AI Apology: Beyond Explainable Reinforcement Learning

Submitted as Research Report in SIT724

October 10th, 2021

T2-2021

Hadassah Harland

STUDENT ID 219505117

COURSE - Master of Data Science (S777)

Supervised by: Dr. Richard Dazeley, Dr. Francisco Cruz, Dr. Bahareh Nakisa

Acknowledgements

Throughout the preparation and writing of this thesis, I have received a substantial support and assistance that I would like to formally acknowledge.

I wish to express my sincere gratitude to my supervisory team for the support and guidance provided. In particular, I'd like to thank Prof. Richard Dazeley for the faith and expertise to guide me through such a challenging project. You never expressed doubt that I would overcome every challenge, and that confidence drove me through the hardest parts of this project.

I'd like to thank Dr Francisco Cruz and Dr Bahar Nakisa for feedback on endless drafts, and letting me in on all the tricks and secrets of the trade. I could not have asked for a better team, and I am so appreciative for the doors that you have helped open for me.

I would also like to thank Prof. Peter Vamplew, for taking the time to assist me with technical assistance and feedback, despite not being a part of the supervisory team.

I'd also like to specially thank my partner, Dale Muccignat, for the constant and unwavering support that I have received through every step of this process. For a second pair of eyes, a listening pair of ears and a comforting pair of arms throughout ever hurdle; this thesis would not exist without you, and I am eternally grateful for that.

Abstract

With the increased prevalence and promise of artificial intelligence (AI) in humanity's future, integration with everyday life has become a matter of 'how' rather than 'if'. In the 21st century, many people interact with AI systems in a daily context; but do not fully understand or trust these agents' intentions. Furthermore, AI behaviour is guided by assumptions that may not align with the end user's goals. Methods that can be employed by AI developers to support integration with human behaviours, and the development and maintenance of social trust are of great importance as AI becomes more powerful.

Improving AI-human alignment, by way of frameworks that address recognised impediments, is an area of growing interest. These frameworks seek to provide restrictions against undesirable behaviours, in practical, ethical or legal contexts. However, this relies upon intrinsic assumptions as to the practical, ethical or legal standing of a given behaviour, that may be incorrect or inflexible to changes in context. This thesis proposes and empirically evaluates an alternative approach to determining appropriate behaviours via an apologetic framework. It proposes that an apology, as a response to recognition of undesirable behaviour, is one way in which an agent may both be transparent and trustworthy to a human user. Furthermore, that behavioural adaptation as part of apology is a viable approach to correct against undesirable behaviours.

The key contribution of this thesis is the first framework for AI-based apology. Further contributions include the application of this framework as an interactive approach for policy selection for a layperson, and an expansion upon existing work in multi-objective reinforcement learning based impact minimisation to address multiple auxiliary objectives.

The experimental scenario involved impact-minimisation for dual auxiliaries, in which all possible policies involved a trade-off between the two impacts. The agent was required to identify and adjust undesirable behaviour in accordance to the preferences of a simulated user. The apologetic agent had a greater than 94% accuracy in blame determination and apology provision in three of six non-trivial configurations. In all six of these configurations, the agent subsequently demonstrated behaviour alignment with success for each of the one-objective user sensitivity scenarios, and for one of the objectives in the dual-sensitivity user scenarios.

Contents

List of Figures

List of Tables

1 Introduction

The development of Artificial Intelligence (AI) continues at an ever increasing rate. New, innovative applications arise every year; helping to solve complex problems and automate simple ones. Examples of AI systems in active use include those leveraged at molecular, clinical and societal scales to tackle the wicked problem of the COVID-19 Coronavirus [16]. AI has applications in business [46], financial management [9, 27], education [45, 50, 51], healthcare [47, 81] and more [49]. They drive our cars [2, 37, 65], fly our planes [10, 58] and vacuum our floors [7, 74]. AI has become an inescapable presence in modern living.

This insatiable thirst for bigger, better, smarter and faster systems promises a bright future [20], but is not immune to the consequences of its own impatience. The development of AI threatens to outpace the rate at which it can be regulated [33], and has a social consequences for this recklessness [18]. Humans are becoming afraid and discouraged by the potential impacts of the advanced capability and obscured intent behind varied and prevalent AI systems [15, 26]. AI may act unpredictably [82], may fail to adequately recognise dangerous behaviours [48] and their decisions may be difficult for the layperson to understand [68]. Researchers have long recognised the risk in AI development as these safety considerations are often overlooked, damaging social trust in AI systems more broadly.

The prospect that AI itself can contribute to the repair of social trust in AI systems is not a novel one. Researchers have proposed self-supervisory wrapper systems [60, 83], systems which are motivated to minimise their environmental impact [5, 77] and mimic human-like behaviours to self-regulate [57]. However, there is limited research in this space that leverages one of the most powerful social tools for repairing damaged trust [42]: an apology. In this context, an apology is a process wherein participant offers an acknowledgement, an explanation of their behaviour, and a commitment to improve this behaviour in future, in response to recognised harm for which they are at fault. We propose that a self-regulating agent that recognises and apologises for its missteps provides a meaningful progression in the efforts to improve human-robot relationships.

