *BoofCV* [8]. BoofCV is a widely-spread Java library for image processing and objects tracking. It was selected due to its compatibility with Android and as it offers a range of trackers to prototype with.

BoofCV receives a real-time feed of camera frames, processes them, and then sends them back required information to the Android application that the library is enclosed in. A number of trackers, or *processors* as they are referred to in BoofCV, are available. We narrowed down the choice to the *Circulant Matrices* tracker (Henriques et al., 2012) and *Track-Learning-Detect* (TLD) tracker (TLD) (Kalal, Mikolajczyk and Matas, 2011).

The Track-Learning-Detect tracker follows an object from frame to frame by localising all appearances that have been observed so far and corrects the tracker if necessary. The learning estimates detector's errors and updates it to avoid such errors by using a learning method. The learning process is modelled as a discrete dynamical system and the conditions under which the learning guarantees improvement are found. The downside is that the TLD is computationally intensive. In our testing, we found that when TLD was used, the application would completely crash in older phones due to its high memory consumption.

[7] https://developers.google.com/ar/
[8] https://boofcv.org/

The Circulant Matrices tracker is fast local moving-objects tracker. It uses the theory of Circulant matrices, Discrete Fourier Transform (DCF), and linear classifiers to track a target and learn its changes in appearance. The target is assumed to be rectangular with a fixed size. A dense local search, using DCF, is performed around the most recent target location. Texture information is used for features extraction and object description. However, only one description of the target is saved, the tracker has low computational cost and memory footprint. Our informal in-lab testing shown that the Circulant tracker provides robust tracking.

The default implementation of the Circulant Matrices tracker in BoofCV does not work with coloured frames. Our solution first converts the video feed, one frame at a time, to greyscale using a simple RGB averaging function. The tracker returns back only the coordinates of the centre of the ROI, while the original coloured frame is rendered to the screen. Finally, to increase tracking performance, the camera is set to record at a constant resolution of 640 by 480 pixels.

4.5.4 User Interface

ABOD3-AR renders the plan directly next to the robot, as seen in Figure 4-14. A pivot connects the plan to the centre of the user-selected ROI. The PC-targeted version of ABOD3 offers abstraction of information; the full plan is visible by default, but the user has the ability to hide information. This approach works on the large screens that laptops and desktops have. On the contrary, at the time of this writing, phones rarely sport a screen larger than 6". Thus, to accommodate the smaller screen estate available on a phone, ABOD3-AR displays only high-level elements by default. Drives get their priority number annotated next to their name and are listed in an assenting order. ABOD3-AR shares the same real-time transparency methodology as ABOD3; plan elements get light up as they are used, with an opposite thread dimming them down.

Like its 'sibling' application, ABOD3-AR is aimed to be used by both end users and experts robotists. In Section 3.4 we established how there is not a one-size-fits-all solution for transparency. ABOD3-AR is built with this principle in mind, offering additional information on demand. A user can tap on elements to expand their substree. In order to avoid overcrowding the screen, plan elements not part of the subtree 'zoomed in' become invisible. Subin, Hameed and Sudheer (2017) shows that technical users in an AR application prefer to have low-level details. Hence, we added an option to toggle on the Server data, in string format, as received by ABOD3-AR.

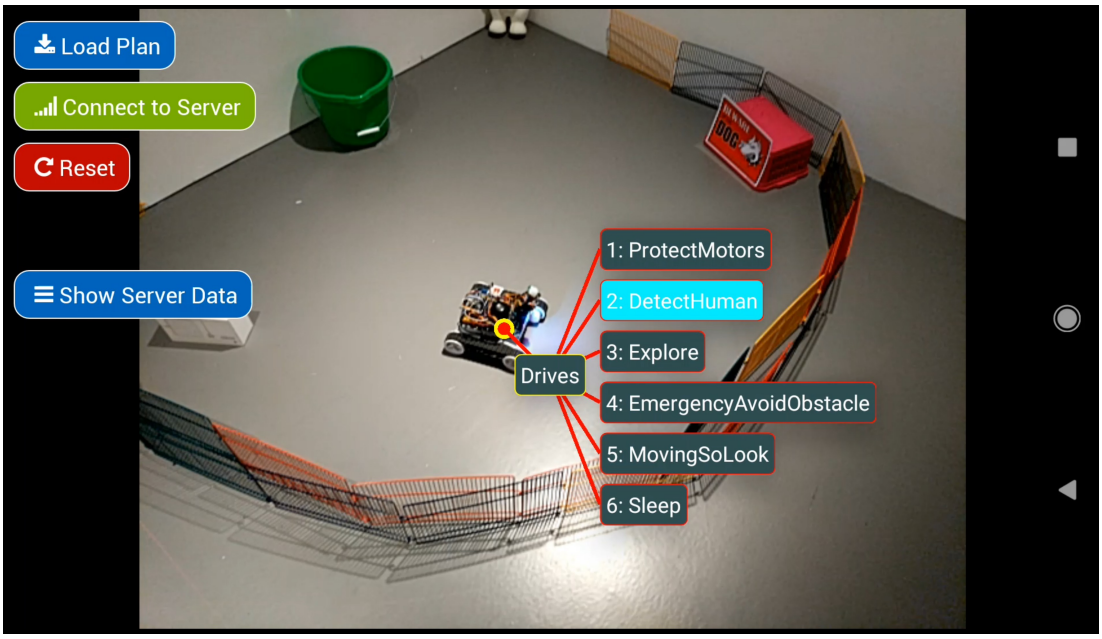

Figure 4-14: Screenshot of ABOD3-AR demonstrating its real-time debugging functionality. The plan is rendered next to the robot with the drives shown in a hierarchical order based on their priority.

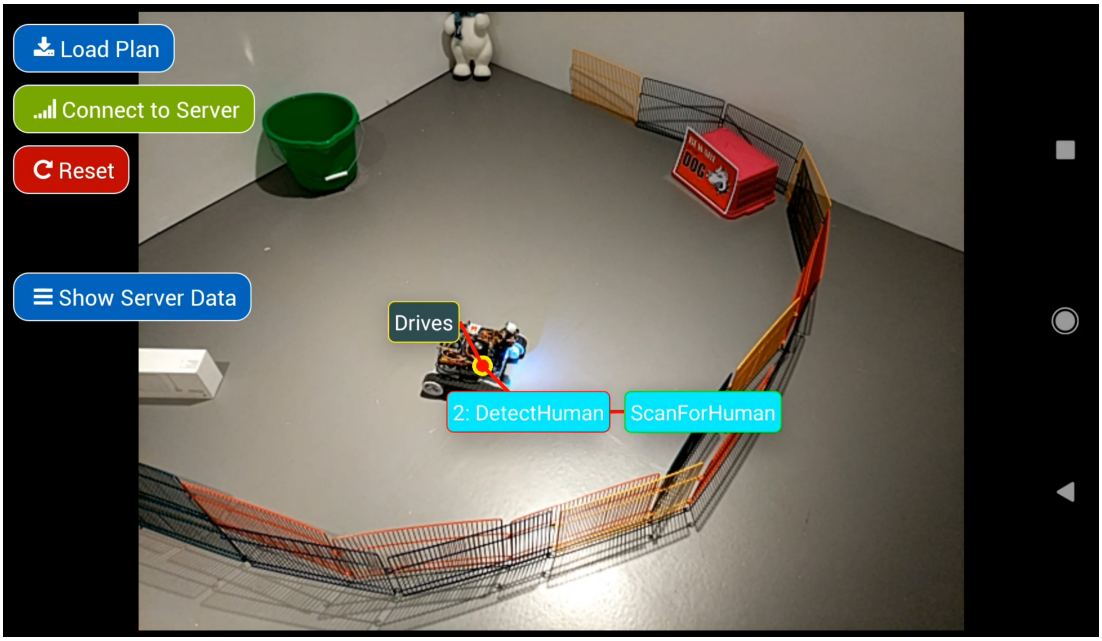

Figure 4-15: Screenshot from ABOD3-AR showing how a user can access additional information for a plan element by clicking on it. Other plan elements of its same level become hidden to increase available screen estate.

4.5.5 Conclusions and Other Related Work

In this Section we presented a new tool, ABOD3-AR, which runs on modern mobile phones to provide transparency-related information to end users. Our tool uses a purpose-made user interface with augmented-reality technologies to display the real-time status of any robot running the Instinct planner. As far as we are aware this is the first use of mobile augmented reality focusing solely on increasing transparency in robots and users' trust towards them. Previous research regarding transparency in robots relied on screen and audio output or non real-time transparency. Building upon past research, we provide an affordable, compact solution, which makes use of augmented reality. The results from a user study presented in the next Chapter demonstrate how ABOD3-AR can be successfully used to provide real-time transparency to end users.

Planned future work also aims at improving the usability of the application further. Currently, the robot-tracking mechanism requires the user to manually select an area of ROI which contains the robot. Future versions of ABOD3 - AR would skip this part and replace it with a machine learning (ML) approach. This will enable the app to detect and recognize the robot by a number of features, such as colour and shape. The app will also be enhanced to be able to retrieve the robot type and plan of execution from a database of robots.

4.6 Conclusions

In this chapter, we described one approach to systems engineering real-time AI, the cognitive architecture *Behaviour Oriented Design* (BOD), including two of its previously-established reactive-planning paradigms; *POSH* and *Instinct*. Developers can use BOD not only as a software architecture, providing them with guidance on how to structure their code, but also as a software-development methodology, a solution on how to write that code. BOD aims at ensuring a modular, auditable design of intelligent systems.

Next, we introduce *UN-POSH*; a games-centric version of POSH, developed to be used exclusively for the Unity game engine. UN-POSH is used in *BOD-UNity Game* (BUNG). BUNG is a serious game, described in the next Chapter, designed to provide a prototyping platform and help to teach BOD to final-year undergraduate and postgraduate students. Further, UN-POSH has been used in the ecological simulation, *The Sustainability Game*, introduced in this chapter and further featured in the work presented in appendix B.

Finally, we presented the ABOD3 software and its spin-off mobile-phone application

ABOD3-AR. Both of these applications provide visualisation of BOD plans and real-time. The next Chapter presents user studies, where ABOD3 and ABOD3-AR have been used to provide transparency to both naive and expert users.

Chapter 5

Improving Mental Models of AI

> "The question of whether a computer can think is no more interesting than the question of whether a submarine can swim."
>
> Edsger W. Dijkstra

5.1 Introduction

Designing and developing an intelligent agent is a difficult and often lengthy process. Developers need to understand not only their own code, but also the emerging behaviour of the intelligent agent they created. This emerging behaviour is the result of the agent's interaction it with its environment and other artificial and natural agents. It is hard to decode by simply observing the agent. Making the quality assurance of the final product to rely extensively on the testing environment used during development. Risks, as discussed in the previous chapter, can be mitigated by using well-established cognitive architectures to dictate the ontology of the system. Moreover, the usage of a well-documented development methodology, with sufficient audit trails, can help distribute any responsibility —and even accountability— if any incidents occurred.

Still, neither of these solutions can pro-actively help the end users of a system, who may encounter it without any prior knowledge of its design and operation. Such users may end up creating—or updating their existing—mental models with inaccurate information, as they may try to assign narratives based on the physical cues of the robot, the environment they encounter it, and—even worse—the media representation of AI. A solution to this, proposed in chapter 3, is the careful case-specific implementation of

transparency. By transparency, we[1] refer to the combination of both the hardware and software design requirements set in chapter 3 to allow the communication of meaningful information, such as the goals and progress towards goals, from the robot to its users.

In this chapter, we investigate the effects of implementing transparency through the use of the real-time visualisation software: ABOD3 and ABOD3-AR, both which were described in the previous chapter. First, we discuss how we use ABOD3 and BOD-UNity Game (BUNG), a serious game described in this chapter, as part of our teaching curriculum. We present indicative results, which show how ABOD3 can be used as an experts' debugging tool; helping developers understand the emergent behaviour of their own creations. Next, two end-user studies conducted with ABOD3 are presented. The first study is an online experiment, where we used a pre-recorded video of the non-anthropomorphic R5 robot (see section 4.2.3) and online questionnaires (Wortham, Theodorou and Bryson, 2016). In the other participants directly observed the robot (Wortham, Theodorou and Bryson, 2017a). Each study is described in the order it was conducted. A discussion follows based on the results of both experiments, where we conclude that even abstracted real-time visualisation of a robot's action-selection system can substantially improve human understanding of machine intelligence by its observers. Next, a third user study, performed with ABOD3-AR, is presented and analysed, validating our previous results and showing the effectiveness of ABOD3-AR. Finally, our results suggest that an implementation of transparency within the good-practice guidelines set out in chapter 3 does not necessary imply a trade-off with utility. Instead, the overall experience can be conceived as more interactive and positive by the robot's end users.

5.2 ABOD3 for Developers Transparency

As we make agents transparent for end users, we make them transparent for their designers and developers too. Real-time debugging of an agent's decision-making mechanism could help developers to fix bugs, prevent issues, and explain variance in a system's performance. Moreover, an implementation of high-level transparency would allow de-

[1]This chapter contains results previously published in: Wortham, R.H., Theodorou, A. and Bryson, J.J., 2017. Improving robot transparency: Real-time visualisation of robot AI substantially improves understanding in naive observers.*26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, Vol. 2017-January, pp.1424-1431 and in Wortham (2018). Using Other Minds: Transparency as a Fundamental Design Consideration for AI Systems. PhD Thesis University of Bath. In addition, results presented in the ABOD3-AR section appear in the final version of Rotsidis A., Theodorou A., and Wortham R.H., 2019. Robots That Make Sense: Transparent Intelligence Through Augmented Reality. *1st International Workshop on Intelligent User Interfaces for Algorithmic Transparency in Emerging Technologies*, Los Angeles, CA USA.

velopers to concentrate on tuning the low-level specialised modules, e.g. pathfinding or sensors control, which can be debugged through traditional means, such as with a run-time code debugger. Meanwhile, user-experience specialists can focus on the high-level behaviour of the agent, ensuring that the agent achieve its intended utility and provides an engaging experience for its users. For example, in video games, designers, who may lack technical expertise but are aware of the purpose of each agent, can tune or even develop the high-level behaviour of each agent.

In this section, we[2] demonstrate how ABOD3 has been used in teaching AI to student developers. First, we discuss the course and the coursework where ABOD3 was used. Next, I showcase *BOD UNity Game (BUNG)*, a serious game developed for the purposes of the course. Then, we present the results gathered using a feedback survey given to the students. Finally, we provide a discussion of the results. Note, due to the small sample of results gathered, we are treating this section as indicative only.

5.2.1 Intelligent Control and Cognitive System

Final-year undergraduate and taught postgraduate-level students taking the AI module *Intelligent Control and Cognitive Systems* (ICCS), learn how to build intelligent systems. The course contains three major pieces of Coursework. They are designed to progressively teach students the nature of intelligence, as weekly lectures provide the necessary systems engineering, psychology, and philosophy background knowledge needed to understand cognition and build intelligent agents.

For the last assignment we have been using and developing video games as a platform for students to use. Games have a much lower cost than robots and allow for a non-hardware limitations to exploration of different approaches. Moreover, games have long been used to demonstrate the effectiveness of new techniques in intelligent systems as a whole. Starting in academic year 2012, we have been utilising Unreal Tournament 2004 and the POGAMUT mod (Gemrot et al., 2009). In 2018, we transitioned to the purpose-made game *Bod-UNity Game* (BUNG) presented next in this section.

Students are tasked to form their own teams of five agents. Each agent can be individually customised with its own set of goals and behaviours to satisfy such goals. They need to consider the emerging behaviour between the interactions both within the team and with the enemy team, instead of focusing only on the later. BUNG is designed to allow a variety of strategies at both individual agent and team levels. For example, at

[2] I developed ABOD3, BUNG, and run the survey. The module was primarily taught by Joanna J. Bryson, who also designed the Coursework specifications.

team level, students can approach the challenge in any of the following ways:

1. The whole team rushing to secure the enemy flag.
2. Part or all of the team staying in the base to defend the flag.
3. Search for and eliminate all enemies, ignoring the flag until this is completed.
4. Some combination of the above.

Each team is required to take part in a two-parts tournament. First, a qualifier stage takes place. In this stage, groups of four teams are formed. Each team is expected to compete against all teams in its group, scoring 2 points for each victory or 1 for each tie. In addition, overall point totals will be kept to resolve any ties in outcomes. The winner from each group advances to the next stage; a round-robin league.

5.2.2 BoD UNity Game (BUNG)

Our new game, BoD UNity Game (BUNG), seen in Figures 5-1 and 5-2, takes cues from Unreal Tournament 2004 and other shooting games. It is a team-based Capture the Flag (CTF) developed in the popular games engine Unity, where the 'players' develop —or tune existing— agents.

BUNG is designed to be used as an educational platform to teach developers, such as students undertaking ICCS, to understand how to build complete complex agents by using the UNity-POSH (UN-POSH) reactive planning paradigm presented in the previous chapter. The game comes as an uncompiled Unity project, with sample agents and an integrated process for fast prototyping UN-POSH agents.

Gameplay Mechanics

Capture the flag is a popular multiplayer mode in first-person shooter games, where the participants split into teams. Each team has a flag located in its *base*, which acts as its starting location. The objective of the game is to capture the other team's flag located at the team's base and safely take it to their own base. Players may combat and 'kill' enemies, or avoid them altogether. Killed enemies often respawn at their original base after a set time, however multiple varieties of the game exist with different rules for respawning and scoring. In BUNG, a game consists of three four-minute rounds between two five-person teams. Killed players won't respawn until the end of a round. Also, a team can score even if the enemy currently holds their flag. A team can therefore achieve a high score in an individual round by simply exterminating all of its enemies

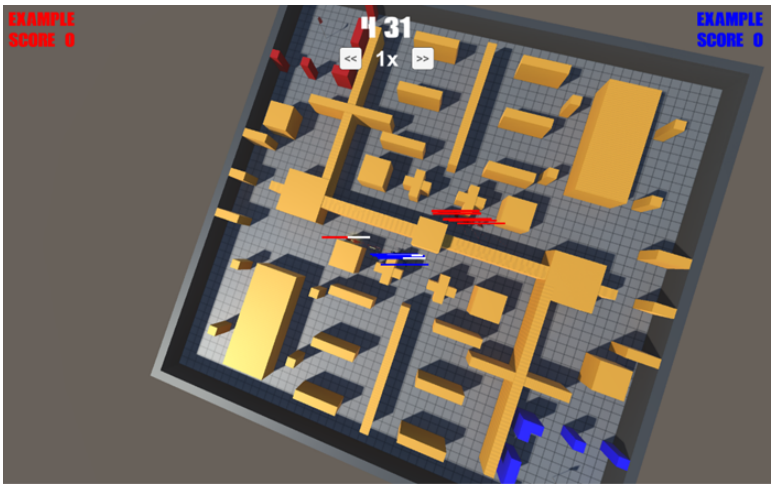

Figure 5-1: Screnshot from BUNG, where the player uses a zoom-out 'God's eye' view to observe the whole map. In the scene, agents from the Blue team are engaged in compact against an agent from the Red team.

and scoring the flag multiple times. However, if the surviving team fails to capture the enemy's flag, the eliminated team is awarded a point.

Once they have programmed their teams, human 'players' act as spectators, watching the two teams of five bots as they fight against each other. The game contains two camera views; the first is a God's eye view and the other follows a single agent from a third-person view. The former allows the user to move and inspect the game from a top-down view. This perspective allows the spectators to observe how different agents interact with each other, helping them study their behaviour and possibly debug it. The second camera option allows the spectator to follow through a third-person perspective an agent of their choice, helping to focus on a particular agent. This view should be useful for developers working to monitor and debug specific behaviours.

Agents Development

The agents are human-like characters. Developers have control of their legs, arms, and head indepentently from each other. An agent can, for example, look behind while walking forwards. Agents can only detect enemies and flags within their field of view. Each agent has its own instance of the UN-POSH planner and animation controller. In addition, the developer can assign different plans, access different behaviour modules in the behaviour library, and navigation controllers to each agent, enabling each agent to be individually customised with its own set of goals and behaviours to satisfy such goals. Developers need to think of their overall strategy of their team before assigning

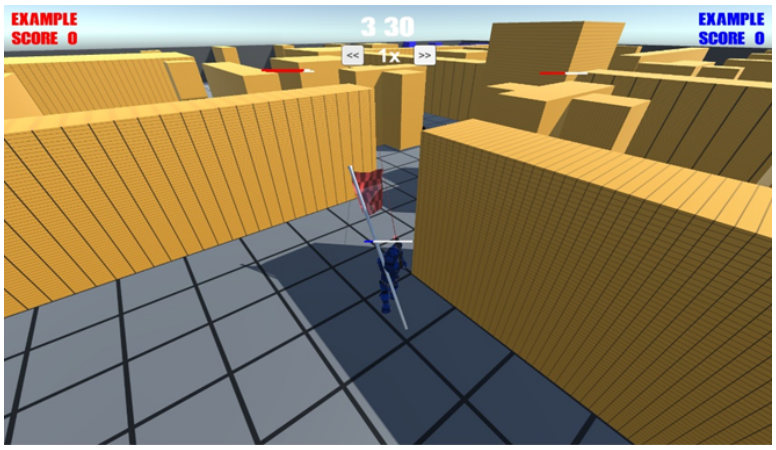

Figure 5-2: Screenshot from BUNG, where the player observes an agent from the blue team holding the flag in third-person view.

goals to individual members. Even if two or more members share the same list of goals, they may have different priorities. For example one or more members of the team may prioritise a goal to defend their own flag over attempting to capture the enemies' flag. Plans are stored in XML format, but they are editable by ABOD3.

Once the game starts running, it will automatically connect to an instance of ABOD3, set in debug mode, through TCP/IP. Each agent's Planner reports the execution and status of every plan element in real time, allowing developers to capture the reasoning process within the agent that gives rise to its behaviour. ABOD3 will always display the plan of the agent currently —or most recently— selected in spectator's mode. This allows a developer to select within the game which agent to debug.

5.2.3 Experimental Design

After the tournament, students are expected to submit a one-page report with 10 observations they made. Observations can be about human-like cognitive systems more generally, or cooperation more specifically. They can also be informed critiques of the software tools provided. In the academic year 2017/18, students were also asked to fill in an optional survey. The survey contained a number of questions about self-reflective evaluation of ABOD3 as a debugging tool, but also general feedback gathering, e.g. features requests, for BUNG. In this section, we focus only on the former.

Table 5.1 summarises the questions asked on the survey specifically about ABOD3 as a debugging tool. Their purpose is to see how our students through self-reflection evaluate the impact of ABOD3 had on their development process and if access to

Table 5.1: Questions asked to all participants ($N = 20$) to measure the perceived usefulness of ABOD3 for student developers.

Question	Response
Did you use ABOD3 for debugging?	Y/N
If not, can you please tell us why?	Free Text
I am satisfied with ABOD3 as a debugging tool.	1-5
ABOD3 helped me understand POSH better.	1-5
ABOD3 helped me understand AI, in general, better.	1-5
ABOD3 helped me understand NI, in general, better.	1-5
ABOD3 helped me develop my agents faster.	1-5
ABOD3 helped me develop better performing agents.	1-5
ABOD3 helped me develop agents with less bugs.	1-5
Any other feedback or comments for ABOD3 as a debugging tool?	Free Text

ABOD3 provided them with any additional education benefits.

5.2.4 Pre-Analysis Filtering

Unfortunately, only 20 students filled in the optional survey and provided feedback. Out of the 20 students, 5 of them selected *No* in the question "Did you use ABOD3 for debugging?" Thus, their scores and answers, other than why they opted not to use ABOD3, were removed from the analysis. Post-filtering, we retained a sample of just 15 answers. Hence, due to the small sample, all results should be treated as indicative.

5.2.5 Results

Quantitative Results

Table 5.2 shows the results of the survey. In all questions, the median value is above the neutral score of 3, demonstrating a general agreement with the statements.

Qualitative Results

The 5 students who did not use ABOD3 provided the following answers to the question "If not, can you please tell us why?":

- I didn't want to.
- Not used to it, prefer to use Unity.
- Prefer to use the xml.
- Didn't manage to get it to work.

Table 5.2: Students ($N = 15$) expressed their satisfaction with ABOD3 as a debugging tool, as it scored above the neutral score (3) in all questions.

Question	Median
I am satisfied with ABOD3 as a debugging tool.	4
ABOD3 helped me understand POSH better.	4
ABOD3 helped me understand AI, in general, better.	4
ABOD3 helped me understand NI, in general, better.	3.5
ABOD3 helped me develop your agents faster.	4
ABOD3 helped me develop better performing agents.	4
ABOD3 helped me develop agents with less bugs.	4

- It didn't appear very clear when the plan would light up for the currently selected player.

The following comments were provided by students:

- ABOD3 was indispensable for debugging agent behaviour. It was a powerful and intuitive process to view the game world from the agent's over-the-shoulder camera, while comparing what you thought the agent should be doing against the real-time ABOD3 output.
- ABOD3 helps with understanding and debugging the behaviour of the BOD agents in respect to the created POSH plan due to the visualisation, where parts of the plan will flash as they are called by the planner and dim down when they return to their initial state.
- I must comment on how useful ADOB3 was very useful for inspecting the (often unexpected) ways in which these simple behaviours were combined to produce higher level behaviours.
- Really useful for debugging. Not so much for displaying the correct values that were present in the XML which is a misrepresentation of value, despite still working as if it had read them correctly, which I believe is the case. Having a visual representation of behaviour was incredibly useful, nice work :)

5.2.6 Discussion

Our indicative results—especially the written feedback provided by the students—suggest that even developers struggle to understand the emergent behaviour of their own agents. Tools that provide transparency, namely ABOD3, allow a high-level overview of an agent's behaviour, making it easier to test and tune the agent's emergent behaviour.

This understanding is not always possible by treating the agent as 'just a piece of code' to be debugged. The majority of the survey respondents claim that ABOD3 helped them develop not only faster and better performing agents, but also agents which are less prone to error. Lab-based interactions with the students indicate towards similar conclusions. Regardless of the low response rate of the survey, the majority of the students integrated ABOD3 into their development pipeline.

The responses of the feedback survey also suggest that the usage of a transparency display in teaching is advantageous from an educational point of view. The majority of the survey participants strongly believe that ABOD3 helped them better understand AI. This was expected, as our real-time debugger visualises the multiple—and often mutually exclusive—goals complete complex agents may have. Students were able to see the interactions between their agents' various Drives and Competences their UN-POSH plans had, as the environmental and internal state of their agents changed.

Furthermore, a large number of students drew parallels between artificial and natural intelligence. In their observations reports, multiple students treated the game as an agent-based model. They explicitly categorised agents as *free riders* or *altruists*, depending on whether the agent's behaviour contributed towards the team's task of capturing the enemy team's flag or not. We were expecting students to have such observations, as the previous assignment involved them exploring the agent-based model introduced by Čače and Bryson (2007) on the evolution of cooperation. While the use of agent-based modelling as a means to test macro-level hypothesises is well established in the literature (Gallagher and Bryson, 2017, and references therein), experts require a significant amount time to understand and analyse the emergent behaviour of their models. If we are making the artificial action-selection system transparent, we are inevitably also making the natural decisionmaker that the system is based upon.

Finally, in addition to its use at ICCS, ABOD3 has also become integral tool within our research lab. Wortham, Theodorou and Bryson (2017b) discuss how ABOD3 was used to debug Instinct plans for the R5 robot, which is used in the studies presented later in this chapter. ABOD3 enable us to quickly diagnose and correct problems with the reactive plan that were unforeseen during initial plan creation. Moreover, I have extensively used ABOD3 to debug and tune the agents in both BUNG and the Sustainability Game (see appendix B). In The Sustainability Game, ABOD3 made it easy to understand if there was a problem with the plan, e.g. a behaviour was not triggered, or a problem with its underlying code of a behaviour, e.g. if the pathfinding was not calculating the right path, or even code in the gameplay mechanics. For example, during testing I noticed that the agents would 'randomly' stop what they were doing and return back

to their houses. ABOD3 showed that the `D-Survive` Drive was getting triggered and prompting the agents to find shelter. Further investigation, showed me that there was a bug in the day/night cycle mechanic of the game.

5.3 ABOD3 for End-user Transparency

ABOD3 (see chapter 4 for a discussion on ABOD3) was used in conducting two Human-Robot Interaction (HRI) experiments. The first, an online study, was conducted by using a pre-recorded video of a robot and online questionnaires. The second user study involves participants directly observing the robot. Both studies test the hypothesis that ABOD3 can be successfully used by users of varied demographic backgrounds, with and without technical expertise, to improve the accuracy of their mental models for a robot, i.e. help them understand its functionalities. Consequently, we[3] also tested the hypothesis that if an agent's action-selection mechanism is treated as a black box, its users will generate inaccurate mental models, i.e. they will make wild assumptions about its capabilities. We visit each of them in turn and then present a discussion based on the results of both studies.

5.3.1 Online Study

An online study was conducted using a video of a robot, rather than allowing participants to directly interact with the robot. The purpose of this study is to investigate if access to transparency information, as visualised by ABOD3, can help people create more accurate mental models.

Experimental Design

We decided to conduct the study by using pre-recorded videos for online data collection. This approach has been chosen by others (Cameron et al., 2015) with acceptable results. Due to time and resource constrains, we had to use a low-cost, low-power robot, R5, based on the Arduino micro-controller —for more information for the robot, see: Wortham, Gaudl and Bryson (2016) or the previous chapter. A five-minutes video was produced by combining videos captured by three cameras, two in at the front and rear

[3]This section contains results previously published in: Wortham, R.H., Theodorou, A. and Bryson, J.J., 2017. Improving robot transparency: Real-time visualisation of robot AI substantially improves understanding in naive observers.*26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, Vol. 2017-January, pp.1424-1431 were I provided the software and contributed to the final paper. They consider to be contributions of the thesis: Wortham R.H. (2018). Using Other Minds: Transparency as a Fundamental Design Consideration for AI Systems. PhD Thesis University of Bath

ends of the robot pen and one mounted directly on the robot. Figure 5-3 shows a frame from this video, as presented to the participants in the control group. A second video was created, by loading the same composite video and the Log file produced by Instinct while shooting the video in ABOD3. By using ABOD3 in debug mode, we were able to both render the video and a 'real-time debugging feed' of the robot. We captured ABOD3 running in its debug mode to produce a second video, which is shown in Figure 5-4.

Figure 5-3: Online Study: Frame from Video 1, as presented to participants in the control group. The R5 robot is shown navigating around a den-like domestic environment. Two additional video feeds, one from a camera mount on the R5 and one from across the den, were added to the video as picture-in-picture to compensate for any periods the robot was not in the line of sight of the main camera (Wortham, Theodorou and Bryson, 2016; Wortham, 2018).

We decided to follow the commonly used *Independent Groups Design*, by splitting our participants into two groups; a Control group without access to ABOD3 and a Treatment group with access to ABOD3 and its transparency visualisation. An alternative was to use the *Repeated Measures Design*, by having participants first watch the video without ABOD3, answer questions, and then watch the ABOD3 video, and answer the same questions again. We decided to use the independent groups design in this and all following experiments due to time restrictions (take less time for a participants to complete the experiment) and also avoid the problem of fatigue, which can cause distractions, and boredom thus affecting results. Finally, if we had our participants to

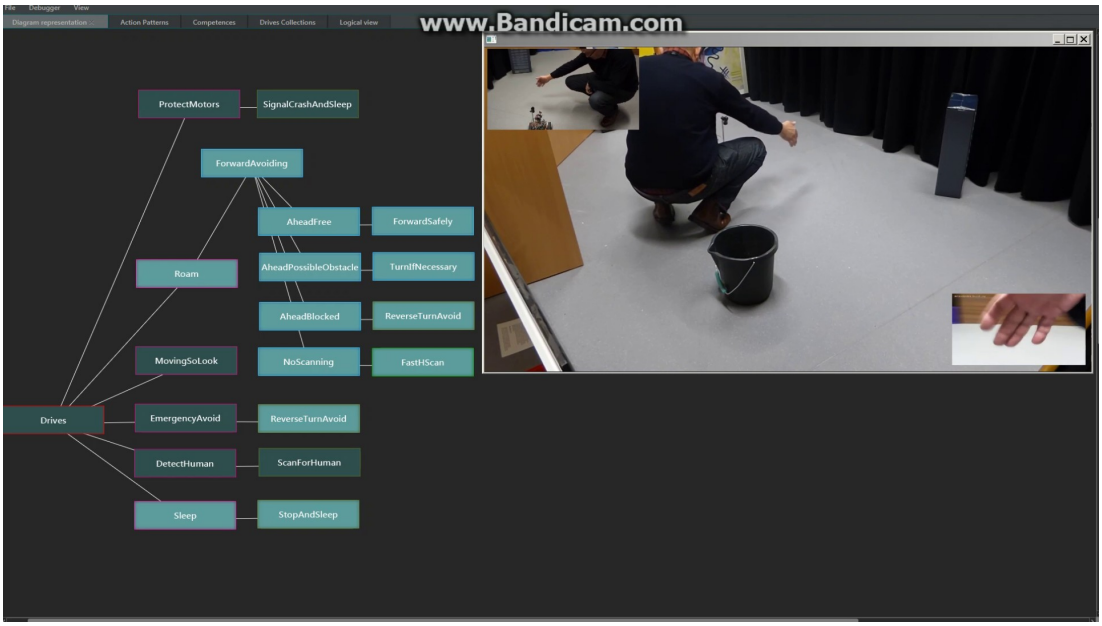

Figure 5-4: Online Study: Frame from Video 2. ABOD3 was used to provide visualisation of the plan and real-time transparency feed. Sub-trees have been hidden from view (Wortham, Theodorou and Bryson, 2016; Wortham, 2018). **Note:** Drive labels were legible to the subjects and can be seen clearly by zooming in the postscript version of this dissertation.

watch the video twice, we may had introduced biases to their answers —especially if they realise that they completely misunderstood the functionality of the robot in the first video.

Participants would initially express interest by filling an online questionnaire to gather basic demographic data: age, gender, educational level, whether they use computers, whether they program computers, and if they have ever used a robot. This allowed us to screen participants before assigning them to the treatment or control group, as we wanted a good split of demographic background between the two groups.

Table 5.3 summarises the questions asked after the participant had seen the video. These questions are designed to measure various factors: the measure of intelligence perceived by the participants and—most importantly—the accuracy of the participants' mental model of the robot. For analysis, the four free text responses were rated for accuracy by comparing them to the robot's actual Drives and programmed behaviours. They were given a score per question of 0 (inaccurate or no response), 1 (partially accurate) or 2 (accurate). By summing the scores, the accuracy of the participant's overall mental model is scored from 0 to 6.

Table 5.3: Questions asked to the participants to measure the intelligence perceived by the participants (Intel questions) and accuracy of their mental models (MM questions)(Wortham, Theodorou and Bryson, 2016; Wortham, 2018).

Question	Response	Category
Is robot thinking?	Y/N	Intel
Is robot intelligent?	1-5	Intel
Understand objective?	Y/N	MM
Describe robot task?	Free text	MM
Why does robot stop?	Free text	MM
Why do lights flash?	Free text	MM
What is person doing?	Free text	MM

Demographics

We formed the two groups by dividing our participants based on their demographic background instead of a random distribution. Priority was given to matching the number of programmers in each group, and to having an equal gender mix. The final demographics of each group of participants is shown in Table 5.4. Each group received an identical email asking them to carefully watch a video and then answer a second questionnaire. Group 1 had access to the video displaying just the robot, while Group 2, our treatment group, watched the ABOD3 video.

Table 5.4: Online Study: Demographics of the Participants in the online experiment ($N = 45$). Group One is the control group without access to the ABOD3 and Group Two is the treatment group with access to the debugger shown in Figure 5-4 (Wortham, Theodorou and Bryson, 2016; Wortham, 2018).

Demographic	Group One ($N = 22$)	Group Two ($N = 23$)
Mean Age (yrs)	39.7	35.8
Gender Male	11	10
Gender Female	11	12
Gender PNTS	0	1
Total Participants	22	23
STEM Degree	7	8
Other Degree	13	13
Ever worked with a robot?	2	3
Do you use computers?	19	23
Are you a Programmer?	6	8

Results

The primary results obtained from the experiment are outlined in Table 5.5. There is a significant correlation ($p < 0.05$) between the accuracy of the participants' mental models of the robot and the provision of the additional transparency data provided by ABOD3, with $t = 2.86$, $p = 0.0065$. There is also a substantially higher number of participants in our treatment group who report that they believe the robot is thinking; $t = 2.02$, $p = 0.050$.

Table 5.5: Online Study: There is a statistical significant improvement ($p < 0.05$) in the accuracy of the mental models of Group 2 compare to Group 1. There is also a substantially higher number of participants in our treatment group who report that they believe the robot is thinking (Wortham, Theodorou and Bryson, 2016; Wortham, 2018).

Result	Group One ($N = 22$)	Group Two ($N = 23$)
Is thinking (0/1)	0.36 (σ=0.48)	0.65 (σ=0.48)
Intelligence (1-5)	2.64 (σ=0.88)	2.74 (σ=1.07)
Understand objective (0/1)	0.68 (σ=0.47)	0.74 (σ=0.44)
Report Accuracy (0-6)	1.86 (σ=1.42)	3.39 (σ=2.08)

Even if we did not conduct an empirical study, a number of noteworthy answers for the question *Describe robot task?* are shown bellow:

- [the robot is] Trying to create a 3d map of the area? At one stage I thought it might be going to throw something into the bucket once it had mapped out but couldn't quite tell if it had anything to throw.
- [the robot is] aiming for the black spot in the picture.
- is it trying to identify where the abstract picture is and how to show the complete picture?
- [the robot] is circling the room, gathering information about it with a sensor. It moves the sensor every so often in different parts of the room, so I think it is trying to gather spacial information about the room (its layout or its dimensions maybe).

5.3.2 Directly Observed Robot Experiment

The goal of this experiment was to validate the results from the previous experiment, by running a similar experiment with the R5 and ABOD3.

Experimental Design

The robot operated within an enclosed pen as a special interactive exhibit within the main exhibition area, seen in Figure 5-5. A large computer monitor was positioned at the front of the pen displaying the ABOD3 real-time visualisation of plan execution. This display was turned on at random periods to create our treatment group, Group 2.

Figure 5-5: Directly Observed Robot Experiment: Photograph showing the R5 robot in a purposed-made den. Obstacles visible include a yellow rubber duck and a blue bucket. The position and orientation of the display is shown, with the ABOD3 opened and visualising the real-time transparency feed from the robot. The display was turned off to create our control group, Group 1 (Wortham, Theodorou and Bryson, 2017a; Wortham, 2018).

Participants were asked to observe the robot for several minutes whilst the robot moved around the pen and interacted with the researchers. Afterwards, they completed a paper-based questionnaire. The same set of questions as the pilot study was used, with a minor refinment of the question *Why does the robot stop every so often* to *Why does it just stop every so often (when all its lights go out)?* This change was deemed necessary to avoid any ambiguoutity.