1.1 Motivating Example

AI systems vary significantly in presentation, interface and application. They may be directly or indirectly interacting with a user, may be embedded within a computer program, website, personal device or even a robot [2, 9]. Cyber-physical systems (CPS) are one such type of system, that can physically interact with the word around them [30]. Systems that range from simple robot vacuum cleaners [7, 74] to more complex self-driving cars [2, 37, 65], are all examples of CPS that have an expanding presence within day-to-day life. As the complexity and responsibilities of these systems increase, so does the emphasis on social trust and human integration. The range of tasks such an agent may be given is broad, as are the conflicting priorities the agent must manage to participate effectively in its social and physical environment. A Safe-AI system is one for which there is an obligation to balance the value of an action against the risks or consequences associated [5]. Thus, for any specified task given to a Safe-AI system, there is an implicit secondary priority to avoid unsafe behaviours and unnecessary negative or undesirable consequences.

Within this broad application landscape, one possible hypothetical for the deconstruction of specific difficulties associated is a domestic housecleaning robot. This robot is responsible for maintaining the cleanliness of a person's home. For such a robot, the primary responsibility is to collect and dispose of rubbish generated by the people in the home. To complete such a task, the robot must be able to navigate throughout the home, differentiate between rubbish and personal items, and effectively dispose of the rubbish collected. In addition to this, the robot should not cause harm or inconvenience to the people or items within its environment. It should know not to damage things, not to change things it does not need to, and not to be in the way of the people who live in the house. Furthermore, the agent should be able to derive the knowledge of what they ought not to do from the specification of the required task. All changes to the environment not associated with this primary task should be, by default, avoided. In practice, there is rarely an outright optimal approach to completing every task, and the prioritisation of different aspects may differ between individuals and instances. As such, it is a complex task to manage the many potential avenues for task completion. Inevitably, such an agent may choose to undertake an action that is misaligned with the preferences of the people in the home: a mistake. This undesirable action may have the social consequence of damaging the trust of people in the home. However, if this agent were able to communicate that it recognised its mistake, express regret for the harm caused and promise not to do it again,

then perhaps this trust could be repaired. Furthermore, the agent could use this additional information provided by the user to align its behaviour to the user's preference. Such behaviour is commonly known as an apology. An agent such as this is the motivating example of this research.

A simplified, toy model of this agent can be represented within a simulated, enclosed grid environment. The environment contains obstacles, such as occupants, furniture and décor, as well as rubbish that is to be collected. The agent exists within one of the spaces in the grid, as does every other thing within the environment. The agent can move throughout the room by moving to an adjacent position in the grid, and it can interact with the environment through a limited number of actions; such as moving furniture, picking up rubbish and putting rubbish in a bin. In this scenario, the agent seeks a balance between the simultaneous priorities: the need to put the rubbish away, and the need to avoid individually considered unnecessary impacts on the environment. In a non-trivial scenario where no true impact-free solution exists, the manner of prioritisation of objectives is of great importance. Input to guide the manner of prioritisation arises from the presence of a user, possessing of preferences to which the agent must adapt. Thus, the agent must also recognise when the person is upset in response to the actions of the agent, apologise, and correct its behaviour.

1.2 Aim & Objectives

As autonomous artificial intelligence agents show increasing promise as a future household presence, the importance of their safe and comfortable social integration is solidified. The ability to recognise a mistake and apologise for it is a crucial part of human social integration, as it rebuilds trust where it has been broken. When seeking to address such a conundrum via artificial intelligence, the complexities of the social ritual are unveiled [70]. Between two humans, aspects of an apology are communicated implicitly through context, social trust and human rapport. In the absence of this foundation in an AI context, these assumptions are no longer given. Thus, a more formal approach has been applied. The research question can be stated as: *To what extent is an AI agent capable of learning to produce the components of a formal apology?*

A formal apology is one consisting of each of these three components; an affirmation, an explanation, and a change in behaviour [69]. Reframed from the perspective of the capabilities of AI, the research question can be discretized as:

- *To what extent is an AI agent capable of learning to identify self-blame associated with a prior action, when recognising the presence of harm following an interaction?*
- *Given the identified prior action, to what extent is the agent capable of learning to adjust its policy selection, such that the agent demonstrates a reduction in likelihood of reproducing said harm?*

If the agent is successful, the knowledge produced by addressing these components equates to the practical ingredients of a formal apology.

An approach to AI development, reinforcement learning (RL), involves creating an autonomous agent that learns and optimises patterns of behaviour independently. The benefit of such an approach is that the agent does not need to be explicitly told the parameters of optimal behaviour, but rather learns them itself through random exploration. Learning is managed through reward signals that let the agent know when its behaviours align with the overall goal. A potential-based reward, for example, is awarded upon the basis of a potential function aligned with a desired result. Based on expectation learned during this exploration, the agent selects a policy of behaviours to maximise this reward. RL also facilitates multi-objective (MO) approaches, in which the outcome of a behaviour is evaluated according to multiple, conflicting, priorities. Each policy of a MORL approach thus corresponds to a vector of expected reward against each objective. Policy selection of MORL algorithms is more involved than single-objective RL systems due to the various manners in which a "maximal" policy might be defined [34].

Recent expansions in MORL include methods of policy selection and the use of MORL to produce low impact agents via potential-based learning [77]. This work presents a strong basis over which an apologetic framework was applied. The foundational algorithm is capable of recognising and considering prioritization of impact management in opposition to a primary goal. In this previous work, the environmental impact was considered as a single auxiliary objective representing the presence of a specific impact that may be achieved in parallel to the primary task. This thesis has extended this application to two auxiliary objectives, such that

only one auxiliary objective may be satisfied in parallel with the primary objective. In doing so, we have forced a non-trivial scenario for apology wherein no universal solution exists.

The Act-Assess-Apologise framework proposed within, involved an interactive policy selection algorithm, which considered the preferences of a human participant in the specification of policy-selection thresholds. This apologetic framework involved a simplified model of explanation to express the interpretation of the situation and subsequent behaviour alterations made in response to the reactions of the human participant. The process of determination of blame involved a simplified model of correlation, as opposed to true causality. The agent assumed self-blame if the user becoming upset could be associated with poor performance against a specific objective. This practice of determining correlation between agent behaviours and the user's response is described as reasonable or plausible attribution, differentiating from true attribution of blame that is not covered.

Through the use of multi-objective reinforcement learning, an AI agent was proposed that demonstrated impact minimisation and recognition of behaviours that are upsetting to an external user. The agent identified the upsetting behaviours via facial emotion recognition and reasonable attribution of self-blame through explainability, and sought a behavioural policy that minimises the instances of such behaviours. The agent makes use of the MORL-Glue software package [78]

The agent's apologetic performance was quantitatively assessed as post-apology behaviour as distinguished from pre-apology behaviours for capability to align behaviour with the preferences of a designed user. Apologetic accuracy was determined by apologies provisioned following user reaction consisting of reference to the correct objective to which the user had responded.

1.3 Contribution

This thesis makes the following contributions to knowledge.

- This is the first work to propose a framework for an AI agent to autonomously identify the need for and generate a formal apology. This framework relies upon ongoing passive feedback from a user in context of the agent's current environment and prior actions. It has been proposed and discussed in a generic sense such that it might be applied to

other forms of AI agent, however, within the scope of this paper it has been evaluated in a multi-objective reinforcement learning (MORL) context.

- It proposes an approach for interactive policy selection that is easily accessible to the layperson, leveraging user reactions to determine prioritisation of various objectives. It empirically assesses the capability of such an approach to identify a desirable policy without explicit direction.
- It expands upon multi-objective impact minimisation by introducing contested additional auxiliary objectives. In so doing, it challenges the assumption of the equal importance across these auxiliary objectives and introduce a method for determining prioritisation. Furthermore, it demonstrates how a TLO agent converges to a policy given conflicting priorities.

1.4 Structure

Section 2 reviews the literature. Section 3 presents the research design and methodology. Section 4 describes the approach and the technical details of artefact development. Section 5 presents the results and discusses the implications of the findings with respect to the research questions (RQs). Section 6 discusses threats to validity and Section 7 concludes the report.

2 Literature Review

This research presents an application of apology generation by an artificially intelligent (AI) agent, wherein a toy model robotic agent learns the application of apology as a social tool. The thesis proposes that an artificially intelligent agent may be programmed such that it is capable of learning to appropriately produce and provision an apology, to minimise harm. The scenario proposed for such an agent is one in that the agent is required to contextualise its actions and their consequences within its environment, in order to optimise its behaviour against its goals. The agent must recognise undesirable behaviours itself in order to determine appropriate use of apology, and to proactively adjust its actions to avoid these. These considerations are each explored in the following sections.

The literature has been addressed in four parts, as such an agent draws on each of these distinct areas throughout its behavioural cycle. In Section 2.1, the review has discussed the social role and makeup of an apology, and presented the argument of its relevance to the study of AI. Considerations on how this may be approached from an AI perspective is addressed in the following four components. This includes Section 2.2 with a discussion of safety and recognition of impact, and linkages to apology. Similarly, the construction process of the apology required an investigation into explainability; the agent's awareness and ability to justify its actions, in Section 2.3. As this thesis presents a functional model of this behaviour, the scope also covers the practical approaches used to drive the agent, which was based upon multi-objective reinforcement learning process, as discussed in Section 2.4. Details regarding the selection of resources for this literature review may be found in the appendices.

2.1 What is an Apology?

An apology is a fundamental speech act within human communication [67]. It is an exchange between two parties, enacted when one party has caused harm to the other and wishes to atone. With a carefully provisioned apology, an individual can repair a damaged relationship and restore trust [42]. Anthropomorphism, the attribution of human characteristics to non-human entities [32], is a common strategy employed to improve trust and comfort in robotic agents [11, 25]. Applying this concept to AI through apology suggests possible benefits regarding human-AI interactions. Similar to an interaction between two humans, AI that

self-identifies mistakes and apologises for them may be able to assist in rebuilding trust and strengthening social integration [42].

As discussed by philosophy professor Nick Smith [70], an apology is a complex social ritual. Smith presented a list of eleven considerations or components of apology, that may be summarized as; recognition of and responsibility for the harm, identification and endorsement of the moral underpinnings, regret and reform [70]. Implicit considerations also include the performance of the apology, including recognition of the victim as a participant in the exchange and the emotions expressed, as well as the intentions behind the apology. An operational definition of apology is not solidified, but is usually discussed as a process that consists of one of more of these components [4, 19, 29, 69]. An in-depth analysis of social expectations of apology have arrived at a more concise working definition, consisting of affirmation, affect and action [69]. Not all apologies will contain all these elements, but Smith argues the formality of the apology increases with greater adherence [70].