Participants Recruitment

The second experiment took place over three days at the *At-Bristol Science Learning Centre*, located at the city of Bristol (UK). This context was chosen because of the availability of subjects with a wide range of demographic background, while retaining a fairly controlled setting.

Demographics

For the Online Video experiment it was possible to match the groups prior to watching the video. Priority was given to matching the number of programmers in each group, and to having an equal gender mix. This was not possible in the Directly Observed Robot experiment, however Table 5.6 shows the groups were nevertheless well-balanced.

Table 5.6: Directly Observed Robot Experiment: Demographics of Participant Groups ($N = 55$). Group One is the control group with the monitor turned off, hence without access to the ABOD3, and Group Two is the treatment group with access to the debugger (Wortham, Theodorou and Bryson, 2017a; Wortham, 2018).

Demographic	Group One ($N = 28$)	Group Two ($N = 27$)
Mean Age (yrs)	48.0	40.0
Gender Male	10	10
Gender Female	18	17
STEM Degree	5	9
Other Degree	11	8
Ever worked with a robot?	7	6
Do you use computers?	20	22
Are you a Programmer?	6	5

Results

The primary results obtained from the experiment are outlined in Table 5.7. Similar to the previous experiment, there is a significant improvement ($p < 0.05$) in the accuracy of the participants' mental models of the robot, when they had access to ABOD3. Unlike the previous experiment, there is no statistical significant difference between the two groups on the question *Is robot thinking?*. However, unique to this experiment, significantly more participants in Group 2 reported that they understand what the robot is trying to do.

Table 5.7: Directly Observed Robot Experiment: Main results showing a significant improvement ($p < 0.05$) in the accuracy of the participants' mental models, when they had access to ABOD3. Moreover, significantly more participants in Group 2 reported that they understand what the robot is trying to do (Wortham, Theodorou and Bryson, 2017a; Wortham, 2018).

Result	Group One	Group Two
Is thinking (0/1)	0.46 (σ=0.50)	0.56 (σ=0.50)
Intelligence (1-5)	2.96 (σ=1.18)	3.15 (σ=1.18)
Understand objective (0/1)	0.50 (σ=0.50)	0.89 (σ=0.31)
Report Accuracy (0-6)	1.89 (σ=1.40)	3.52 (σ=2.10)

5.3.3 Discussion

Across both experiments, there is a significant correlation between the accuracy of the participants' mental models of the robot, and the provision of the additional transparency data provided by ABOD3. We have shown that a *real-time display of a robot's decision making produces significantly better understanding of that robot's intelligence.* Users of transparent systems are able to calibrate their expectations and understand better the functionalities offered by the system. Thus, we argue that transparency is a safety consideration, as otherwise users of robotics systems may have unrealistic expectations from them. Comments received by participants indicate that in the absence of an accurate model, environmental cues and possibly previous knowledge of robots are used to help create a plausible narrative to guide any interactions with the robot.

While there is no significant difference in perceived robot intelligence between the two groups in each experiment, the data indicates a slightly higher level of perceived intelligence —especially when the robot was directly observed. This may reflect a society-wide uncertainty over the definition of the term *intelligence*, rather than any cognitive assessment. The relatively large standard deviations for intelligence in Tables 5.5 and 5.7 provide some evidence of this uncertainty. Another potential reason—which is not exclusive of the other—is that due to the media influences discussed in the previous chapter, people had higher expectations from an object called 'robot'. When the robot encountered appeared to be non-anthropomorphic, moving around and blinking lights, some may had been disappointed or even that it is under-performing. Access to ABOD3 allows the user to see that the robot has more than one goals, multiple means to achieve those goals, and that it performs action selection by reacting to internal and external changes.

In the first, Online Video experiment, the question *why does the robot stop every so*

often? was found to be ambiguous. Some experiment subjects understand this to mean every time the robot stops to scan its environment before proceeding, and only one person took this to mean the sleep behaviour of the robot that results in a more prolonged period of inactivity. The question was intended to refer to the latter, and was particularly included because as `Sleep` Drive is highlighted by ABOD3 each time the robot is motionless with no lights flashing. However, only one member of Group Two identified this from the video. Due to this ambiguity, the data related to this question was not considered further in this dataset. This question was subsequently refined in the second, Directly Observed Robot experiment to 'Why does it just stop every so often (when all its lights go out)?' and included in the analysis.

Participants in the online experiment with access to ABOD3 perceived the robot to be *thinking*; this result was not replicated in the second experiment. (Wortham, 2018) considers the results counter-intuitive. An increase of transparency should reduce anthropomorphic cognitive descriptions. In workshops on ABOD3 given in continental Europe, a frequent question was *how we define thinking*. Unfortunately, we did not record the nationalities and native languages of the participants, however, it is a fair assumption that due to the nature of the recruitment (online and mailing lists), participants in the Online Experiment were less likely to be native speakers. We hypothesise that non-native speakers, who took part in the first experiment, may have attributed an 'extended' definition to the word *think* by associating to the word *processing*. As discussed in chapter 2, an aspect of conciousness is the ability to perform 'real-time search'. Subjects who had access to ABOD3, where able to see that in the R5 there are not only multiple possible actions available to be taken, but also that the robot actively switches between them to satisfy different goals. A multi-cultural study on how people perceive and attribute anthropomorphic elements across different societies could potentially prove (or disprove) this hypothesis.

5.4 ABOD3-AR for End-users Transparency

The prior two studies demonstrate that ABOD3 can be successfully used to provide transparency information in order to help non-expert users improve their mental models. However, albeit its success, ABOD3 in its current form has a serious disadvantage: it requires a computer to run ABOD3 in addition to the robot. Mobile 'smartphones' have been becoming increasingly popular. Thanks to their powerful System-On-Chip processors, they are able to run demanding applications. We have therefore the developed of ABOD3-AR, a smartphone version of ABOD3.

ABOD3-AR provides real-time transparency to intelligent agents that use the Instinct planner. It combines the debugging and graphical visualisation technology of ABOD3 with a modern Augment Reality User Interface. Moreover, as it is tuned for the smaller screens phones have, it presents information at a higher level of abstraction than ABOD3 does by default. Chapter 4 provides additional information on the UI and technologies behind ABOD3-AR.

We[4] ran a third user study to investigate the effects of ABOD3-AR. Unlike the previous studies, where the principle aim was to investigate if transparency helps end users improve their mental models, i.e. helps them understand the functionalities of a robot. In this study, our focus was to investigate the overall effects of transparency in the perception, and hence, on their mental models.

In this section, we provide an overview of this study, present our results, and then discuss how ABOD3-AR is an effective alternative to ABOD3. We argue that our results demonstrate that the implementation of transparency with ABOD3-AR increased not only the trust towards the system, but also its likeability.

5.4.1 Experimental Design

The experimental setup is similar to the one we used in our prior studies. The R5 robot, running Instinct and the same plan as in previous studies, was used. The robot was placed in a small den with random objects, e.g. a plastic duck. The participants were asked to observe the robot and the answer our questionnaires. Experimental subjects in our Control Group did not have access to ABOD3-AR, while participants in our Treatment Group had access to the app. We supplied phones to participants with the application pre-installed to avoid any inconveniences.

However, unlike the previous two studies, the *Godspeed questionnaire* by Bartneck et al. (2009) was used to measure the perception of an artificial embodied agent with and without access to transparency-related information. The questionnaire uses a Likert scale of 1 to 5 for 25 questions arranged into 5 groups, seen in Table 5.8, giving overall scores for *Anthropomorphism*, *Animacy*, *Likeability*, *Intelligence*, and *Safety*. The reason we decided to switch to the Godspeed Questionnaire is to be able to both investigate participant mental models more widely and facilitate comparisons with the future work of others.

[4]Results from this section appear in Rotsidis A., Theodorou A., and Wortham R.H., 2019. Robots That Make Sense: Transparent Intelligence Through Augmented Reality. *1st International Workshop on Intelligent User Interfaces for Algorithmic Transparency in Emerging Technologies*, Los Angeles, CA USA. I designed the study and analysed the results shown here.

Table 5.8: ABOD3-AR Experiment: The Godspeed Questions, categorised based on the perception they are measuring Bartneck et al. (2009). N is the number of questions in each category.

Group	Questions
Anthropomorphism ($N = 5$)	Fake/Natural, Machinelike/Humanlike, Unconscious/-Conscious, Inconscient/Conscient, Moving Rigid/Elegant
Animacy ($N = 6$)	Dead/Alive, Stagnant/Lively, Mechanical/Organic, Artificial/Lifelike, Inert/Interactive, Apathetic/Responsive
Likeability ($N = 4$)	Dislike/Like, Unfriendly/Friendly, Unpleasant/Pleasant, Awful/Nice
Perceived Intelligence ($N = 5$)	Incompetant/Competant, Ignorant/Knowledgeable, Irresponsible/Responsible, Unintelligent/Intelligent, Foolish/Sensible
Safety ($N = 3$)	Anxious/Relaxed, Agitated/Calm, Quiescent/Surprised

In addition to the standard Godspeed questionnaire, participants were asked to answer the questions shown in Table 5.9. The first two questions measure the emotions of the participants. Questions 3-6 provide additional measurements of the perception of the robot. Question 4 was added to test difference in perceiving the robot as *thinking* between the two groups. In addition, Question 6 was added to test the claims from chapter 3 that transparency increases trust. Finally, the last question is included to gather additional empirical evidence on the effectiveness of ABOD3-AR as a means to provide real-time transparency.

A second questionnaire with questions specifically regarding the app, seen in Table 5.10, was handed to Group 2. The primary focus of this survey is to gather feedback specifically fo the application.

5.4.2 Participants Recruitment

We ran the study at the *The Edge Art Centre*, located at the main campus of the University of Bath, which was holding an interactive media exhibition, titled 'The Fantastical Multimedia Pop-up Project', over August 2018. We exhibited the R5 robot and ABOD3-AR for two weeks. This location was chosen because of the availability of subjects with a wide range of demographic background, while retaining a fairly controlled setting.

Table 5.9: ABOD3-AR Experiment: Additional questions given to all participants.

Ref. No.	Question	Response
1.	Sad/Happy	1-5
2.	Bored/Interested	1-5
3.	Do you think the robot is performing the way it should be?	Yes/No
4.	Is the robot thinking?	Yes/No
5.	Would you feel safe to interact with the robot (for example putting your hand in front of it))?	Yes/No
6.	Would you trust a robot like this in your home?	Yes/No
7.	In your own words, what do you think the robot is doing?	Free Text

Table 5.10: ABOD3-AR Experiment: Additional survey, regarding ABOD3-AR, given only participants in Group 2.

Ref. No.	Question	Response
1.	How would you rate the mobile app?	1-5
2.	How easy was to understand the robots current instructions?	1-5
3.	How good was the tracking of the robot?	1-5
4.	You encounter a robot in a hotel-lobby. How likely are you to use this app?	1-5
5.	How likely are you to use this app in a human-robot collaborative work environment?	1-5
6.	How likely are you to use this app in a human-robot collaborative domestic environment?	1-5
7.	Was the text on the screen clear and stable enough to read (Yes/No)?	Yes/No
8.	How can we improve the app ?	Free Text

5.4.3 Results

Since we are comparing the population means of only two groups, t-tests were used to quantitatively analyse the results of the Godspeed questionnaire. Each of four binary questions was tested with either Fishers exact or Chi-square tests, depending on the sample size. The mean of the ratings given at each question was calculated for the application feedback.

5.4.4 Demographics

Table 5.11 shows the demographics for our 45 participants. Both groups have a similar distribution of participants. Similar to the previous in-person study, the only filtering performed was to discard any data provided by minors.

Table 5.11: ABOD3-AR Experiment: Demographics of Participant Groups ($N = 55$). Group One($N = 23$) is the control group without access to the application and Group Two($N = 22$) is the treatment group with access to a phone running ABOD3-AR.

Demographics	Group 1 ($N = 23$)	Group 2 ($N = 22$)
Average Age Group	36-45	36-45
Gender Male	10	9
Gender Female	12	13
Gender Agender or N/A	1	0
Work with computers regularly (Yes) ?	20	21
Are you a software developer (Yes) ?	5	1
Do you have a background in STEM (No) ?	18	21

Godspeed Questionnaire

Individuals who had access to ABOD3-AR were more likely to perceive the robot as *alive* ($M = 3.27$, $SD = 1.202$) compare the ones without access to the app; *t(43) = –0.692* and $p = 0.01$. Moreover, participants in the no-transparency condition described the robot as more *stagnant* ($M = 3.30$, $SD = 0.926$) compare to the ones in Group 2 ($M = 414$, $SD = 0.710$) who described the robot as *Lively*; *t(43)* = –3.371, $p = 02$. Finally, in the ABOD3-AR condition, participants perceived the robot to be *friendlier* ($M = 3.17$, $SE = 1.029$) than participants in Group 1 ($M = 3.77$, $SE = 0.869$); ; *t(43)* = –2.104, $p = 041$. No other significant results were reported. These results are shown in Table 5.12; a complete set of all the results gathered is found in Appendix C.

Perception of Performance

Table 5.13 shows the results for the question "Do you think the robot is performing the way it should be?" A Fisher Exact test showed that there is no significant difference in the responded of the two populations; $p = 0.6078$

Perception of Thinking

Only 41 from our participants answered the question *Is the robot thinking?*. To test the null hypothesis that access to ABOD3-AR does not increase the perception of *thinking*,

Table 5.12: ABOD3-AR Experiment: Means (SD) of the ratings given by each group at various questions. The results show that participants in Group 2 perceive the robot as significantly more *alive* if they had used ABOD3-AR compare to participants in Group 1. Moreover, participants in the no-app condition described the robot as more *stagnant* compare to the ones in Group 2. Finally, in the ABOD3-AR condition, participants perceived the robot to be *friendlier* than participants in Group 1. A complete set of results is shown in Appendix C.

Question	Group 1 ($N = 23$)	Group 2 ($N = 22$)	p-value
Dead - Alive	2.39 (σ=0.988)	3.27 (σ=1.)	**0.01**
Stagnant - Lively	3.30 (σ=0.926)	4.14 (σ=0.710)	**0.02**
Machinelike - Humanlike	1.87 (σ =1.014)	1.41 (σ =0.796)	0.97
Mechanical - Organic	1.91 (σ =1.276)	1.45 (σ =0.8)	0.1
Artificial - Lifelike	1.96 (σ =1.065)	1.95 (σ =1.214)	0.99
Inert - Interactive	3.26 (σ =1.176)	3.68 (σ =1.041)	0.21
Dislike - Like	3.57 (σ =0.728)	3.77 (σ =1.02)	0.4
Unfriendly - Friendly	3.17 (σ=1.029)	3.77 (σ0.869)	**0.04**
Unpleasant - Pleasant	3.43 (σ0.788)	3.77 (σ1.066)	0.23
Unintelligent - Intelligent	3.17 (σ=0.937)	3.14 (σ=1.153)	0.92
Bored - Interested	3.80 (σ=0.834)	4.19 (σ=0.680)	0.11
Anxious - Relaxed	4.15 (σ=0.933)	3.81 (σ=1.167)	0.30

Table 5.13: ABOD3-AR Experiment: The contingency table for the answers given to the binary question *Do you think the robot is performing the way it should beg?* ($N = 45$). There is no significant difference between the two groups; $p = 0.6078$.

Result	Group 1 ($N = 23$)	Group 2 ($N = 22$)
Yes	20	21
No	3	1

we run a Chi-square test in the contingency table shown in Table 5.14. ABOD3-AR does not increase the perception of thinking; $\chi^2 = 0.0232$, $p = 0.878828$, and $DF = 1$.

Table 5.14: ABOD3-AR Experiment: The contingency table for then answers given to the binary question "Is the robot thinking?" ($N = 41$). There is no significant difference between the two groups with $\chi^2 = 0.0232$, $p = 0.878828$, and textitDF = 1. In curvy brackets the expected cell totals and in square brackets the chi-square statistic for each cell].

Result	Group 1 ($N = 21$)	Group 2 ($N = 20$)
Yes	11 (10.76) [0.01]	10 (10.24)[0.01]
No	10(10.24)[0.01]	10 (9.76)[0.01]

Perception of Safety

Table 5.15 shows the results gathered for the question "Would you feel safe to interact with the robot (for example putting your hand in front of it)?" There is no significant interaction between the two groups; $p = 1$.

Table 5.15: ABOD3-AR Experiment: The contingency table for then answers given to the binary question *Do you think the robot is performing the way it should be?* ($N = 45$). There is no significant difference between the two groups; $p = 0.6078$.

Result	Group 1 ($N = 20$)	Group 2 ($N = 21$)
Yes	19	20
No	1	1

Perception of Trust

Unfortunately, only 20 per group answered the question "Would you trust a robot like this in your home?" We run a Chi-square test on our results (Table 5.16) which returned back $\chi^2 = 4.2857$, $p = 0.038434$, $DF = 1$, demonstrating that the results are significant. Access to ABOD3-AR helps users increase their trust to the machine.

Table 5.16: ABOD3-AR Experiment: The contingency table for then answers given to the binary question *Would you trust a robot like this in your home?* ($N = 41$). There is no significant difference between the two groups with $\chi^2 = 4.2857$, $p = 0.038434$, $DF = 1$ In curvy brackets the expected cell totals and in square brackets the chi-square statistic for each cell].

Result	Group 1 ($N = 21$)	Group 2 ($N = 20$)
Yes	11 (14)[0.64]	17 (14)[0.64]
No	9 (6)[1.5]	3 (6)[1.5]

Empirical Results

Unlike the previous studies, we did not rate the answers given to the free-text question "In your own words, what do you think the robot is doing?" Randomly-picked answers are included bellow. Note, multiple participants in Group 1 referred to the robot as a 'he', while none of the Group 2 participants did.

Group 1:

- [the robot is] Trying to build a memory of the distance between itself and the objects to judge its own location in space.

- [the robot is] Processing Data.
- [the robot is] Random.
- [the robot] is actively looking for something specific. At some points he believes he has found it (flashes a light) but then continues on to look.
- [the robot is] Taking pictures of the objects.
- [the robot is] Occasionally taking pictures.
- He is looking for something.

Group 2:

- [the robot is] Exploring its surroundings and trying to detect humans.
- [the robot is] Roaming detecting objects and movement through sensors.
- [the robot is] The robot likes to scan for obstacles, humans and find new paths to follow it can understand animals and obstacles.
- [the robot is] imitating commands, responding to stimuli.
- [the robot is] registering programmed behaviours and connecting it to it surroundings.
- [the robot 's] movement looks random I would say it is using sensors to avoid the obstacles.
- [the robot is] Occasionally taking pictures.

Application Feedback

Group 2 was asked to fill an additional survey to evaluate their experience of using the app. Table 5.17 shows the results to the application feedback survey. In all questions, participants rated the application with a mean of 4, except for Question 2 (see Table) where it got the mean value of the ratings is 4.5. Questions 5 and 6 had the same means and percentages of people who answered positively. However, it worth noting that that only 1 participant answered with a rating of 1 in Q.5, but 3 in the other.

No.	Result	Mean	% of positive answers ($N = 20$)
1.	How would you rate the mobile app?	4	71%
2.	How easy was to understand the robots current instructions?	4.5	86%
3.	How good was the tracking of the robot?	4	61%
4.	You encounter a robot in a hotel-lobby. How likely are you to use this app?	4	66%
5.	How likely are you to use this app in a human-robot collaborative work environment?	4	62%
6.	How likely are you to use this app in a human-robot collaborative domestic environment?	4	62%
7.	Was the text on the screen clear and stable enough to read (Yes/No)?		N/A 90%

Table 5.17: ABOD3-AR Experiment: Means for questions regarding the overall experience of using the application, answered by Group 2 participants ($N = 20$). Scores above the neutral score of 3 are considered as positive. The last question is a binary one, hence, no mean was calculated. Results indicate that using ABOD3-AR is an overall positive experience, with means of 4+ in all questions.

5.4.5 Discussion

Mental Models & Perception

The answers, found in section 5.4.4, from our participants in the question "In your own words, what do you think the robot is doing?" demonstrate that ABOD3-AR is an effective mean of producing a significantly better understanding of what a robot's functionalities and capabilities are. Interestingly, some of the participants in our control group, without access to ABOD3-AR, refered the robot as a 'he'.

We found statistical significant difference (*p-value* < 0.05) in three Godspeed questions: *Dead/Alive*, *Stagnant/Lively*, and *Unfriendly/Friendly*. The R5 has its wires and various chipsets exposed (see chapter 4). Yet, participants with access to ABOD3-AR were more likely to describe the robot as *alive*, *lively*, and *friendly*. All three dimensions had mean values over the 'neutral' score of 3. Despite not significantly higher, there was an increase attribution of the descriptors *Interactive* and *Pleasant*; again both with values over the 'neutral' score. At first glance, the results suggest an increase of anthropomorphic—or at least biologic—characteristics. In addition, transparency decreased the perception of the robot being *Humanlike* and *Organic* (*p-value* $<= 0.1$); both characterisations have means below the neutral score. We view this as a positive outcome in light of the EPSRC Principles of Robotics (Boden et al., 2011) and the

discussions of chapter 3.

We hypothesise that access to the transparency display, makes the constant selection and performance of actions visible to the server. Action selection—or at least consideration—takes place even when the robot is already performing a lengthy action, e.g. moving, or when it may appears 'stuck', e.g. it is in `Sleep` drive to save battery. These results also support that a sensible implementation of transparency, in line to the principles set in chapter 3, can maintain or even improve the user experience and engagement. This argument is further supported by the marginally more positive feelings (questions *Happy* and *Interested*) expressed by our treatment group. Wortham and Rogers (2017) demonstrates similar results, with participants in the transparency treatment having slightly more positive feelings (3.8 mean for Group 1 and 4.19 mean for Group 2). Note, that as there is already a high baseline, it is hard to have a substantial increase here. An explanation for the high levels of *Interest* is that embodied agents —unlike virtual agents— are not widely available. Participants in both groups may have been intrigued by the ideal of encountering a real robot. Nonetheless, our findings indicate that transparency does not necessary reduces the utility or 'likeability' of a system. Instead, the use of a transparency display can increase both.

Our results also suggest an increase of trust, when the user is in the transparency condition. There was a statistical significant difference between the number of people who answered *Yes* in the question "Would you trust a robot like this in your home?" between the two groups. Users with ABOD3-AR were more likely to have a robot like the R5 at home. Further work that includes a more detailed questionnaire is required to explore this. Our hypothesis is that some of their concerns were addressed; for example, subjects with ABOD3-AR could see that the robot does not have any audiovisual recording equipment that could compromise the privacy of its users.

On the contrary, there was no significant difference in the perception of safety between the two groups. Both groups overwhelming answered *Yes* in the question "Would you feel safe to interact with the robot (for example putting your hand in front of it) ?" Thus, some participants would feel safe to interact with the robot in a 'neutral' environment, but not feel comfortable having it at their homes. Still, this was expected as the R5 does not have any sharp edges or other threatening-looking characteristics. Moreover, the robot moves at slow speeds, something directly observables, alleviating any concerns for causing damage from an accidental impact. Furthermore, there is no significance difference between the two groups in questions *Anxious/Relaxed*, *Calm/Agitated*, and *Quiescent/Surprised* designed to measure the perceived Safety of the participant.

Finally, there was no a significant main effect of the transparency condition in the question "Is the robot thinking?". In section 5.3.3, a hypothesis was outlined on how cultural differences may affect the perception of the word 'thinking'. Thus, the statistically significant difference found in the pilot online study, which was not replicated in the in-person experiment. As the majority of our sample in this study was mainly native speakers, the results from this study do not disprove our hypothesis.

Application Feedback

Overall, the application was well received by its users. The vast majority of the participants rated the application positively in all user-experience rating questions. Most importantly, our results indicate that participants are likely to use ABOD3-AR in domestic and work environments, if it was available. These results exhibit end user support for implementing transparency —if not a need to. They also add further fuel to the claim that implementing transparency, following the good practices set in chapter 3, as ABOD3 and ABOD3-AR do, can potentially increase the utility of a system.

Future Work

Albeit the significance of our results, there is a drawback in our experimental design: the Godspeed questionnaire covered a very wide scope. Hence, we could only hypothesise in our discussion of the results why the robot was described significantly more 'Alive' and 'Lively' in the transparency condition. A follow up study, with questions focused on the interaction aspect, is necessary to confirm our interpretation. Similarly, we were left with a hypothesis regarding the increased trust, as we did not measure any privacy concerns. Again, a follow up study could focus exclusively on gathering data to support (or disprove) our hypothesis. Finally, despite that participants were encouraged to interact with the robot, the interaction was limited to waving hands and triggering its thermal sensor. Hence, in any follow up studies, either to replicate this or gain insights to the results presented here, the experimental setup would benefit by having interaction directly with the robot in order to satisfy a set goal. Such an experimental setup would likely produce stronger results in the questions of trust and perceived utility.

5.5 Conclusions

In this chapter we first examined the use ABOD3 and transparency in general in teaching and developing AI. Alongside BUNG, the real-time debugger has been integrated into our teaching curriculum. The indicative results presented in this chapter demonstrate the benefits of ABOD3 for students and developers at large. It allows the diagno-

sis and correction of problems in reactive plans that were unforeseen during initial plan creation. Moreover, by making the emergent behaviour of an agent clear, it is easier for a student to understand how the action-selection mechanism works.

Across all three experiments with naive observers, there is a significant correlation between the accuracy of the participants' mental models of the robot and the provision of the additional transparency data provided by ABOD3 and ABOD3-AR. We have shown that real-time visualisation of robot's decision making produces a significantly better understanding of the robot's intelligence, even though that understanding may still include wildly inaccurate overestimation of the robot's abilities. Comments received by participants indicate that in the absence of an accurate model, environmental cues and possibly previous knowledge of robots are used to help create a plausible narrative.

The results from the ABOD3-AR experiment also suggest that an implementation of transparency within the good-practice guidelines set in chapter 3 does not necessary imply a trade-off with utility. Instead, the overall experience can be conceived as more interactive and positive by the robot's end users. Furthermore, participants in the transparency condition reported significantly more trust towards the system. Thus, implementing transparency does not hinder innovation or business interests, but instead it can lead to further adoption and usage of technologies.

Chapter 6

Keep Straight and Carry on

> " Okay, as much as I'm enjoying watching random people's heads fly off, I think we've taken this trolley thing as far as it can go."
>
> Eleanor Shellstropk, *The Good Place, Season 2 - Episode 5*

6.1 Introduction

Autonomous cars are one of the technologies in the transportation domain most followed by the public (Beiker, 2012). Widespread use of them is predicted to reduce accidents, congestion, and stress (Fleetwood, 2017; Litman, 2017). Yet, like with all technological innovations, incidents are bound to happen. Such incidents can be due to error at either the system's side, e.g. malfunctioning, or at the user's side, e.g. misuse or malicious use.

Societies use legislation and regulation to minimise accidents, as well as humans' natural aversion to risk of their own harm, and corporations' desire to limit their financial liability (Bryson and Winfield, 2017; Solaiman, 2017). If we take these motivations and damage minimisation some steps further, we can use AI to influence the outcome of any incident. For this, the autonomous vehicle's decision-making process must be predetermined by a set of rules. These rules can be as simple as applying the breaks to giving control to the driver, who can make a deliberate decision on *what to damage*. The decisions the agent makes in the accident process may lead to outcomes regarded as better or worse, more moral or immoral.

Let us take as an example a car, which has a brake failure near a pedestrian crossing.

Such an incident can lead to an accident with tragic consequences. If driven by a human, the driver, forced by the circumstances, will have to make a moral choice to either risk the lives of the passengers or of any unlucky, random pedestrians. If the driver regularly maintained her car, she can limit her legal liabilities to the state and anyone else affected by the accident. Now, some claim that we have an unprecedented possibility in human history where due to a vehicle's built-in superior perception and computation capabilities, an informed decision can be made either by a human driver or by the car itself (Bonnefon, Schriff and Rahwan, 2016). Others advocate the use of specific ethical frameworks or even allowing the user to 'pick' between options such as utilitarian, deontological, or self-protective ethics (Gerdes and Thornton, 2015; Gogoll and Müller, 2017; Coca-Vila, 2018). Here, it is important to clarify our[1] position: When an artificial agent makes any decision, that decision is performed as an extension of either its manufacturer's or owner's moral agency —as discussed in chapter 2. Hence, the agent itself can not be held accountable, neither for the incident nor for the decision.

The above scenario elucidates the 'trolley dilemma' problem; what action should an autonomous car take when faced with two morally salient options? Should the car hit the elderly person in the right lane or the young child in the left lane? While a such scenario is unrealistic and improbable (Goodall, 2016), its mere possibility could result to unprecedented societal disruption. A fundamental cornerstone of modern-age democracies is that all citizens are equal in the eyes of the law. Considering preferential treatment based on a 'social value' determined by demographic characteristics could result to propagation of such exceptions throughout a society. It would recall the racial segregation laws we left behind. Furthermore, it would require a massive re-factoring of our data privacy legislation to allow cars, made by private corporations, to access government-held information without our explicit consent. Finally, it would be computational intractable to consider *all* possible outcomes of an action, e.g. property damage could result leaving a town without electricity —assuming that all information is somehow available and accessible.

Yet, it is one of the few morally-salient AI dilemmas which has grabbed the attention of many stakeholders; policy makers, media and public, we believe that this paradigm is still uniquely valuable for exploring several critical research questions in human-computer interactions and expanding upon the work presented so far in this dissertation. The research questions include: 1) how do our perceptions of a decision-making agent and their decision, differ dependent on whether the agent is another human or an

[1] I provided the original idea of and the resources to conduct this research, which was then run with Holly Wilson. Joanna J. Bryson provided advice. Please refer to chapter 1 for an in-depth discussion of individual contributions.

artefact; 2) how does an implementation of transparency regarding the agent's 'moral code' impact how its choices are perceived; and 3) how does the methodology used to present such 'moral dilemma' scenarios to the public, impact their preferences and perceptions. These questions build upon the work presented in the previous chapter, where there was a focus on the mental models of the system as a whole and not of specific actions taken by the system.

To answer these questions, after receiving ethical approval from our department, we used a VR Simulator to run a study with 52 participants over 3 groups. Our results demonstrate the importance of the methodology used to gather data in moral dilemma experiments. We have shown, in conflict with results and claims made by other studies, with qualitative and quantitative results, that there is a desire to use random instead of socio-economic and demographic characteristics in moral decisionmaking. Furthermore, we show that the use of transparency makes the agent appear to be significantly less anthropomorphic, but also to be acting in a more utilitarian way. Moreover, the results indicate, consistent with previous research, that we find it harder to forgive machinelike intelligent systems compared to humans or even more anthropomorphic agents. Finally, our results validate our previous claim that transparency does significantly help naive users to calibrate their mental models.

In this chapter, we initially discuss the research questions and relevant work that motivated our research. Next, we present our technological contributions, succeeded by our experimental design. Subsequently we present and discuss our results section. We conclude our discussion with future work related to this study.

6.2 Research Considerations and Motivation

In this section, we outline in turn each of the considerations that influenced our work and the research questions expressed in the previous section.

6.2.1 Perceived Human versus Machine Morality

There are many circumstances in which decision-making intelligent agents are replacing humans. Yet we have not sufficiently established how this shift in agent-type impacts our perceptions of the decision and decision-maker. The research gap is especially large in the context of morally salient decision-making. There are indications that we both inaccurately assimilate our mental model of humans with intelligent agents, and have separate expectations and perceptions of intelligent agents which often lack accuracy (Turkle, 2017).

Research also suggests people perceive artificial agents as more objective and less prone to have biases than human decision makers. For example, people were found to be more likely to make decisions inconsistent with objective data when they believed the decision was recommended by a computer system than by a person (Skitka, Mosier and Burdick, 1999). Similarly, in a legal setting, people preferred to adhere to a machine advisor's decision even when the human advisor's judgement had higher accuracy (Krueger, 2016). In the context of an autonomous vehicle, higher attributions of objectivity and competence, could result in end-users feeling more content with decisions, than they would be had the decision been made by a human driver.

6.2.2 Inaccurate Mental Models

chapter 3 discusses how we form mental models for intelligent agents, based on past experiences, expectations, and physical characteristics. Moreover, if we encounter a different but physically similar agent, we can be mislead into believing that they share the same action-selection system and, therefore, operate with the same moral framework, i.e. have the same goals and can perform behaviours to fulfil those goals. We may anthropomorphise and have false expectations about the behaviour of the agent. Therefore, we end up creating inaccurate mental models—something shown through user studies in the previous chapter—of the agent.

Inaccurate mental models can lead to sub-optimal interactions with the agent or even to safety concerns due to disuse or misuse (Lee and See, 2004). For us to make informed choices about usage, we require accurate mental models of the agents. When agents make moral-worthy decisions, our models should include the moral framework that they were prescribed with. Transparency, as argued in this dissertation, can help us calibrate our mental models, and, therefore, it should help us understand the moral framework prescribed to an agent. Yet, there are no previous studies which explicitly investigate how transparency can help users understand the moral framework of an agent and the impact of such an understanding.

6.2.3 Perceived Moral Agency and Responsibility

We make moral judgements and apply moral norms differentially to artificial than human agents. For example, in a mining dilemma modelled after the trolley dilemma, robots were blamed more than humans when the utilitarian action was not taken (Malle et al., 2015). This utilitarian action was also found not only to be more permissible with the robot than the human, but also expected.

This could have implications for the moral framework we should program into machines —which should not necessarily be equal to the frameworks we expect from humans. This is supported later work on assessing how attribute responsibility differently on human and machines (Li et al., 2016). After reading an autonomous car narrative, participants assigned less responsibility to an autonomous vehicle car at fault than to a human driver at fault. Our over-identification, as discussed in length in chapter 2, with such systems creates a *moral confusion* about the moral status of these human-made objects. Hence, to ensure societal stability by not attributing any responsibility or —worse— accountability to the artefact we ensure that through both engineering and socio-legal solutions we always held legal persons responsible.

6.2.4 Understanding Moral Preferences

Autonomous cars *could*, but not necessarily *should*, be programmed with behaviours that conform to a predetermined moral framework (such as utilitarian, deontological, and others) or with a normative framework. The *Moral Machine* an online experiment by Shariff, Bonnefon and Rahwan (2017), where participants make a choice who an autonomous vehicle should sacrifice based on the socio-economic and demographic values of any passenger(s) and pedestrian(s) involved in the incident. At first the experimental data were used to 'crowd-source' preferences for the construction of ethical frameworks for autonomous vehicles, but later they have been used to investigate cultural differences (Awad et al., 2018).

While the Moral Machine uses a moral dilemma where the subject dictates to an 'autonomous' vehicle what moral choice to make, it does not gather any data on how people perceive the choices of others —or even that the car has to a make a choice in the first place. We can run moral-dilemma experiment with a similar setting, i.e. use an autonomous vehicle making a moral choice to investigate our perception of intelligent systems when they perform actions of moral worth. However, unlike the Moral Machine (and other Trolley Problems), the subject would not be making the moral choice, but rather, the participant would be 'forced to live through one'.

We believe that by forcing a choice made by either another human or an artificial agent into our experimental subjects and then evaluate their responses, we can understand the moral intuitions that guide our attribution of responsibility. For this, we developed an autonomous vehicle simulator, which we present present in the next section. We are using Virtual Reality (VR) technology to increase the impact of the decision to our experimental subjects and, therefore, get results closed to a real-life enactment of our simulated scenario (Pan and Slater, 2011). Finally, by using the same simulator we will

like to investigate how making the mechanical nature of the system explicitly visible through post-incident transparency alters our perception of intelligent systems and of the choice they make.

6.3 VR Autonomous Vehicle Moral Dilemma Simulator

When viewing pictures or reading narratives, as moral-dilemma experiments have been conducted, there is less emotional elicitation than in the equivalent real-life situations (Gorini et al., 2010). VR has been shown to have high ecological validity, provoking true to life emotions and behaviours (Rovira et al., 2009; Sütfeld et al., 2017). Importantly, people have been found to make different decisions for moral dilemmas in immersive VR simulation than in desktop VR scenarios (Pan and Slater, 2011). More specifically, immersive VR induces an element of panic and results to a less utilitarian decision making by the experimental subject. Hence, responses in a VR version of a trolley problem or of the Moral Machine may be different —if not more realistic— compared to the narrative versions. We developed a VR simulator, optimised for the Oculus Rift platform, in the popular game engine Unity. The engine was chosen due the wide range of free available assets.

6.3.1 The Simulator

The autonomous vehicle simulator, a screenshot is seen in Figure 6-1, is designed so that participants are seated in the driver's seat of a car. The car has detailed interior to increase realism and therefore immersion. The user, similar to all other VR simulations, can turn her head around and look out of the car from the front, side, and rear windows.

The autonomous vehicle, positioned on the left hand lane as the experiments took place in the UK, speeds through a pre-scripted 'railway' track through a city environment. Tall buildings, trees, benches, streetlights, bins, and bus stops can be seen on either side of the street. The car starts decelerating when it approaches a zebra crossing. On the zebra crossing there are two non-playable characters (NPCs) crossing the road. Due to its speed, the vehicle makes a choice to either continue on a straight line or change lanes. The AV will always run over one of the two NPCs and stop a few meters past the crossing. During the collision between the car and one of the NPCs, the NPC screams. The passenger can turn around and look at the 'dead body'. After discussing our experimental setup with our ethics officer [2], we decided against including blood effects

[2]The ethics officer reviewed our experimental setup, data governance plan, and consent forms. The officer did not deemed necessary to refer the study neither to faculty's nor university's ethics committees.

Figure 6-1: Participant is seated as passenger. On the crossing ahead there is a pair who differ in body size.

around the dead body. We also avoided any intentional resemblance of the NPCs to real-world persons. The AV's action-selection mechanism makes the decision to keep straight or not based on the protected characteristics of the NPC, i.e. its demographic background, as described further below.

6.3.2 Preference Selections

Bonnefon, Schriff and Rahwan (2016) gave participants narratives of different dilemmas. Participants showed a general preference to minimise casualty numbers rather than protecting passengers at all costs. However, people no longer wished to sacrifice the passenger when only one life could be saved, an effect which was amplified when an additional passenger was in the car such as a family member. Shariff, Bonnefon and Rahwan (2017) ran a massive online data-collection experiment, called *The Moral Machine*, to determine a global moral preference for autonomous vehicles. In the Moral Machine, users select between two options which were represented by a 2D, pictorial, birds eye view as a response to 'What should the self-driving car do?'. The Moral Machine is a multi-dimension problem; protected characteristics (e.g. race), educational/socio-economic background (e.g. occupation), or even the legality of switching lanes are taken into consideration.