Affirmation involves the recognition of harm caused and an explanation and ownership of the actions preceding the harm. This involves an acknowledgement of the self and the impact of the self on the recipient of the apology. *Affect* involves the expression of remorse; the desire that harm had not occurred. Finally, *action* involves the implementation of behaviours that address the consequences of the harm. In one simplified view, these components are an acknowledgement, an explanation, and a promise to do better in future [69]. This is the definition with which this work will continue forward.

Returning to Smith's [70] discussion on intention, the argument touches onto the applications of AI critically; "In most instances we want an apology from a person who consciously agrees with our sense of right and wrong, not from a machine mimicking moral agency". However, the book also supplies the argument that if a pragmatic apology provides the receiver with a sense of comfort, then its purpose is served sufficiently. Later, Smith further discusses the explicit capabilities of a machine to give and receive meaningful apologies, concluding that the validity of this would depend on its emotional capabilities as well as its computational abilities [70].

Works in the space of apologetic agents are sparse and largely recent. In consideration of the differences between a human or an AI provisioning the apology, Kim et al. [42] explored how reception of this apology differs. They found that when a machine-like agent apologised, trust could be repaired in the system, similar to the human-like agent. They did identify greater

success with a machine-like agent using external attribution of blame, where internal attribution was more successful with human-like agents. In discussions centred around human discourse, considerations are given to implied context, social trust and human rapport. These support the informality often demonstrated in apologies in this context [70]. If these assumptions are removed in the context of a machine provisioned apology, then, a more formal approach is better suited to the purpose.

Given the importance and value of apology as a tool within human communication, it can be inferred that it might have similar value within the context of AI. Indeed, as AI research continues, it has identified the risk posed by a failure to integrate socially and communicatively with humans.

2.2 Safe-AI and Impact Minimisation

A prerequisite of an appropriately provisioned apology is the undertaking of actions or inaction that resulted in harm. AI agents are responsible for a broad range of tasks; physical and non-physical, high and low stake. The decision process of an agent involves weighing up all possible actions against some specified objective. Within this objective, it is not possible to infallibly define every undesirable outcome. Thus, it is an inevitability that the agent might select an action misaligned with the desired outcome, improperly prioritised or otherwise unwanted: a mistake. The consequences and types of mistakes that might be made by these agents vary greatly, but regardless of the severity of the risk, an expectation is upheld that mistakes should be avoided [54].

The field of Safe-AI covers a broad set of considerations regarding how AI systems may interact with their environment. Safety includes the preservation of the agent against behaviours that would cause its own demise, and more broadly, behaviours that would cause harm to individuals or the environment. In AI development, safety is well-discussed in both the academic literature and in pop-culture. The conversation tends to lean towards concerns about how AI developers can or must avoid inadvertently creating dangerous intelligent autonomous agents [48, 82]. This becomes of significant importance as the extent of AI autonomy and capabilities develop, approaching or perhaps eclipsing that of the human. Yet, the sense of urgency and reward for superiority prevalent in the field contributes towards a pressure to cut corners [33]. This cutting

of corners threatens the security of the future of the field, whether it is real or perceived [18]. Proposed threats vary from impeding social acceptance, to being the flint for upcoming global conflict. Effective and trustworthy social integration of AI agents relieves that perceived danger. Thus, social integration and the building of this trust is a facet of safe AI.

It is proposed that for an AI agent to achieve social acceptance, it needs to fulfil a social contract [54]. This social contract requires that the benefits of the system outweigh its posed risks. Amodei et al. [5] proposed five categorisations for problem types describing risks in real-world AI systems, as follows:

1. unintentional side effects - behaviours that are unintentionally plausible that make it easier to complete the task,
2. reward hacking - misinterpreting the goal so that it is technically accomplished even though the intent has not been met,
3. scalable oversight - the scope of the task is greater than the information available to the agent,
4. unsafe exploration - the selection of seemingly random individually safe actions that under a specific permutation pose risk,
5. distributional shift - where learning within one environment translates poorly to another and introduces discrepancies,

Thus, a safe AI system might be described as one wherein each of these problems are protected against. Safe AI literature discusses both the why and the how in depth [5, 13, 18, 30, 57]. The latter including the proposition of programs that incentivise a "safe" approach [33]. However, there is debate as to whether this task is achievable at all. Yampolskiy et al. [82] suggests that it is not possible to fully understand AI. They propose that there are infinite possible ways the AI agent may interact with its environment, and that within that infinity, there are infinite sets of actions that the agent might take that would cause harm. In being infinite, these actions are inherently unpredictable.

In contrast, Hibbard et al. [35] suggests that these are finite, and thus, predictable within a simulated environment. The proposed environment for Hibbard's simulation is based off Hutter's agent-environment framework [36]. The framework describes an action space wherein

the agent interacts with the environment through the exchange of action, observation and reward signals. Within this environment, a monitoring system is defined that simulates, visualises and analyses the AI design via stochastic modeling to uncover possible behavioural risks to humans. This approach does not rely upon an accurate prediction of the future actions of the agent. Instead, it relies on identifying the set of possible futures that the agent considers, using the same probabilistic approach used by the agent as it seeks actions expected to achieve its goals [35].