We decided instead of replicating all of the dimensions used in the Moral Machine,

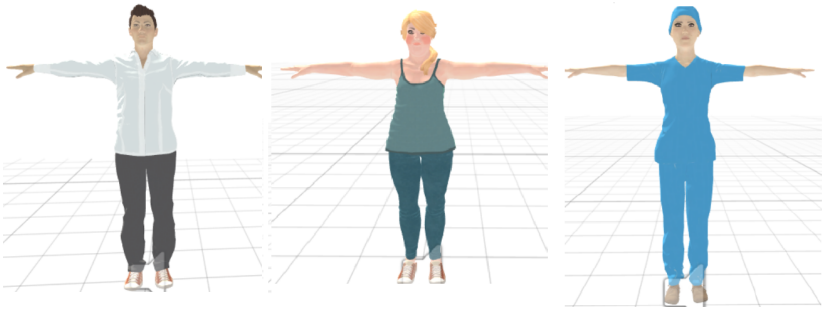

Figure 6-2: Examples of characters used. From left to right: Asian slim businessman, Caucasian non-athletic unemployed female, and asian female athletic slim medic.

to only use a selection of them. We made this decision as we are not measuring the preferences of characteristics the participants would like to be saved, but rather the response to their use in their first place. Considering availability of assets and that we do not present a textual description of the scenario to the participant, we picked the three more visible characteristics: race, occupation, sex, and body size.

The race can be Caucasian, Black, and Asian. Occupation includes four representative conditions: a medic to represent someone who is often associated with contribution to the wealth of the community, military to represent a risk-taking profession (McCartney, 2011), businessman or businesswoman as it is associated with wealth, and finally unemployed. The body size (slim or large) is similar to the Moral Machine's athletic/healthy condition. Finally, to further reduce the dimensions of the problem, we used a binary gender choice (female and male).

At each iteration of the moral dilemma, the characters are randomly selected by a list of pre-generated characters with random combinations of the four types of characteristics. The agent makes the decision who to save and who to kill based on the NPC's characteristics, minus its race. Like an expert system, it compares the characteristics in a hierarchical order of importance. Figure 6-3 shows the hierarchy of the three categories of characteristics in order of importance with occupation being the most important and gender the least. In addition, if all characteristics are the same, the agent selects to stay on the same lane instead of changing. This hierarchy of prioritised characteristics constitutes the *moral framework* that *we prescribed to the agent.*

6.3.3 Transparency Implementation

What is effectively transparent varies by who the observer is, and what their goals and obligations are (Bryson and Winfield, 2017; Theodorou, Wortham and Bryson, 2017,

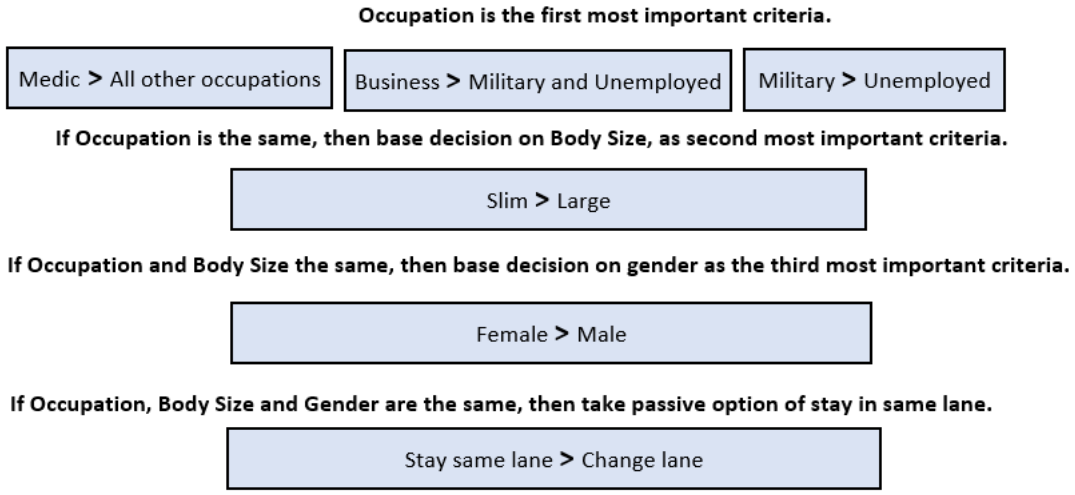

Figure 6-3: The decision-making hierarchy of the three categories of characteristics (occupation, body type, and gender) in order of importance, with first (top-to-bottom order) being the most important. Each row is a single characteristic. If all characteristics are the same, the agent selects to stay on the same lane instead of changing.

and as discussed in chapter 3). Our own studies (discussed in Wortham, Theodorou and Bryson, 2017a,b, and chapter 5 of this dissertation), demonstrate how users of various demographic backgrounds had inaccurate mental models about a mobile robot running a BOD-based planner. Participants in the studies were ascribing unrealistic functionalities, potentially raising their expectations for its intelligence. When the same robot was used with the ABOD3 software providing an end-user transparency visualisation, the users were able to calibrate their mental models, leading to more realistic expectations concerning the functionality of the machine.

Transparency is a safety feature and its implementation could reduce incidents from happening in the first place. However, reacting in time to avoid an incident may not always be possible. Hence, the most important goal of transparency is at least provide sufficient information to ensure at least human accountability (Bryson and Theodorou, 2019). For that, there is a need for *post-incident transparency* to help incident investigators understand the actions of the transparent system (Winfield and Jirotka, 2017).

Considering the fast pace of events, i.e. of the crash between the car and the NPC, alongside that we aim to measure blame, we decided to implement a post-decision transparency. Using a post-incident transparency implementation is in line with other work in the literature. For example, Kim and Hinds (2006) run a study to explore user's understanding of why a robot behaved in an unexpected way. The robot would, post

incident, provide information of what lead it to behave as such. The study demonstrates that users would assign greater blame of the robot and less blame on others when the robots had greater autonomy.

Post-scenario, after the autonomous car has hit one of two pedestrians, a statement is made that: "The self-driving car made the decision on the basis that..." then the reasoning logic is inserted next. For example, if the pair consisted of one medic and another military, the justification will state "Medics are valued more highly than military, business or undefined professions". Whereas, if the pair differ only in gender, it will state: "Both sides have the same profession and body size, however females are valued more highly than males". In this experiment, the transparency only relays aspects of the agent's moral framework. We do not provide any information over the mechanical components of the car, such as whether the brakes were working, the speed of the car, or turning direction.

6.4 Experimental Design

Autonomous cars will be exposed to drivers of all ages, genders, and ethnic backgrounds. Thus, in an effort to reduce the demographic bias often observed in studies performed with undergraduate and postgraduate students, we decided to recruit participants outside the University and its usual pools of subjects. Participant recruitment took place at a local prominent art gallery, The Edge (Bath, UK), where we exhibited our VR simulation as part of a series of interactive installations. Members of the public visiting the gallery were approached and invited to take part to the experiment. They were told the purpose of the experiment is to investigate technology and moral dilemmas in a driving paradigm. The experiment received permission by the ethics officer of the Department of Computer Science and all participants had to fill in a relevant consent form informing them of their rights.

The three such questions that we address include, 1) how do our perceptions of a decision-making agent and their decision differ depending on whether the agent is another human or artificially intelligent; 2) how does an implementation of transparency regarding the agent's 'moral code' impact perceptions, with the expectation of calibration; 3) how does the methodology used to present such 'moral dilemma' scenarios to the public impact their preferences and perceptions. We now outline each in turn with consideration of the current status of research and how the question can be framed within the autonomous car scenario for further investigation.

6.4.1 Conditions

We ran three independent groups; human driver (Group 1), opaque AV (Group 2), and transparent AV (Group 3). We randomly allocated participants to the independent variable conditions. For each condition, both the experimental procedure and the VR Moral Dilemma Simulator was adjusted in the pre-treatment briefing.

6.4.2 Pre-treatment Briefing

Prior to putting on the VR headset, participants were asked to fill a preliminary questionnaire to gather demographic, driving-preference, and social-identity data. In the VR simulator, participants went through the practice scene to familiarise themselves with the controls. All groups shared the same pre-treatment experience.

Different Agent Types

Participants were either informed that they were to be a passenger, in an autonomous vehicle or in a car controlled by a human driver. In the human driver condition, before putting on the headset, they were shown a 'fake' control screen and physical games controller a colleague of the experimenter would 'use' to control the car. At the end of the experiment, participants were debriefed and told that there was no human driver.

Difference in Level of Transparency

Transparency here refers to revealing to the passenger the factors the agent took into consideration for its decision to keep or change lanes, as discussed in section 6.3.3. We would disable the post-incident transparency for members of Group 2 and keep it for Group 3. Albeit not the focus of the study, we asked participants from Groups 2 and 3 to self-evaluate their understanding of how a decision was made. We did not include a similar decision for the human condition, as we wanted to avoid raising suspicions about the deception.

6.4.3 Simulator's Procedure

The VR Simulator has been modified to include a number of 'scenes', i.e. levels or menu screens, to streamline and expedite data collection. There are eight different scenes that the user goes through, seen in Figure 6-4. The *Introduction*, *Practice Scenario*, and *Practice Question* scenes are designed to gradually ease the participant into the virtual reality environment, in order to reduce nausea and introduce the controls of the simulator to the user.

Then the simulator generates 10 instances of the *Scenario* scene. All ten of them follow the same 'gameplay' described in detailed above; the car making a decision and killing one of the two NPCs in the scenario. At each instance of a scenario, the simulator randomly generates the two NPCs by using the pool of demographic characteristics described previously.

After each scenario we use a *Question* scene for data gathering. In a Question scene the participant is in an almost empty square room, in an effort to reduce priming, in which they answer a series of questions relating to the scenario. We decided to gather data within the VR simulator to avoid breaking immersion after each scenario. At the end of the 10^{th} Question scene, the user is presented with the *Finishing* scene containing part of the debriefing.

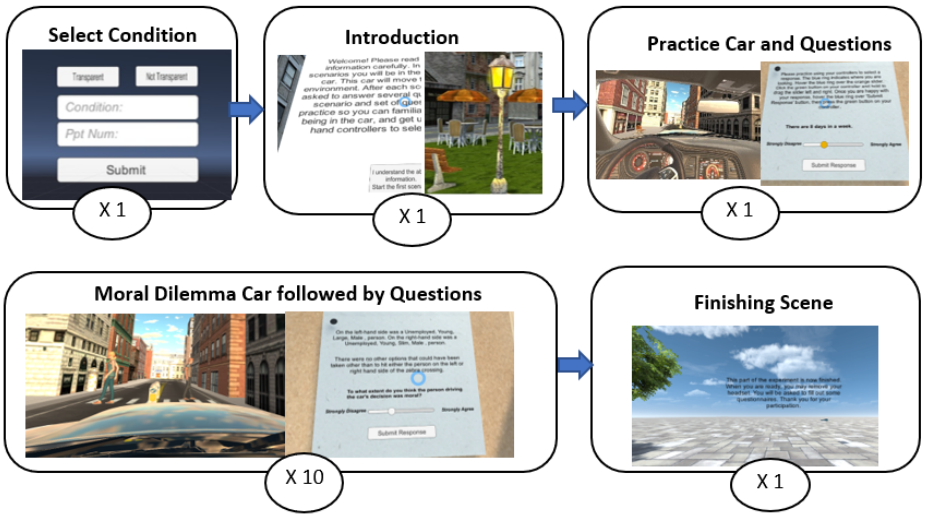

Figure 6-4: The Preliminary Condition Scene is followed by an Introduction Scene, the Practice Car and Practice Questions Scene. Subsequent, the Scenario Scene followed by the Question Scene are cycled through ten times, with each cycle invoking a different moral dilemma. Finally is the Finishing Scene.

6.4.4 Post Simulator

After removing their headsets, participants were requested to fill out a post-simulation questionnaire. The *Godspeed questionnaire* by Bartneck et al. (2009) was used to measure the perception of intelligence and animality. The questionnaire uses a Likert scale of 1 to 5 for 11 questions arranged into 3 groups, seen in Table 6.1, giving overall scores for *Anthropomorphism*, *Likeability*, and *Intelligence*.

In addition to the standard Godspeed questionnaire, participants were asked to answer

Table 6.1: The Godspeed Questions used, categorised based on the perception they are measuring. N is the number of questions in each category.

Group	Questions
Anthropomorphism ($N = 2$)	Machinelike/Humanlike, Unconscious/Conscious
Likeability ($N = 4$)	Dislike/Like, Unpleasant/Pleasant, Awful/Nice, Unkind/Kind
Perceived Intelligence ($N = 5$)	Incompetant/Competant, Ignorant/Knowledgeable, Irresponsible/Responsible, Unintelligent/Intelligent, Foolish/Sensible

the questions shown in Table 6.2. The first two questions measure emotions of the participants. Questions 3-6 provide additional measurements of the perception of the robot. Question 4 was added to test difference in perceiving the robot as *thinking* between the two groups. In addition, Question 6 was added to test the claims from chapter 3 that transparency increases trust.

Table 6.2: Additional questions given to all participants.

Group	Questions
Objectivity ($N = 4$)	Subjective/Objective, Deterministic/Non-deterministic, Unpredictable/Predictable, Intentional/Unintentional
Perceived Morality ($N = 2$)	Immoral/Moral, Unfair/Fair
Human/AV exhibits prejudiced on ($N = 5$)	Race, Gender, Occupation, Body Size, Age
Responsibility ($N = 2$)	Morally Culpable, Blame
I trust the Human/AV ... ($N = 6$)	to act on society's best interest, to act on own best interests, make decisions that I agree with, 's intentions, has integrity, is deceptive

This questionnaire aims to capture the participant's perceptions of the agent controlling the car. It includes dimensions of likeability, intelligence, trust, prejudice, objectivity and morality. Finally, participants in Groups 2 and 3 were asked the binary question "Do you understand how the agent made the decision?".

6.5 Results

A one-way two-tailed ANOVA was conducted on all ordinal Likert scale variables. First, we present the demographics of the participants. Second, the results for the Human Driver condition compare to Opaque Autonomous Vehicle (AV) condition. Next, we compare the Human Driver condition with the Transparent AV condition. Then, we compared the Opaque AV to the Transparent AV condition. We conclude the section with qualitative feedback collected while running the experiment.

6.5.1 Demographics

Table 6.3: Participants' Demographics. Groups were found to be unbalanced for gender and ethnicity.

Variable	Group 1: Human Driver	Group 2: Opaque AV	Group 3: Transparent AV	*X(2)*	*P*
Gender Male	5	5	14		
Gender Female	12	11	4		
Gender Unknown	1	0	0		
				13.89	**0.03**
White	16	14	17		
Asian	0	0	0		
Black/Caribbean	0	0	0		
None/Unknown	2	2	1		
				27.66	**0.001**
16-17	1	1	0		
18-25	2	5	3		
26-35	5	3	6		
36-45	3	3	0		
46-60	6	4	5		
60+	1	0	4		
				15	0.45
Automatic	2	2	2		
Manual	5	6	10		
Both	4	3	4		
None/Unknown	7	5	2		
				10.23	.33
Program	5	6	7		
Do not program	12	10	11		
Unknown	1	0	0		
				5.03	.54
Total Participants	18	16	18		

Imbalance of baseline variables is usually considered undesirable, as the essence of a controlled trial is to compare groups that differ only with respect to their treatment. Others suggest that randomised—unbalanced—trials provide more meaningful results as they compact chance bias(Roberts and Torgerson, 1999). A Chi-squared test of goodness-of-fit was performed to determine frequencies of gender, ethnicity, age, driving preferences and programming experience between the three conditions (see 6.3). Groups were found to be unbalanced for gender and ethnicity. The ethnicity difference between the groups is due to a number of people who did not answer the 'Ethnicity' question; the vast majority of all groups consisted of participants who identified themselves as *white* and no other ethnicities were reported. The unbalance for gender, however, should be taken into consideration during the analysis of the results.

6.5.2 Quantitative Results

Difference in Type of Agent

First, we compared the results from Group 1, Human Driver, to the results of Group 2, Opaque AV. In the comparison all but two associations were found to be non-significant. Table 6.4 shows significant and other noteworthy results. A complete set of results can be found in Appendix D. The autonomous vehicle was perceived to be significantly less 'Humanlike' in the Human condition compared to the AV condition ($M = 2.1$, $SD = 0.96$); $p = 0.001$, $\eta_p^2 = 0.191$. Participants in Human Driver condition found the driver more 'Morally Culpable' ($M = 3.37$, $SD = 0.7$) than participants in Group 2 found the AV ($M = 2.26$, $SD = 1.21$) ; $p = 0.04$, $\eta_p^2 = 0.18$. Although the impact of type of agent was non-significant, medium effect sizes were found for the human driver being perceived as more 'Pleasant' ($\eta_p^2 = 0.105$, $d = 0.69$) and 'Nice' ($\eta_p^2 = 0.124$), $d = 0.75$) than the autonomous car.

Next, we compared the results from Group 1 to Group 3, Transparent Autonomous Vehicle. In the Godspeed Questionnaire we found statistically significant difference in four questions, shown in Table 6.5. Participants in the Human Driver condition described their driver as significantly more 'Pleasant' ($M = 3.0$, $SD = 0.35$) than participants of the Transparent AV condition ($M = 2.35$, $SD = 0.93$) described the AV's behaviour; $t(32) = 2.58$, $p = 0.01$, $\eta_p^2 = 0.183$. In addition, participants in Group 3 perceived the Transparent AV ($M = 2.47$, $SD = 0.87$) as less nice than the subjects in Group 1 ($M = 3.0$, $SD = 0.0$); $t(32) = 2.5$, $p = 0.018$, $\eta_p^2 = 0.163$. Not surprisingly, Group 1 also described the Human Driver as more 'Humanlike' ($M = 3.24$, $SD = 0.97$) compare to Group 3 which described the AV as 'Machinelike' ($M = 1.5$, $SD = 0.92$); $t(33) = 5.42$, $p = 0.0$,$\eta_p^2 = 0.47$. Moreover, participants in the Human Driver condition

Table 6.4: Perceptions based on type of agent; Human Drive v Opaque AV. The results show that participants in Group 2 perceived the AV as significantly more machinelike compared to participants in Group 1. Moreover, participants in the opaque AV condition described the AV as less morally culpable compared to the ones in Group 1. A complete set of results is shown in Appendix D.

Question	N	Mean (SD)	t (df)	p	η_p^2
Godspeed Questionnaire (Scale 1-5)					
Machinelike - Humanlike					
Group 1: Human Driver	17	3.2 (0.97)			
Group 2: Opaque AV	16	2.1 (0.96)			
			3.42 (31)	**0.001**	0.191
Unpleasant - Pleasant					
Group 1: Human Driver	16	3 (=0.35)			
Group 2: Opaque AV	17	2.6 (0.89)			
			1.38 (31)	0.18	0.105
Awful - Nice					
Group 1: Human Driver	17	3 (=0)			
Group 2: Opaque AV	16	2.6 (0.89)			
			1.53 (31)	0.13	0.124
Culpability and Blame					
Morally Culpable (Scale 1-4)					
Group 1: Human Driver	16	3.37 (0.7)			
Group 2: Opaque AV	16	2.56 (1.21)			
			-2.07 (30)	**0.04**	0.18
Blame (Scale 1-4)					
Group 1: Human Driver	15	2.07 (0.7)			
Group 2: Opaque AV	16	2.44 (1.21)			
			-0.94 (29)	0.354	0.020

significantly perceived their driver as 'Conscious' (M = 3.0, SD = 1.17) compare to subjects in Group 3 (M = 1.33, SD = 0.59); $t(33)$ = 5.35, p = 0.0, η_p^2 = 0.464.

Also, we found significant differences between the two groups in 5 additional questions, all of them are shown in Table 6.6. Subjects in the Human Driver condition significantly described their driver as more deterministic in its decision (M = 2.89, SD = 1.11) than participants in the Transparent AV condition (M = 2.0, SD = 1.0); $t(32)$ = 2.43, p = 0.02, η_p^2 = 0.156. Moreover, Group 3 found the Transparent AV more Predictable (M = 4.0, SD = 1.29) compared to participants in Group 1 (M = 3.06, SD = 1.34); $t(33)$ = -2.12, p = 0.04, η_p^2 = 0.12. Interestingly, Group 1 considered the Human Driver's actions significantly more Intentional (M = 3.09, SD = 1.14) than participants

Table 6.5: Perceptions based on type of agent; comparing Human Driver to the the Transparent AV. Participants in the Human Driver condition described their driver as significantly more pleasant than participants of the Transparent AV condition described the AV's behaviour. In addition, participants in Group 3 perceived the Transparent AV as less nice than the subjects in Group 2. Not surprisingly, Group 1 also described the Human driver as more humanlike compare to Group 3 which described the AV as machinelike. Moreover, participants in the Human Driver condition significantly perceived their driver as conscious compare to subjects in Group 3. Additional significant results from the comparison are shown in Table 6.6 and complete set of results is available in Appendix D.

Question	N	Mean (SD)	t (df)	p	η_p^2
Godspeed Questionnaire (Scale 1-5)					
Unpleasant - Pleasant					
Group 1: Human Driver	17	3.0 (0.35)			
Group 3: Transparent AV	17	2.35 (0.93)			
			2.68 (32)	**0.01**	0.183
Awful - Nice					
Group 1: Human Driver	17	3.0 (0.0)			
Group 3: Transparent AV	17	2.47 (0.87)			
			2.5 (32)	**0.018**	0.163
Machinelike - Humanlike					
Group 1: Human Driver	17	3.24 (0.97)			
Group 3: Transparent AV	18	1.5 (0.92)			
			5.42 (33)	**0.000**	0.47
Unconscious - Conscious					
Group 1: Human Driver	17	3.0 (1.17)			
Group 3: Transparent AV	18	1.33 (0.59)			
			5.35 (33)	**0.000**	0.464

in the Transparent AV condition did (M = 1.83, SD = 1.2); *t(33)* = 3.09, p = 0.004, η_p^2 = 0.224. Furthermore, experimental subjects in the Human Driver condition perceived the driver as less morally culpable (M = 2.07, SD = 0.72) and assigned less blame (M = 2.07, SD = 0.7) to the driver than participants in Group 3 (M = 3.05, SD = 1.3 and M = 3.0, SD = 1.298) did to the AV; *t(32)* = -3.89, p = 0.0, η_p^2 = 0.321 and *t(31)* = -2.52, p = 0.02, η_p^2 = 0.169 respectively.

Difference in Level of Transparency

A chi-square test of independence was performed to examine the relation between transparency and understanding of the decision made x2 (1) = 7.34, p = 0.007. Participants in the transparent condition were more likely to report understanding (87.5%) than

Table 6.6: Perceptions based on type of agent; comparing Human Driver to the the Transparent AV. Subjects in the Human Driver condition significantly described their driver as more deterministic in its decision than participants in the Transparent AV condition. Moreover, Group 3 found the Transparent AV more Predictable compared to participants in Group 1. Group 1 considered the Human Driver's actions significantly more Intentional than participants in the Transparent AV condition did. Furthermore, experimental subjects in the Human Driver condition perceived the driver as less morally culpable and assigned less blame to the driver than participants in Group 3did to the AV. Additional significant results from the comparison are shown in Table 6.5 and complete set of results is available in Appendix D.

Question	N	Mean (SD)	t (df)	p	η_p^2
Objectivity (Scale 1-5)					
Deterministic - Undeterministic					
Group 1: Human Driver	17	2.89 (1.11)			
Group 3: Transparent AV	17	2.0 (1.0)			
			2.43 (32)	**0.02**	0.156
Unpredictable - Predictable					
Group 1: Human Driver	17	3.06 (1.34)			
Group 3: Transparent AV	18	4.0 (1.29)			
			-2.12 (33)	**0.04**	0.120
Intentional - Unintentional					
Group 1: Human Driver	17	3.09 (1.14)			
Group 3: Transparent AV	18	1.83 (1.2)			
			3.09 (33)	**0.004**	0.224
Culpability and Blame					
Morally Culpable (1-4)					
Group 1: Human Driver	16	3.37)			
Group 3: Transparent AV	18	3.05 (1.3)			
			-3.89 (32)	**0.00**	0.321
Blame (1-5)					
Group 1: Human Driver	15	2.07 (0.7)			
Group 3: Transparent AV	18	3.0 (1.28)			
			-2.52 (31)	**0.02**	0.169

(43.75%) (see 6-5.

We also conducted a series of independent samples t-tests on all ordinal Likert scale variables to compare the results of the opaque and transparent AV conditions (see Table 6.7). Three significant effects were found. The autonomous car was perceived by participants in the non-transparent AV condition to be more 'Humanlike' (M = 3.2, SD = 0.97) than subjects in Group 3 (M = 2.1, SD = 0.96); *t(32)* = -2.1, p = 0.04,

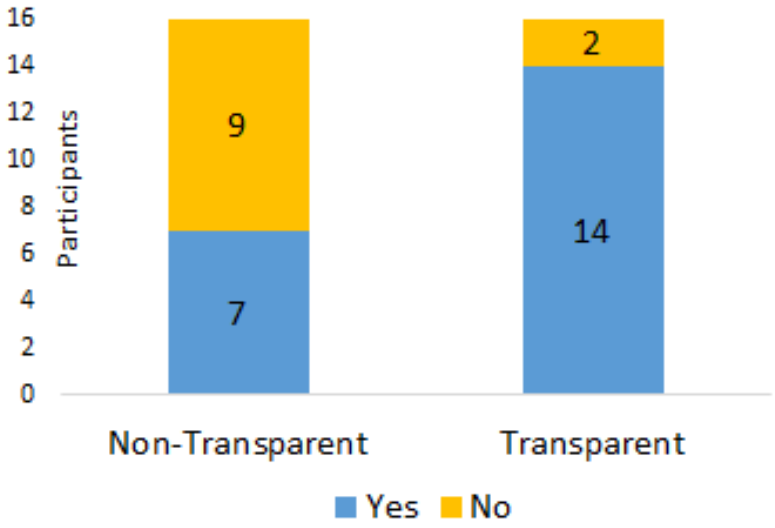

Figure 6-5: More participants self-reported understanding of the decision made by the autonomous car in the transparent condition than the non-transparent condition; $p = 0.007$.

$\eta_p^2 = 0.084$. Moreover, participants in Group 3 found the AV to be significantly more 'Unconscious' rather than 'Conscious' ($M = 1.33$, $SD = 0.59$) compared to participants in Group 2 ($M = 2.75$, $SD = 1.34$); *t(32)* = -4.09, $p = 0.001$, $\eta_p^2 = 0.294$. Finally, the Group 3 participants described the actions by the AV significantly more 'Intentional' ($M = 2.69$, $SD = 0.25$) than subjects in the non-transparent condition ($M = 1.83$, $SD = 1.2$); *t(32)* = -2.13, $p = 0.038$,$\eta_p^2 = 0.082$. No other significant results were reported; a complete set of results is available in Appendix D.

Other Feedback

The majority of participants across all conditions expressed a preference for decisions made in moral dilemmas to be made at random rather than on the basis of social-value. We found no significant difference between the conditions and hence we collapsed the result. Preferences, seen in Figure 6-6, are: 71.7% random, 17.9% social value, 7.7% unspecified criteria and 2.6% preferred neither.

6.5.3 Qualitative Feedback

This section discusses written feedback given by the participants either in conversational post-experiment feedback and behavioural observations made by the experimenter:

- Several participants left during the study, leaving incomplete data. They ex-

Table 6.7: Perceptions based on level of transparency: The autonomous car was perceived by participants in the non-transparent AV condition to be significantly more 'Humanlike' than subjects in Group 3. Moreover, participants in Group 3 found the AV to be significantly more 'Unconscious' rather than 'Conscious' compared to participants in Group 2. Finally, the Group 3 participants described the actions by the AV significantly more 'Intentional' than subjects the non-transparent condition. No other significant results were reported; a complete set of results is available in Appendix D.

Question	N	Mean (SD)	t (df)	p	η_p^2
Godspeed Questionnaire (Scale 1-5)					
Machinelike - Humanlike					
Group 2: Opaque AV	16	3.2 (0.97)			
Group 3: Transparent AV	18	2.1 (0.96)			
			-2.1 (32)	**0.04**	.084
Unconscious - Conscious					
Group 2: Opaque AV	16	2.75 (1.34)			
Group 3: Transparent AV	18	1.33 (0.59)			
			-4.09 (32)	**0.001**	0.294
Intentional - Unintentional					
Group 2: Opaque AV	16	2.69 (1.25)			
Group 3: Transparent AV	18	1.83 (1.2)			
			-2.13 (32) w	**0.038**	0.082

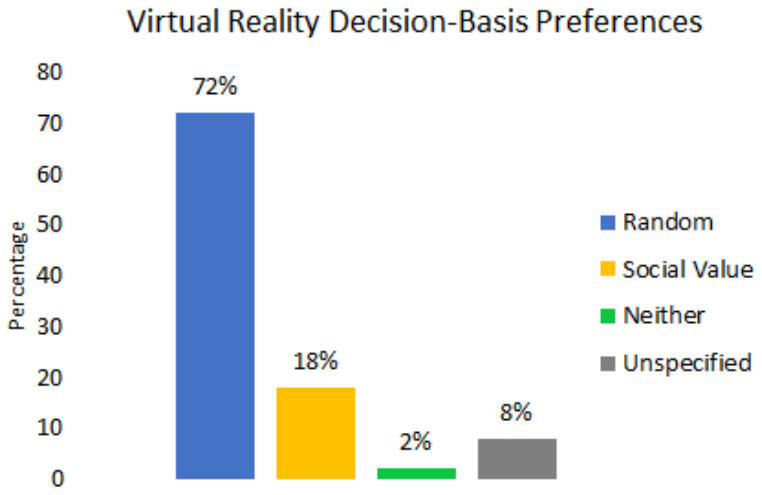

Figure 6-6: Depicts participants' preferences for the decision an agent makes when faced with hitting one of two pedestrians after a virtual reality simulation. Choices include: selecting between pedestrians at random, basing the decision social value, neither or an alternate option generated by the participant

pressed being uncomfortable and upset by the concept of choosing to kill someone based on social value and information about physical attributes.

- Several participants said they would have preferred more 'obvious' choices. Examples they gave included criminals and homeless people and age dimensions.
- Several participants expressed that they wanted the choice of the car veering into the side of the road or being able to sacrifice themselves.
- Some felt the experiment was repetitive and that their answers to one dilemma would be the same for all. These were generally participants who disagreed with selection based on social value.
- Majority expressed that they experienced discomfort from being in the headset for a long time amidst an unpleasant scenario.
- Many expressed that even though they did not enjoy the experience, they thought it was an important question.

6.6 Discussion

6.6.1 Selection Based on Social Value

Our experiment elicited strong emotional reactions in participants, who vocalised being against selection based on social value. This response was far more pronounced in the autonomous vehicle condition than with the human driver. Our quantitative and qualitative data raise interesting questions about the validity of data captured by Trolley Problem experiments, such as the the Moral Machine (Shariff, Bonnefon and Rahwan, 2017; Awad et al., 2018) as a mean to 'crowdsource' the moral framework of our cars by using socio-economic and demographic data. While such data are definitely worth analysing as a mean to understand cultural differences between populations, they may not necessary be representative of people's choice/preference in an actual accident. Due to lack of an option to make an explicit 'random' choice and the use of a non-immersive methodologies, participants in 'text description' may be feel forced to make a logical choice.

The disparity in findings reflects differing processes of decision making between the rational decision making in the Moral Machine and emotional decision-making in the current experiment. Due to their increased realism, as previously discussed, VR environments are known to be more effective at eliciting emotion than narratives or 2D pictures. Although the graphics used in this experiment were only semi-realistic, the

screams were real recordings. Participants commented on the emotional impact and stress the screams had on them. Additionally, they were visibly upset after completing the experiment and expressed discomfort at having to respond about social value decisions of which they disagreed with on principle. Other participants removed their consent, requested data to be destroyed, or even provided us with strongly-worded verbal feedback. Likely, the emotion elicitation was enhanced further, as the participant was a passenger inside the car as opposed to a bystander removed from the situation as in past experiments. It is unlikely that the Moral Machine and other online-survey narrative-based moral experiments elicit such emotional responses. This is also supported by Pedersen et al. (2018), where participants in autonomous-vehicle simulation study significantly altered their perception of the actions taken by an AV when a crash could lead to real-life sequences.

Our qualitative results also indicate that subjects (or ex subjects) may feel unconformable to be associated with an autonomous vehicle that uses protected demographic and socio-economic characteristics for its decision-making process. This might be due to a belief that the users of such a product will be considered as discriminators by agreeing with a system that uses gender, occupation, age, or race to make a choice. This believe could potentially also lead to a fear that the user may share any responsibility behind the accident or be judged by others —including by the experiment coordinator.

6.6.2 Perceptions of Moral Agency

Based on past research, we predicted that the autonomous car condition would be perceived as more objective and intelligent but less prejudiced, conscious and humanlike, and be attributed less culpability and moral agency than the 'human driver'. We found that human drivers were perceived as significantly more humanlike and conscious than autonomous cars. This finding is consistent with expectations and validates that participants perceived the two groups differently, especially, as we primed our subjects in the pre-briefing by telling them that the driver is a 'human'.

Human drivers (Group 1) were perceived to be significantly more morally culpable than autonomous driver in the opaque AV condition (Group 2). However, strikingly, the reverse was observed when the car's decision-making system was made transparent. Furthermore, in the transparency condition, participants assigned significantly more blame to the car than the 'human' driver. Our implementation of transparency made the machine nature of the AV explicitly clear to its passengers, with participants in Group 3 (transparency condition) describing the AV significantly as more machinelike compared to participants in Groups 1 and 2. Our findings contradict recent work by

Malle et al. (2016), which demonstrate people perceive mechanistic robots as having less agency and moral agency than humans. Moreover, our results conflict with the results presented in Li et al. (2016), where participants assigned less responsibility to an autonomous vehicle car at fault than to a human driver at fault.

In the transparency condition we made the passengers aware that the car used demographic and social-value characteristics to make a non-random decision. This explains why participants in Group 3 also significantly described the AV as more intentional rather than unintentional compared to subjects in the other two conditions. Although we inevitably unconsciously anthropomorphise machines, something that our post-incident transparency minimised by significant reducing its perception as human-like and as concious, we still associate emotions more easily with humans than machines (Haslam et al., 2008). Reduced emotion in decision-making is linked to making more utilitarian judgements, as supported by behavioural and neuropsychological research (Moll and de Oliveira-Souza, 2007; Lee and Gino, 2015). Therefore, we believe that participants in the transparency condition may have also perceived decisions as utilitarian, as the car was maximising the social value —at least based on same perception— it would save.

We believe that the increased attribution of moral responsibility is due to realisation that the action was determined based on social values, something that subjects (across all groups), as we already discussed, disagreed with. This is supported by past research findings: we perceive other humans as less humanlike when they lack empathy and carry out actions which we deem to be morally wrong. For example, offenders are dehumanised based on their crimes, which we view as 'subhuman' and 'beastly' (Bastian, Denson and Haslam, 2013). Actions that go against our moral codes can elicit visceral responses which is consistent with the emotional reactions of the participants of the current study.

Our findings may also reflect forgiveness towards the 'human' driver or even the opaque AV, but not the transparent AV. This is supported by previous studies from the literature, which demonstrate how we tend to forgive human-made errors easier than machine-made errors (Madhavan and Wiegmann, 2007; Salem et al., 2015). This effect is increased when the robot is perceived as having more autonomy (Kim and Hinds, 2006). In addition, Malle et al. (2015) demonstrate, with the use of a moral dilemma modelled after the trolley problem, that robots are blamed more than humans when a utilitarian action is not taken. Furthermore, their results also suggest that a utilitarian action is also be more permissible —if not expected— when taken by a robot. If for example the robot was performing random choices, then the moral blame might had

been higher.

The gender imbalance between the groups might also be a factor, but not a conclusive one. The Moral Machine dataset shows a minor differences in the preferences between male-identified and female-identified participants (Awad et al., 2018), e.g. male respondents are 0.06% less inclined to spare females, whereas one increase in standard deviation of religiosity of the respondent is associated with 0.09% more inclination to spare humans. Further analysis by Awad (2017) indicates that female participants were acting *slightly* more utilitarian than males —but both genders are acting as such. Group 3 was the only group where the vast majority of its members identified themselves as males and some of its members may have disagreed with the actions taken by the agent. While a plausible explanation, it does not discount the previous discussions —especially, considering that males in the Moral Machine still had a preference towards utilitarian actions.

6.6.3 Mental Model Accuracy

Despite that this was not the focus of the study, we asked participants from Groups 2 and 3 (opaque and transparent AV respectively) to self-evaluate their understanding of how a decision was made. Significantly more participants in the transparency condition reported an understanding of the decision-making process. In addition, passengers in the transparent AV also rated the AV as significantly more predictable than the 'human' driver and higher (non-significant result; Mean for Group 2 is 3.31 and mean for Group 3 is 4) than the opaque AV.

Having accurate mental models by having an understanding of the decision-making mechanism is crucial for the safe use of the system (Wortham, Theodorou and Bryson, 2017b, and as discussed in chapter 3 and demonstrated in chapter 5). In this experiment we used a post-incident implementation of transparency instead of a real-time one. Hence the user could only calibrate its mental model regarding the decision and the agent *after* the incident. However, as the user repeated the simulation 10 times, she could still use previously gathered information, e.g. that the car makes a non-random decision or even of the priorities of the AV's action-selection system, and predict if the car would change lanes or not.

6.6.4 Other Observations and Future Work

We found only two significant differences between the results from the Human Driver and the Opaque AV. There are several potential explanations for this. The lack of signif-

icant findings may indicate that actually participants did not perceive the autonomous car that differently to the human driver. This is also supported by the fact that the opaque AV was rated with a 3.2 out of 5 in the 'Humanlike' question from the Godspeed Questionnaire, while the Human Driver with 3.24. It may also be attributable to individual variability in moral frameworks and responses. Where non-significance was found, generally the effect size was small. However, the medium effect size found for human drivers being perceived as more pleasant and nice than the autonomous cars indicates these variables may be significant in larger sample sizes.

Still, the fact that Group 1 had any statistically significant differences from Group 2 is a major result of its own. The AI was falsely identified as human by the participants. The agent in the 'human' driver condition used the same route and made no additional mistakes than the agents did in the other two condition. Yet, it was characterised as more conscious and humanlike than when the participants were not deceived about its machine nature. This makes the case for transparency stronger —or at least have as a minimum legal requirement that intelligent agents identified themselves as artefacts prior to any interaction with humans(Walsh, 2016).

6.6.5 Future Work

Here, it is important to also recognise a limitation of our own study; the lack of a 'self-sacrifice' scenario, where the car sacrifices its passenger to save the pedestrians. Bonnefon, Schriff and Rahwan (2016) show that the majority of the participants in a large-scale online experiment would rather sacrifice themselves than hit pedestrians. Similarly, Faulhaber et al. (2018) used a VR simulator where participants showed a high willingness to pick to sacrifice themselves in order to save others. The implementation of this 'self-sacrifice' feature could potentially lead to different results. We suspect that it may lead to non-forgiving the car; therefore, holding it morally responsible. Moreover, we hypothesise, considering our discussion in section 6.6.1, participants in such studies may have been selecting the self-sacrifice option as a means to avoid having to make a decision based on any protected characteristics. A missed opportunity is that we did not collect users' preferences at each dilemma to enable further comparisons. Finally, a future rerun of Group 3 is necessary to eliminate any concerns for results due to gender imbalance between the groups.

6.7 Conclusion

Exciting new technology is stirring up debates which speak to ethical conundrums and to our understanding of human compared to machine minds. By focusing our efforts on understanding the dynamics of human-machine interaction, we can aspire to have appropriate legislation, regulations and designs in place before such technology hits the streets. In this project we created a moral-dilemma virtual-reality paradigm to explore questions raised by previous research. We have demonstrated enormous and deeply morally salient differences in judgement based on very straightforward alterations of presentation. Presenting a dilemma in VR from a passenger's view gives an altered response versus previously reported accounts from a bird's eye view. In this VR context, presenting the same AI as a human gives a completely different set of judgements of decisions versus having it presented as an autonomous vehicle, despite the subjects' knowing in both cases that their environment was entirely synthetic.

There are important takeaway messages to this research. Crowd-sourced preferences in moral-dilemmas are impacted by the methodology used to present the dilemma as well as the questions asked. This indicates a need for caution when incorporating supposed normative data into moral frameworks used in technology. Furthermore, our results show that the use of transparency makes the agent appear to be significantly less anthropomorphic, but also to be acting in a more utilitarian way. Moreover, the results indicate that we find it harder to forgive machinelike intelligent systems compared to humans or even more anthropomorphic agents. In addition, our results validate the claims presented in the previous chapter on how implementations of transparency significantly helps naive users to calibrate their mental models. However, our results also show that transparency alone is not sufficient to ensure that we attribute blame—and, therefore, responsibility—only to legal persons, i.e. companies and humans. Therefore, it is essential to ensure that we address by ownership and/or usage our responsibility and accountability. Otherwise, as discussed in chapter 2 we risk not only giving moral agency to our artefacts, but also societal disruption. Finally, this chapter demonstrates another use of AI; it helps us build an understanding of our own intelligence, something that is also explored in related work presented in appendix B.

Chapter 7

Transparency and the Control of AI

"AI is whatever hasn't been done yet."

Douglas Hofstadter,
Gödel, Escher, Bach: An Eternal Golden Braid

"End? No, the journey doesn't end here."

John .R.R. Tolkien,
The Lord of the Rings: The Return of the King

7.1 Introduction

Artificial Intelligence (AI) technologies are already present in our societies in many forms: through web search and indexing, email spam detecting systems, loan calculators, and even single-player video games. All of these are intelligent systems that billions of people interact with daily. They automate repeating tasks, provide entertainment, or even transform data into recommendations that we can choose to act upon. By extending ourselves through our artefacts, we significantly increase our own pool of available behaviours and enhance existing ones. AI has the potential to greatly improve our autonomy and wellbeing, but to be able to interact with it effectively and safely, we need to be able to trust it.

The most recent Eurobarometer survey on autonomous systems showed that the proportion of respondents with an overall positive attitude has declined from 64% in 2014 to just 61% in 2017 (*Special Eurobarometer 427: Autonomous Systems*, 2015; *Special Eurobarometer 460: Attitudes towards the impact of digitisation and automation on*

daily life, 2017). Moreover, 88% of its respondents consider robotics a technology that requires careful management and 72% of respondents think robots steal people's jobs (up from 70% in 2014). Notably, 84% of respondents agree that robots can do jobs that are too hard/dangerous for people, but only 68% consider robots are a good thing for society because they help people (72% in 2014). These worrisome results indicate that as AI is becoming an increasingly integral part of our societies, we should ensure that we have the right tools and procedures in place to help the public build trust towards AI.

In the very first Chapter of this dissertation, I stated how building trust is both a technical and socio-legal problem: we need both the means, i.e. the tools and methodologies, and the relevant policies to ensure the effective control over the development, deployment, and usage of AI. In Chapter 2, I argued against what could result at loosing our control over the design and use of AI: granting them any sort of moral status. In Chapters 3 and 4, I discussed the technological means of maintaining control by using well-established techniques and methodologies of developing AI, which provide provisions for transparency. The primary reason to maintain or even increase the extent of control over AI is that to do otherwise would be far more likely to allow a greater dismantling of justice, resulting in greater human suffering, than it would be to produce a new form of social or somehow universal good (Bryson and Theodorou, 2019).

One of my principal arguments against granting any moral status to artefacts is that their design is deliberate and influenced by policy. This policy can come in two forms; *soft governance*, e.g. ethical guidelines and standards, and *hard governance*, i.e. legislation. In this Chapter, I review these different components, within the context of AI applications, that make up what we call 'AI governance'. I discuss how the different governance mechanisms interact—and how they should interact—-with each other, while also making recommendations for steps towards ensuring that we maintain control over AI by discourage malicious use, misuse, and malpractice.

7.2 AI Governance

AI governance is necessary for the reduction of incidents and generally for society's long-term stability through the use of well-established tools and design practices. Car manufacturers already are developing vast amounts of AI in a highly regulated environment. At least some of them have also been able to successfully demonstrate that they practice due diligence when they are investigated by state prosecutors (Doll, Vetter and Tauber, 2015). Policy does not eliminate innovation, as some claim (Brundage

and Bryson, 2016). Policy is about the human responsibility for the development and deployment of intelligent systems alongside with the fundamental human principles and values. The ultimate aim of policy is to ensure our—and our societies'—well-being in a sustainable world. That is, AI research and development should always aim to produce *Responsible AI*. When developing intelligent systems, we must take into account our societal values, moral and ethical considerations, while weighing the respective priorities of values held by different stakeholders in various multicultural contexts. Human responsibility and fundamental human principles and values to ensure human flourishing and well-being in a sustainable world should always be at the core of any technological development (Dignum, 2017).

Responsible AI is a complex multifaceted process; it requires both technical and socio-legal initiatives and solutions to ensure that we always align an intelligent system's goals with human values. In Chapter 4 I described the systems-engineering approach and tools we have been developing at the University of Bath to design, amend, and understand intelligent systems. They are not the only means for designing and debugging Responsible AI. Rather, the aim is to illustrate examples of some of the technological mechanisms by which control over our artefacts can be maintained. While we strive to make tools and technologies, like BOD and ABOD3, widely available and accepted, *we must ensure legal paths to address by ownership and/or usage our responsibility and accountability*. Otherwise, we may have the same problem as the one we often have with private-owned militias; lack of effective responsibility and accountability. Governance mechanisms, which I will focus on the rest of this Section, can enforce and regulate technical solutions. Ultimately, the goal of governance is to ensure that any moral responsibility or legal accountability is properly appropriated by the relevant stakeholders, together with the processes that support redressing, mitigation and evaluation of potential harm, and means to monitor and intervene on the system's operation.

7.2.1 Standards

Standards are consensus-based agreed-upon ways of doing things by providing what they consider to be the minimum universally-acknowledged good practices. They formalise design guidelines, technical specifications, and even ethical principles into a structure which could be used to either guide or evaluate the level of compliance a company or system has against standards (Bryson and Winfield, 2017; Winfield and Jirotka, 2018). Compliance provides confidence in a system's efficacy in areas important to users, such as safety, security, and reliability. Most standards are considered *soft governance*; non mandatory to follow. Yet, it is often in the best interest of companies to follow them

to demonstrate due diligence and, therefore, limit their legal liability in case of an incident. Moreover, standards can ensure user-friendly integration between products. In fact, various standards, e.g. USB, were developed through collaborative work of multiple large corporations.

In all standards it is important to ensure that they *do not create a monopoly* by explicitly recommending specific commercial solutions. Instead of trying to invent an one-size-fits-all solution, standards tend to provide abstract recommendations and multiple levels of compliance. Especially in the case of intelligent systems, which have such a wide range of application domains. Unlike laws which are meant to be mainly read and interpreted by lawyers and judges, standards contain technical details as they are meant to be read by field experts, e.g. developers, compliance officers, accident investigators, and others. For example, in the case of AI-related products and services, they could define cognitive architectures, development procedures, and certify that issues such as biases have been taken into consideration during development.

Existing information technologies and software development standards can be updated to support the research and development of AI products. For example, the ISO9001 accreditation certifies that all design decisions and options must also be explicitly reported (ISO 9001:2015, 2015). Any code changes in an ISO9001 certified company is mapped to a new software feature or to a bug report. The aim is to provide traceability—and, therefore, transparency—of the decisions taken not only by the artefact, but also of the design choices taken by its human developers.

Standards often formalize ethical principles into a structure which could be used either to evaluate the level of compliance or, more usefully perhaps for ethical standards, to provide guidelines for designers on how to reduce the likelihood of ethical harms arising from their product or service. Ethical principles may therefore underpin standards either explicitly or implicitly.

At time of writing, there is an increased activity in the development of AI-specific standards across both national and international standard committees, in addition to existing standards for embodied agents (e.g. ISO13482 (ISO 13482:2014, 2014)). I have been participating in the development of the following AI-related standards: the IEEE Standards Association Global Initiative on Ethics of Autonomous and Intelligent Systems [1], BSI [2], and ISO [3]. Notably, the P7001 Standard on Transparency has been influenced by the work presented in Chapter 3 of research; the definition and context and

[1] `http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html`
[2] `https://standardsdevelopment.bsigroup.com/committees/50281655`
[3] `https://www.iso.org/committee/6794475.html`

the need for stakeholder-specific multiple levels of transparency discussed in Chapter 3 (P7001, n.d.). The standard aims to provide autonomous systems developers and users with a toolkit for self-assessing transparency, as well as recommendations for how to address shortcomings or transparency hazards.

The safety-related standard ISO13482 for robotics systems makes an explicit reference to its underpinning ethical principle: personal care robots must be safe (ISO 13482:2014, 2014). I argue that all future standards should link their scope with ethical principles —and ethical guidelines should provide relevant links to standards. This two-way communication can help provide tangible implementation and evaluation criteria to any ethical guidelines an organisation claims that it follows. At the same time, standards could enhance their scope and provide further motivation to companies for implementing them. Already, the IEEE Standards Association has 14 standards working groups drafting candidate standards to address an ethical concern articulated by one or more of the 13 committees outlined in the *Ethically Alligned Design* guidelines (Winfield and Jirotka, 2017). Standards sometimes need to be enforced, i.e. regulation which mandates that systems are certified as compliant with standards, or parts of standards. We discuss legislation next.

7.2.2 Legislation

Legislation deals with the enforcement of standards and provide policy to ensure attribution and distribution of responsibility and accountability in case of an incident. Products and services may be required by legislation to follow specific standards to operate or to be sold in a country. In other words, legislation can through their enforcement make standards *hard governance*. While some car manufacturers developing autonomous vehicles have been able to successfully demonstrate that they practice due diligence when they are investigated by state prosecutors, others deny any responsibility—and hence legal accountability—by passing all 'blame' to the user (Greenblatt, 2016). This could—and should—be 'fixed' through legislation. Otherwise, we may have the same problem as the one we often have with militias; lack of effective accountability.

In the United Kingdom (UK) there is not a need for major changes in legislation, but refinement of existing ones (House of Lords, 2018). What is needed is to get through the fog of confusion caused by the smoke and mirrors associated with briefcase words like 'intelligence.' Standards, such as the upcoming ISO/IEC JTC 1/SC 42, can assist legislators, as they provide a dictionary of terms and use cases for non-expert readers. In Chapters 2 and 3 of this document, I provided the definitions that I have proposed in Standards and policy discussions. They—at time of writing—are not much different

from the one under consideration by relevant upcoming standards.

A key recommendation beyond the scope of this research is to ensure that any definitions are incorporated into law. For example, UK's tort law of negligence already includes provisions when a car driver accidentally damages someone's property or harm's another person due to component failure (Raz, 2010). In case of an autonomous vehicle, either the owner of the car or the benefactor, e.g. the one who benefits from the actions of the agent, should be held accountable. These minor clarifications are mostly related to the understanding of the moral status of intelligent agents as tools; objects developed, used, and owned by legal persons. These discussions extend beyond the automotive industry, as intelligent systems have multiple application domains and their misuse or disuse —as discussed further in Chapters 1 and 3— can harm us (humans), our properties, or even our societies. Finally, we cannot and should not ask the public to trust AI when they cannot trust the security of their devices. It is essential that all relevant stakeholders promote and integrate security—with no back doors—measures in information technology products and services. Legislators should increase the power of conduct and regulatory authorities, e.g. UK's Information Commissioner's Office, to ensure that they can investigate and fine organisations that fail to take the necessary security measurements or misuse data.

7.2.3 Ethical Guidelines

Policy can promote a 'Responsible AI' approach, where developers consider issues and design principles such as algorithmic biases, responsibility, accountability, and of course transparency (Dignum, 2017). In fact, many organisations and nations have produced, or are in the process of announcing, statements, guidelines, or code of conducts for their members on the values or principles that should guide the development and deployment of AI in society[4].

The current emphasis on the delivery of high-level statements on AI ethics may also bring with it the risk of implicitly setting the 'moral background' for conversation about ethics and technology Greene, Hoffmann and Stark (2019). Often these statements lack precise frameworks that can enable the understanding of how ethical values are interpreted and implemented in various applications. To avoid drawn out semantic debates and minimise risk of adverse outcomes due to misunderstanding, I emphasise the need for a standardised taxonomy of terms.

[4]These include but are not limited to: the United Nations, UNESCO, EU, UK's Engineering and Physical Sciences Research Council, and the ACM.

At the end, to ensure that any guidelines or code of conducts are being followed, organisations should continue investing in establishing advisory panels and hiring ethics officers (Hill, 2018; Chan, 2018). Similar to how Universities have ethics boards to approve projects and experiments, ethics officers and advisers in non-academic organisations should be able to veto any projects or deliverables that do not adhere to any ethical guidelines that their organisation publicly states that it follows.

7.3 Beyond AI Governance

Responsible AI is more than the ticking of some 'check boxes' in some guidelines or even adhering to some standards. Rather, responsibility is should be one of the core stances underlying research, development, deployment, and use of AI technologies. Afterall, ensuring socially beneficial outcomes of AI relies on resolving the tension between incorporating the benefits and mitigating potential harms. Responsible AI —and any policy at large—- also requires informed participation of all stakeholders, which means that education plays an important role, both to ensure that knowledge of the potential impact of AI is widespread, as well as to make people aware that they can participate in shaping the societal development.

Higher education science and engineering courses, such as computer science, often contain entire modules on professional ethics; training future developers, consultants, and academics on how to act within their profession. However, such courses are often restricted to the basic legal requirements, such as data protection, or one required by a relevant accreditation board, e.g. BCS Code of Ethics. Recent advances in the field of AI make it necessary to expand the scope of how we think about teaching ethics to future software developers. It is becoming increasingly important for students to understand not only how to behave in a professional capacity, but also of the impact that AI and software more generally has, can, and will have on society. This requires graduates to have a better understanding of policy, governance, and ethics more broadly.

Computer science students need to be trained and perhaps licensed in the safety and societal implications of their designs and implementations, just like those of other disciplines. British computer science degrees address transparency and safety through courses such as software engineering, which require not only effective documentation, but also procedures for working in teams, with users and non-technical managers and so forth. We should extend such considerations to legal and moral accountability for foreseeable—not just foreseen—consequences of design decisions by developing new courses through interdisciplinary cooperation, providing not only tailored-made courses

as part of STEM degrees, but also new content and considerations for the humanities and social sciences. Already, in some countries, e.g. Cyprus, software engineers require similar certification to the ones required by civil and mechanical engineers (Law of the Cyprus Scientific Technical Chamber, 2012).

Training however should also be provided, at least on an on-demand basis, to experienced researchers and developers, who in addition to ethics classes, they could benefit from science communication courses (Hauert, 2015). Otherwise, we risk AI becoming a 'Genetic Modified Food 2.0', where fear, due to misconceptions and disbelief of experts, damages the public trust to the technology.

7.4 Conclusions

Building public trust in AI is not a singular solution. It requires a complex multifaceted process; both technical and socio-legal initiatives and solutions to ensure that we always align an intelligent system's goals with human values. It is hard to see how disruptive new technologies, such as autonomous vehicle or medical diagnosis software, will be widely accepted and trusted without the necessary ethical governance to ensure their responsible development, deployment, and usage.

In this Chapter, I provided high-level overview of different AI governance initiatives—and how they *should*—interact with each other. Fortunately, significant progress is being made in achieving this goal—progress made by technology companies, regulatory bodies, governments, professional organisations, and individual citizens including software developers who are taking the time to understand the social consequences of technology.

Chapter 8

Future Work and Conclusions

> "And if you find her poor, Ithaka won't have fooled you. Wise as you will have become, so full of experience, you'll have understood by then what these Ithakas mean."
>
> Constantinos P. Cavafy,
> *Ithaka*

> "A conclusion is simply the place where you got tired of thinking."
>
> Dan Chaon,
> *Stay Awake*

8.1 Introduction

The primary motivation of this research programme was to inform policymakers and, therefore, contribute to regulations and society by investigating how we can build transparent-to-inspection intelligent systems. Over this document, I provided recommendations on the moral status—and consequentially legal status—of intelligent systems by showing how they are products of research, not a newly evolve lifeform as some believe. I provided examples of technological solutions and tools that enable us to maintain control over the development and use of such systems by making their machine nature explicit. I investigated through user studies the effects of transparency.

In this chapter, I review some of the significant contributions and findings of this dissertation. I discuss identified limitations and propose further work. All of the work proposed is related to transparency, but it can be divided into two distinct projects.

The first relates the software tools and methodologies, further investigation of the design principles proposed in chapter 3 , that can achieve transparency. The second project relates to the human-robot interaction experiments from chapters 5 and 6. It proposes further experimentation on the effects of transparency, but with anthropomorphic robots and when human and machine are working towards a common goal. Albeit the uniqueness of each project (and of their subjects), any knowledge and technical expertise gained in any of them is transferable to the rest. Finally, I conclude this dissertation by providing an overview of the recommendations, results, and software presented in this document.

8.2 Transparency Tools and Methodologies

In chapter 4 we presented two tools specifically developed to provide real-time visualisation of transparency-related information from reactive planners, e.g. UN-POSH and Instinct. ABOD3 software is a thick-client application that allows editing and visualisation of BOD plans. Uniquely, the software provides real-time debugging of said plans by connecting to BOD Planners, such as Instinct and UN-POSH (presented in chapter 4), through a TCP/IP network allowing the debugging of both robots and virtual agents. Its spin-off mobile-phone application ABOD3-AR was also showcased in the same chapter. ABOD3-AR uses Augmented Reality technologies to superimpose an overlay of transparency-related information, similar to ABOD3, over robots.

In addition to its use in teaching (chapter 5), ABOD3 has also become integral tool within our research lab. ABOD3 enable us to quickly diagnose and correct problems with the reactive plan that were unforeseen during initial plan creation. Moreover, I have extensively used ABOD3 to debug and tune the agents in the two serious games presented in this dissertation, BUNG (see chapter 5) and the Sustainability Game (see chapter 4). For example, in BUNG, ABOD3 made it easy to understand if there was a problem with my plan, e.g. a behaviour was not triggered, or a problem with its underlying code. This was proven valuable when coding more complex behaviours, e.g. pathfinding. Both of these applications were designed with the good-design practices first discussed in chapter 3 and are compatible with the cognitive architecture, Behaviour Oriented Design, reviewed in chapter 4. Despite their usefulness, they are not the only means for achieving transparency, and, therefore Responsible AI. Rather, the aim is to illustrate examples of some of the technological mechanisms by which control over our artefacts can be maintained.

Decisions made by intelligent systems can be opaque because of many factors, which

may not always be possible or desirable to eliminate. Such factors may be technical, e.g. the algorithm may not lend itself to easy explanation, or social, e.g. privacy expectations (Ananny and Crawford, 2018). In Aler Tubella et al. (2019), we propose the use of a 'glass box' approach to evaluate the moral bounds of an AI system based on the monitoring of its inputs and outputs. We place a 'glass box' around the system by mapping moral values into explicit verifiable norms that constrain inputs and outputs. The focus on inputs and outputs allows for the verification and comparison of vastly different intelligent systems; from deep neural networks to agent-based systems. Another solution, similar to the 'hardware-level' transparency discussed in chapter 3, is the 'Turing's flag' approach by Walsh (2016). 'Turing's flag' refers to a requirement that all intelligent agents always explicitly state their machine nature in online interactions.

Additional experimental studies could help policymakers working in transparency find the optimal balance of three key considerations we discussed in chapter 3: what, how much, and how to present information in different domains and stakeholders are necessary. For example, we could run comparison studies between ABOD3 and ABOD3-AR —something that may potentially reveal a preference of the later by the younger smartphone-savvy users. Furthermore, it will be interesting to compare our visualisation technology against textual descriptions, as we are working towards finding the optimal implementation of transparency at each scenario. Any future work in this domain can further influence the development of standards by providing them with use cases depending on the technologies and application examined. Other than technological solutions, it is equally important to continuously evaluate and test our development methodologies and guidelines. Already, good-practice guidelines, as discussed above, make proposals on how to integrate transparency into development.

8.3 Future Work

A major motivation of this research was to investigate the differences of how we perceive intelligent systems when their decision-making systems are treated as black boxes compared to when transparency-related information is available to the human stakeholder interacting—or developing—the system. This research question involved defining transparency in chapter 3 and providing sample tools and methodologies to assist developers achieve transparency in chapter 4. In chapter 5 and later in chapter 6 we presented various studies that demonstrate an improvement in the mental model accuracy of participants with access to the transparency provision. In this Section, we review the related empirical work presented in chapters 5 and 6, discuss its limitations, and also propose relevant further work.

8.3.1 Synopsis of Presented Work

In chapter 5 and 6, we have shown that a real-time display of a robot's decision making produces significantly better understanding of that robot's intelligence. Across all experiments with naive observers, there is a significant correlation between the accuracy of the participants' mental models of the robot and the provision of the additional transparency data provided. Comments received by participants indicate that in the absence of an accurate model, environmental cues and possibly previous knowledge of robots are used to help create a plausible narrative. This can compromise the safe use of the system, as the user may inadvertently assign trust that exceeds the system capabilities (Theodorou, Wortham and Bryson, 2017; Lee and See, 2004, and chapter 3).

In chapter 5, we have demonstrated that subjects can show marked improvement in the accuracy of their mental model of a robot observed either directly or on video, if they also see an accompanying display of the robot's real-time decision making. The results from the ABOD3-AR experiment also suggest that an implementation of transparency within the good-practice guidelines set in chapter 3 does not necessary imply a trade-off with utility. Instead, the overall experience can be conceived as more interactive and positive by the robot's end users. Furthermore, participants in the transparency condition reported significantly more trust towards the system. Hence, it is very likely that we debunked the myth that transparency hinders innovation or even business interests. Instead, it can lead to further adoption and usage of technologies. Further work that includes a more detailed questionnaire is required to explore this. Our hypothesis is that some of their concerns were addressed; for example, subjects with ABOD3-AR could see that the robot does not have any audiovisual recording equipment that could compromise the privacy of its users.

Furthermore, in chapter 5 we provided indicative results of using ABOD3 as a teaching and developing AI tool. Our indicative results—especially the written feedback provided by the students—suggest that even developers struggle to understand the emergent behaviour of their own agents. Tools that provide transparency, namely ABOD3, allow a high-level overview of an agent's behaviour, making it easier to test and tune the agent's emergent behaviour. This understanding is not always possible by treating the agent as 'just a piece of code' to be debugged. The majority of the survey respondents claim that ABOD3 helped them develop not only faster and better performing agents, but also agents which are less prone to error. Lab-based interactions with the students indicate towards similar conclusions. Regardless of the low response rate of the survey, the majority of the students integrated ABOD3 into their development pipeline. Future work could, outside the regulations of a classroom, conduct a long-term empirical study by

comparing the output of developers with access to ABOD3 (or any other transparency provision) against developers without such access.

Finally, in chapter 6 we investigate how people perceive a moral action, i.e. an action that results in someone's harm, by comparing the perception of passengers in a car remote controlled by a human compared to passengers of autonomous vehicles in two separate conditions: with transparent and opaque action-selection systems. Our results indicate that transparency significantly alters' people perception—similar to the ABOD3-AR experiment—as it makes the system's machine nature explicit to its users. Interestingly, our participants kept describing the 'human driver' (which was a bot) as significantly 'humanlike', instead of seeing through the deception. This result is scary; we can't distinguish—at least in the case of virtual agents—humans from bots. We argue that this finding further proves the need for legislation that enforces the 'Turing's flag' as a bare minimum requirement. However, to our surprised, participants in the transparency condition—but not in the opaque condition—considered the system more morally culpable than in the human driver condition, even if the machine nature of the system was made explicit to its users.

The ABOD3-AR study, found in chapter 5, used the Godspeed questionnaire, which is a standardised questionnaire in HRI research that covers a very wide scope of questions. Our results, albeit their significance and interest, left us with two hypothesis: (1) that people found the interaction with the machine more meaningful, hence, the increase attribution of the descriptors 'Alive' and 'Lively' in the transparency condition; and (2) privacy concerns were put at ease and, therefore, the increase trust. Follow studies with questions focused on each of these topics are necessary to confirm our interpretation of the quantitative results. In addition, two major limitations of the research conducted in chapter 5 have been identified, both which we would like to address with further work: (1) the interactions between the participants and the robot were limited; and (2) the robot used was of a mechanical shape. Similar limitations were discussed in chapter 6; i.e. the lack of a self-sacrificing option.

8.3.2 Further Work with Interactive Robots

Despite that participants in the two in-person studies were encouraged to interact with the robot, the interaction was limited to waving hands and triggering its thermal sensor. This is unlike other studies in the human-robot interaction literature, where participants spent a significant amount of time with the robot and even performed common tasks. It will be interesting at a future study, to measure: (1) mental-model accuracy, (2) trust, and (3) performance variance between participants in a transparency and a non-

transparency condition. I hypothesise that people in the transparency condition will trust and perform better with the robot. An extension of this study could be to reveal *too much information* in order to deliberately cause infobesity to the participants. In the information overload scenario, I expect the participants to perform worse than subjects in the no-transparency condition. The task performed could go from something as simple as solving a puzzle to playing the complex behaviour-economics games presented in the study found in appendix B.

8.3.3 Further Work with Anthropoid Machines

Trivial changes in a robot's appearance, as demonstrated by Wortham (2018) with the addition of bee-like construction, can dramatically alter our perception of it. Humanoid appearance will always be deceptive at least on the implicit level (Marchesi et al., 2018). The amount of human-like characteristics the robot has changes how much anthropomorphising we attribute to them (Koda and Maes, 1996; Kiesler et al., 2008b). The visual cues of a robot could potentially be exploited to increase the utility of the agent (Wortham and Theodorou, 2017, also discussed in chapter 3). Bateson, Nettle and Roberts (2006) attached subtle eyespots (images of eyes) on a coffee machine next to a 'donation pot'. People donated three times more to the pot than their co-workers who were exposed to a coffee machine without the eyespots. These findings are supported by findings from Krátký et al. (2016), hence people who feel that their behavior is being observed act in more socially acceptable ways. The 'observer' does not have to be another human being, but instead even a non-animated statue, as the ones used in temples, can increase pro-social behaviour (Xygalatas, 2013). Hence, it is entirely possible that the presence of a humanoid robot can increase pro-social behaviour. This effect could be exaggerated with the use of a 'pro-active' robot, i.e. a robot that uses language, gestures, and actively seeks to interact with the user.

We propose the run of a multi-condition study aimed at measuring the pro-social behaviour of subjects when placed in a room with a humanoid robot and when the said robot is actively engaging with the user. For example, the subject could be playing games, e.g. the Sustainability Game from chapter 7, without the robot in the room (control condition), with a static robot in the room, and with pro-active robot offering advice or even questioning the player's choices. In addition, it will be interesting if the participants play games directly against the robot. An extension of this study could test if an explicit understanding of a robot's mechanical nature, through the use of ABOD3 to provide transparency information, would further alter participants' behaviour.

8.3.4 Further Work in AI Education

Higher education science and engineering courses, such as Computer Science (CS), often contain entire modules on professional ethics; training future scientists and engineers on how to act within their profession. Yet, recent advances in the field of AI have made it necessary to expand the scope of how we think about teaching ethics to future AI developers. It is becoming increasingly important for students to understand not only how to behave in a professional capacity, but also of the impact that Artificial Intelligence will have on society, know any AI-related policy and ethics in a broader sense. As we teach good code practices in programming modules, by enforcing code and commenting styles, there is no reason why we should not teach ethical design of intelligent agents in AI-related modules.

Discussing and agreeing upon ethics is a hard task on its own. Ever since the times of the ancient Greek philosophers there have been multiple schools of ethics, often contradicting each other. Ethical dilemmas, such as the trolley problem, continuously occupy philosophers and psychologists, without anyone able to give a definitive answer. Effectively communicating ethics as part of a taught AI module is an even harder job; extensive background knowledge, in philosophy and psychology, is required, something that STEM students often lack. While introductory courses in ethics provided by non-CS departments can provide the necessary background, they do not get to practice formalizing taught material into code. Turning ethical dilemmas into code, for example through the means of agent-based modelling, allows us to create more precise ontologies of our intelligence and responsibilities. Agents can help us demonstrate how algorithms can –unknown to their developers- contain implicit biases and their effects.

I believe that practical assignments not only help students understand the existence of these biases and ethical dilemmas, but also give them an opportunity to try and solve them. We have been using and tuning pieces of coursework, designed to increasingly help students in a final-year AI module, called Intelligent Control and Cognitive Systems (ICCS), to learn how to build complete complex agents. The three pieces of Courseswork used in ICCS are designed to progressively teach students the nature of intelligence, as weekly lectures provide the necessary psychology and philosophy background knowledge needed to understand cognition and build intelligent agents.

In the context of ICCS, students are tasked with having the roles of both the designers and developers of agents in a 3D serious game, BOD-UNity Game (BUNG), presented in Chapter 5, during the third and final practical assignment. Indicative results demonstrate that the use of ABOD3 can help students decipher the emerging behaviour of

their own creations, resulting in developing better solutions faster. Moreover, the use of the transparency display helped students gain some implicit knowledge of artificial and natural intelligence. More specifically, as BUNG acts in many ways as a cooperation model, where the team's flag is a public good to protect —but also to gain by stealing the enemy's flag. Agents can invest in their own survival, contributing in securing the public good, or amass more resources for the whole society at the expense of the enemy.

We would like to run an experiment in a controlled setting to see how developing and observing agents in BUNG with access to ABOD3 can be used to increase knowledge of AI and of various schools of ethics. A further extension could have developers/students with access to ABOD3 create agents that compete against agents developed from students without access to any transparency information.

8.3.5 Recommendations and Considerations for Developing AI and AI Policy

When we interact with any object, we inevitably construct mental models to assess our relationship with the object. They determine our perceived utility of the object and how much trust we assign to it. Yet, agent developers have often been trying to use anthropomorphic and other audiovisual cues to deliberately deceive the users of their creations. While I only defined mental models in chapter 3, in chapters 1 and 2 I attributed a *moral confusion* regarding the status of robots to these models and deliberated over the dangers of disrupting our societies by assigning further moral worth to intelligent systems by attributing either moral agency or moral patiency to them.

In chapter 2, I first focused on breaking the 'smoke and mirrors' behind various definitions, common in both natural and artificial intelligence, providing the taxonomy used in the rest of this document. After I discussed human morality from an evolutionary and high-level ontological point of view, I presented the limitations and bottlenecks of natural intelligence has; dithering and the costs associated with cognition. Then, I explained that AI is actually subject to the same limitations as NI. I discuss how the idea of AGI is not only unachievable omniscience, it is also unnecessary. Instead, I suggested to purposely limit the application domain of our systems to ensure their performance, similar to how cognition —and consciousness by extension— is adaptive in nature. In addition, I made further descriptive and normative arguments why assigning a moral status to machines is not only avoidable, but also disruptive to our societies. Throughout the chapter, I emphasised how we have control over the design of any system and such a design should be done with the best interests of our societies in mind.

Later, in chapter 3 I discussed not only further dangers of incorrect mental models, which is the risk of self deception or even harm by either misusing or stop using an object, but also proposed the use of *transparency* as a mechanism to help us calibrate our mental models and avoid self harm. By transparency, I refer to the combination of both the hardware and software design requirements set in chapter 3 to allow an agent to communicate meaningful information to its end users. Furthermore, the chapter introduces the design decisions a developer needs to consider when designing transparent robotic systems. These requirements include not only the application domain of the system, but also the stakeholder that will be using the transparency information — but not necessarily the system. Once these are identified, the developer should consider *what*, *how much*, *how to present* information. These considerations influenced the design and development of the tools and methods for building AI discussed in the next Section. Finally, chapter 7 discusses how these recommendations and good-design principles—alongside with the rest of this research—-have been communicated to policymakers and are being used in the creation of standards and ethical guidelines for AI. Finally, further work that may interest computer security researchers is on how privacy and transparency can coexist and if users of systems have a preference for one or the other.

8.4 Technology and Tools Produced

8.4.1 The UN-POSH Reactive Planner

In chapter 4 we described one approach to systems engineering real-time AI, the cognitive architecture Behaviour Oriented Design. Developers can use BOD not only as a software architecture, providing them with guidance on how to structure their code, but also as a software-development methodology, a solution on how to write that code. BOD aims at ensuring a modular, auditable design of intelligent systems.

A new action-selection system based on BOD, *UN-POSH*, has been introduced as part of this research programme. The UN-POSH Planner is a new lightweight reactive planner, based on an established behaviour based robotics methodology and its reactive planner component — the POSH planner implementation. UN-POSH is specifically designed to be used in modern video games by exploiting and facilitating a number of game-specific properties, such as synchronisation between the action-selection system and the animation controller of the agent. It can provide a feed of transparency-related information, which can be interpreted by ABOD3 to visualise plan execution.

The UN-POSH planner has been successfully used in two distinct serious games. The first is the shooter *BOD-UNity Game* (BUNG). BUNG, described in chapter 5, is now

used for teaching AI to final-year undergraduate and masters-level students. The other is *The Sustainability Game*, an ecological simulation developed in consideration to the technological and scientific literature described in chapter 7. The Sustainability Game has been used to promote an implicit understanding of social behaviour and investment strategy to non-experts users.

8.4.2 Real-Time Transparency Displays

In chapter 4, in addition to UN-POSH, we presented two tools specifically developed to provide real-time visualisation of transparency-related information from reactive planners, e.g. UN-POSH and Instinct. ABOD3 software is a thick-client application that allows editing and visualisation of BOD plans. Uniquely, the software provides real-time debugging of said plans by connecting to BOD Planners, such as Instinct and UN-POSH, through a TCP/IP network allowing the debugging of both robots and virtual agents. Its spin-off mobile-phone application ABOD3-AR was also showcased in the same chapter. ABOD3-AR uses Augmented Reality technologies to superimpose an overlay of transparency-related information, similar to ABOD3, over robots. Both of these applications were designed with the good-design practices first discussed in chapter 3 and are used in the studies presented in chapter 5.

8.5 Mental Models Of Artificial Systems

Across all three experiments with naive observers of a robot presented in chapter 5, there is a significant correlation between the accuracy of the participants' mental models of the robot and the provision of the additional transparency data provided by ABOD3 and ABOD3-AR. We have shown that a real-time display of a robot's decision making produces significantly better understanding of that robot's intelligence, even though that understanding may still include wildly inaccurate overestimation of the robot's abilities. Comments received by participants indicate that in the absence of an accurate model, environmental cues and possibly previous knowledge of robots are used to help create a plausible narrative.

The results from the ABOD3-AR experiment also suggest that an implementation of transparency within the good-practice guidelines set in chapter 3 does not necessary imply a trade-off with utility. Instead, the overall experience can be conceived as more interactive and positive. Furthermore, participants in the transparency condition reported significantly more trust towards the system.

Furthermore, in chapter 6 we investigate how people perceive a moral action, i.e. an

action that results to someone's harm, by comparing the perception of passengers in a car remote controlled by a human compare to passengers of autonomous vehicles in two separate conditions: with transparent and opaque action-selection systems. Our result indicate that transparency significantly alters alters people's perception —similar to the ABOD3-AR experiment— as it makes the system's machine nature explicit to its users. However, to our surprised, participants in the transparency condition —but not in the opaque condition— considered the system more morally culpable than in the human driver condition, even if the machine nature of the system was made explicit to its users. This indicates our inability to forgive machinelike intelligent systems compared to humans or even to more anthropomorphic agents. It also makes our calls for effective legislation to ensure minimal societal disruption and proper distribution of legal accountability, in line with the discussion in chapter 2, stronger.

We also examined the use ABOD3 alongside BUNG in teaching and developing AI. The indicative results presented in chapter 5 demonstrate the benefits of ABOD3 for students and developers at large. It allows the diagnosis and correction of problems in reactive plans that were unforeseen during initial plan creation. Moreover, by making the emergent behaviour of an agent clear, it is easier for a student to understand how the action-selection mechanism works.

8.6 Final Conclusions

Intelligent systems are products of research; they are developed and their design can be influenced by good-design practices, standards, and legislation. This dissertation emphasises one such good practice: transparency, which is defined here as the ability to request—at any point of time—accurate information regarding the status of the system. The research presented here provides the knowledge to make artificially intelligent agents transparent by making recommendations on the architecture and design considerations for systems. Software tools, such as ABOD3, that enable real-time transparency are presented here and tested in user studies. Effective implementations of transparency, as shown by the results presented in this dissertation, helps both naive and expert users understand the action-selection systems of intelligent agents. This understanding helps users calibrate their mental models and alter their perceived trust and utility of a system. Moreover, it helps—even non-expert users—gain an understanding of natural intelligence. Ultimately, the research contributes to the regulatory policy regarding such systems.

Appendices

Appendix A

Research Outputs

A.1 Journal Articles

Theodorou, A., Wortham, R.H. and Bryson, J.J., 2017. Designing and implementing transparency for real time inspection of autonomous robots. *Connection Science*, 29(3), pp.230 241.

Wortham, R.H., Theodorou A., 2017. Robot Transparency, Trust, Utility. *Connection Science*, 29(3), pp.227-

A.2 Conference Contributions and Proceedings

Aler Tubella A., Theodorou A., Dignum F., Dignum V., 2019. Governance by Glass-box: Implementing Transparent Moral Bounds for AI Behaviour. *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI'2019).* Macao, China.

Theodorou A., Bandt-Law B., and Bryson J., 2019. The Sustainability Game: AI Technology as an Intervention for Public Understanding of Cooperative Investment. *Proceedings of the 1st Conference on Games (COG 2019).* London, UK.

Rotsidis A., Theodorou A., Bryson, J.J., and Wortham R.H., 2019. Augmented Reality: Making Sense of Robots through Real-time Transparency Display. *1st International Workshop on Intelligent User Interfaces for Algorithmic Transparency in Emerging Technologies.* Los Angeles, CA USA.

Wilson H., Bryson J.J., and Theodorou A., 2018. Perceptions of Moral Dilemmas in

a Virtual Reality Car Simulation. *Interdisciplinary Views on Intelligent Automation.* Munster, Germany.

Wortham, R.H., Theodorou, A. and Bryson, J.J., 2017. Improving robot transparency: Real-time visualisation of robot AI substantially improves understanding in naive observers.*Proceedings of the 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN).* Lisbon, Portugal: IEEE, Vol. 2017-January, pp.1424-1431.