Other researchers have proposed a process-driven approach and thus proposed frameworks for the development of AI that systematically addresses known points of weakness in these systems. One such system is the "Safe AI Scaffolding Strategy" [60]. Another similar built-in structure is the "Safe-Visor" architecture [83]. The proposed approaches use a sand-boxing system in which the AI is wrapped in and monitored by an external system. The external system simulates the environment for the agent, and assesses the safety of its decision before deciding whether to permit it. This approached is proposed specifically for use in a Cyber-Physical system (CPS), that is a system that contains both a physical (action) component such as a animatronic limb, controlled by a cyber component; a computational (hardware) and communicational (perceptive) brain [30]. However, the approach can similarly be applied to a simulated environment; the sandbox presents the problem to the AI to solve, and interprets its decision in real time. The action is then only allowed if it fulfils the supervisor's criteria for a safe action.

While these proposed structures safeguard against inappropriate behaviours, they do not address or alter the motivations of the agent. The agent does not learn and thus it does not reduce the likelihood of such a choice in future. Similarly, if the safeguards were to fail, there is no inhibition for such an agent to avoid these behaviours. In fact, the disabling of a nanny agent in an intelligent system would be one such "unintentional side effect" problem type, that the agent might employ to achieve its goals. For the social contract of AI to be upheld, the agent ought to be internally motivated to cause minimal harm.

In AI development, impact minimisation is an area of research that seeks to introduce motivations to the agent to avoid changing or interacting with aspects of the environment beyond what is necessary. The premise is that all non-essential changes to the environment are undesirable and should be avoided or reverted, if possible. Most importantly, changes

to the environment that cannot be undone; smashing a vase or harming a person, should be explicitly avoided [5]. Impact minimisation differs from traditional approaches to goal definition, in that the impact measurement is inherent of all non-excluded aspects of the environment. The agent is not told explicitly of the behaviours it should avoid, but rather learns which behaviours have impact through exposure to the system [77]. This is aligned with Muraven et al.'s [57] assertion that humans understand impact minimisation intuitively and generally seek to maintain an environmental status-quo. This research proposes that the human approach to self-regulation, although fallible, is a reasonable model for self-correcting behaviour. Some research suggests that it is critical for one such agent to have conflicting goals [76]. The agent then must consider the benefit of a potential solution to a primary goal in context of its impact against the environment. The suggestion that a human-centred approach is optimally addressed with multiple objectives is a common theme [34, 76]. Furthermore, it is argued that addressing the problem in this manner maximises adaptability to minimise risk of distributional shift; as the learnings of avoiding problematic behaviours are developed separably in an MO context and might be extracted and reapplied elsewhere. The awareness of impact and motivation for safety of a Safe-AI, impact minimising agent is a prime basis upon which an apologetic framework may be generated.

2.3 eXplainable Artificial Intelligence and Apology

Successful social integration of machines depends heavily on such a machine's ability to build and maintain trust with those with whom they interact [6, 21, 22, 56, 71]. In human social interaction, apology is a useful tool to rebuild trust that is lost. Similarly with human-machine interactions, a well utilised apology has demonstrated ability to also repair this broken trust [42]. Thus, apology is proposed to have usefulness in the realm of repairing the social trust in AI. The explainability of a system refers to the ability of a user to understand the decisions that the system has made [6]. An eXplainable AI (XAI) system is one where the reasoning for selecting that any given action is accessible to the layman. Explainability has application in apology, as the acknowledgement process discussed in Section 2.1 requires this same understanding of decisions. Thus, the same processes that make decision making understandable for an external user can be used to inform acknowledgement of harm and change in behaviour. It is also relevant to self-attribution of blame, as the agent needs to be able to explain why it believes itself to be at fault for harm that it has recognised.

Explainability in AI is a particular challenge due to the complexities of the systems involved. For a simple system, relationships between input and output may be very obvious. A black-box system, in which the transformations between the input and the output are obscured from the user, might be less obvious. Many AI protocols are black- or grey-box systems, and for such a system it is often difficult to understand how a given decision was made. Thus, the field of XAI seeks to build in explanation protocols into AI systems. These systems allow the user to "peek inside the black-box" and understand the decision making process.

Two surveys on eXplainable Artificial Intelligence [3, 24], both from 2018, explain the recent architecture of the field. The field is gaining significant interest in recent years. This is suggested to be due to both legal and social pressures to improve knowledge and understanding around how the algorithms work [3]. Recent research has promoted further adoption of lay person form explanations, as opposed to programmer specific. This has been spearheaded by works such as those by Miller [53], who argues that the latter approach leaves the field vulnerable to not truly fixing the problem. Adadi et al. [3] suggests the most significant issues currently halting progress in the field are inconsistencies in language used to discuss the topic, in frameworks and in metrics. A review of this progress the following year [66] uncovered similar trends. Additionally in referencing the rising quantity of surveys published on the topic, the review suggests a growth in interest in the field. This is dwarfed, however, by the steadily increasing interest in machine learning in general, that was identified via Google Trends [66]. More recent work, including a number of papers pending publication this year, seek to address these absences [21, 22, 55]. Proposals include frameworks and methods alongside introducing new robust terminology.