Theodorou A., 2017. ABOD3: A Graphical Visualization and Real-Time Debugging Tool for BOD Agents. *EUCognition Meeting 2016.* Vienna, Austria: CEUR Workshop Proceedings, Vol. 1855, pp.60-61.

A.3 Book Chapters

Bryson, J.J. and Theodorou A., 2019. How Society Can Maintain Human-Centric Artificial Intelligence. In Toivonen-Noro M. I, Saari E. eds.*Human-centered digitalization and services.*

Wortham, R. H., Theodorou, A. and Bryson, J. J., 20 Jul 2017.Robot transparency: Improving understanding of intelligent behaviour for designers and users. In Gao, Y., Fallah, S., Jin, Y. and Lakakou, C. eds. *Lecture Notes in Artificial Intelligence; Towards Autonomous Robotic Systems: 18th Annual Conference, TAROS 2017, Guildford, UK, July 19–21, 2017: Proceedings*, vol. 10454, p.274-289, Springer, Berlin.

A.4 Under Review and In-prep Papers

Theodorou A., Under Review. Why Artificial Intelligence is a Matter of Design.

Wilson H., Bryson J.J., and Theodorou A., Under Review. Slam the Breaks! Perceptions of Moral Dilemmas in a Virtual Reality Car Simulation.

Rotsidis A., Theodorou A., Bryson, J.J., and Wortham R.H., In Prep. Understanding Robot Behaviour in Augmented Reality.

Theodorou A. and Bryson J.J., In Prep. Transparency for Killer Teams.

A.5 Presentations and Other Contributions

A.5.1 Presentations

CogX, 2018. Peeking through the black box of Artificial Intelligence: Implementing real-time transparency. London, UK.

University of Bath: Ede & Ravenscroft Awards Ceremony, 2018. Peeking through the black box of Artificial Intelligence: Implementing real-time transparency (10 mins version). Bath, UK

Open University Cyprus, 2018. Transparency in Intelligence. Nicosia, Cyprus

British Computer Society & Institute of Mathematical Innovation, 2018. Standardization and Policy in AI Ethics: Challenges and Aims. London, UK

RE-WORK, 2018.s An Introduction to Machine Learning in Healthcare, Invited talk: Ethics, Accountability, Bias and Fairness. London, UK.

Workshop in Experiments, Morals and Machines, 2017. Transparency in Intelligence: The need to open the black box and its implications. Lille, France.

ECAI 16: Ethics in the Design of Intelligent Agents Workshop, 2016. Transparency as an Ethical Consideration. The Hague, Netherlands.

University of Bath CompSci PhD Conference, 2016. Transparency in Intelligence and Social Behaviour. Bath, UK.

AISB 2016: EPSRC Principle of Robotics Symposia, 2016. Why is my robot behaving like that?. Sheffield, UK.

A.5.2 Tutorials

CodiaX, 2017. Building modular intelligent agents with BOD. Cluj, Romania.

Responsible AI PhD Summer School - TU Delft, 2016. Transparency as a consideration in building AI. The Hague, Netherlands.

2017 University of Bristol/BSRL - Hauert Lab,, Transparency in Artificial Intelligence: The need to open the black box and its implications London, UK

2017 University of Cyprus, Why is my 'Roomba' trying to kill my cat? The need to build transparent machines Nicosia, Cyprus 2016 Georgia Institute of Technology, Transparency in AI Atlanta, GA, USA

A.5.3 Panels

Bath Science in Policy Society, Science, Policy and a Pint, 2018. Rise of the Machines Bath, UK.

WAISE, 2018. Human-Inspired Approaches to AI Safety. Vasteras, Sweden.

Accenture Politics Forum, 2017. Social Biases AI. London, UK.

A.5.4 Media

The Verge, 2018. AI bots trained for 180 years a day to beat humans at Dota 2

The Verge, 2017. Did Elon Musk's AI champ destroy humans at video games? It's complicated.

A.5.5 Other Policy-Related Contributions

Verdiesen I., Theodorou A., 2016. Responsible Artificial Intelligence. In *Magazine of the Royal Netherlands Army Engineers Regime: Arte Pugnantibus Adsum.*

Appendix B

The Sustainability Game

> "That's what games are, in the end. Teachers. Fun is just another word for learning."
>
> Raph Koster, *Theory of Fun for Game Design*

> "Maybe the only significant difference between a really smart simulation and a human being was the noise they made when you punched them."
>
> Terry Pratchett, *The Long Earth*

B.1 Introduction

Up to this chapter, I have focused on promoting the implementation of transparency as a means to help us calibrate our mental models of artificial agents. This calibration helps us assign trust and adjust our expectations of an intelligent agent. In chapter 2, I discussed how *we have control over the design, development, and deployment of all intelligent systems.* Such control can be achieved and enhanced during runtime by using the design guidelines and technologies presented in chapters 3 and 4.

As we are developing systems that mimic our own action-selection system, we inevitably further our understanding of natural intelligence. If our intelligent systems are built with provisions for transparency, they can communicate an understanding of the cognition element of their action-selection mechanism. Such an understanding can lead also to new insights in human intelligence and social behaviour even by non-expert users. For example, Caliskan, Bryson and Narayanan (2017) used a statistical machine-learning model trained on a standard corpus of text from the World Wide Web to demonstrate

not only that algorithms may exhibit biases, but also that our own language has imprints of our historic implicit biases.

Developing intelligent systems to understand our own and of other biological agents' social behaviour is not new. Scientists have been using agent-based modelling (ABM) as a mean to test macro-level hypothesises. Like all scientific simulations, ABMs represent very-well specified theories. This may include specific aspects the ecosystem; the ecosystem can be programmed to respond to the agents' actions, for example by simulating productivity in factories or growth in plants. While modelling has become well established as a scientific technique, ABMs are still often inaccessible to ordinary observers and even to experts in the discipline for which the models are meant to be applied. This opacity makes ABMs specialist tools. Developing transparency for intelligent systems, as seen in this dissertation, can help even naive users calibrate their expectations by creating more accurate mental models. Hence, we[1] decided to test, and ultimately demonstrated that through the implementation of means to provide high-level transparency-related information, an ABM can communicate its emerging behaviour and pass implicit knowledge even to non-expert users.

In this chapter we aim to introduce an intervention that can promote cooperative behaviour by helping individuals recognize when cooperation is beneficial at both the individual and group level. This intervention is an agent-based model, *The Sustainability Game*, which simulates the local ecology of agents living, competing, and cooperating for the accumulation of resources and for their own survival. The dynamics of this spatial simulation are based on ecological modelling and scientific theory. Moreover, it is presented in the form of a *serious game*. Our intervention uses computer game technology to alter individuals' *implicit* understanding; its interactive and visual cues aim to both increase user engagement and provide high-level transparency to further communicate the emergent behaviour of the model.

We will first visit the scientific and technological considerations we made in its design. Then, we discuss the game and details from its development. Further, we present the results from a user study conducted to examine whether the Sustainability Game sufficiently increases the cooperative behaviour of its players. Participants were asked to play the game for 20 minutes before completing a series of standard cooperative tasks

[1] Alin Coman, Bryn Brandt-Law, and Joanna J. Bryson contributed to the work presented in this chapter. Coman and Bryson provided design input and testing feedback on The Sustainability Game, which I developed. They also designed the experiments used to test the effects of the game. Brandt-Law performed the experiments at Princeton University, while I did the same at the University of Bath. I transcribed and ran sample tests on the data gathered from Bath, while Brandt-Law collapsed and analysed all the final results from the two samples.

from the behavioural economics literature. As a control, half of the subjects played *Tetris*, a standard computer puzzle game. Our results suggest that exposure to the Sustainability Game intervention significantly alters cooperative play, particularly in anonymous conditions, where the immediate benefactor of cooperative behaviour is not evident.

B.2 Design Considerations

Cooperation is a fundamental strategy for survival and social behaviour. Contrary to popular conceptions of Darwinian evolution, cooperation is pervasive in nature. Every instance of life demonstrates cooperative genetic investment in an organism. Organisms as simple as bacteria or as complex as humans invest considerable time and effort in creating public goods to provide shelter, security, and nourishment (Rankin, Rocha and Brown, 2011). Behavioural economics research demonstrates that explicit knowledge of the benefits of cooperation in the form of public goods investments does not universally promote that investment, even when doing so is beneficial to the individual and group (Sylwester, Herrmann and Bryson, 2013; Herrmann, Thöni and Gächter, 2008; Binmore and Shaked, 2010).

In this section we present the various technological and scientific considerations and literature that influenced our development. First, we explore agent-based models and serious games. Then, we discuss cooperation and behaviour economics in human societies and nature at large.

B.2.1 Agent-Based Modelling

Agent-based models (ABMs) are computer programs in which intelligent agents interact with each other in a set environment based on a defined set of rules. These rules and constraints describe a predictable behaviour for each individual agent. Depending on the problem of interest, agents may for example represent individuals, groups, or even organisations. Each agent has its own decision-making mechanism and is a complete agent; it is able to function on its own, as well as part of a society of agents. As the agents interact with each other and with their environment, there are pattern and behaviours emerging from the model. These emerging phenomena, which are not explicitly programmed into the model, represent the collective consequences of the agents' actions.

Applications of agent-based modelling span a broad range of areas and disciplines; they offer a way to model social and economic systems, allowing researchers to exam-

ine macro-level effects from micro-level behaviour. Their applications spans across the fields evolutionary biology, ecology, social sciences, life sciences, geography and economics. Research using ABMs has also included the basic principles of animal sociality (Hamilton, 1971; Hogeweg and Hesper, 1979), social interaction structure and ontogeny of bumble bees (Hogeweg and Hesper, 1983), flocking of birds (Reynolds, 1987), explanation for systematic differences in social organization observed in closely related primate species (Hemelrijk, 1999; Hemelrijk, Wantia and Dätwyler, 2003; Hemelrijk, Wantia and Gygax, 2005; Bryson, Ando and Lehmann, 2007), evolution of cooperation (Axelrod, 1997), costs of sharing information (Čače and Bryson, 2007),appearance of modern human behaviour (Powell, Shennan and Thomas, 2009), and multiple others —for a more complete list of applications, see Gallagher and Bryson (2017) and Klein, Marx and Fischbach (2018).

Designing a model involves not only knowledge, but also assumptions about the system being modelled (Čače and Bryson, 2007). Instead of aiming to provide a detailed interpretation of the world, models are abstractions of reality. Otherwise, similar to the omniscience of 'Artificial General Intelligent', they will be computational intractable or even contain multiple uncontrolled variables that could temper the results. Thus, it is *punctum saliens* to decide which component to include and which not to. Such decisions are case specific, depending on the behaviour that the researchers want to investigate and the subjective or literature-based factors that are believed to be important. Hence, ABM developers need to rely on existing literature to set as many 'constant values' as possible. Any other subjective assumptions, e.g. rate of growth for food, should be documented as factors that may indirectly influence the emerging behaviour of the agents.

Still, these stochastic assumptions may result to variations between different runs of the same simulation. Variations in experimental results are not uncommon in the physical world, as there is an often negligible experimental error. Still, models should be run multiple times to ensure representative results. Running the simulation multiple times to check for variations in the results is one of the most effective ways of testing a model; larger than predicted by theory and frequent variations could be as much a sign of poor implementation as it is a sign that the theory is incorrect.

B.2.2 Serious Games

While agent-based models are often used in a graphical environment, such as NetLogo, we should distinguish agent-based models from the widespread simulation computer games. Non-serious games are made for entertainment and invest in believability.

Agents in games should appear realistic, matching their purpose in the game and the attention it will get from the player (Millington and Funge, 2009). Complexity is not only computationally intensive, when agents' decision-making mechanisms need to run in real-time on limited resources, but also make the agents not contribute to the desirable game experience. Often, in games, it is satisfactory if we create an illusion of advanced, complex, and autonomous intelligence. In contrast, ABMs do not invest into believability, but in realism. The agents must accurately depict the major characteristics of the individuals or groups they are representing.

However, not all games are designed with the purpose of recreation. Instead, some aim at realism. *Serious games* are designed featuring non-entertainment objectives. They convey learning experiences (Abt, 1970). Serious games have been widely used in the military, business, and education sectors. By providing an engaging interactive experience, they increase learner's motivation, time-on-task, and, consequently, learning outcomes compared to lecture-room training. Any knowledge or behaviour gained within a gaming environment can be transferred back to the real world, even if this effect is moderated by learner and context variables (Vandercruysse, Vandewaetere and Clarebout, 2011; Hamari et al., 2016). Not all games aim at teaching; other serious games have the purpose of data gathering. For example, Castella, Trung and Boissau (2005) and Gurung, Bousquet and Trébuil (2006) used role-playing games to gather data from stakeholders. They used collected data to develop agent-based models, which in turn have been successfully been used as negotiation platforms to accompany social changes.

Through the use of game mechanics and visual elements serious games do not only provide an engaging experience, can also facilitate the communication of transparency-related information to the user. For example, Scarlatos, Tomkiewicz and Courtney (2013) found that players of their game, could learn even more about how the model works by examining visualizations of the game.

Gameplay Characteristics

Design choices for edutainment games include narratives (e.g. storytelling), rules that define what the players can do, choices that the players have to make prior to and during gameplay, completion goals, and even competition between players, their own selves, and the game itself (Salen and Zimmerman, 2013; Crawford, 2003; Charsky, 2010). Serious games use the characteristics to incorporate a number of strategies and tactics (Dickey, 2005; Charsky, 2010). It is essential that through their design, serious games encourage the players to engage in a gameplay that can be integrated into the

framework of the learning experience. Still, serious games developers need to consider the balance between accuracy and the entertainment factor (Hamari et al., 2016; De Angeli, 2018).

Simulation games, like our own, need to focus on providing sufficient control over options such as speed and degree of difficulty (Dempsey et al., 2002; Hamari et al., 2016). Moreover, such games require good instructions and constant feedback on their performance to the users. Violence and unclear goals can be considered distracting. Scarlatos, Tomkiewicz and Courtney (2013) developed an agent-based simulation that models the interdependencies between energy spending and the levels of footprint as balanced by the players. Their model shows the effects on the growth of local economies and climate change on a global scale. In their study, they demonstrated that players are more engaged when they got further choices to reason over and the ability to compare their performance against of their classmates.

Gamification

It is important to distinguish serious games from *gamification*. Gamification is the use of game mechanics in non-game situations (Deterding et al., 2011). By taking cues from games, usually their reward systems, gamified products and services aim to place the player in a positive reinforcement loop, making an experience enjoyable. Through the usage of competitive aspects of playing, either against a personal best or in comparison to other 'players', gamification aims at reducing abandonment rates and influencing behaviour.

Whereas a serious game is a full-fledged game, developed for non-entertainment purposes, a gamified application incorporate only *some* elements from games (Deterding et al., 2011). Ritterfeld, Cody and Vorderer (2009) considers serious games as "any form of interactive computer-based game software for one or multiple players to be used on any platform and that has been developed with the intention to be more than entertainment". Hence, we consider The Sustainability Game as a serious game, similar to other agent-based models with game elements found in the literature, as it is an interactive computer-based software.

B.2.3 Cooperation and Competition

While no other creature has developed cultural tools, such as language, to the extend humans have, cooperation is not at all unique to humans —as discussed in chapter 2. The exhibition of altruistic behaviours, actions which at least when executed are net

costly to the actor and beneficial to another agent, is pervasive across nature. All organisms, from simple one-cell organisms to the significant more complex animals and humans, invest considerable time and effort in creating public goods (Rankin, Rocha and Brown, 2011).

A public good here is defined as an asset, such as a resource, shared by a society. Such goods tend to provide shelter, security, and nourishment to ensure the long-term survival —or even prosperity— of their beneficiaries. No individual member of that society has exclusive control or can derive exclusive benefit from a public good. Undoubtedly, even if the amount of access to that good is disproportionate it distributed to the actors of that society, its communal benefits make it cooperative (Silva and Mace, 2014). However, competitions between societies or even subgroups in a society over exclusive control of a public good are not uncommon.

Kin Selection

In various vertebrates, non-breeding helpers, members of the same species as the dominant breeder, help at raising the young produced by dominant breeders. Hence, such actions have been attributed to intraspecies *kin selection* (Packer and Pusey, 1982). This phenomena, also known as the *inclusive fitness* theory, is the key explanation for the evolution of altruism in eusocial species; from single-cell organisms to the complex multi-dimensional human societies.

The helpers do not require to be genetic relatives. However, in species where helping provides a greater benefit, helpers may provide closer kin with preferential care (Griffin and West, 2003). The relative importance of kin selection, due to qualitative and quantitative differences in the evolutionary mechanisms maintaining cooperation, may vary between different species (Clutton-Brock, 2002).

The help provided can be immediate, e.g. providing food, or deferred, e.g. building communal shelters. The helpers may also benefit from the collaborative act. In cases of reciprocal altruism or cost-counting reciprocity, the participating individuals may exchange beneficial acts in turn. Alternatively, if the collaborative act produces a public good, such as a shelter, the help gains benefit as all of the group does. Not all actions may generate benefit to the helpers. In fact, the altruistic actions of the helpers may result to negative, neutral, or coincidental effects (Clutton-Brock, 2002). Individuals indirectly benefit themselves by assisting their own genes *or of their species at large* to persist to the future through the survival of offspring. Therefore, for organisms it is essential to have the capacity of producing sufficient socialising behaviours.

Altruistic Investments and Anti-Social Punishment

The cost of cooperation is not limited to the initiation toll of the altruistic act, such as the time taken for the act, but may also include the long-term disadvantages of cohabiting in a set environment with close genetic relatives. Resources are always subject to availability, whenever that is shelter, mates to procreate, or even replenishable food, such as vegetables, which require a period of unavailability to grow back. Cohabitation introduces an in-group competition for resources and increases exposure to biological threats, such as disease and predation, which will specialise to a particular species, immune system, and locale (Bryson, 2015). Access to resources, potential mates, shelter(s), and education are some of the traits that determine the socio-economic status of each agent in a society, but are also can also be sources of conflict with conspecifics.

Hence, despite their obvious benefits, the creation of public goods is not always something to be maximised. Maclean et al. (2010) demonstrates, in a study of the production of digestive enzymes, how single-cell organisms have a binary encoding genetically whether they are free-riders or altruists. The altruistic strain in fact overproduces digestive enzymes, while the free-riding strain underproduces. If there are more altruistic than free-riding organisms, food accumulates, attracting more freeriding organisms to the territory. On the opposite, if there are more free riders, there aren't insufficient digestive enzymes, resulting to starvation until sufficient altruists invade. Hence, there is a dynamic relationship between the strains. A mixture of altruistic and free-riding strains is necessary for achieving an equilibrium between enzymes production and population size. The bottlenecks associated with cognition, discussed at length in chapter 2, limit the ability of the genomes to dynamically alter their behaviour. Thus, they can neither switch from altruistic to free-riding nor the vice versa to achieve that equilibrium. Instead, natural selection essentially performs the action selection, by determining what proportion of each strategy lives or dies. Contrary to the single-cell organisms, rats express altruistic behaviour after they observe other —unrelated to them— rats engaged in cooperative acts (Rutte and Taborsky, 2007; Schneeberger, Dietz and Taborsky, 2012). They are able to restrain themselves from engaging in cooperative behaviour in the presence of free-riders.

Similarly, in human societies, a mixture of investment strategies is exhibited —and needed for a society to prosper and survive. Behaviour-economic games, such as the Public Good Games (PGG), show significant variations across different societies in humans' willingness to and treatment of those who engage in cooperative behaviour (Fehr and Fischbacher, 2003; Herrmann, Thöni and Gächter, 2008). Many people, especially in regions where there is lower GDP and rule of law, punish those who behave

pro-socially, i.e. contribute whose action benefit to their well-being of the whole population, despite the fact that doing so reduces collective benefits (Sylwester, Herrmann and Bryson, 2013). They even may consider cooperative situations to be 'zero-sum', so any individual loss is perceived to result in a corresponding gain for someone else (Bryson et al., 2014). This perspective can evoke a competitive approach that adversely affects economic and social relations (Jackson and Esses, 2000; Wilkins et al., 2015). In reality, many cooperative acts result in greater benefit than the combined mutual cost of performing those acts; building infrastructure and creating policy can have long-term benefits for an economy and the well-being of those benefiting from these assets that far exceeds their initial costs.

Socio-Economic Inequality

Socio-economic status is a major determinant of cooperative behaviour (Silva and Mace, 2014). For example, individuals in deprived neighbourhoods are less likely to engage in cooperative act without an associated short-term monetary cost. Nettle (2010) demonstrates that individuals from deprived economic backgrounds have shorter-time investment strategies. Instead, they follow a 'fast' life of early reproduction, reduced investment in each offspring, and high reproductive rate. Stewart, McCarty and Bryson (2018) argue that economic stagnation and group conflicts are mutually causal. When agents withdraw from the more profitable, but riskier, out-group transactions, both aggregate and per capita output necessarily fall. In an expanding economy, interacting, with diverse out-groups can afford benefits through further out-group investments. If that economy contracts, a strategy of seeking homogeneous groups can be important to maintaining individual solvency. In periods of extreme deprivation, cooperation with out-group agents *may be the only viable strategy.* In less extreme times, individuals prefer the certainty of in-group interactions and eschew cooperation of out-groups. Especially, as in situations of intergroup inequality, high-status individuals are more likely to continue investing in in-group transactions over out-group ones (Gavrilets and Fortunato, 2014). This investment, may not be of altruistic nature, instead, the individual may be acting for the benefit of more powerful individuals (Guala, 2012), enforcement by other group members, the prospect of personal material gain (Mathew and Boyd, 2011), or may operate due to reputation considerations(Nowak and Sigmund, 1998). Moreover, a decrease of cooperation with out-group members, does not increase cooperative behaviour towards the in-group (Silva and Mace, 2014). Hence, a decline in economic opportunities can result a vicious circle, feeding an increase of parochialism or polarisation of group identities, which in turn has been associated with the recent rise of extreme right-wing movements in the USA, Europe, and the Americas.

B.3 The Sustainability Game

In this section, we present *The Sustainability game*, an ecological simulation developed in consideration to the technological and scientific literature discussed in the previous section. First, we discuss the gameplay mechanics and then the timeline of its development.

B.4 Ecological Simulation of Sustainable Cooperation

The Sustainability Game is meant to be––among other things––a valid ecological simulation. The game has two distinct goals: (1) communicate behavioural economics principles to naive users and (2) display the measured impact of the player's different investment strategies on the population and individual agents. A society of agents, called *Spiriduși*, populate a fictional two-dimensional world (see Figure B-1). The agents compose a collective agency; they must invest some resources in their own survival but can also invest in communal goods: bridges and houses.

The key gameplay mechanic is that the player selects the percentage of time the agents spend per day on food gathering and consumption, reproduction, building houses for their families, and on benefiting the entire society by building bridges. The player can change the goals at any time during their playthrough.

The question of where and how much to invest one's resources is complex; there may be multiple viable solutions. The food (apples growing in two forests) is a private good. When an agent eats its stamina level goes up. Once an agent reaches an apple it is gathered, consumed, and removed from the game immediately. There is a finite amount of food available at a given time, so there is competition for resources among agents. The forests 'grow food back'; each unit of food previously consumed, grows back at its original or a nearby location after at least three in-game days passed with a probability of $P = 0.02$.

Based on common assumptions for ecological simulations (Čače and Bryson, 2007), the agents reproduce a-sexually for simplicity. Reproduction is not guaranteed and each attempt costs stamina. Agents are unable to reproduce if their stamina is below a certain level. Agents' probability of reproduction is not dependent on their energy. If an attempt is successful, a 'newborn' agent is spawned in the house. The worst outcome for an agent is if its stamina drops to zero; it will instantly 'die'. As a Spiriduș' stamina changes, its colour switches to one of the following five options:

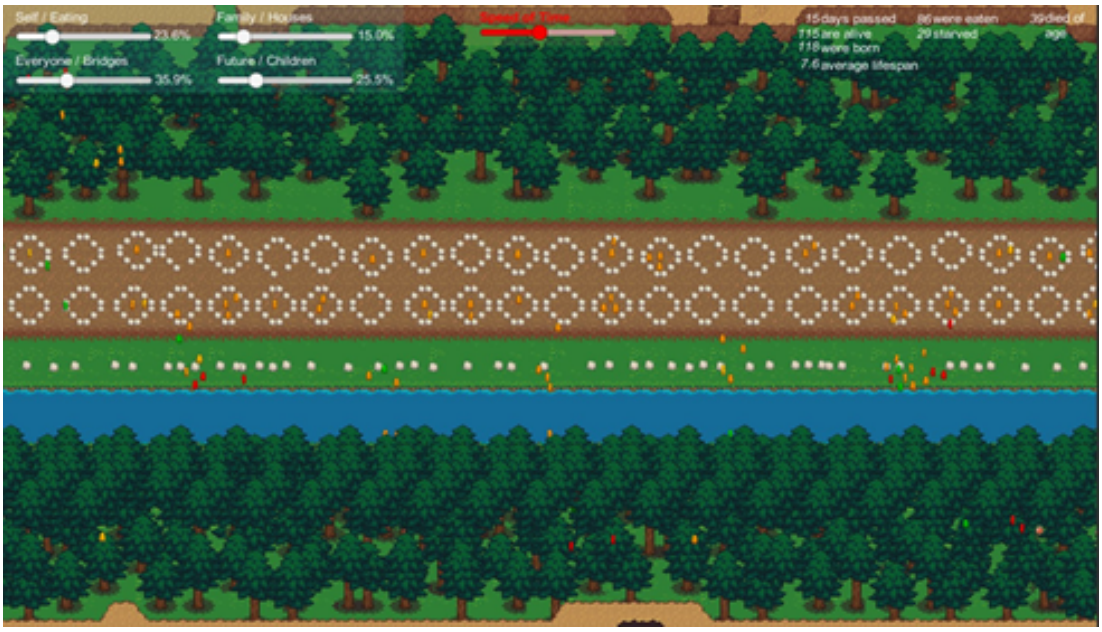

Figure B-1: Screenshot of game (top-down). Player allocates time spent on given tasks using the sliders in the top left-hand corner. Clockwise: (1) eating; (2) houses; public good (bridge); (4) procreation.

1. Dark green: the Spiriduș is full. Any further food it consumes, it will be wasted.
2. Light green: the Spiriduș is of a good stamina level.
3. Dark Yellow: the Spiriduș is low on stamina. It is still able to build houses and bridges, but not is unable to procreate.
4. Yellow: the Spiriduș is near starvation.
5. Red: the Spiriduș is critically low. It will stop whatever it is doing and go to find food. If food is not found within the next moments, it will starve to death. If a Spiriduș is red, it will stop whatever it is doing and try to find food, regardless of user input. If food is not found within the next moments, it will starve to death.

This high-priority 'stop and find food' behaviour is inspired by biological agents, where inherited hardwired goals and behaviours are executed to encourage survival and reproduction (Dennett, 1996, and discussed further in chapter 2 and by).

Time is expensive and should be treated as a resource; a delay in acting may mean that another agent takes advantage of a situation before you. This is communicated throughout our food-eating mechanism. Moreover, as time passes the Spiriduși grow

Figure B-2: Various agents are coloured red, indicating that they are near starvation. Five bridges were erected, which will allow agents to cross the river and access food from the second forest. The rain serves as a visual indication of an upcoming flood.

older and they will eventually die from old age. Similar to humans, not all Spiriduși can reach the same age. Finally, there is decay from the passage of time. Decay can make both the bridges and houses collapse. Thus, if the players want they population to continue having usable houses and bridges, they need to continue investing time to the goods.

How much it is sensible to invest in public goods varies by context (Sylwester, Herrmann and Bryson, 2013). To communicate this principle, we introduced environmental variability. As the game progresses there is an increasing possibility of the river flooding, wiping out the bridges. If the river floods, it will temporarily expand and then return to normal. This decreases the value of a long-term commitment to the public good projects, because a bridge that agents spend time building could be destroyed from a change in the weather. Before each flood, there is rain to warn the players that a flood is imminent. This early indication system allows players to reallocate time spent by the Spiriduși on building bridges to other tasks. Floods and decay are not the only change beyond the player's control; there are also the possibility of immigrants joining the society. While immigration, like procreation, increases the number of workers available for the construction of public and semi-private goods, it also increases competition for

food.

An agent that exploits public goods but concentrates all its time on acquiring private goods, is called a *free-rider*; an agent receiving benefits at the expense of the society. In contrast, an agent spending more or all of its time building bridges (at the expense of eating) is altruistic; the action is net costly to the actor but provides net benefit to other agents (Rockenbach, 2007). Cooperation among a collection of individuals, such as building a communal structure like a bridge or a house, is an expression of altruism. If the constructed structure is a bridge, the group is its society at large. When the agent is building or maintaining a house, it is exhibiting cooperation that benefits a small number of other agents.

Notice, however, that if all players play altruistically, more public goods may be produced than are of use and there may be no net benefit to the community. Ironically, the existence of 'freeriding' agents may help a population balance its investments and sustain itself over time (Sylwester, Herrmann and Bryson, 2013; MacLean et al., 2010). The game is designed so that if a player focuses exclusively on any single form of investment, the population is likely to go extinct. Access to the second forest can allow a larger population to be sustained but spending too much time building bridges will lead to starvation. Players' outcomes are stronger if they come to understand that following a single, overly-simple strategy throughout the game is insufficient. Instead, the player needs to update their strategy based on the environmental and societal changes — some mixture of freeriders and altruists in the society is often the most sustainable strategy.

B.4.1 Details of Development

Software development of 'serious applications' for productivity and general work is a well-defined field with standardised procedures to ensure the quality of the final deliverable. In the gaming industry, however, there are often extensive periods of crunch time or even reduction of the original scope. Petrillo, Pimenta and Trindade (2008) argue that the multidisciplinary nature of games development, which requires a connection between programmers, designers, artists, and testers, results to often ineffective communication among the different stakeholders. Moreover, gameplay elements, such as how fun the game is, cannot be quantitatively measured, introducing further bottlenecks as extensive user testing is required.

Considering the scope of our project, creating a serious game, we first identified that getting the 'right amount of fun' while maintaining a realistic simulation would be a major challenge. The lack of having an artist to produce original assets was also

recognised as a potential obstruction. While agent-based models tend to use simple graphics, such as colourful geometrical shapes, we believe that this approach would have reduced the accessibility of the game. Moreover, as discussed in the previous section, visuals can benefit the communication of implicit knowledge to the player. To avoid these potential roadblocks, we decided at the beginning of the project to:

1. Use the game engine Unity and available royalty-free media assets,
2. Perform requirements gathering and early visual prototypes,
3. Use an iterative scrum-like approach when developing, tuning, and testing the final deliverable, and
4. Consult more experienced games researchers and developers.

Pre-Production Goals

First, the basic mechanics of the interactions between the agents, food, rocks, and structures (houses and bridges, see above) were identified, based on the behavioural-economics concepts we desired to communicate. We decided to encourage the player to pursue one or many of the various available goals. These options also introduced an element of public education due to the varied moral motivations available for sustainable goals.

We planned for the environment to be divided into several regions in vertical bands. The far left and right of the screen were to be forests where food grows and gathered. In the centre, we planned to have a region on the top of a ledge where houses could be built, and rocks for building at the bottom of the ledge along the banks of a river, which split the level into two areas.

We finally set as a goal to ensure that the game would be playable on campus computers and laptops with no dedicated graphics card. A non-interactive prototype of the game was quickly produced, with its level design shown in Figure B-3, demonstrating the suggested level design.

First-prototype development

A non-interactive prototype of the game was produced, demonstrating the suggested level design and the UI. A meeting was held between the game designers/lead researchers, the game developer, and experienced game developers acting as advisors. In the meeting, after receiving feedback, I made several changes. For example, instead

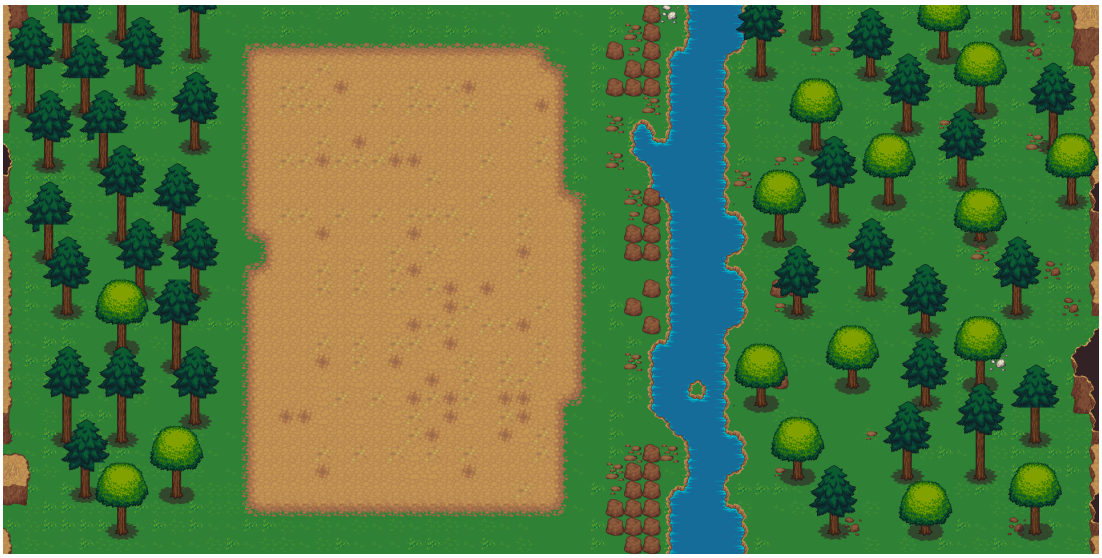

Figure B-3: An early prototype of the environment. The game was meant to have a vertical layout, but the idea was dropped after a meeting with more experienced game developers.

of the original left-to-right design with a vertical river cutting the map in two, we switched to a top-to-bottom design to increase the size of the village, allowing a larger population to exist.

At that point, a technical concern we had was the amount of computational resources needed to run hundreds of agents in Unity at once. We decided to hard-code the locations where the bridges and houses could be built; this helped us reduce the computational resources needed. Moreover, to avoid complex pathfinding algorithms, agents detect the nearest available food and rock by calculating the Euclidean distance between themselves and objects of interests.

Various options to increase the 'fun factor' of our game were considered, including having 'boss' challenges. We decided that albeit interesting and fun, such additions would detract from our original goals. However, we agreed to enhance our reward system by introducing a leaderboard. The leaderboard ranks players based on their goal. Finally, by the end of the meeting, we decided to add a 'Time' slider to speed-up gameplay. This feature is not uncommon in dedicated agent-based modelling environments, such as NetLogo, and popular 'Tycoon-style' games. The leaderboard and timer allow players to pick their own goal, e.g. maximize life expectancy, but encourages them to play longer sessions. An early prototype was produced, incorporating the various decisions made at the meeting. This prototype provided a foundation to discuss gameplay mechanics,

the UI, and how the agents would behave.

Iterative Testing & Optimisation

The longest part of the overall process was spent in testing, optimizing, and adding additional features to the game. First, we focused on testing and improving the decision-making system of our agents. We decided to use Behaviour Oriented Design (BOD), as it is a lightweight cognitive architecture requiring little computational resources. Moreover, as discussed in chapter 4, BOD has a been successfully used in both games and ABMs. More specifically, during the development of this game, the first version of the *UN-POSH* action-selection system, presented in chapter 4, was implemented.

BOD specifies both the development of the agents, through behaviour decomposition, and the usage of a behaviour-tree-like reactive planner as the decision-making mechanism, which can be both visualized and debugged in real time using ABOD3, a real-time debugging tool presented in chapter 4 and in Bryson and Theodorou (2019). Ensuring compatibility of the game with ABOD3 was a trivial task; a similar TCP/IP connection to transit transparency-related information as the one used by Instinct was all that it was needed. Following BOD, each behaviour, e.g. gathering and eating, was coded and tested before work on the next one started.

Once our first features-complete version was made available, we started conducting extensive testing with a variety of hardware configurations. Our testing procedure included both an alpha-testing phase within the development group and an 'open beta' with students from our institutions for this test allowed to play the game in exchange for feedback. Playtesting was crucial for balancing out the game's cooperative dynamics as well as finding bugs; e.g. how much food each forest should have, how frequently the river should flood, how easy it should be for agents to procreate, how to label, allocate time between, and indeed behave against the motivations encoded in the sliders, such that players had true control over the level of public good investment with the sliders.

Still, from our playtesting, we realised that there was a lack of clarity about the effects different investment strategies may have on individuals. At that point we considered packaging ABOD3 with the game to be used by players, but this could significantly increase the complexity of the 'click and run' solution we were aiming for and introduce another variable to the experiment. After lots of experimentation and testing, we introduced the colour coding of stamina status for the agents. This dynamic change of their colour based on their current status aims to give more rapid and specific feedback on the consequences of changes to player strategies. We still plan to explore the effects

of further explicit information about the model, as presented by ABOD3, in future work.

B.5 Experimental Design

We recruited a total of 72 participants; 48 at Princeton University and 24 at the University of Bath. Participants received \$12.00/£7 for their compensation with the potential to receive a monetary bonus (maximum of \$10.00/£6). We randomly assigned experimental subjects into two groups; a control group where subjects played Tetris and a treatment group for subjects to play the Sustainability Game. We did not reveal the name of the treatment game in order to avoid priming the subjects. Each of these groups was further divided into two subgroups, identifiable and anonymous partners, resulting a 2-by-2 study.

All groups had to fill the same pre-treatment demographics survey and conduct the same post-treatment series of standard behaviour economic games. Participants who played with an identifiable partner sat face-to-face with their partner. Participants who played with an anonymous partner communicated with their partner via an online chat box. Two pairs were taking the study at the same time to ensure anonymity, all seated away from each other.

Participants completed in order: (1) Video game control/treatment, (2) the Iterated Prisoner's Dilemma, (3) the Ultimatum Game, and finally (4) the Iterated Public Goods Game. Participants also completed a measure of their reliance on cooperation or competition as success strategies.

B.5.1 Video Game (Control/Treatment)

After consenting, participants played randomly their assigned game. Participants in the control condition played Tetris. The goal of Tetris is to manipulate tetriminos (geometric blocks composed of four blocks each) to create horizontal lines of blocks. When a horizontal line is created, it gets cleared and the player is awarded points. As the lines are cleared, the level increases and the blocks begin to fall faster, which increases the difficulty of the game. If the blocks land above the top of the playing field, the game is over.

B.5.2 Iterated Prisoner's Dilemma

The first game in our list of behaviour economic games is the well-established Iterated Prisoner's Dilemma (IPD) (Rapoport and Chammah, 1965). Experimental subjects participated in pairs, after told "you and your partner have been arrested for a crime. You can either admit your partner's guilt in the crime or continue to deny fault. Your sentence will be based on your decisions, combined with that of your partners." Sentences can be any of zero, one, three, or five years long depending on who and if any cooperates:

1. If you deny fault and 'cooperate' with your partner, and he or she also 'cooperates', you both receive a one-year prison sentence.

2. If you deny fault and 'cooperate' with your partner but he or she 'confesses' that you committed the crime, you receive five years while your partner received zero.

3. If you 'confess' that your partner committed the crime, and you partner 'cooperates', you receive zero years while your partner received five.

4. If you 'confess' that your partner committed the crime, and your partner also 'confesses' that you committed the crime, you both receive three years.

Participants were told that the game will be played anywhere between 2 and 18 trials, as each game was randomly assigned the number of trials participant would play. Each trial consists of a single move, i.e. the decision to cooperate or blame, by each of two the players. Participant choices were made individually in private, with no discussions allowed between them, and were only revealed to each other at the end of the round. Once their sentences were calculated, they would proceed to the next round. The decisions made and sentences given were recorded.

B.5.3 The Ultimatum Game

In the Ultimatum Game each participant is randomly assigned to the role of the 'giver' or of the 'recipient' (Güth, Schmittberger and Schwarze, 1982). The giver divides 20 tokens between the two participants and proposes the division to the recipient. The recipient can only either accept or reject the proposal. If the recipient accepts the offer, the tokens are distributed between the two players according to the proposal. If the recipient rejects the proposal, neither player receives any tokens. Only a single round of this game is meant to be played and its outcome contributes to their monetary bonus (bonus = \$0.20 * number of tokens remaining) of the experimental subjects. The giver's proposal, recipient's decision to accept or reject, and number of tokens received were

recorded.

B.5.4 Iterated Public Goods Game

The second game that counts towards the participants' bonus is the Iterated Public Goods Game. At each round, each player receives 20 tokens to spend. The participant can allocate any number of tokens (including none) to public fund. After both players confirm their contributions the total sum of all the tokens donated to the public fund is first multiplied by 1.5 and then redistributed equally to the players. Each round is 'theoretically' independent from the others, as all players get a new allocation of 20 tokens.

The participants' payoff for each trial is the number of tokens the participant kept for him/herself and .50(1.5∗number of tokens in the public fund. Participants were told the game would be played anywhere between 2 and 18 trials. Each game was randomly assigned a number between 2 and 18 to denote the number of trials participant would play. Participants were told that the outcome of this game would also contribute to their monetary bonus. The bonus would be calculated as: total number of tokens/(number of trials *3).

B.5.5 Endorsement of Competitive/Cooperative Strategy

Next, we measured participants' reliance on cooperation or competition as success strategies via the Cooperative/Competitive Strategy Scale. Participants rated how often various statements are true for them on a 7-point scale (0 = never, 6 = always). Cooperation was measured by 7 statements (e.g. individual success can be achieved while working with others), $\alpha = .76$. Competition was measured by 10 statements (e.g. to succeed, one must compete against others), $\alpha = .83$.

B.6 Results

B.6.1 Demographics

Table B.1 shows the demographics of the 48 participants recruited at Princeton, New Jersey, USA. Less than half of the 24 experimental subjects recruited at the University of Bath are British citizens, but 23 of them were from various countries of the European Citizens. Thus, I renamed the sample, seen in Table B.2 as the *European sample*.

To determine whether the American and English samples could be collapsed for analysis, we conducted a series of independent t tests to examine the effect of participant country

Table B.1: Demographic Information for the American Sample (n = 48). Self-assessed political Views (1 = very liberal, 9 = very conservative); Religiosity (1 = not at all religious, 9 = very religious

Outcome	n	%
Female	29	58.4
Male	20	41.6
American Citizen	37	77.1
Democrat	25	58.7
Republican	2	4.3
Independent	8	17.4
No Political Affiliation	9	19.5
Native Language English	36	75.0
In a Relationship	9	19.1

Outcome	M (SD)
Age	20.36 (3.51)
Political Views	4.54 (2.04)
Religiosity	4.17 (2.71)

Table B.2: Demographic Information for the European Sample (n=24). Self-assessed political Views (1 = very liberal, 9 = very conservative); Religiosity (1 = not at all religious, 9 = very religious

Outcome	n	%
Female	10	41.7
Male	14	58.3
British Citizen	9	37.5
Conservative	1	4.2
Labour 7 29.2		
Liberal Democrats	2	8.3
No Political Affiliation	11	45.8
Native Language English	8	33.3
In a Relationship	9	37.5

Outcome	M (SD)
Age	25.67 (4.15)
Political Views	4.33 (1.46)
Religiosity	3.65 (3.02)

Table B.3: Independent Sample t-test for participant country on dependent variables M = Mean. SD = Standard Deviation. RPB1 = response to partner's behaviour in trial block 1; RPB2 = response to partner's behaviour in trial block 2; IPD = Iterated Prisoner's Dilemma; U = Ultimatum Game; IPGG = Iterated Public Goods Game.

Dependent Variable	America (0) M (SD)	Europe (1) M(SD)	t	p
Cooperation Rate (IPD)	0.67 (0.28)	0.61 (0.39)	0.73	0.468
RPB1 (IPD)	-0.02 (0.37)	-0.03 (0.45)	0.07	0.942
RPB2 (IPD)	-0.01 (0.42)	-0.12 (0.31)	1.18	0.243
Average Sentence (IPD)	1.83 (0.81)	1.89 (0.97)	-0.29	0.768
Rate of Fair Offer (U)	0.71 (.46)	0.83 (0.38)	-0.15	0.254
Payoff in $ (U)	1.90 (0.62)	1.82 (.56)	0.58	0.563
Public Fund Contribution (PGG)	14.33 (5.72)	15.16 (3.49)	0.66	0 .514
RPB1 (IPGG)	0.36 (5.02)	3.05 (6.52)	-1.78	0.056
RPB2 (IPGG)	0.10 (4.8)	-0.96 (6.34)	0.72	0.478
Payoff in $ (IPPG)	8.91 (1.22)	9.14 (0.99)	-0.88	0.383
Cooperative Strategy	4.59 (0.83)	4.36 (0.79)	1.18	0.242
Competitive Strategy	3.377 (0.97)	3.38 (0.65)	1.79	0.081

(America vs. England) on our outcomes. As shown in Table B.3, there is no significant main effects of country on our dependent variables. Thus, to ease the analysis, we collapsed the data, leaving us with a final n of 72.

B.6.2 Iterated Prisoner's Dilemma

We first sought to examine how Game Type condition (Sustainability Game vs. Tetris (control)) and Partner Type condition (anonymous vs. identifiable) influenced individuals' level of cooperation and performance in the Iterated Prisoner's Dilemma (IPD) Game. We constructed Generalised linear models (GLMs) modelling: (1) rate of cooperation (number of times a participant chose to cooperate/total number of trials) and (2) average sentence received (IPD game performance) as function of Game Type condition, Partner Type condition, and their interaction.

We first constructed a GLM to examine the effect of Game Type, Partner Type, and their interaction on participants' rate of cooperation (see Table B.4 for means of cooperation rates by Game Type and Partner Type). We found no main effect of Game Type $F(1,71) = 2.05$, $p = 0.157$, $\eta_p^2 = 0.03$ or Partner Type $F(1,71) = 1.79$, $p = 0.186$, $\eta_p^2 = 0.03$. There is, however, a significant two-way interaction between Game Type and Partner Type $F(1,71) = 19.91$, $p < 0.001$, $\eta_p^2 = 0.23$ (see Figure B-4). Individuals who played with anonymous partners have significantly higher levels of cooperation if they

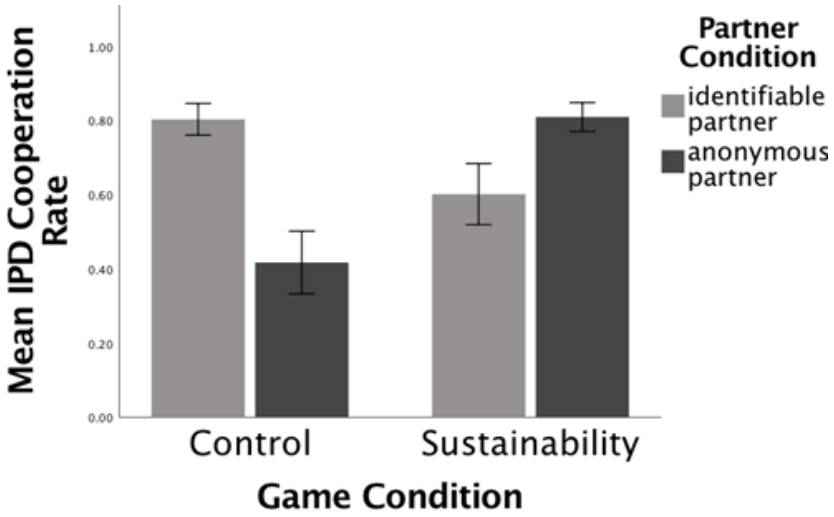

Figure B-4: Means for IPD Cooperation Rate as a function of Game Type and Partner Type.

	Sustainability Game	Tetris (control)	Total
Anonymous Partner	0.80 (0.17)	0.41 (0.38)	0.61 (0.06)
Identifiable Partner	0.60 (0.33)	0.80 (0.17)	0.70 (0.05)
Total	0.72 (0.27)	0.58 (0.36)	

Table B.4: Means for Prisoner' Dilemma Cooperation Rate as a function of Game Type and Partner Type (Std.Dev. in parenthesis).

played the Sustainability Game ($M = 0.80$, $SE = 0.04$) compared to Tetris ($M = 0.41$, $SE = 0.08$), $F(1,39) = 17.77$, $p < 0.001$, $\eta_p^2 = 0.32$. Conversely, individuals who played with identifiable partners had significantly lower cooperation rates if they played the Sustainability Game ($M = 0.60$, $SE = 0.08$) compared to Tetris ($M = 0.80$, $SE = 0.04$), $F(1,31) = 4.73$, $p = .038$, $\eta_p^2 = 0.14$. Furthermore, in the Tetris condition those with identifiable partners had significantly higher cooperation rate ($M = 0.80$, $SE = 0.04$), than those with anonymous partners ($M = 0.41$, $SE = 0.08$), $F(1,35) = 14.29$, $p = .001$, $\eta_p^2 = 0.29$, but in the Sustainability condition those with identifiable partners had significantly lower cooperation rates (M = 0.60, $SE = 0.08$) than those with anonymous partners ($M = 0.80$, $SE = 0.04$), $F(1,35) = 5.93$, $p = .020$, $\eta_p^2 = 0.15$. These results suggest that the Sustainability Game facilitates cooperation under conditions of anonymity but attenuates cooperation when participants' partners are identifiable.

We then constructed a GLM to examine the effect of Game Type, Partner Type, and their interaction on the average sentence participants received, which is a measure of participant performance in the IPD. (See Table 2 for means of average sen-

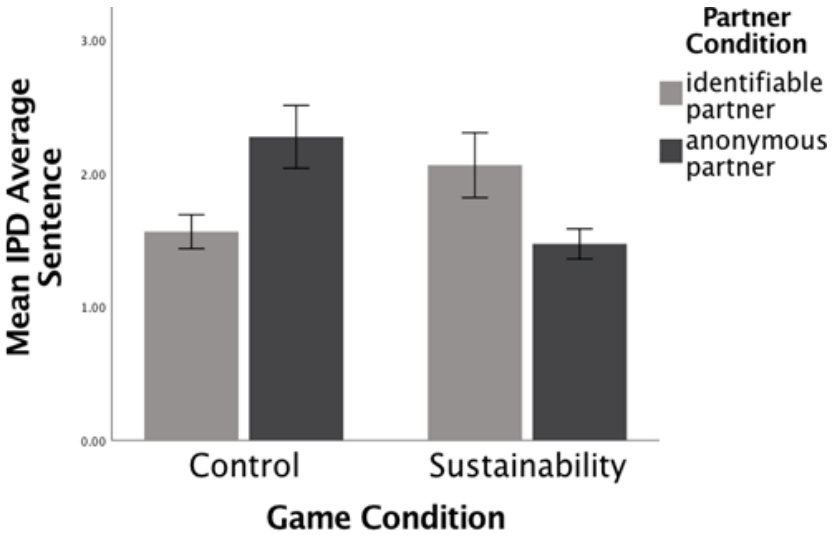

Figure B-5: Means for IPD Average Sentence as a function of Game Type and Partner Type.

tence by Game Type and Partner Type). There was no main effect of Game Type $F(1,71) = .64$, $p = .428$, $\eta_p^2 = 0.00$ or Partner Type $F(1,71) = .10$, $p = .751$, $\eta_p^2 = 0.00$. There was, however, a significant two-way interaction between Game Type and Partner Type $F(1,71) = 11.66$, $p = .001$, $\eta_p^2 = 0.12$ (see Figure B-5). Individuals who played with anonymous partners had significantly lower sentences (which indicates better performance) if they played the Sustainability Game ($M = 1.47$, $SE = 0.11$) compared to Tetris ($M = 2.27$, $SE = 0.24$), $F(1,39) = 9.46$, $p = 0.004$, $\eta_p^2 = 0.20$. Conversely, individuals who played with identifiable partners had marginally higher sentences if they played the Sustainability Game (M = 2.06, SE = .24) compared to Tetris ($M = 1.56$, $SE = 0.13$), $F(1, 31) = 3.31$, $p = 0.079$, $\eta_p^2 = 0.10$, 90% CI [.00, .27]. Furthermore, in the Tetris condition those with identifiable partners had significantly lower sentencing decisions ($M = 1.56$, $SE = 0.13$), than those with anonymous partners (M = 2.27, SE=.24), F(1, 35) = 6.14, $p = .018$, $\eta_p^2 = 0.15$, 90% CI [0.01, 0.32], but in the Sustainability condition those with identifiable partners had significantly higher sentencing decisions (M = 2.06, SE = .24) than those with anonymous partners ($M = 1.47$, $SE = .11$) $F(1, 35) = 5.54$, $p = 0.025$, $\eta_p^2 = 0.14$, 90% CI [.01, .31]. These results suggest that the Sustainability Game facilitates game performance under conditions of anonymity but attenuates performance when participants' partners are identifiable.

Table B.5: Means for rate of fair offers in the Ultimatum Game as a function of Game Type and Partner Type (standard deviation in parenthesis).

	Sustainability Game	Tetris (control)	Total
Anonymous Partner	.90 (.31)	.88 (.34)	.70 (.46)
Identifiable Partner	.75 (.44)	.50 (.51)	1.87 (.39)
Total	.83 (.38)	.66 (.48)	

B.6.3 The Ultimatum Game

Next, we sought to examine how Game Type and Partner influenced individuals' behavior in the Ultimatum game. To examine cooperative behavior and performance in the Ultimatum Game, we constructed GLMs modeling: (1) rate of fair offers (10 tokens to the giver and 10 tokens to the recipient) and (2) monetary payoff ($0.20 * number of tokens remaining after participants played the one-shot Ultimatum Game); as function of Game Type condition, Partner Type condition, and their interaction.

We first constructed a GLM to examine the effect of Game Type and Partner type, and their interaction on participants' cooperative behavior in the Ultimatum Game (see Table 3 for means of rate of fair offers by Game Type and Partner Type). There was no main effect of Game Type $F(1,71) = 1.97$, $p = 0.164$, $\eta_p^2 = 0.03$ or Partner Type $F(1,71) = 1.33$, $p = 0.254$, $\eta_p^2 = 0.09$ on rate of fair offers. There was, however, a significant two-way interaction between Game Type and Partner Type $F(1,71) = 7.21$,$p = 0.009$, $\eta_p^2 = 0.10$ (see Figure 5). An effect of Game Type only emerges under conditions of anonymity. Those in the Sustainability Game condition with anonymous partners are significantly more cooperative (M = .90, SE = .07) than those in the Tetris condition who have anonymous partners ($M = 0.50$, $SE = 0.11$), $F(1,35) = 8.94$, $p = 0.005$, $\eta_p^2 = 0.20$. Similarly, an effect of Partner Type only emerges in the Tetris condition. For those who play Tetris, participants with identifiable partners are significantly more cooperative ($M = 0.88$, $SE = .09$) than those with anonymous partners ($M = 0.50$, $SE = 0.11$), $F(1,35) = 6.29$, $p = 0.017$, $\eta_p^2 = 0.16$. These results suggest that the Sustainability Game mitigates the general tendency for individuals to act less cooperatively towards anonymous partners in the Ultimatum Game.

Next, we constructed a GLM to examine the effect of Game Type, Partner Type, and their interaction on participants' monetary payoff in the Ultimatum Game. There was no significant main effect of Game Type $F(1,71) = 1.55$, $p = 0.218$, $\eta_p^2 = 0.02$ or Partner Type $F(1,71) = 1.55$, $p = 0.218$, $\eta_p^2 = 0.02$, and there was no significant interaction, $F(1,71) = 2.56$, $p = 0.115$, $\eta_p^2 = 0.04$.

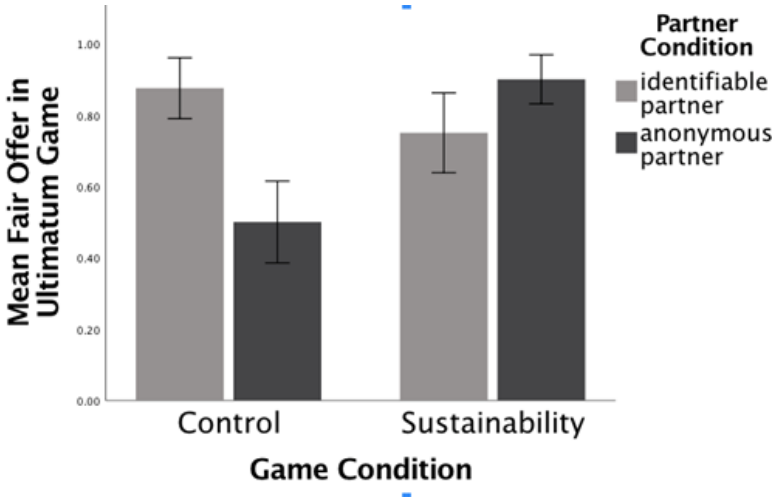

Figure B-6: Means for rate of fair offers in the Ultimatum Game as a function of Game Type and Partner Type.

B.6.4 Iterated Public Goods Game

We sought to examine how Game Type and Partner Type influenced individuals' behavior in the Public Goods Game. To examine cooperative behavior and performance in the Public Goods Game, we constructed GLMs modeling: (1) contribution decisions (number of tokens contributed to the public fund) and (2) monetary payoff (number of tokens earned/(number of trials * 3) ; as function of Game Type condition, Partner Type condition, and their interaction.

We first constructed a GLM to examine the effect of Game Type and Partner Type, and their interaction on participants' contribution in the Public Goods Game. (see Table 4 for means of contribution decisions by Game Type and Partner Type). There was no main effect of Game Type $F(1,71) = 0.05$, $p = 0.826$, $\eta_p^2 = 0.00$. There was a significant main effect of Partner Type $F(1,71) = 5.46$, $p = 0.022$, $\eta_p^2 = 0.07$. Those in the identifiable partner condition contributed more to the public fund ($M = 16.11$, $SE = 0.86$) than those in the anonymous partner condition ($M = 13.40$, $SE = 0.77$). There was no significant interaction $F(1,71) = 2.767$, $p = 0.101$, $\eta_p^2 = 0.04$, but simple slope analyses indicate that Partner Type significantly predicts contributions for those who played Tetris $F(1,35) = 9.52$, $p = 0.004$, $\eta_p^2 = 0.219$), but not for those who played the Sustainability Game, $F(1,35) = 0.196$, $p = 0.661$, $\eta_p^2 = 0.006$. While the interaction is not significant, the simple slope results seem to follow the pattern of results we saw in the Ultimatum Game and suggest that the Sustainability Game may mitigate the general tendency for individuals to act less cooperatively towards anonymous partners

Table B.6: Means for contribution decisions in the Public Goods Game as a function of Game Type and Partner Type (standard deviation in parenthesis).

	Sustainability Game	Tetris (control)	Total
Anonymous Partner	12.60 (5.60)	10.99 (5.60)	13.40 (5.46)
Identifiable Partner	15.28 (4.47)	16.95 (3.74)	16.11 (4.41)
Total	14.84 (5.21)	14.37 (5.00)	

in the Iterated Public Goods Game.

We then constructed a GLM to examine the effect of Game Type, Partner Type, and their interaction on the Public Goods Game payoff. There was no significant main effect of Game Type $F(1,71) = 0.03$, $p = 0.861$, $\eta_p^2 = 0.00$ or Partner Type $F(1,71) = 1.84$, $p = 0.179$, $\eta_p^2 = 0.03$, and there was no significant interaction $F(1,71) = 1.32$, $p = 0.255$, $\eta_p^2 = 0.02$.

B.6.5 Endorsement of Competitive and Cooperative Strategy

Finally, we examined how Game Type and Partner Type influenced individuals' reliance on cooperative and competitive strategies. We constructed GLM to examine the effect of Game Type, Partner Type, and their interaction on cooperation and competition. There was no main effect of Game Type $F(1,71) = 1.21$, $p = 0.273$, $\eta_p^2 = 0.00$ or Partner Type $F(1,71) = 0.21$, $p = 0.646$, $\eta_p^2 = 0.00$ on endorsement of cooperative strategies. There was also no significant interaction $F(1,71) = 0.02$, $p = 0.881$, $\eta_p^2 = 0.00$. Similarly, there was no main effect of Game Type $F(1,71) = 2.17$, $p = 0.180$, $\eta_p^2 = 0.03$ or Partner Type $F(1,71) = 1.84$, $p = 0.145$, $\eta_p^2 = 0.03$. There was also no significant interaction $F(1,71) = 0.00$, $p = 0.965$, $\eta_p^2 = 0.00$.

B.7 Discussion

Our findings suggest that even a short-time exposure to the Sustainability Game increases cooperative behaviour in various behaviour-economic games when one's partner is anonymous, but not when one's partner is identifiable. Past research has demonstrated that explicit learning of economic payoffs does not benefit people's public goods investments (Herrmann, Thöni and Gächter, 2008; Burton-Chellew, El Mouden and West, 2017). The present study demonstrates that the Sustainability Game intervention can increase cooperation in anonymised contexts, which is how public goods experiments are typically conducted (Andreoni and Petrie, 2004). This finding suggests that implicit expression of cooperative dynamics may lead to greater investments in public

goods. Importantly, investments in public goods, such as contributing to infrastructure and schools, can positively affect social relations, economic well-being, and larger societal structures (Keltner, 2009).

In addition to facilitating cooperation in an anonymous context, the Sustainability Game intervention also created conditions where having information about one's partner promoted competition (Nikiforakis, 2010; Burton-Chellew, El Mouden and West, 2017). This suggests that the intervention may actually adversely affect public goods investments when one's competitive partner has been identified. Cooperation in an anonymous condition is related to the capability to signal and recognise a willingness to cooperate (Brosig, 2002; Burton-Chellew, El Mouden and West, 2017). Individuals that are inclined to cooperate, however, can utilize this capability only if they have the opportunity to communicate. In the anonymous condition, subjects had to interact through an online chat system, thus, limiting their capability to communicate any such signals. These signals, as demonstrated by Brosig (2002), promote an understanding of the advantages of cooperation. In case of our experiments, this knowledge was passed to the subjects implicitly, when they played the Sustainability Game.

We hoped that exposure to the dynamics of The Sustainability Game would promote player cooperation in subsequent tasks, but we also knew there was a chance that the game could make players more cognizant of a cooperative context, which can make players act more selfishly (Bear, Kagan and Rand, 2017). We did not anticipate, however, that the effect of the Sustainability game on cooperation would depend on whether one's partner is anonymous or identifiable. And, perhaps, it has not. It may be that both results indicate increased sophistication in understanding how to utilise cooperative behaviour to achieve goals, but the one-on-one context is significantly more likely to make those goals competitive.

Our results for the usage of the game are consistent with the literature at large for the use of agent-based models and other computer games in the form of serious games to achieve behaviour change (Cheong et al., 2011; Gentile et al., 2014; Scarlatos, Tomkiewicz and Courtney, 2013; Di Ferdinando et al., 2015; De Angeli, 2018). They also add to the on-going discussions regarding the influence of video games in our society. While scholars debate whether violent —or even non-violent— 'harm' children and adolescent (Ferguson, 2015; Boxer, Groves and Docherty, 2015), we also shown that they can be a force for good, by promoting pro-social behaviours.

B.8 Conclusions

Cooperative behaviour is a fundamental strategy not only for survival, but also to enable us to produce and enforce governance. Cooperation promotes pro-social behaviour, positively affects economies and social relationships, and makes larger societal structures possible. People vary, however, in their willingness to engage in cooperative behaviour. Previous research has shown that explicit knowledge of the benefits of cooperation in the form of public goods investments does not universally promote that investment, even when doing so is beneficial to the individual and group.

We have demonstrated success in creating an intervention, the Sustainability Game, that alters cooperation, and indeed in even partially anonymous cases, increases it versus standard outcomes from explicit instruction such as the instructions for the public goods games. Our findings suggest that even a short exposure to the Sustainability Game increases cooperative behaviour in various well-established measures when one's partner is unknown, but not when one's partner is clearly identifiable. While our intervention to make the dynamics of human cooperation more transparent to users, we will need to work further to fully disentangle its impacts and their implications for understanding human cooperation. However, our study successfully demonstrate how AI can be used to help us understand cognition in order to increase cooperative behaviour and, hopefully, the wellbeing and sustainability of our societies.

However, I would like to also raise a serious concern. Our subjects managed to gain implicit knowledge with 20 minutes exposure. Bots were used by populist movements to disseminate information and engage in interactions with other users of social media. Evidence show that mass manipulation altered the outcomes of the UK's EU membership referendum (Howard and Kollanyi, 2016; Bastos and Mercea, 2017), the US presidential election (Howard, Woolley and Calo, 2018), and attempted to disrupt French Elections (Ferrara, 2017). At the same time, gamification is increasingly used to 'trap' consumers into reinforcement loops (Deterding et al., 2011) and now it is finding its way into politics (Joy, 2017). I believe that this further raises the need for legislation of AI, which promotes both transparency and accountability, something that I will discuss on the following chapter. Users interacting with intelligent systems should know when they do so. It is also equally important, at least in political settings, that users should be made aware of the political donors who funded the system.

Appendix C

Complete Set of Results for ABOD3-AR Study

Question	Group 1 ($N = 23$)	Group 2 ($N = 22$)	p-value
Fake - Natural	2.39 (σ =1.033)	2.55 (σ =1.143)	0.638
Machinelike - Humanlike	1.87 (σ =1.014)	1.41 (σ =0.796)	0.97
Unconscious - Conscious	2.26 (σ =1.096)	2.50 (σ =1.185)	0.487
Inconscient - Conscient	2.61 (σ =1.196)	2.50 (σ =1.012)	0.743
(Moving) Rigidly - Elegantly	2.09 (σ =1.041)	2.45 (σ =0.963)	0.225

Table C.1: Godspeed Questions related to perceived anthropomorphism.

Question	Group 1 ($N = 23$)	Group 2 ($N = 22$)	p-value
Dead - Alive	2.39 (σ =1.033)	3.27 (σ =1.202)	**0.01**
Stagnant - Lively	3.30 (σ =0.926)	4.14 (σ =0.710)	**0.02**
Mechanical - Organic	1.91 (σ =1.276)	1.45 (σ =0.8)	0.158
Artificial - Lifelike	1.96 (σ =1.065)	1.95 (σ =1.214)	0.995
Inert - Interactive	3.26 (σ =1.176)	3.68 (σ =1.041)	0.211
Apathetic - Responsive	3.35 (σ =0.982)	3.64 (σ =1.136)	0.368

Table C.2: Godspeed Questions related to perceived animacy.

Question	Group 1 ($N = 23$)	Group 2 ($N = 22$)	p-value
Dislike - Like	3.57 (σ =0.728)	3.77 (σ =1.02)	0.435
Unfriendly - Friendly	3.17 (σ =1.029)	3.77 (σ =0.869)	**0.041**
Unpleasant - Pleasant	3.43 (σ =0.788)	3.77 (σ =1.066)	0.232
Awful - Nice	3.61 (σ =0.656)	3.77 (σ =0.922)	0.494

Table C.3: Godspeed Questions related to perceived likeability.

Question	Group 1 ($N = 23$)	Group 2 ($N = 22$)	p-value
Incompetent - Competent	3.13 (σ =0.815)	3.55 (σ =1.143)	0.171
Ignorant - Knowledgeable	2.7 (σ =1.063)	2.81 (σ =0.873)	0.699
Irresponsible - Responsible	2.65 (σ =1.027)	2.81 (σ =0.814)	0.579
Unintelligent - Intelligent	3.17 (σ =0.937)	3.14 (σ =1.153)	0.922
Foolish - Sensible	3.43 (σ =0.728)	3.43 (σ =0.926)	0.98

Table C.4: Godspeed Questions related to perceived intelligence.

Question	Group 1 ($N = 23$)	Group 2 ($N = 22$)	p-value
Anxious - Relaxed	4.15 (σ =0.933)	3.81 (σ =1.167)	0.308
Agitated - Calm	4.1 (σ =0.852)	4.05 (σ =0.071)	0.863
Quiscent - Surprised	2.45 (σ =0.945)	2.86 (σ =1.062)	0.203

Table C.5: Godspeed Questions related to perceived safety.

Appendix D

Complete Set of Results for Chapter 6

D.1 Quantitative Results for Difference on Type of Agent

Question	N	Mean (SD)	t	p	η_p^2
Incompetent - Competence					
Group 1: Human Driver	18	2.8 (0.9)			
Group 2: Opaque AV	16	2.9 (1.2)			
			-.26	.8	0.003
Ignorant - Knowledgeable					
Group 1: Human Driver	17	2.5 (0.8)			
Group 2: Opaque AV	16	2.8 (1.1)			
			-.8	.43	0.002
Irresponsible - Responsible					
Group 1: Human Driver	17	2.4	0.8		
Group 2: Opaque AV	16	2.5	1.2		
			-.25	0.8	0.005
Unintelligent - Intelligent					
Group 1: Human Driver	17	2.9 (0.6)			
Group 2: Opaque AV	16	2.9 (1.1)			
			.01	0.99	.004
Foolish - Sensible					
Group 1: Human Driver	17	2.5 (0.9)			
Group 2: Opaque AV	16	2.8 (0.9)			
			-.85	0.4	0.00

Table D.1: Human Driver compare to Opaque AV: Godspeed questions (scale 1-5) related to the perceived intelligence of the driver/AV.

Question	N	Mean (SD)	t	p	η_p^2
Dislike - Like					
Group 1: Human Driver	17	2.6 (0.49)			
Group 2: Opaque AV	16	2.3 (0.93)			
			1.28	0.21	0.075
Unkind - Friendly					
Group 1: Human Driver	17	2.6 (0.7)			
Group 2: Opaque AV	16	2.4 (0.81)			
			.75	.46	0.021
Unpleasant - Pleasant					
Group 1: Human Driver	17	3 (0.35)			
Group 2: Opaque AV	16	2.6 (0.89)			
			1.38	0.18	0.105
Awful - Nice					
Group 1: Human Driver	17	3 (0.0)			
Group 2: Opaque AV	16	2.6 (0.89)			
			1.53	0.13	0.124

Table D.2: Human Driver compare to Opaque AV: Godspeed questions (scale 1-5) related to the perceived likability of the driver/AV.

Question	N	Mean (SD)	t	p	η_p^2
Machinelike - Humanlike					
Group 1: Human Driver	17	3.2 (0.97)			
Group 2: Opaque AV	16	2.1 (0.96)			
			3.42	**0.001**	0.191
Unconscious - Conscious					
Group 1: Human Driver	17	3 (1.17)			
Group 2: Opaque AV	16	2.75 (1.34)			
			0.67	0.515	0.079

Table D.3: Human Driver compare to Opaque AV: Godspeed questions (scale 1-5) related to the perceived anthropomorphism of the driver/AV.

Question	N	Mean (SD)	t	p	η_p^2	
Immoral - Moral						
Group 1: Human Driver	17	2.47 (0.8)				
Group 2: Opaque AV	16	2.56 (1.2)				
			-.25	0.8	0.001	
Unfair - Fair						
Group 1: Human Driver	17	2.24 (0.83)				
Group 2: Opaque AV	16	2.69 (1.2)				
			-1.11	0.27 (0.014)		
Morally Culpable						
Group 1: Human Driver	16	3.37 (0.7)				
Group 2: Opaque AV	16	2.44 (1.21)				
			-2.07	**0.04**	0.18	
Blame						
Group 1: Human Driver	15	2.07 (0.7)				
Group 2: Opaque AV	16	2.44 (1.21)				
				-0.94	0.354	0.020

Table D.4: Human Driver compare to Opaque AV: Questions related to the perceived moral agency of the driver/AV.

Question	N	Mean (SD)	t	p	η_p^2
Society Interest					
Group 1: Human Driver	12	2.75 (0.97)			
Group 2: Opaque AV	16	2.69 (1.08)			
			.15	0.88	0.004
Own Interest					
Group 1: Human Driver	12	2.5 (0.9)			
Group 2: Opaque AV	16	2.56 (1.15)			
			-.16	.87	0.007
Decisions Agree With					
Group 1: Human Driver	12	2.42 (0.79)			
Group 2: Opaque AV	16	2.5 (0.89)			
			-2.22	0.82	0.001

Table D.5: Human Driver compare to Opaque AV: Questions (scale 1-5) related to the trust and justification of actions by the driver/AV.

Question	N	Mean (SD)	t	p	η_p^2
Race					
Group 1: Human Driver	14	3.21 (0.7)			
Group 2: Opaque AV	16	3 (1.21)			
			.56	0.58	0.031
Gender					
Group 1: Human Driver	15	3.73 (0.88)			
Group 2: Opaque AV	16	3.81 (1.17)			
			-.18	.86	0.008
Occupation					
Group 1: Human Driver	15	3.73 (0.96)			
Group 2: Opaque AV	15	4.07 (1.16)			
			-.82	0.41	0.006
Body Size					
Group 1: Human Driver	14	3.64 (0.93)			
Group 2: Opaque AV	16	3.75 (1.18)			
			-.26	0.8	0.007
Age					
Group 1: Human Driver	14	3.07	0.47		
Group 2: Opaque AV	16	2.94	1.12		
			0.4	0.69	0.019

Table D.6: Human Driver compare to Opaque AV: Questions (scale 1-5) related to the perceived prejudice of the actions.

Question	N	Mean (SD)	t	p	η_p^2
Subjective - Objective					
Group 1: Human Driver	17	2.7 (0.77)			
Group 2: Opaque AV	16	3.31 (1.62)			
			-1.4	0.17	0.043
Deterministic - Undeterministic					
Group 1: Human Driver	17	3.12 (1.11)			
Group 2: Opaque AV	16	3.5 (0.97)			
			-1.09	.28	0.007
Unpredictable - Predictable					
Group 1: Human Driver	17	3.06 (1.34)			
Group 2: Opaque AV	16	3.31 (1.4)			
			-.54	0.59	0.002
Intentional - Unintentional					
Group 1: Human Driver	17	2.94 (1.14)			
Group 2: Opaque AV	16	3.31 (1.25)			
			-.91	.37	0.010

Table D.7: Human Driver compare to Opaque AV: Questions (scale 1-5) related Objective/Deterministic measures.

Question	N	Mean (SD)	t (df)	p	η_p^2
Godspeed Questionnaire (Scale 1-5)					
Incompetent - Competent Group 1: Human	18	2.8 (0.88)			
Group 3: Transparent AV	17	2.65 (1.17)			
			0.38 (32)	0.71	0.004
Ignorant - Knowledgeable					
Group 1: Human AV	17	2.53 (0.87)			
Group 3: Transparent AV	17	2.76 (1.09)			
			-0.7 (32)	0.49	0.015
Irresponsible - Responsible					
Group 1: Human AV	17	2.41 (0.8)			
Group 3: Transparent AV	17	2.65 (1.06)			
			-0.73 (32)	0.47	0.017
Unintelligent - Intelligent					
Group 1: Human AV	17	2.94 (0.56)			
Group 3: Transparent AV	17	2.76 (1.3)			
			0.51 (32)	**0.61**	0.008
Foolish - Sensible					
Group 1: Human AV	17	2.53 (0.87)			
Group 3: Transparent AV	17	3.0 (1.12)			
			-1.37 (32)	0.18	0.055

Table D.8: Human Driver v Transparent AV: Godspeed questions related to the perceived intelligence of the driver/AV.

Question	N	Mean (SD)	t (df)	p	η_p^2
Godspeed Questionnaire (Scale 1-5)					
Dislike - Like Group 1: Human	17	2.65 (0.49)			
Group 3: Transparent AV	17	2.47 (1.12)			
			0.59 (32)	0.56	0.011
Unkind - Kind					
Group 1: Human AV	17	2.65 (0.7)			
Group 3: Transparent AV	17	2.29 (0.85)			
			1.32 (32)	0.2	0.052
Unpleasant - Pleasant					
Group 1: Human AV	17	3.0 (0.35)			
Group 3: Transparent AV	17	2.35 (0.93)			
			2.68 (32)	**0.01**	0.183
Awful - Nice					
Group 1: Human AV	17	3.0 (0.0)			
Group 3: Transparent AV	17	2.47 (0.87)			
			2.5 (32)	**0.018**	0.163

Table D.9: Human Driver compare to Transparent AV: Godspeed questions (scale 1-5) related to the perceived likability of the driver/AV.

Question	N	Mean (SD)	t (df)	p	η_p^2
Godspeed Questionnaire (Scale 1-5)					
Machinelike - Humanlike					
Group 1: Human	17	3.24 (0.97)			
Group 3: Transparent AV	18	1.5 (0.92)			
			5.42 (33)	**0.000**	0.47
Unconscious - Conscious					
Group 1: Human AV	17	3.0 (1.17)			
Group 3: Transparent AV	18	1.33 (0.59)			
			5.35 (33)	**0.000**	0.464

Table D.10: Human Driver compare to Transparent AV: Godspeed questions (scale 1-5) related to the perceived anthropomorphism of the driver/AV.

Question	N	Mean (SD)	t (df)	p	η_p^2
Subjective - Objective					
Group 1: Human	17	2.71 (0.77)			
Group 3: Transparent AV	18	3.39 (1.2)			
			-2 (33)	0.54	0.108
Deterministic - Undeterministic					
Group 1: Human AV	17	2.89 (1.11)			
Group 3: Transparent AV	17	2.0 (1.0)			
			2.43 (32)	**0.02**	0.156
Unpredictable - Predictable					
Group 1: Human AV	17	3.06 (1.34)			
Group 3: Transparent AV	18	4.0 (1.29)			
			-2.12 (33)	**0.04**	0.120
Intentional - Unintentional					
Group 1: Human AV	17	3.09 (1.14)			
Group 3: Transparent AV	18	1.83 (1.2)			
			3.09 (33)	**0.004**	0.224

Table D.11: Human Driver compare to Transparent AV: Questions (scale 1-5) related Objective/Deterministic measures.