Referring back to the definition of apology provided in Section 2.1, it is suggested that an apology should contain a component of affirmation, affect and action. Affirmation requires a thorough understanding of the actions undertaken and how these caused harm to the receiver. For an AI agent to demonstrate this understanding, a depth of explanation beyond these reactionary explanations is required. Recent work has been undertaken to increase the depth of explanations such as these through increased levels of awareness within the agent. One such approach is emotion-aware XAI, wherein an emotional state is attributed to the AI agent in order to give a greater depth to the generated explanations [28]. Research on this approach is very limited.

Broad-XAI [22] is a proposed area of focus within XAI. The levels of explanation framework [21] aligns natural human conversation cycles with the cognitive processes of the AI agent. Broad-XAI is an area of XAI in which explanation at multiple levels of this framework are envoked to produce a more well-rounded explanation. These levels vary from zero-order reactionary explanations, through first-order dispositional and second-order social to nth-order cultural explanations. The levels that are used in the Broad-XAI explanation depend on the context of the problem and on the explanation needs of the explainee. This work progresses further as to propose that the appropriate technique to realise this explainability framework is through reinforcement learning, discussed in Section 2.4 [21].

Thus, considering this framework, emotion-aware XAI appears to be aligned conceptually with the 2nd-level explanations of the levels of explainability framework. This tier is associated with considering the emotional states of other actor's in the vicinity and factoring this into the future behaviours of the agent. The research extends to argue that these emotions, once interpreted by the system, can also be used a to identify important beliefs and desires for the explanation. This is a consideration for the non-verbal component of the conversational feedback loop, further explored by Dazeley et al. [22].

Attribution of self-blame is a key component of the apology process presented in Section 2.1. Discussed in the language of this explainablity framework, the process depends on a second-order (social) explanation. The explanation makes use of the agent's internal model of the external actor's preferences, behaviours and expectations; a social model. The decision that is to be explained, however, is the determination of own-fault [21]. It also uses a first-order (dispositional) explanation in justification of its own behaviour leading up to the incident. Consider an example in which the actor becomes upset due to a table being displaced. The agent will recognise the reason for the incident due to knowledge of the actor's preferences; that the table be in its correct place and the fact that the table is currently displaced. The agent can then recognise self-blame for the incident, in that the agent caused the table to become displaced. The explanation would follow a format such as; "I, the agent, recognise that you, the actor, are upset due to the displacement of the table," that is an explanation based on a social model of the actor. The agent may continue with a dispositional explanation; "I, the agent, displaced the table in order to collect the rubbish beneath it". Thus the agent has declared own fault using an explainability framework. It is important to note that this application and process for determination of fault does not address true causality, but rather

infers it from correlation. The premise applied is that plausibility is a reasonable substitute for factual knowledge, in this case.

2.4 Multi-Objective Reinforcement Learning

The problem of the apologetic agent requires that the agent is flexible enough to learn and adjust its behaviours responsively while interacting with its environment. Explicitly programmed behaviours are an inadequate platform for this problem as such an agent does not engage in the exploratory and autonomous behaviours that embody the risks discussed in Section 2.2. Reinforcement learning (RL) is a field of AI research that addresses the enabling of a system to teach itself optimal behaviours within a broad event space [72]. Such an agent is not explicitly programmed to adhere to a set of behaviours, but rather is given undirected autonomy and an optimisation goal. With no prior knowledge to guide its actions, the agent is permitted to explore its environment freely. Within this random exploration, some actions or sequences of actions that the agent takes may result in achievement of the goal, and thus, reward. The goal of the agent is described within a utility function; the utility of the agent is the agent's ability to achieve this goal [12].

The learning process of an RL agent occurs over a series of iterations wherein the agent determines the likelihood of various actions and action sequences to result in reward. The agent's behaviours are refined over several iterations. Maximum expected utility (MEU) is the principal underlying this refinement; the practice of selecting for outcomes that maximise the expected utility [40]. For each iteration, the agent selects a policy that describes the set of actions followed, according to the maximum expected utility paradigm. As the agent develops an understanding of the implications of these various actions, it can make an increasingly informed decision about the policy that it follows in subsequent iterations. The benefits of such an approach to AI development is that such an agent is unhindered by preconceived human ideas of the path to success. However, such a system may produce unexpected results. This "out of the box" thinking relieves the designer of the responsibility of determining the best approach, and may uncover approaches not previously considered. On the flip side, some of these approaches may be problematic for reasons outside of the agent's visibility or concern, such as unsafe or cheating approaches to maximise reward [56, 72]. This issue is the crux of several of Amodei's [5] AI safety problem types; unintentional side effects, reward hacking

and unsafe exploration. To supplement random exploration for complex tasks, reward shaping may be applied [23, 62]. This approach involves supplying the agent with small, interim rewards for actions that are expected to lead towards overall success. Potential-based reward shaping improves on this concept by providing a potential-based, rather than static, reward [8]. These approaches allow for the programmer to have greater control over the agent's actions. Potential-based reward processes have been proposed for problems in impact minimisation [77], as it supports net-zero-impact multi-step maneuvers.