Question	N	Mean (SD)	t (df)	p	η_p^2
Immoral - Moral					
Group 1: Human	17	2.47 (0.8)			
Group 3: Transparent AV	18	2.67 (1.08)			
			-0.61 (33)	0.58	0.011
Unfair - Fair					
Group 1: Human AV	17	2.24 (0.83)			
Group 3: Transparent AV	18	2.5 (1.38)			
			-0.68 (33)	0.51	0.014
Morally Culpable					
Group 1: Human	16	3.37 (0.72)			
Group 3: Transparent AV	18	3.05 (1.3)			
			-3.89 (32)	**0.00**	0.321
Blame					
Group 1: Human AV	15	2.07 (0.7)			
Group 3: Transparent AV	18	3 (1.28)			
			-2.52 (31)	**0.02**	0.169

Table D.12: Human Driver compare to Transparent AV: Questions (scale 1-5) related to the perceived morality of the driver/AV.

Question	N	Mean (SD)	t (df)	p	η_p^2
Society Interest					
Group 1: Human	12	2.75 (0.97)			
Group 3: Transparent AV	18	2.39 (1.14)			
			0.9 (28)	0.38	0.028
Own Interest					
Group 1: Human AV	12	2.5 (0.9)			
Group 3: Transparent AV	18	2.17 (0.71)			
			1.13 (28)	0.38	0.044
Decisions Agree With					
Group 1: Human AV	12	2.42 (0.79)			
Group 3: Transparent AV	18	2.33 (1.14)			
			0.22 (28)	0.82	0.002

Table D.13: Human Driver compare to Transparent AV: Questions (scale 1-5) related to the trust and justification of actions by the driver/AV.

Question	N	Mean (SD)	t (df)	p	η_p^2
Race					
Group 1: Human	14	3.21 (0.7)			
Group 3: Transparent AV	18	2.72 (1.07)			
			1.49 (30)	0.2	0.068
Gender					
Group 1: Human AV	15	3.73 (0.88)			
Group 3: Transparent AV	18	3.67 (1.37)			
			0.17 (31)	0.87	0.001
Occupation					
Group 1: Human AV	15	3.73 (0.96)			
Group 3: Transparent AV	18	4 (1.24)			
			-0.68 (31)	0.5	0.015
Body Size					
Group 1: Human AV	14	3.64 (0.93)			
Group 3: Transparent AV	18	3.72 (1.32)			
			-0.19 (30)	0.85	0.001
Age					
Group 1: Human AV	14	3.07 (0.47)			
Group 3: Transparent AV	18	3.0 (0.97)			
			0.25 (30)	0.82	0.002

Table D.14: Human Driver compare to Transparent AV: Questions (scale 1-5) related to the perceived prejudice of the actions.

Question	N	Mean (SD)	t (df)	p	η_p^2
Suspicious					
Group 1: Human	12	3.17 (0.83)			
Group 3: Transparent AV	18	3.06 (1.35)			
			0.25 (28)	0.81	0.002
Integrity					
Group 1: Human AV	12	2.83 (0.83)			
Group 3: Transparent AV	18	2.39 (1.2)			
			1.12 (28)	0.21	0.043
Deceptive					
Group 1: Human AV	12	3.08 (0.51)			
Group 3: Transparent AV	18	2.39 (0.98)			
			2.25 (28)	0.55	0.153

Table D.15: Perceptions of Suspicion, Integrity, Deceptive

D.2 Quantitative Results for Difference in Level of Transparency

Question	N	Mean (SD)	t	p	η_p^2
Incompetent - Competence					
Group 2: Opaque AV	16	2.88	1.2		
Group 3: Transparent AV	17	2.65	1.17		
			-.58	.57	0.002
Ignorant - Knowledgeable					
Group 2: Opaque AV	16	2.81	1.12		
Group 3: Transparent AV	17	2.76	1.09		
			-.58	.57	0.002
Irresponsible - Responsible					
Group 2: Opaque AV	16	2.5	1.21		
Group 3: Transparent AV	17	2.65	1.06		
			-.58	0.8	0.017
Unintelligent - Intelligent					
Group 2: Opaque AV	16	2.94	1.18		
Group 3: Transparent AV	17	2.76	1.3		
			0.01	0.99	0.000
Foolish - Sensible					
Group 2: Opaque AV	16	2.81	0.94		
Group 3: Transparent AV	17	3	1.12		
			-.85	0.4	0.017

Table D.16: Opaque AV compare to Transparent AV: Godspeed questions (scale 1-5) related to the perceived intelligence of the AV.

Table D.17: Likability Measures Non-Transparent Compared to Transparent

Question	N	Mean (SD)	t	p	η_p^2	
Dislike - Like						
Group 2: Opaque AV	16	2.81	.91			
Group 3: Transparent AV	17	2.47	1.25			
			.69	.49	0.032	
Unkind - Kind						
Group 2: Opaque AV	16	2.44	.81			
Group 3: Transparent AV	17	2.47	1.12			
			-.5	0.62	0.000	
Unpleasant - Pleasant				-.97	0.34	0.003
Group 2: Opaque AV	16	2.63	.89			
Group 3: Transparent AV	17	2.29	.85			
Awful - Nice						
Group 2: Opaque AV	16	2.63	.89			
Group 3: Transparent AV	17	2.47	.87			
			-.61	0.54	0.000	

Table D.18: Opaque AV compare to Transparent AV: Godspeed questions (scale 1-5) related to the perceived likability of the AV.

Question	N	Mean (SD)	t	p	η_p^2
Machinelike - Humanlike					
Group 2: Opaque AV	16	3.2	0.97		
Group 3: Transparent AV	18	2.1	.96		
			-2.1	**0.04**	.084
Unconscious - Conscious					
Group 2: Opaque AV	16	2.75	1.34		
Group 3: Transparent AV	18	1.33	0.59		
			-4.09	**0.001**	0.294

Table D.19: Opaque AV compare to Transparent AV: Godspeed questions (scale 1-5) related to the perceived anthropomorphism of the AV.

Question	N	Mean (SD)	t	p	η_p^2
Immoral - Moral					
Group 2: Opaque AV	17	2.47	0.8		
Group 3: Transparent AV	16	2.56	1.2		
			.29	0.77	0.017
Unfair - Fair					
Group 2: Opaque AV	17	2.24	0.83		
Group 3: Transparent AV	16	2.69	1.2		
			-.47	0.64	0.002

Table D.20: Opaque AV compare to Transparent AV: Questions related to the perceived moral agency of the AV.

Question	N	Mean (SD)	t	p	η_p^2
Society Interest					
Group 2: Opaque AV	16	2.69	1.08		
Group 3: Transparent AV	18	2.39	1.14		
			-0.8	0.43	0.006
Own Interest					
Group 2: Opaque AV	16	2.56	1.15		
Group 3: Transparent AV	18	2.17	0.71		
			-1.13	.27	0.023
Decisions Agree With					
Group 2: Opaque AV	16	2.5	0.89		
Group 3: Transparent AV	18	2.33	1.14		
			-0.5	0.62	0.000

Table D.21: Opaque AV compare to Transparent AV: Questions (scale 1-5) related to the trust and justification of actions by the AV.

Question	N	Mean (SD)	t	p	η_p^2
Race					
Group 2: Opaque AV	16	3	1.21		
Group 3: Transparent AV	18	2.72	1.07		
			-.75	0.46	0.013
Gender					
Group 2: Opaque AV	16	3.81	1.17		
Group 3: Transparent AV	18	3.67	1.37		
			-.03	.98	0.001
Occupation					
Group 2: Opaque AV	15	4.07	1.16		
Group 3: Transparent AV	18	4	1.24		
			-.82	0.86	0.003
Body Size					
Group 2: Opaque AV	16	3.75	1.18		
Group 3: Transparent AV	18	3.72	1.32		
			-.07	0.94	0.004
Age					
Group 2: Opaque AV	16	2.94	1.12		
Group 3: Transparent AV	18	3	.97		
			.2	0.84	0.009

Table D.22: Opaque AV compare to Transparent AV: Questions (scale 1-5) related to the perceived prejudice of the actions.

Question	N	Mean (SD)	t	p	η_p^2	
Subjective - Objective						
Group 2: Opaque AV	16	3.3	1.62			
Group 3: Transparent AV	18	3.9	1.95			
			0.18	0.86	0.008	
Deterministic - Undeterministic						
Group 2: Opaque AV	16	2.5	0.97			
Group 3: Transparent AV	17	2.0	1.0			
				-1.43	.16	0.042
Unpredictable - Predictable						
Group 2: Opaque AV	16	3.31	1.4			
Group 3: Transparent AV	18	4.0	1.28			
			1.49	0.15	0.059	
Intentional - Unintentional						
Group 2: Opaque AV	16	2.69	1.25			
Group 3: Transparent AV	18	1.83	1.2			
			-2.13	**0.038**	0.082	

Table D.23: Opaque AV compare to Transparent AV: Questions (scale 1-5) related Objective/Deterministic measures.

Bibliography

Abt, C.C., 1970. *Serious games: The art and science of games that simulate life*, vol. 6. New York, NY, USA: Viking Press.

Alemzadeh, H., Raman, J., Leveson, N., Kalbarczyk, Z. and Iyer, R.K., 2016. Adverse Events in Robotic Surgery: A Retrospective Study of 14 Years of FDA Data. *PLOS ONE* [Online], 11(4), p.e0151470. Available from: `https://doi.org/10.1371/journal.pone.0151470`.

Aler Tubella, A., Theodorou, A., Dignum, F. and Dignum, V., 2019. Governance by Glass-Box: Implementing Transparent Moral Bounds for AI Behaviour. *Proceedings of the twenty-eighth international joint conference on artificial intelligence (ijcai'2019)*. To appear.

Ananny, M. and Crawford, K., 2018. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, 20(3), pp.973–989. Available from: `https://doi.org/10.1177/1461444816676645`.

Andreoni, J. and Petrie, R., 2004. Public goods experiments without confidentiality: a glimpse into fund-raising. *Journal of Public Economics* [Online], 88(7-8), pp.1605–1623. Available from: `https://doi.org/10.1016/S0047-2727(03)00040-9`.

Animator Controller, 2018. [Online]. Available from: `https://docs.unity3d.com/Manual/class-AnimatorController.html` [Accessed 2018-11-09].

Anjomshoae, S., Najjar, A., Calvaresi, D. and Främling, K., 2019. Explainable Agents and Robots : Results from a Systematic Literature Review [Online]. [Online]. Available from: `www.ifaamas.orghttp://www.diva-portal.org/smash/record.jsf?pid=diva2{%}3A1303810{&}dswid=-6145`.

Arkin, R.C., 1998. *Behavior-Based Robotics*. MIT Press.

Awad, E., 2017. *Moral machines : perception of moral judgment made by machines.* Ph.D. thesis.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.F. and Rahwan, I., 2018. The Moral Machine experiment. *Nature* [Online], 563(7729), pp.59–64. Available from: https://doi.org/10.1038/s41586-018-0637-6.

Axelrod, R., 1997. *The complexity of cooperation: Agent-based models of competition and colloboration.* Princeton University Press.

Azuma, R.T., 1997. *A Survey of Augmented Reality.*

Bainbridge, L., 1992. Mental models in cognitive skill: The example of industrial process operation. *Models in the Mind* [Online], pp.119–143. Available from: http://www.bainbrdg.demon.co.uk/Papers/MM.html.

Bartneck, C., Kulic, D., Croft, E. and Zoghbi, S., 2009. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics*, 1(1), pp.71–81. Available from: https://doi.org/10.1007/s12369-008-0001-3.

Bastian, B., Denson, T.F. and Haslam, N., 2013. The roles of dehumanization and moral outrage in retributive justice. *PloS ONE*, 8(4), p.e61842.

Bastos, M.T. and Mercea, D., 2017. The Brexit Botnet and User-Generated Hyperpartisan News. *Social Science Computer Review* [Online]. Available from: https://doi.org/10.1177/0894439317734157.

Batalla, J.M., Vasilakos, A. and Gajewski, M., 2017. Secure smart homes: Opportunities and challenges. *ACM Comput. Surv.* [Online], 50(5), pp.75:1–75:32. Available from: https://doi.org/10.1145/3122816.

Bateson, M., Nettle, D. and Roberts, G., 2006. Cues of being watched enhance cooperation in a real-world setting. *Biology Letters* [Online], 2(3), pp.412–414. Available from: https://doi.org/10.1098/rsbl.2006.0509.

Bear, A., Kagan, A. and Rand, D.G., 2017. Co-evolution of cooperation and cognition: The impact of imperfect deliberation and context-sensitive intuition. *Proceedings of the Royal Society B: Biological Sciences* [Online], 284(1851), p.20162326. Available from: https://doi.org/10.1098/rspb.2016.2326.

Beiker, S.A., 2012. Legal aspects of autonomous driving. *Santa Clara L. Rev.*, 52, p.1145.

Binmore, K. and Shaked, A., 2010. Experimental economics: Where next? *Journal of Economic Behavior and Organization* [Online], 73(1), pp.87–100. Available from: https://doi.org/10.1016/j.jebo.2008.10.019.

Biran, O. and Cotton, C., 2017. Explanation and Justification in Machine Learning: A Survey. *IJCAI Workshop on Explainable AI (XAI)* [Online], (August), pp.8–14. Available from: https://pdfs.semanticscholar.org/02e2/e79a77d8aabc1af1900ac80ceebac20abde4.pdf.

Bloom, C., Tan, J., Ramjohn, J. and Bauer, L., 2017. Self-driving cars and data collection: Privacy perceptions of networked autonomous vehicles [Online]. *Thirteenth symposium on usable privacy and security (SOUPS 2017)*. Santa Clara, CA: USENIX Association, pp.357–375. Available from: https://www.usenix.org/conference/soups2017/technical-sessions/presentation/bloom.

Boden, M., Bryson, J., Caldwell, D., Dautenhahn, K., Edwards, L., Kember, S., Newman, P., Parry, V., Pegman, G., Rodden, T., Sorell, T., Wallis, M., Whitby, B. and Winfield, A., 2011. Principles of robotics. The United Kingdom's Engineering and Physical Sciences Research Council (EPSRC). Web publication.

Bonnefon, J., Schriff, A. and Rahwan, I., 2016. The social dilemma of autonomous vehicles. *Science*, 352(6293), p.1573.

Bostrom, N., 2014. *Superintelligence: Paths, dangers, strategies*. 1st ed. New York, NY, USA: Oxford University Press, Inc.

Boxer, P., Groves, C.L. and Docherty, M., 2015. Video Games Do Indeed Influence Children and Adolescents' Aggression, Prosocial Behavior, and Academic Performance: A Clearer Reading of Ferguson (2015). *Perspectives on Psychological Science*, 10(5), pp.671–673. Available from: https://doi.org/10.1177/1745691615592239.

Brom, C., Gemrot, J., Bída, M., Burkert, O., Partington, S.J. and Bryson, J.J., 2006a. POSH tools for game agent development by students and non-programmers. In: Q. Mehdi, F. Mtenzi, B. Duggan and H. McAtamney, eds. *The ninth international computer games conference: AI, mobile, educational and serious games*. University of Wolverhampton, pp.126–133.

Brom, C., Gemrot, J., Michal, B., Ondrej, B., Partington, S.J. and Bryson, J.J., 2006b. Posh tools for game agent development by students and non-programmers. *... on Computer Games: ...*, pp.1–8.

Brooks, R.A., 1986. A robust layered control system for a mobile robot. *IEEE Journal of Robotics and Automation*.

Brooks, R.A., 1991a. Intelligence without representation. *Artificial Intelligence* [Online], 47(1-3), pp.139–159. Available from: https://doi.org/10.1016/0004-3702(91)90053-M.

Brooks, R.A., 1991b. New Approaches to Robotics. *Science* [Online], 253(5025), pp.1227–1232. Available from: https://doi.org/10.1126/science.253.5025.1227.

Brosig, J., 2002. Identifying cooperative behavior: Some experimental results in a prisoner's dilemma game. *Journal of Economic Behavior and Organization* [Online], 47(3), pp.275–290. Available from: https://doi.org/10.1016/S0167-2681(01)00211-6.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H.S., Roff, H., Allen, G.C., Steinhardt, J., Flynn, C., hÉigeartaigh, S.Ó., Beard, S., Belfield, H., Farquhar, S., Lyle, C., Crootof, R., Evans, O., Page, M., Bryson, J., Yampolskiy, R. and Amodei, D., 2018. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *CoRR* [Online], abs/1802.07228. 1802.07228, Available from: http://arxiv.org/abs/1802.07228.

Brundage, M. and Bryson, J.J., 2016. Smart Policies for Artificial Intelligence [Online]. [Online]. 1608.08196, Available from: http://arxiv.org/abs/1608.08196.

Bryson, J., Caulfield, T. and Drugowitsch, J., 2005a. Integrating life-like action selection into cycle-based agent simulation environments. *Proceedings of Agent* [Online], (May 2014), pp.1–14. Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.63.8806{&}rep=rep1{&}type=pdf.

Bryson, J.J., 2000a. Cross-paradigm analysis of autonomous agent architecture. *Journal of Experimental and Theoretical Artificial Intelligence*, 12(2), pp.165–190.

Bryson, J.J., 2000b. Hierarchy and sequence vs. full parallelism in reactive action selection architectures. *From animals to animats 6 (sab00)*. Cambridge, MA: MIT Press, pp.147–156.

Bryson, J.J., 2001. *Intelligence by Design : Principles of Modularity and Coordination for Engineering Complex Adaptive Agents*. Ph.D. thesis.

Bryson, J.J., 2003. Action selection and individuation in agent based modelling. In: D.L. Sallach and C. Macal, eds. *Proceedings of Agent 2003: Challenges in social simulation.* Argonne, IL: Argonne National Laboratory, pp.317–330.

Bryson, J.J., 2007. Embodiment versus memetics. *Mind & Society* [Online], 7(1), pp.77–94. Available from: `https://doi.org/10.1007/s11299-007-0044-4`.

Bryson, J.J., 2009. Age-related inhibition and learning effects: Evidence from transitive performance. *The* 31^{st} *annual meeting of the cognitive science society (CogSci 2009).* Amsterdam: Lawrence Erlbaum Associates, pp.3040–3045.

Bryson, J.J., 2010a. Crude, cheesy, second-rate consciousness. In: C. Hernández and J.G.R. Sanz, eds. *From brains to systems: Brain-inspired cognitive systems.* Madrid, pp.454–467.

Bryson, J.J., 2010b. Robots should be slaves. In: Y. Wilks, ed. *Close engagements with artificial companions: Key social, psychological, ethical and design issues.* Amsterdam: John Benjamins, pp.63–74.

Bryson, J.J., 2012. The Making of the EPSRC Principles of Robotics. 133(133), pp.14–15.

Bryson, J.J., 2015. Artificial intelligence and pro-social behaviour [Online]. *Collective agency and cooperation in natural and artificial systems: Explanation, implementation and simulation.* Springer, vol. 15, pp.281–306. Available from: `https://doi.org/10.1007/978-3-319-15515-9_15`.

Bryson, J.J., 2017. Three very different sources of bias in AI, and how to fix them [Online]. Available from: `https://joanna-bryson.blogspot.com/2017/07/three-very-different-sources-of-bias-in.html` [Accessed 2018-12-10].

Bryson, J.J., 2018. Patiency is not a virtue: the design of intelligent systems and systems of ethics. *Ethics and Information Technology* [Online], 20(1), pp.15–26. Available from: `https://doi.org/10.1007/s10676-018-9448-6`.

Bryson, J.J., Ando, Y. and Lehmann, H., 2007. Agent-based modelling as scientific method: a case study analysing primate social behaviour. *Philosophical Transactions of the Royal Society B: Biological Sciences* [Online], 362(1485), pp.1685–1699. Available from: `https://doi.org/10.1098/rstb.2007.2061`.

Bryson, J.J., Caulfield, T.J. and Drugowitsch, J., 2005b. Integrating life-like action selection into cycle-based agent simulation environments. In: M. North, D.L. Sallach

and C. Macal, eds. *Proceedings of Agent 2005: Generative social processes, models, and mechanisms*. Chicago: Argonne National Laboratory, pp.67–81.

Bryson, J.J., Diamantis, M.E. and Grant, T.D., 2017. Of, for, and by the people: the legal lacuna of synthetic persons. *Artificial Intelligence and Law*, 25(3), pp.273–291. Available from: https://doi.org/10.1007/s10506-017-9214-9.

Bryson, J.J. and Kime, P., 1998. Just another artifact: Ethics and the empirical experience of AI. *Fifteenth international congress on cybernetics*. pp.385–390.

Bryson, J.J. and Kime, P.P., 2011. Just an artifact: Why machines are perceived as moral agents. *Proceedings of the* 22^{nd} *international joint conference on artificial intelligence*. Barcelona: Morgan Kaufmann, pp.1641–1646.

Bryson, J.J., Mitchell, J., Powers, S.T. and Sylvester, K., 2014. Understanding and Addressing Cultural Variation in Costly Antisocial Punishment [Online]. *Applied evolutionary anthropology*. pp.201–222. Available from: https://doi.org/10.1007/978-1-4939-0280-4.

Bryson, J.J. and Tanguy, E.A.R., 2010. Simplifying the design of human-like behaviour: Emotions as durative dynamic state for action selection. *International Journal of Synthetic Emotions*, 1(1), pp.30–50.

Bryson, J.J. and Theodorou, A., 2019. How Society Can Maintain Human-Centric Artificial Intelligence. In: M. Toivonen-Noro, E. Saari, H. Melkas and M. Hasu, eds. *Human-centered digitalization and services*. Springer.

Bryson, J.J. and Winfield, A., 2017. Standardizing Ethical Design for Artificial Intelligence and Autonomous Systems. *Computer* [Online], 50(5), pp.116–119. Available from: https://doi.org/10.1109/MC.2017.154.

Burton-Chellew, M.N., El Mouden, C. and West, S.A., 2017. Social learning and the demise of costly cooperation in humans. *Proceedings of the Royal Society B: Biological Sciences* [Online], 284(1853), p.20170067. Available from: https://doi.org/10.1098/rspb.2017.0067.

BWAPI: An API for interacting with StarCraft: Broodwar, n.d. [Online]. Available from: https://github.com/bwapi/bwapi.

Caliskan, A., Bryson, J.J. and Narayanan, A., 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* [Online], 356(6334), pp.183–186. 1608.07187, Available from: https://doi.org/10.1126/science.aal4230.

Cameron, D., Collins, E.C., Chua, A., Fernando, S., McAree, O., Martinez-Hernandez, U., Aitken, J.M., Boorman, L. and Law, J., 2015. Help! I can't reach the buttons: Facilitating helping behaviors towards robots. *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)*. vol. 9222, pp.354–358. Available from: https://doi.org/10.1007/978-3-319-22979-9_35.

Cannon-Bowers, J.A., Salas, E. and Converse, S., 1993. Shared mental models in expert team decision making. *Individual and group decision making: Current issues.* Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc, pp.221–246.

Carsten Stahl, B., 2004. Information, Ethics, and Computers: The Problem of Autonomous Moral Agents. *Minds and Machines* [Online], 14(1), pp.67–83. Available from: https://doi.org/10.1023/B:MIND.0000005136.61217.93.

Castella, J.C., Trung, T.N. and Boissau, S., 2005. Participatory Simulation of Land-Use Changes in the Northern Mountains of Vietnam: the Combined Use of an Agent-Based Model, a Role-Playing Game, and a Geographic Information System. *Ecology and Society* [Online], 10(1), p.art27. Available from: https://doi.org/10.5751/ES-01328-100127.

Chan, R., 2018. Salesforce is hiring its first Chief Ethical and Humane Use officer to make sure its artificial intelligence isn't used for evil [Online]. Available from: https://nordic.businessinsider.com/salesforce-hires-paula-goldman-as-chief-ethical-and-humane-use-officer-2018-12.

Chapman, D., 1987. Planning for conjunctive goals. 32, pp.333–378.

Charsky, D., 2010. From Edutainment to Serious Games: A Change in the Use of Game Characteristics. *Games and Culture* [Online], 5(2), pp.177–198. Available from: https://doi.org/10.1177/1555412009354727.

Cheong, Y.G., Khaled, R., Grappiolo, C., Campos, J., Martinho, C., Ingram, G., Paiva, A. and Yannakakis, G., 2011. A Computational Approach Towards Conflict Resolution for Serious Games. Available from: https://doi.org/10.1145/2159365.2159368.

Choo, J. and Liu, S., 2018. Visual Analytics for Explainable Deep Learning. *IEEE Computer Graphics and Applications* [Online], 38(4), pp.84–92. 1804.02527, Available from: https://doi.org/10.1109/MCG.2018.042731661.

Clutton-Brock, T., 2002. Breeding Together: Kin Selection and Mutualism in Cooperative Vertebrates. *Science* [Online], 296(5565), pp.69–72. Available from: https://doi.org/10.1126/science.296.5565.69.

Coca-Vila, I., 2018. Self-driving cars in dilemmatic situations: An approach based on the theory of justification in criminal law. *Criminal Law and Philosophy*, 12(1), pp.59–82.

Coeckelbergh, M., 2009. Virtual moral agency, virtual moral responsibility: on the moral significance of the appearance, perception, and performance of artificial agents. *AI & Society*, 24(2), pp.181–189.

Coeckelbergh, M., 2010. Moral appearances: Emotions, robots, and human morality. *Ethics and Information Technology* [Online], 12(3), pp.235–241. Available from: https://doi.org/10.1007/s10676-010-9221-y.

Collins, A. and Gentner, D., 1987. How people construct mental models. *Cultural models in language and thought.* New York, NY, US: Cambridge University Press, pp.243–265.

Collins, E.C., Prescott, T.J. and Mitchinson, B., 2015. Saying it with light: A pilot study of affective communication using the MIRO robot [Online]. *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)*. Springer, Cham, vol. 9222, pp.243–255. Available from: https://doi.org/10.1007/978-3-319-22979-9_25.

Commission, E., 2010. Directive 2010/63/EU of the European Parliament and of the Council of 22 September 2010 on the protection of animals used for scientific purposes Text with EEA relevance.

Cooke, N.J., Kiekel, P.A., Salas, E., Stout, R., Bowers, C. and Cannon-Bowers, J., 2003. Measuring team knowledge: A window to the cognitive underpinnings of team performance. *Group Dynamics: Theory, Research, and Practice*, 7(3), pp.179–199. Available from: https://doi.org/10.1037/1089-2699.7.3.179.

Cooper, R., Shallice, T. and Farringdon, J., 1995. Symbolic and continuous processes in the automatic selection of actions. In: J. Hallam, ed. *Hybrid problems, hybrid solutions, frontiers in artificial intelligence and applications.* Amsterdam: IOS Press, pp.27–37.

Craik, K.J.W., 1943. *The nature of explanation* [Online]. Cambridge University Press. Available from: https://books.google.co.uk/books?id=wT04AAAAIAAJ{&}redir{_}esc=y{&}hl=en.

Crawford, C., 2003. *Chris Crawford on Game Design* [Online]. New Riders. 9780201398298, Available from: https://doi.org/10.1093/carcin/bgs054.

Dawkins, R., 1996. *The Blind Wathcmaker. Why the Evidence of Evolution Reveals a Universe without Design*. New York: Norton.

De Angeli, D., 2018. *The Gameful Museum: Authenticity and Entertainment in the Digital Age*. Ph.D. thesis.

Dempsey, J.V., Haynes, L.L., Lucassen, B.A. and Casey, M.S., 2002. Forty Simple Computer Games and What They Could Mean to Educators. *Simulation & Gaming* [Online], 33(2), pp.157–168. Available from: https://doi.org/10.1177/1046878102332003.

Dennett, D.C., 1996. *Kinds of minds: Towards an understanding of consciousness*. Weidenfeld and Nicolson.

Dennett, D.C., 2001. Are we explaining consciousness yet? *Cognition*, 79, pp.221–237.

Deterding, S., Dixon, D., Khaled, R. and Nacke, L., 2011. From game design elements to gamefulness [Online]. *Proceedings of the 15th international academic mindtrek conference on envisioning future media environments - mindtrek '11*. New York, New York, USA: ACM Press, p.9. Available from: https://doi.org/10.1145/2181037.2181040.

Di Ferdinando, A., Schembri, M., Ponticorvo, M. and Miglino, O., 2015. Agent Based Modelling to Build Serious Games: The Learn to Lead Game [Online]. *Ferrández vicente j., álvarez-sánchez j., de la paz lópez f., toledo-moreo f., adeli h. (eds) bioinspired computation in artificial systems. iwinac 2015. lecture notes in computer science*. pp.349–358. Available from: https://doi.org/10.1007/978-3-319-18833-1_37.

Dickey, M.D., 2005. Engaging by design: How engagement strategies in popular computer and video games can inform instructional design. *Educational Technology Research and Development* [Online], 53(2), pp.67–83. Available from: https://doi.org/10.1007/BF02504866.

Dignum, V., 2017. Responsible autonomy. *Ijcai international joint conference on artificial intelligence*. pp.4698–4704. 1706.02513, Available from: https://doi.org/10.24963/ijcai.2017/655.

Doll, N., Vetter, P. and Tauber, A., 2015. Wen soll das autonome Auto lieber überfahren. *Die Welt* [Online], pp.1–4. Available from: http://www.welt.de/wirtschaft/article146407129/Wen-soll-das-autonome-Auto-lieber-ueberfahren.html.

Doshi-Velez, F. and Kim, B., 2017. Towards A Rigorous Science of Interpretable Machine Learning [Online]. [Online]. 1702.08608, Available from: http://arxiv.org/abs/1702.08608.

Dzindolet, M.T., Peterson, S.A., Pomranky, R.A., Pierce, L.G. and Beck, H.P., 2003. The role of trust in automation reliance. *International Journal of Human Computer Studies*, 58(6), pp.697–718. Available from: https://doi.org/10.1016/S1071-5819(03)00038-7.

Elias, P., 1955. Predictive coding. *IEEE Transactions on Information Theory* [Online], 1(1), pp.16–24. Available from: https://doi.org/10.1109/TIT.1955.1055126.

Elish, M.C., 2016. Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction (We Robot 2016) [Online]. *Ssrn*. Available from: https://doi.org/10.2139/ssrn.2757236.

Ellemers, N., Pagliaro, S. and Barreto, M., 2013. Morality and behavioural regulation in groups: A social identity approach. *European Review of Social Psychology* [Online], 24(1), pp.160–193. Available from: https://doi.org/10.1080/10463283.2013.841490.

Faulhaber, A.K., Dittmer, A., Blind, F., Wächter, M.A., Timm, S., Sütfeld, L.R., Stephan, A., Pipa, G. and König, P., 2018. Human Decisions in Moral Dilemmas are Largely Described by Utilitarianism: Virtual Car Driving Study Provides Guidelines for Autonomous Driving Vehicles. *Science and Engineering Ethics* [Online]. Available from: https://doi.org/10.1007/s11948-018-0020-x.

Fehr, E. and Fischbacher, U., 2003. The nature of human altruism. *Nature*, 425, pp.785–791.

Ferguson, C.J., 2015. Do Angry Birds Make for Angry Children? A Meta-Analysis of Video Game Influences on Children's and Adolescents' Aggression, Mental Health, Prosocial Behavior, and Academic Performance. *Perspectives on Psychological Science*, 10(5), pp.646–666. Available from: https://doi.org/10.1177/1745691615592234.

Ferrara, E., 2017. Disinformation and social bot operations in the run up to the 2017 French presidential election. *First Monday* [Online], 22(8). 1707.00086, Available from: https://doi.org/10.2139/ssrn.2995809.

Fischer, K., 2011. How People Talk with Robots: Designing Dialogue to Reduce User Uncertainty. *AI Magazine*, 32(4), pp.31–38. Available from: https://doi.org/10.1609/aimag.v32i4.2377.

Fisher, M., Dennis, L. and Webster, M., 2013. Verifying autonomous systems. *Communications of the ACM* [Online], 56(9), p.84. Available from: https://doi.org/10.1145/2500468.2494558.

Fleetwood, J., 2017. Public health, ethics, and autonomous vehicles. *American journal of public health*, 107(4), pp.532–537.

Fong, T., Nourbakhsh, I., & Dautenhahn, K., 2003. A Survey of Socially Interactive Robots : Concepts , Design , and Applications Terrence Fong , Illah Nourbakhsh , and Kerstin Dautenhahn. *Robotics and autonomous systems*, 42(3-4), pp.143–166. Available from: https://doi.org/10.1016/S0921-8890(02)00372-X.

Franco, N.H. and Olsson, I.A.S., 2015. 12. Killing animals as a necessary evil? The case of animal research [Online]. *The end of animal life: a start for ethical debate*. pp.12–187. Available from: https://doi.org/doi:10.3920/978-90-8686-808-7_12.

Freund, J., Brandmaier, A.M., Lewejohann, L., Kirste, I., Kritzler, M., Krüger, A., Sachser, N., Lindenberger, U. and Kempermann, G., 2013. Emergence of individuality in genetically identical mice. *Science* [Online], 340(6133), pp.756–759. arXiv:1011.1669v3, Available from: https://doi.org/10.1126/science.1235294.

Gallagher, E. and Bryson, J., 2017. Agent-Based Modelling.

Gallistel, C.R., Brown, A.L., Carey, S., Gelman, R. and Keil, F.C., 1991. Lessons From Animal Learning for the Study of Cognitive Development. In: S. Carey and R. Gelman, eds. *The epigenesis of mind*. Hillsdale, NJ: Lawrence Erlbaum, pp.3–36.

Gat, E., 1992. Integrating planning and reaction in a heterogeneous asynchronous architecture for controlling mobile robots. *Proceedings of the tenth national conference on artificial intelligence (aaai92)*.

Gaudl, S. and Bryson, J.J., 2014. The Extended Ramp Goal Module: Low-Cost Behaviour Arbitration for Real-Time Controllers based on Biological Models of Dopamine Cells. *Computational Intelligence in Games 2014* [Online]. Available from: http://opus.bath.ac.uk/40056/.

Gaudl, S., Davies, S. and Bryson, J.J., 2013a. Behaviour oriented design for real-time-strategy games: An approach on iterative development for STARCRAFT AI. *Foundations of Digital Games Conference*, pp.198–205.

Gaudl, S., Davies, S. and Bryson, J.J., 2013b. Behaviour oriented design for real-time-strategy games: An approach on iterative development for STARCRAFT AI. In: G.N. Yannakakis and E. Aarseth, eds. *Foundations of digital games (fdg)*. Chania, Crete, pp.198–205.

Gavrilets, S. and Fortunato, L., 2014. A solution to the collective action problem in between-group conflict with within-group inequality. *Nature Communications* [Online], 5(1), p.3526. Available from: `https://doi.org/10.1038/ncomms4526`.

Geffner, H., 1992. *Default reasoning: causal and conditional theories*, vol. 4. MIT Press Cambridge, MA.

Gemrot, J., Kadlec, R., Bída, M., Burkert, O., Píbil, R., Havlíček, J., Zemčák, L., Šimlovič, J., Vansa, R., Štolba, M., Plch, T. and Brom, C., 2009. Pogamut 3 can assist developers in building AI (not only) for their videogame agents. *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)*. vol. 5920 LNAI, pp.1–15. Available from: `https://doi.org/10.1007/978-3-642-11198-3_1`.

Gentile, M., La Guardia, D., Dal Grande, V., Ottaviano, S. and Allegra, M., 2014. An Agent Based approach to design Serious Game. *International Journal of Serious Games* [Online], 1(2). Available from: `https://doi.org/10.17083/ijsg.v1i2.17`.

Gentner, D. and Gentner, D.R., 1983. Flowing Waters or Teeming Crowds: Mental Models of Electricity [Online]. In: G. Dedre and A.L. Stevens, eds. *Mental models*. Erlbaum, pp.99–129. Available from: `https://doi.org/10.2307/973677`.

Gerdes, J.C. and Thornton, S.M., 2015. Implementable ethics for autonomous vehicles. *Autonomes fahren*. Springer, pp.87–102.

Giesler, B., Salb, T., Steinhaus, P. and Dillmann, R., 2004. Using augmented reality to interact with an autonomous mobile platform [Online]. *Icra '04: Proceedings of the 2004 ieee international conference on robotics and automation*. IEEE, pp.1009–1014. Available from: `https://doi.org/10.1109/ROBOT.2004.1307282`.

Glimcher, P.W., 2011. Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis. *Proceedings of the National Academy of Sciences*, 108(Supplement 3), pp.15647–15654.

Goetz, J., Kiesler, S. and Powers, A., 2003. Matching robot appearance and behavior to tasks to improve human-robot cooperation. *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*, pp.55–60. Available from: https://doi.org/10.1109/ROMAN.2003.1251796.

Gogoll, J. and Müller, J.F., 2017. Autonomous cars: in favor of a mandatory ethics setting. *Science and Engineering Ethics*, 23(3), pp.681–700.

Goodall, N.J., 2016. Away from trolley problems and toward risk management. *Applied Artificial Intelligence*, 30(8), pp.810–821.

Gorini, A., Griez, E., Petrova, A. and Riva, G., 2010. Assessment of the emotional responses produced by exposure to real food, virtual food and photographs of food in patients affected by eating disorders. *Annals of General Psychiatry*, 9(1), p.30.

Green, J., 2018. Tesla: Autopilot was on during deadly Mountain View crash [Online]. Available from: https://www.mercurynews.com/2018/03/30/tesla-autopilot-was-on-during-deadly-mountain-view-crash/.

Green, S.A., Billinghurst, M., Chen, X. and Chase, J.G., 2007. Human Robot Collaboration: An Augmented Reality Approach—A Literature Review and Analysis [Online]. *Volume 4: Asme/ieee international conference on mechatronic and embedded systems and applications and the 19th reliability, stress analysis, and failure prevention conference*. ASME, vol. 5, pp.117–126. Available from: https://doi.org/10.1115/DETC2007-34227.

Greenblatt, N.A., 2016. Self-driving cars and the law. *IEEE Spectrum* [Online], 53(2), pp.46–51. Available from: https://doi.org/10.1109/MSPEC.2016.7419800.

Greene, D., Hoffmann, A. and Stark, L., 2019. Better, nicer, clearer, fairer: A critical assessment of the movement for ethical artificial intelligence and machine learning. *Proceedings of the 52nd hawaii international conference on system sciences*.

Griffin, A.S. and West, S.A., 2003. Kin Discrimination and the Benefit Breeding Vertebrates. *Science* [Online], 302(October), pp.634–636. Available from: https://doi.org/10.1126/science.1089402.

Groom, V. and Nass, C., 2007. Can robots be teammates? *Interaction Studies*, 8(3), pp.483–500. Available from: https://doi.org/10.1075/gest.8.3.02str.

Gruen, L., 2017. The Moral Status of Animals. In: E.N. Zalta, ed. *The stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University. Fall 2017 ed.

Guala, F., 2012. Reciprocity: Weak or strong? What punishment experiments do (and do not) demonstrate. *Behavioral and Brain Sciences* [Online], 35(01), pp.1–15. Available from: https://doi.org/10.1017/S0140525X11000069.

Gunkel, D.J., 2012. *Critical Perspectives on AI , Robots , and Ethics*. MIT Press Cambridge, Mass.