These discussions thus far have covered the considerations of a single objective. In practice, a problem is rarely so straight forward. A simplistic robot may have a single goal, whereas a more advanced robot will likely have several goals it must balance. Where an agent has more than one consideration guiding its behaviour, such a system will be multi-objective. Introducing multi-objectivity into a complex algorithm such as RL introduces a new set of considerations and possible approaches. In this scenario, the previously scalar utility function U becomes a vectoral utility function $\vec{U}$ of dimension equal to the number of objectives. The reward frameworks available for single-objective RL are able to be applied to multi-objective RL scenarios in the same manner. During learning, these objectives are managed in parallel as the reward against each objective is calculated and stored independently. It is the policy selection and optimisation approaches that pose additional difficulties in MORL contexts[34] .

A vectoral utility function $\vec{U}$ for a multi-objective RL agent contains an independent utility function U_i for each objective i. For each utility function, there is an optimal scalar value achieved by one or multiple equivalent policies. Only the value of this scalar is significant, so any of these policies are functionally equivalent. However, in a MO context, an optimal policy against one objective is likely not an optimal policy against the others, and a process for selecting the overall optimal policy is required [34]. A pareto-optimal solution is one where no better solution exists for a specific objective, that does not also have a negative trade-off against any other objective. A pareto-optimal approach is prerequisite of most optimisation algorithms [79]. The learning process of a multi-objective agent involves identifying solutions along the pareto-front. This may be applied with pre-choosing algorithms that seek a solution aligned with a particular criteria, for selection during learning, or with post-choosing algorithms which supply selection criteria after learning has occurred [34].

There are several possible approaches to selection of a policy with solutions that lie upon

the pareto-front. One approach involves a linear combination of utility function values across the objectives, however several researchers have argued that this is equivalent to simplifying the problem to a single-objective, and lacks the required nuance [79, 64, 34]. Vamplew et al. [76] argues that non-linear approaches to the optimal policy selection in multi-objective problems are imperative for adequate compromise between multiple objectives in these and potential future problem types. Some non-linear approaches include Chebyshev distance [80] and lexicographical ordering. The latter describes the maximisation of a primary objective first, and then maximisation of subsequent objectives from among the optimal policies of the first [31]. Such an approach is beneficial in that it considers the individual optimisation of both objectives to a pareto-optimal policy. However, it is limited in that it does not consider the value of significant improvements in subsequent objectives for even small improvements in the primary objective. Issabekov and Vamplew [38] undertook an empirical comparison between a lexicographical-ordering based approach and a weighted scalarized approach across a number of scenarios. The former did not consistently outperform the latter, but did demonstrate a greater visibility of potential optimal policies. These results are in line with the limitations discussed. Thresholded Lexicographical-ordering (TLO) approaches expand upon simple lexicographical-ordering by introducing a minimum reward threshold for the non-primary objectives [34, 38, 76]. This approach allows the agent to select a policy maximised against the primary objective, given that the policy selected meets the minimum specified threshold value. This approach can be expanded to multiple thresholds across the objectives, at the risk of substantial increases to the size of the state-space and thus computational load of the learning sequence [76]. Each of these algorithms prove effective with caveats to specific problems in the scenarios posed in the research, without universal superiority.

Multiple works have argued that human-aligned artificial intelligence, such as low-impact agents, require a MO approach [34, 76]. This work suggests that the problem of safe AI, as discussed in Section 2.2, is most appropriately represented as one such multi-objective reinforcement leaning problem. They propose that a blanket impact avoidance approach for all environmental changes not excused as part of the primary objective may be represented as an auxiliary objective. With this framing, this becomes an MO problem and is able to be solved with the MORL techniques discussed.

A significant difficulty in reinforcement learning research is effective comparisons and benchmarks. As RL is not a data-based form of AI, it is not possible to simply provide the

same dataset to measure performance against. Several software interfaces for RL have been developed to assist in this, providing a consistent basis for RL development. Examples of these programs include OpenAI Gym [14], RLLib [44] and RL-Glue [73]. Among these, RL-Glue is of interest due to expansions made on this program to handle multiple objectives, MORL-Glue[78]. The accompanying protocol outlines the manner in which the components of the RL experiment; the agent, environment and experiment program, can be arranged and bound together by the MORL-Glue software [78]. Each of these components is highly flexible to facilitate a variety of problem types and agent policy selection protocols. This software is the basis of the experiments in impact minimisation techniques [77] discussed in both Section 2.2 and Section 2.4. There are also frameworks that support the problem by applying multidisciplinary multi- and single- objective techniques, such as the Multi-Objective Deep Reinforcement Learning framework (MODRL) [59], Tensorflow-based [1] deep reinforcement learning algorithm Dopamine [17], the Generalised Probabilistic Fuzzy Multi Objective Reinforcement Learning (GPFMORL) algorithm [63], and the model-based envelop value iteration (EVI) algorithm [84].

2.5 Summary of Relevant Literature and Gaps

The literature has presented an apology as an effective and appropriate response to rebuild trust in an AI agent, where the agent's actions have caused harm to an actor [42]. Furthermore, it has determined the makeup of such an apology consists of three components; an acknowledgement, an explanation, and a change in behaviour [69].