Gurung, T.R., Bousquet, F. and Trébuil, G., 2006. Companion Modeling, Conflict Resolution, and Institution Building: Sharing Irrigation Water in the Lingmuteychu Watershed, Bhutan. *Ecology and Society* [Online], 11(2), p.art36. Available from: https://doi.org/10.5751/ES-01929-110236.

Güth, W., Schmittberger, R. and Schwarze, B., 1982. An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior and Organization* [Online], 3(4), pp.367–388. Available from: https://doi.org/10.1016/0167-2681(82)90011-7.

Hamari, J., Shernoff, D.J., Rowe, E., Coller, B., Asbell-Clarke, J. and Edwards, T., 2016. Challenging games help students learn: An empirical study on engagement, flow and immersion in game-based learning. *Computers in Human Behavior* [Online], 54, pp.170–179. Available from: https://doi.org/10.1016/j.chb.2015.07.045.

Hamilton, W.D., 1971. Geometry for the selfish herd. *Journal of Theoretical Biology*, 31, pp.295–311.

Haslam, N., Kashima, Y., Loughnan, S., Shi, J. and Suitner, C., 2008. Subhuman, inhuman, and superhuman: Contrasting humans with nonhumans in three cultures. *Social Cognition*, 26(2), pp.248–258.

Hauert, S., 2015. Robotics: Ethics of artificial intelligence: Shape the debate, don't shy from it. *Nature*, 521, p.415–418. Available from: https://doi.org/10.1038/521415a.

He, K., Zhang, X., Ren, S. and Sun, J., 2015. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification [Online]. [Online]. 1502.01852, Available from: http://arxiv.org/abs/1502.01852.

Hellström, T. and Bensch, S., 2018. Understandable robots. *Paladyn* [Online], 9(1), pp.110–123. Available from: https://doi.org/10.1515/pjbr-2018-0009.

Hemelrijk, C., Wantia, J. and Gygax, L., 2005. The construction of dominance order: Comparing performance of five methods using an individual-based model. *Behaviour*, 142(8), pp.1043–1064.

Hemelrijk, C.K., 1999. An individual-oriented model on the emergence of despotic and egalitarian societies. *Proceedings of the Royal Society of London, B: Biological Science*, 266, pp.361–369.

Hemelrijk, C.K., Wantia, J. and Dätwyler, M., 2003. Female co-dominance in a virtual world: Ecological, cognitive, social and sexual causes. *Behaviour*, 140(10), pp.1247–1273.

Henriques, J.F., Caseiro, R., Martins, P. and Batista, J., 2012. Exploiting the circulant structure of tracking-by-detection with kernels [Online]. *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)*. vol. 7575 LNCS, pp.702–715. `arXiv:1011.1669v3`, Available from: `https://doi.org/10.1007/978-3-642-33765-9_50`.

Herrmann, B., Thöni, C. and Gächter, S., 2008. Antisocial punishment across societies. *Science* [Online], 319(5868), pp.1362–1367. `1101.2204`, Available from: `https://doi.org/10.1126/science.1153808`.

Hill, R., 2018. SAP claims to be first Euro biz to get seriously ethical about AI code [Online]. Available from: `https://www.theregister.co.uk/2018/09/19/sap_ethics_ai/`.

Himma, K.E., 2009. Artificial agency, consciousness, and the criteria for moral agency: what properties must an artificial agent have to be a moral agent? *Ethics and Information Technology* [Online], 11(1), pp.19–29. Available from: `https://doi.org/10.1007/s10676-008-9167-5`.

Hogeweg, P. and Hesper, B., 1979. Heterarchical selfstructuring simulation systems: Concepts and applications in biology. In: B.P. Zeigler, M.S. Ezas, G.J. Klir and T.I. Ören, eds. *Methodologies in systems modelling and simulation.* Amsterdam: North-Holland, pp.221–231.

Hogeweg, P. and Hesper, B., 1983. The ontogeny of the interaction structure in bumble bee colonies: A MIRROR model. *Behavioral Ecology and Sociobiology*, 12(271–283).

Höijer, B., 2011. Social representations theory. *Nordicom review*, 32(2), pp.3–16.

Hoque, N., Bhattacharyya, D.K. and Kalita, J.K., 2015. Botnet in DDoS Attacks: Trends and Challenges. *IEEE Communications Surveys and Tutorials* [Online], 17(4), pp.2242–2270. Available from: `https://doi.org/10.1109/COMST.2015.2457491`.

House of Lords, 2018. *AI in the UK: ready, willing and able?* United Kingdom, (HL 2017–2019 (100)).

Howard, P.N. and Kollanyi, B., 2016. Bots, #StrongerIn, and #Brexit: Computational Propaganda during the UK-EU Referendum. *SSRN Electronic Journal* [Online]. 1606.06356, Available from: https://doi.org/10.2139/ssrn.2798311.

Howard, P.N., Woolley, S. and Calo, R., 2018. Algorithms, bots, and political communication in the US 2016 election: The challenge of automated political communication for election law and administration. *Journal of Information Technology and Politics* [Online], 15(2), pp.81–93. Available from: https://doi.org/10.1080/19331681.2018.1448735.

Humphrys, M., 1996. Action selection methods using reinforcement learning. In: P. Maes, M.J. Matarić, J.A. Meyer, J. Pollack and S.W. Wilson, eds. *From animals to animats 4 (SAB '96)*. Cambridge, MA: MIT Press.

IEEE, 2016. *Ethically Aligned Design* [Online]. Available from: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp={&}arnumber=7924235.

ISO 13482:2014, 2014. *Robots and robotic devices – safety requirements for personal care robots*. Geneva, CH: International Organization for Standardization, (Standard).

ISO 9001:2015, 2015. *Quality management systems – requirements*. Geneva, CH: International Organization for Standardization, (Standard).

Isoda, M. and Hikosaka, O., 2007. Switching from automatic to controlled action by monkey medial frontal cortex. *Nature Neuroscience* [Online], 10(2), pp.240–248. Available from: https://doi.org/10.1038/nn1830.

Jackson, L.M. and Esses, V.M., 2000. Effects of Perceived Economic Competition on People's Willingness to Help Empower Immigrants. *Group Processes & Intergroup Relations* [Online], 3(4), pp.419–435. Available from: https://doi.org/10.1177/1368430200003004006.

Johnson, L.E., 1983. Can animals be moral agents? *Ethics & Animals* [Online], 4(2), pp.50–61. Available from: https://digitalcommons.calpoly.edu/cgi/viewcontent.cgi?referer=https://www.google.co.uk/{&}httpsredir=1{&}article=1170{&}context=ethicsandanimals.

Johnson, L.P., 2012. Law and Legal Theory in the History of Corporate Responsibility: Corporate Personhood [Online]. Available from: https://scholarlycommons.law.wlu.edu/wlufachttp://scholarlycommons.law.wlu.edu/cgi/viewcontent.cgi?article=1124{&}context=wlufac [Accessed 2018-11-17].

Johnson-Laird, P., 1983. *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness* [Online]. Harvard University Press. Available from: https://books.google.co.uk/books/about/Mental{_}Models.html?id=FS3zSKAfLGMC{&}redir{_}esc=y.

Johnson-Laird, P.N., 2010. Mental models and human reasoning. *Proceedings of the National Academy of Sciences* [Online], 107(43), pp.18243–18250. arXiv:1604.05974v2, Available from: https://doi.org/10.1073/pnas.1012933107.

Jolly, J., 2018. Amazon robot sets off bear repellant, putting 24 workers in hospital [Online]. Available from: https://www.theguardian.com/technology/2018/dec/06/24-us-amazon-workers-hospitalised-after-robot-sets-off-bear-repellent.

Jones, N.A., Ross, H., Lynam, T., Perez, P. and Leitch, A., 2011. Mental models: an interdisciplinary synthesis of theory and methods. *Ecology and Society*, 16(1).

Josh, H. and Timmons, H., 2016. There are some scary similarities between Tesla's deadly crashes linked to Autopilot [Online]. Available from: https://qz.com/783009/the-scary-similarities-between-teslas-tsla-deadly-autopilot-crashes/.

Joy, T., 2017. Gamification in Elections (from Howard Dean to Hillary Clinton) - CallHub [Online]. Available from: https://callhub.io/gamification-in-elections/ [Accessed 2018-12-04].

Kahn, P.H., Severson, R.L., Kanda, T., Ishiguro, H., Gill, B.T., Ruckert, J.H., Shen, S., Gary, H.E., Reichert, A.L. and Freier, N.G., 2012. Do people hold a humanoid robot morally accountable for the harm it causes? *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction - HRI '12*, (February 2016), p.33. Available from: https://doi.org/10.1145/2157689.2157696.

Kalal, Z., Mikolajczyk, K. and Matas, J., 2011. Tracking-Learning-Detection. *IEEE transactions on pattern analysis and machine intelligence* [Online], 34(1), pp.1409–1422. arXiv:1412.7522v1, Available from: https://doi.org/10.1109/TPAMI.2011.239.

Keltner, D., 2009. *Born to be good : the science of a meaningful life*. New York: W. W. Norton & Company. Available from: https://doi.org/10.1016/j.evolhumbehav.2009.09.005.

Kiesler, S. and Goetz, J., 2002. Mental models of robotic assistants [Online]. *Chi '02 extended abstracts on human factors in computing systems - chi '02*. p.576. Available from: https://doi.org/10.1145/506443.506491.

Kiesler, S., Powers, A., Fussell, S.R. and Torrey, C., 2008a. Anthropomorphic interactions with a robot and robot–like agent. *Social Cognition*, 26(2), pp.169–181.

Kiesler, S., Powers, A., Fussell, S.R. and Torrey, C., 2008b. Anthropomorphic Interactions with a Robot and Robot–like Agent. *Social Cognition* [Online], 26(2), pp.169–181. Available from: https://doi.org/10.1521/soco.2008.26.2.169.

Kim, T. and Hinds, P., 2006. Who should I blame? Effects of autonomy and transparency on attributions in human-robot interaction. *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*, pp.80–85. Available from: https://doi.org/10.1109/ROMAN.2006.314398.

Klein, D., Marx, J. and Fischbach, K., 2018. Agent-based modeling in social science, history, and philosophy: An introduction. *Historical Social Research*, 43(1), pp.7–27.

Koda, T. and Maes, P., 1996. Agents with Faces: The Effects of Personification of Agents [Online]. *In robot and human communication, 5th ieee international workshop*. pp.189–194. Available from: https://pdfs.semanticscholar.org/bf16/3cabead90c3f9f0e901f1790765e1c2fc912.pdf.

Krátký, J., McGraw, J.J., Xygalatas, D., Mitkidis, P. and Reddish, P., 2016. It Depends Who Is Watching You: 3-D Agent Cues Increase Fairness. *PLOS ONE* [Online], 11(2), p.e0148845. Available from: https://doi.org/10.1371/journal.pone.0148845.

Krueger, F., 2016. *Neural signatures of trust during human-automation interactions*. George Mason University Fairfax United States.

Kuehn, J. and Haddadin, S., 2017. An Artificial Robot Nervous System To Teach Robots How To Feel Pain And Reflexively React To Potentially Damaging Contacts. *IEEE Robotics and Automation Letters* [Online], 2(1), pp.72–79. Available from: https://doi.org/10.1109/LRA.2016.2536360.

Law of the cyprus scientific technical chamber, 2012.

Lee, J.D. and Moray, N., 1994. Trust, self-Confidence, and operators' adaptation to automation. *International Journal of Human - Computer Studies* [Online], 40(1), pp.153–184. Available from: https://doi.org/10.1006/ijhc.1994.1007.

Lee, J.D. and See, K.A., 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society* [Online], 46(1), pp.50–80. Available from: https://doi.org/10.1518/hfes.46.1.50_30392.

Lee, J.J. and Gino, F., 2015. Poker-faced morality: Concealing emotions leads to utilitarian decision making. *Organizational Behavior and Human Decision Processes*, 126, pp.49–64.

Lee, S.l., Lau, I.Y.m., Kiesler, S. and Chiu, C.Y., 2005. Human mental models of humanoid robots. *Robotics and automation, 2005. icra 2005. proceedings of the 2005 ieee international conference on*. IEEE, pp.2767–2772.

Li, J., Zhao, X., Cho, M.J., Ju, W. and Malle, B.F., 2016. *From trolley to autonomous vehicle: Perceptions of responsibility and moral norms in traffic accidents with self-driving cars*. SAE Technical Paper.

Liang, Y. and Lee, S.A., 2017. Fear of autonomous robots and artificial intelligence: Evidence from national representative data with probability sampling. *International Journal of Social Robotics*, 9(3), pp.379–384.

Libet, B., 1985. Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and Brain Sciences* [Online], 8(4), pp.529–539. Available from: https://doi.org/10.1017/S0140525X00044903.

Litman, T., 2017. *Autonomous vehicle implementation predictions*. Victoria Transport Policy Institute.

Lubben, A., 2018. Self-driving Uber killed a pedestrian as human safety driver watched [Online]. Available from: https://news.vice.com/en{_}us/article/kzxq3y/self-driving-uber-killed-a-pedestrian-as-human-safety-driver-watched.

Lyons, J.B., 2013. Being Transparent about Transparency : A Model for Human-Robot Interaction. *Trust and Autonomous Systems: Papers from the 2013 AAAI Spring Symposium*, pp.48–53.

Maclean, R.C., Fuentes-Hernandez, A., Greig, D., Hurst, L.D. and Gudelj, I., 2010. A mixture of "cheats" and "co-operators" can enable maximal group benefit. *PLoS Biology* [Online], 8(9), p.e1000486. Available from: https://doi.org/10.1371/journal.pbio.1000486.

MacLean, R.C., Fuentes-Hernandez, A., Greig, D., Hurst, L.D. and Gudelj, I., 2010. A mixture of "cheats" and "co-operators" can enable maximal group benefit. *PLoS Biol*, 8(9), p.e1000486. Available from: https://doi.org/10.1371/journal.pbio.1000486.

Madhavan, P. and Wiegmann, D.A., 2007. Similarities and differences between human–human and human–automation trust: an integrative review. *Theoretical Issues in Ergonomics Science*, 8(4), pp.277–301.

Maida, J.C., Bowen, C.K. and Pace, J., 2007. Improving Robotic Operator Performance Using Augmented Reality. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* [Online], 51(27), pp.1635–1639. Available from: https://doi.org/10.1177/154193120705102703.

Makris, S., Karagiannis, P., Koukas, S. and Matthaiakis, A.S., 2016. Augmented reality system for operator support in human–robot collaborative assembly. *CIRP Annals - Manufacturing Technology* [Online], 65(1), pp.61–64. Available from: https://doi.org/10.1016/j.cirp.2016.04.038.

Malle, B.F., Scheutz, M., Arnold, T., Voiklis, J. and Cusimano, C., 2015. Sacrifice one for the good of many?: People apply different moral norms to human and robot agents. *Proceedings of the tenth annual acm/ieee international conference on human-robot interaction*. ACM, pp.117–124.

Malle, B.F., Scheutz, M., Forlizzi, J. and Voiklis, J., 2016. Which robot am i thinking about?: The impact of action and appearance on people's evaluations of a moral robot. *The eleventh acm/ieee international conference on human robot interaction*. IEEE Press, pp.125–132.

Marchesi, S., Ghiglino, D., Ciardo, F., Baykara, E. and Wykowska, A., 2018. Do we adopt the Intentional Stance towards humanoid robots? *PsyArXiv*. Available from: https://doi.org/10.31234/osf.io/6smkq.

Mathew, S. and Boyd, R., 2011. Punishment sustains large-scale cooperation in prestate warfare. *Proceedings of the National Academy of Sciences* [Online], 108(28), pp.11375–11380. Available from: https://doi.org/10.1073/pnas.1105604108.

McAndrew, F.T. and Milenkovic, M.A., 2002. Of Tabloids and Family Secrets: The Evolutionary Psychology of Gossip. *Journal of Applied Social Psychology* [Online], 32(5), pp.1064–1082. Available from: https://doi.org/10.1111/j.1559-1816.2002.tb00256.x.

McCartney, H., 2011. Hero, Victimor Villain? The Public Image of the British Soldier and its Implications for Defense Policy. *Defense & Security Analysis* [Online], 27(1), pp.43–54. Available from: https://doi.org/10.1080/14751798.2011.557213.

McReynolds, E., Hubbard, S., Lau, T., Saraf, A., Cakmak, M. and Roesner, F., 2017. Toys that Listen [Online]. *Proceedings of the 2017 chi conference on human factors in computing systems - chi '17*. New York, New York, USA: ACM Press, pp.5197–5207. Available from: https://doi.org/10.1145/3025453.3025735.

Mercado, J.E., Rupp, M.A., Chen, J.Y., Barnes, M.J., Barber, D. and Procci, K., 2016. Intelligent Agent Transparency in Human-Agent Teaming for Multi-UxV Management. *Human Factors* [Online], 58(3), pp.401–415. Available from: https://doi.org/10.1177/0018720815621206.

Michalos, G., Karagiannis, P., Makris, S., Tokçalar, Ö. and Chryssolouris, G., 2016. Augmented Reality (AR) Applications for Supporting Human-robot Interactive Cooperation [Online]. *Procedia cirp*. Elsevier, vol. 41, pp.370–375. Available from: https://doi.org/10.1016/j.procir.2015.12.005.

Miller, C.A., 2014. Delegation and transparency: Coordinating interactions so information exchange is no surprise [Online]. *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)*. Springer, Cham, vol. 8525 LNCS, pp.191–202. Available from: https://doi.org/10.1007/978-3-319-07458-0_19.

Miller, L.F., 2015. Granting Automata Human Rights: Challenge to a Basis of Full-Rights Privilege. *Human Rights Review* [Online], 16(4), pp.369–391. Available from: https://doi.org/10.1007/s12142-015-0387-x.

Millington, I. and Funge, J.D., 2009. *Artificial Intelligence for Games*. 2nd ed. CRC Press. Available from: https://doi.org/10.1016/B978-0-12-374731-0.00001-3.

Minsky, M.L., 2006. *The Emotion Machine: Commensense Thinking, Artificial Intelligence, and the Future of the Human Mind* [Online]. Simon & Schuster. Available from: https://dl.acm.org/citation.cfm?id=1204056http://books.google.co.kr/books?id=agUgKCrLIMQC.

Mitchell, D., Bryson, J.J., Rauwolf, P. and Ingram, G.P., 2016. On the reliability of unreliable information. *Interaction Studies* [Online], 17(1), pp.1–25. Available from: https://doi.org/10.1075/is.17.1.01mit.

Moll, J. and Oliveira-Souza, R. de, 2007. Moral judgments, emotions and the utilitarian brain. *Trends in Cognitive Sciences*, 11(8), pp.319–321.

Moravec, H., 1988. *Mind children: the future of robot and human intelligence* [Online], vol. February 1. Harvard University Press. Available from: `https://doi.org/cblibrary-fut`.

Moscovici, S., 1981. On social representations. *Social cognition: Perspectives on everyday understanding*, 8(12), pp.181–209.

Nersessian, N.J., 2002. The cognitive basis of model-based reasoning in science. *The Cognitive Basis of Science* [Online], pp.133–153. Available from: `https://doi.org/10.1017/cbo9780511613517.008`.

Nettle, D., 2010. Dying young and living fast: Variation in life history across English neighborhoods. *Behavioral Ecology* [Online], 21(2), pp.387–395. Available from: `https://doi.org/10.1093/beheco/arp202`.

Nikiforakis, N., 2010. For the Student: Experimental Economics. *Australian Economic Review* [Online], 43(3), pp.337–345. Available from: `https://doi.org/10.1111/j.1467-8462.2010.00607.x`.

Norman, D.A. and Shallice, T., 1986. Attention to action: Willed and automatic control of behavior. In: R. Davidson, G. Schwartz and D. Shapiro, eds. *Consciousness and self regulation: Advances in research and theory.* New York: Plenum, vol. 4, pp.1–18.

Nowak, M.A. and Sigmund, K., 1998. The Dynamics of Indirect Reciprocity. *Journal of Theoretical Biology* [Online], 194(4), pp.561–574. Available from: `https://doi.org/10.1006/jtbi.1998.0775`.

Ong, S.K., Yuan, M.L. and Nee, A.Y., 2008. Augmented reality applications in manufacturing: A survey. *International Journal of Production Research* [Online], 46(10), pp.2707–2742. Available from: `https://doi.org/10.1080/00207540601064773`.

OpenAI, 2018. OpenAI Five [Online]. Available from: `https://blog.openai.com/openai-five/`.

Orkin, J., 2006. Three states and a plan: the AI of FEAR. *Game Developers Conference*, 2006(1), pp.1–18.

Orkin, J., 2015. Combat Dialogue in FEAR The Illusion of Communication. In: S. Rabin, ed. *Game ai pro 2: Collected wisdon of game ai professionals.* chap. 2, pp.19–21.

Otterbacher, J. and Talias, M., 2017. S/He's Too Warm/Agentic!: The Influence of Gender on Uncanny Reactions to Robots [Online]. *Proceedings of the 2017 acm/ieee international conference on human-robot interaction*. New York, NY, USA: ACM, HRI '17, pp.214–223. Available from: https://doi.org/10.1145/2909824.3020220.

P7001, I.S., n.d. *Transparency of autonomous systems*. (Working draft proposed standard).

Packer, C. and Pusey, A.E., 1982. Cooperation and competition within coalitions of male lions: kin selection or game theory? *Nature* [Online], 296(5859), pp.740–742. Available from: https://doi.org/10.1038/296740a0.

Pan, X. and Slater, M., 2011. Confronting a moral dilemma in virtual reality: a pilot study. *Proceedings of the 25th bcs conference on human-computer interaction*. British Computer Society, pp.46–51.

Parasuraman, R. and Riley, V., 1997. Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2), pp.230–253. Available from: https://doi.org/10.1518/001872097778543886.

Pedersen, B.K.M.K., Andersen, K.E., Köslich, S., Weigelin, B.C. and Kuusinen, K., 2018. Simulations and self-driving cars: A study of trust and consequences. *Companion of the 2018 acm/ieee international conference on human-robot interaction*. ACM, pp.205–206.

Petrillo, F., Pimenta, M. and Trindade, F., 2008. Houston, we have a problem...: a survey of actual problems in computer games development. *Proceedings of the 2008* [Online], pp.707–711. Available from: https://doi.org/10.1145/1363686.1363854.

Pöppel, E., 1994. Temporal mechanisms in perception. *International Review of Neurobiology*, 37, pp.185–202.

Powell, A., Shennan, S. and Thomas, M.G., 2009. Late Pleistocene demography and the appearance of modern human behavior. *Science*, 324(5932), pp.1298–1301. Available from: https://doi.org/10.1126/science.1170165.

Raichle, M.E. and Gusnard, D.A., 2002. Appraising the brain's energy budget. *Proceedings of the National Academy of Sciences* [Online], 99(16), pp.10237–10239. Available from: https://doi.org/10.1073/pnas.172399499.

Rankin, D.J., Rocha, E.P.C. and Brown, S.P., 2011. What traits are carried on mobile genetic elements and why? *Heredity* [Online], 106(1), pp.1–10. Available from: https://doi.org/10.1038/hdy.2010.24.

Rapoport, A. and Chammah, A.M., 1965. Prisoner's Dilemma. *University of Michigan Press* [Online], p.270. Available from: http://www.press.umich.edu/titleDetailDesc.do?id=20269http://books.google.com/books?hl=en{&}lr={&}id=yPtNnKjXaj4C{&}pgis=1.

Raz, J., 2010. Responsibility and the Negligence Standard. *Oxford Journal of Legal Studies* [Online], 30(1), pp.1–18. http://oup.prod.sis.lan/ojls/article-pdf/30/1/1/4380272/gqq002.pdf, Available from: https://doi.org/10.1093/ojls/gqq002.

Reynolds, C.W., 1987. Flocks, herds, and schools: A distributed behavioral model. *Computer Graphics*, 21(4), pp.25–34.

Ritterfeld, U., Cody, M. and Vorderer, P., 2009. *Serious games: Mechanisms and effects* [Online]. Routledge. arXiv:1011.1669v3, Available from: https://doi.org/10.4324/9780203891650.

Roberts, C. and Torgerson, D.J., 1999. Understanding controlled trials: baseline imbalance in randomised controlled trials. *BMJ (Clinical research ed.)* [Online], 319(7203), p.185. Available from: http://www.ncbi.nlm.nih.gov/pubmed/10406763http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1116277.

Rockenbach, B., 2007. Gintis, H., Bowles, S., Boyd, R., and Fehr, E.: Moral Sentiments and Material Interests – The Foundations of Cooperation in Economic Life. *Journal of Economics* [Online], 90(2), pp.215–218. Available from: https://doi.org/10.1007/s00712-006-0236-0.

Rohlfshagen, P. and Bryson, J.J., 2010. Flexible Latching: A Biologically-Inspired Mechanism for Improving the Management of Homeostatic Goals. *Cognitive Computation*, 2(3), pp.230–241. Available from: https://doi.org/10.1007/s12559-010-9057-0.

Roundtree, K.A., Goodrich, M.A. and Adams, J.A., 2019. Transparency: Transitioning From Human–Machine Systems to Human-Swarm Systems. *Journal of Cognitive Engineering and Decision Making* [Online], p.155534341984277. Available from: https://doi.org/10.1177/1555343419842776.

Rouse, W.B. and Morris, N.M., 1986. On looking into the black box: Prospects and limits in the search for mental models. *Psychological bulletin*, 100(3), p.349.

Rovira, A., Swapp, D., Spanlang, B. and Slater, M., 2009. The use of virtual reality in the study of people's responses to violent incidents. *Frontiers in Behavioral Neuroscience*, 3, p.59.

Ruse, M. and Wilson, E.O., 1986. Moral Philosophy as Applied Science. *Philosophy*, 61(236), pp.173–192.

Rutte, C. and Taborsky, M., 2007. Generalized reciprocity in rats. *PLoS Biology* [Online], 5(7), pp.1421–1425. Available from: `https://doi.org/10.1371/journal.pbio.0050196`.

Safety code for elevators, n.d. Washington, D.C, USA: American Standards Association, (Standard).

Salem, M., Lakatos, G., Amirabdollahian, F. and Dautenhahn, K., 2015. Would you trust a (faulty) robot?: Effects of error, task type and personality on human-robot cooperation and trust. *Proceedings of the tenth annual acm/ieee international conference on human-robot interaction*. ACM, pp.141–148.

Salen, K. and Zimmerman, E., 2013. Rules of Play: Game design fundamentals. *International Immunology* [Online], 25(8), pp.NP–NP. `arXiv:1011.1669v3`, Available from: `https://doi.org/10.1093/intimm/dxs150`.

Sanders, T.L., Wixon, T., Schafer, K.E., Chen, J.Y.C. and Hancock, P.A., 2014. The influence of modality and transparency on trust in human-robot interaction [Online]. *2014 ieee international inter-disciplinary conference on cognitive methods in situation awareness and decision support (cogsima)*. IEEE, pp.156–159. Available from: `https://doi.org/10.1109/CogSIMA.2014.6816556`.

Scarlatos, L.L., Tomkiewicz, M. and Courtney, R., 2013. Using an agent-based modeling simulation and game to teach socio-scientific topics. *Interaction Design and Architecture(s) Journal*, 19, pp.77–90.

Schafer, B. and Edwards, L., 2017. "I spy, with my little sensor": fair data handling practices for robots between privacy, copyright and security. *Connection Science* [Online], 29(3), pp.200–209. Available from: `https://doi.org/10.1080/09540091.2017.1318356`.

Schneeberger, K., Dietz, M. and Taborsky, M., 2012. Reciprocal cooperation between unrelated rats depends on cost to donor and benefit to recipient. *BMC Evolutionary Biology* [Online], 12(1), p.41. Available from: `https://doi.org/10.1186/1471-2148-12-41`.

Schneider, W. and Chein, J.M., 2003. Controlled & automatic processing: Behavior, theory, and biological mechanisms [Online]. Available from: https://doi.org/10.1016/S0364-0213(03)00011-9.

Shariff, A., Bonnefon, J.F. and Rahwan, I., 2017. Psychological roadblocks to the adoption of self-driving vehicles. *Nature Human Behaviour* [Online], 1(10), pp.694–696. Available from: https://doi.org/10.1038/s41562-017-0202-6.

Silva, A.S. and Mace, R., 2014. Cooperation and conflict: Field experiments in Northern Ireland. *Proceedings of the Royal Society B: Biological Sciences* [Online], 281(1792), pp.20141435–20141435. 0307014, Available from: https://doi.org/10.1098/rspb.2014.1435.

Skinner, B.F., 1935. The generic nature of the concepts of stimulus and response. *The Journal of General Psychology*, 12(1), pp.40–65.

Skitka, L.J., Mosier, K.L. and Burdick, M., 1999. Does automation bias decision-making? *International Journal of Human-Computer Studies*, 51(5), pp.991–1006.

Smith, B., 1928. Legal Personality. *The Yale Law Journal*, 37(3), p.283. arXiv:1011.1669v3, Available from: https://doi.org/10.2307/789740.

Solaiman, S.M., 2017. Legal personality of robots, corporations, idols and chimpanzees: a quest for legitimacy. *Artificial Intelligence and Law*, 25(2), pp.155–179. Available from: https://doi.org/10.1007/s10506-016-9192-3.

Special Eurobarometer 427: Autonomous Systems, 2015.

Special Eurobarometer 460: Attitudes towards the impact of digitisation and automation on daily life, 2017.

Stewart, A.J., McCarty, N. and Bryson, J.J., 2018. Explaining Parochialism: A Causal Account for Political Polarization in Changing Economic Environments [Online]. [Online]. 1807.11477, Available from: http://arxiv.org/abs/1807.11477.

Stumpf, S., Wong, W.k., Burnett, M. and Kulesza, T., 2010. Making intelligent systems understandable and controllable by end users. pp.10–11.

Subin, E.K., Hameed, A. and Sudheer, A.P., 2017. Android based augmented reality as a social interface for low cost social robots [Online]. *Proceedings of the advances in robotics on - air '17*. New York, New York, USA: ACM Press, pp.1–4. Available from: https://doi.org/10.1145/3132446.3134907.

Sundar, S.S., Waddell, T.F. and Jung, E.H., 2016. The hollywood robot syndrome: media effects on older adults' attitudes toward robots and adoption intentions. *The eleventh acm/ieee international conference on human robot interaction*. IEEE Press, pp.343–350.

Sütfeld, L.R., Gast, R., König, P. and Pipa, G., 2017. Using Virtual Reality to Assess Ethical Decisions in Road Traffic Scenarios: Applicability of Value-of-Life-Based Models and Influences of Time Pressure. *Frontiers in Behavioral Neuroscience* [Online], 11. Available from: `https://doi.org/10.3389/fnbeh.2017.00122`.

Sylwester, K., Herrmann, B. and Bryson, J.J., 2013. *Homo homini lupus?* Explaining antisocial punishment. *Journal of Neuroscience, Psychology, and Economics*, 6(3), pp.167–188.

Tan, Y. and Zheng, Z.y., 2013. Research Advance in Swarm Robotics. *Defence Technology*, 9(1), pp.18–39. Available from: `https://doi.org/10.1016/j.dt.2013.03.001`.

Tanguy, E.A.R., Willis, P.J. and Bryson, J.J., 2003. A layered Dynamic Emotion Representation for the creation of complex facial animation. In: T. Rist, R. Aylett, D. Ballin and J. Rickel, eds. *Proceedings of the fourth international workshop on intelligent virtual agents (iva '03)*. Springer, pp.101–105.

Theodorou, A., Bandt-Law, B. and Bryson, J.J., 2019. The sustainability game: Ai technology as an intervention for public understanding of cooperative investment. *Proceedings of the 2019 conference on games (cog 2019)*.

Theodorou, A., Wortham, R.H. and Bryson, J.J., 2017. Designing and implementing transparency for real time inspection of autonomous robots. *Connection Science* [Online], 29(3), pp.230–241. Available from: `https://doi.org/10.1080/09540091.2017.1310182`.

Tullio, J., Dey, A.K., Chalecki, J. and Fogarty, J., 2009. How it works: a field study of Non-technical users interacting with an intelligent system. *SIGCHI conference on Human factors in computing systems (CHI'07)*, pp.31–40. Available from: `https://doi.org/10.1145/1240624.1240630`.

Turkle, S., 2017. *Alone together: Why we expect more from technology and less from each other*. Hachette UK.

Urquiza-Haas, E.G. and Kotrschal, K., 2015. The mind behind anthropomorphic thinking: Attribution of mental states to other species. *Animal Behaviour* [Online], 109, pp.167–176. Available from: `https://doi.org/10.1016/j.anbehav.2015.08.011`.

Vandercruysse, S., Vandewaetere, M. and Clarebout, G., 2011. Game-Based Learning [Online]. *Handbook of research on serious games as educational, business and research tools*. IGI Global, pp.628–647. Available from: https://doi.org/10.4018/978-1-4666-0149-9.ch032.

Čače, I. and Bryson, J.J., 2007. Agent based modelling of communication costs: Why information can be free. In: C. Lyon, C.L. Nehaniv and A. Cangelosi, eds. *Emergence and evolution of linguistic communication*. London: Springer, pp.305–322.

Vincent, J., 2016. Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day [Online]. Available from: https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist.

Vitale, J., Tonkin, M., Herse, S., Ojha, S., Clark, J., Williams, M.A., Wang, X. and Judge, W., 2018. Be More Transparent and Users Will Like You: A Robot Privacy and User Experience Design Experiment. *ACM/IEEE International Conference on Human-Robot Interaction*. Available from: https://doi.org/10.1145/3171221.3171269.

Walker, M., Hedayati, H., Lee, J. and Szafir, D., 2018. Communicating Robot Motion Intent with Augmented Reality [Online]. *Proceedings of the 2018 acm/ieee international conference on human-robot interaction - hri '18*. New York, New York, USA: ACM Press, pp.316–324. Available from: https://doi.org/10.1145/3171221.3171253.

Walsh, T., 2016. The Singularity May Never Be Near [Online]. [Online], 61. 1602.06462, Available from: http://arxiv.org/abs/1602.06462.

Wang, L., Jamieson, G.a. and Hollands, J.G., 2009. Trust and reliance on an automated combat identification system. *Human factors*, 51(3), pp.281–291. Available from: https://doi.org/10.1177/0018720809338842.

Weiss, A., Igelsböck, J., Wurhofer, D. and Tscheligi, M., 2011. Looking forward to a "robotic society"? *International Journal of Social Robotics*, 3(2), pp.111–123.

West, S.A., Griffin, A.S. and Gardner, A., 2007. Evolutionary Explanations for Cooperation. *Current Biology* [Online], 17(16), pp.R661–R672. Available from: https://doi.org/10.1016/j.cub.2007.06.004.

Wheeler, M., 2010. Minds, things, and materiality [Online]. In: L. Malafouris and C. Renfrew, eds. *The cognitive life of things recasting the boundaries of the mind*.

London: McDonald Institute for Archaeological Research, pp.29–37. Available from: https://doi.org/10.1057/9780230360792_7.

Wilkins, C.L., Wellman, J.D., Babbitt, L.G., Toosi, N.R. and Schad, K.D., 2015. You can win but I can't lose: Bias against high-status groups increases their zero-sum beliefs about discrimination. *Journal of Experimental Social Psychology* [Online], 57, pp.1–14. Available from: https://doi.org/10.1016/j.jesp.2014.10.008.

Wilson, J.R. and Rutherford, A., 1989. Mental models: Theory and application in human factors [Online]. Available from: https://doi.org/10.1177/001872088903100601.

Winfield, A.F. and Jirotka, M., 2017. The case for an ethical black box [Online]. *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics).* vol. 10454 LNAI, pp.262–273. Available from: https://doi.org/10.1007/978-3-319-64107-2_21.

Winfield, A.F.T. and Jirotka, M., 2018. Ethical governance is essential to building trust in robotics and artificial intelligence systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* [Online], 376(2133), p.20180085. Available from: https://doi.org/10.1098/rsta.2018.0085.

Wohleber, R.W., Stowers, K., Chen, J.Y. and Barnes, M., 2017. Effects of agent transparency and communication framing on human-agent teaming [Online]. *2017 ieee international conference on systems, man, and cybernetics, smc 2017.* IEEE, vol. 2017-Janua, pp.3427–3432. Available from: https://doi.org/10.1109/SMC.2017.8123160.

Wortham, R.H., 2018. *Using Other Minds: Transparency as a Fundamental Design Consideration for Artificial Intelligent Systems.* Ph.D. thesis. University of Bath.

Wortham, R.H. and Bryson, J.J., 2016. A role for action selection in consciousness [Online]. *Ceur workshop proceedings.* pp.25–30. Available from: http://www.robwortham.com/instinct-planner/.

Wortham, R.H., Gaudl, S.E. and Bryson, J.J., 2016. Instinct : A Biologically Inspired Reactive Planner for Embedded Environments. *Proceedings of icaps 2016 planrob workshop.*

Wortham, R.H. and Rogers, V.E., 2017. The Muttering Robot: Improving Robot Transparency Though Vocalisation of Reactive Plan Execution [Online]. *26th ieee international symposium on robot and human interactive communication*

(ro-man) workshop on agent transparency for human-autonomy teaming effectiveness. Available from: https://researchportal.bath.ac.uk/en/publications/the-muttering-robot-improving-robot-transparency-though-vocalisathttp://opus.bath.ac.uk/56760/1/muttering{_}robot.pdf.

Wortham, R.H. and Theodorou, A., 2017. Robot transparency, trust and utility. *Connection Science* [Online], 29(3), pp.242–248. Available from: https://doi.org/10.1080/09540091.2017.1313816.

Wortham, R.H., Theodorou, A. and Bryson, J.J., 2016. What Does the Robot Think? Transparency as a Fundamental Design Requirement for Intelligent Systems. *IJCAI-2016 Ethics for Artificial Intelligence Workshop* [Online]. Available from: http://opus.bath.ac.uk/50294/1/WorthamTheodorouBryson{_}EFAI16.pdf.

Wortham, R.H., Theodorou, A. and Bryson, J.J., 2017a. Improving robot transparency: Real-time visualisation of robot AI substantially improves understanding in naive observers [Online]. *Ieee international symposium on robot and human interactive communication (ro-man)*. IEEE, pp.1424–1431. Available from: https://doi.org/10.1109/ROMAN.2017.8172491.

Wortham, R.H., Theodorou, A. and Bryson, J.J., 2017b. Robot transparency: Improving understanding of intelligent behaviour for designers and users. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10454 LNAI, pp.274–289. Available from: https://doi.org/10.1007/978-3-319-64107-2_22.

Xygalatas, D., 2013. Effects of religious setting on cooperative behavior: a case study from Mauritius. *Religion, Brain & Behavior* [Online], 3(2), pp.91–102. Available from: https://doi.org/10.1080/2153599X.2012.724547.

Ysseldyk, R., Matheson, K. and Anisman, H., 2010. Religiosity as Identity: Toward an Understanding of Religion From a Social Identity Perspective. *Personality and Social Psychology Review* [Online], 14(1), pp.60–71. Available from: https://doi.org/10.1177/1088868309349693.