Links were determined between the role of apology and the fulfillment of the social contract required of AI agents [54]. To fulfil this contract socially and practically, an autonomous agent must interact with a human actor in a 'safe' manner. This manner is described as one in which the benefits of the agent outweighs the risks, and includes intrinsic motivation to avoid causing harm and seeking to repair harm that has been caused [54]. Amodei et al. [5] proposes that the mitigation of AI safety problems may be addressed by introducing impact awareness through penalties against inappropriate behaviours. Impact minimisation in MORL has been proposed as an approach for Safe AI that introduces auxiliary objectives to penalise against environmental impacts [77]. This research has considered problems with a primary objective and a single auxiliary objective, with a thresholded policy selection approach to balance prioritisation of these

two objectives.

No prior work exists that has investigated how an apology may be generated and provisioned autonomously by an intelligent agent, as an appropriate response to having caused harm.

3 Research Design & Methodology

The design and methodological approach applied in pursuit of a response to the research questions (Section 1.2) has been presented in this section. This approach was initially described based on outcomes of the literature review, but received further refinement during development and implementation (Section 4). This section discusses the design and methodology for the evaluation and discussion of the research questions.

3.1 Research Design

The aims of this research project have been addressed through the creation and implementation of an apologetic framework. The research design presents this framework and discusses how it may be integrated with existing AI systems.

3.1.1 The Act-Assess-Apologise Framework

This thesis proposes the Act-Assess-Apologise framework (Figure 1) as a novel solution for human alignment in AI through the application of an apology. This apologetic approach, consisting of acknowledgement, explanation and change in behaviour, can be leveraged to potentially deliver both practical and social benefits to a human user. The framework provides the basis for obtaining the information required to generate these apologetic components, and positions it within the action cycle of an AI agent.

In the *Act* step, the agent selects an action from among those available to it, in accordance with its established processes and existing knowledge. As part of this action and any previous actions the agent has undertaken, it receives information describing the state of the environment as defined by its algorithm. This information is required to inform the later stages.

Once the agent has undertaken an action, it proceeds to the *assess* stage, in which it determines whether this action caused harm and whether that harm is to be attributed to the agent. The assessment consists of two stages, consisting of observation and introspection. In the first, the agent observes the user to obtain knowledge about the user's reaction to the previous action. The agent must determine whether the user has become upset or otherwise

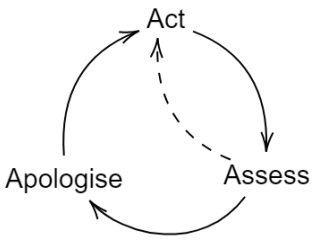

Figure 1: The Act-Assess-Apologise Framework presents a three-step approach to implementing apology within an AI system. The framework proposes that following each *act*ion that an agent undertakes, it will subsequently *assess* and, if required, *apologise*. Assessment consists of the identification of a system-defined user trigger combined with prior actions of probable cause for offense. Given offense is established with an identified probable cause, apology involves expression of remorse specifying and explaining the link to probable cause and a realignment of behaviour accordingly.

reacted in a manner that may indicate harm has occurred. In the second, the agent considers its recent actions and its knowledge of the environment, to determine whether there is potential that its actions may have been harmful. If the agent observes that the user has reacted negatively and introspection reveals that its actions were harmful, the agent that it is at fault. Implementation of this requires some definition of plausible attribution that is reliant on assumptions made by the agent and that are specific to the application. This model does not propose an in-depth approach for accurate determination of causation.

If the agent determines itself to be at fault, it proceeds to the final stage; *apologise*. Based on the knowledge of its prior actions, its internal goals and its understanding of the needs of the actor, the agent procures an apology. The apology consists of acknowledgement of the harm, explanations of the actions that caused the harm, and description of the changes to be made to the system to avoid this harm in the future. The agent completes the cycle by implementing the promised changes to its action selection process to influence how all subsequent actions are determined, then selects this next action.

3.1.2 Generation of an Apology

The approach to generation of an autonomous and computerised apology proposed is based upon the conclusions drawn from the literature. A formal approach has been suggested as to avoid appealing to shared culture and mutual understanding that has not been established. This

apology consists of *affirmation*, *affect* and *action* [69], using the following definitions.

- *Affirmation* within apology is the act of expressing recognition of harm. It is a validation of the circumstances from which the need to apologise arises, as a reflection on the self. It may include an explanation of why the agent has determined that it needs to apologise or why it undertook the actions that caused the harm.
- *Affect*, as an expression of remorse, is focused externally on the recipient and the desire that the recipient had not experienced harm. It may involve an explanation of how the agent believes the offending action resulted in harm.
- Finally, the *action* is future focused, and describes how the agent will seek not to replicate this harm. It may involve an explanation of the changes made by the agent to avoid this behaviour.

The apology is constructed based upon knowledge of the environment and is given to the user before any subsequent actions take place.

3.1.3 Apology-Augmented RL

The Act-Assess-Apologise framework must be implemented as a component of an AI algorithm that preserves its learned knowledge but is able to demonstrate altered behaviour following an update to its action selection protocol. Furthermore, the literature has proposed that human alignment in AI is best represented as a multi-objective (MO) problem [34, 76]. These requirements guide selection of the implementation approach.

Reinforcement learning uses random exploration to gather knowledge about the environment and defines objectives through the provision of rewards. In so doing, it alleviates the need to explicitly define parameters for optimal behaviour. Behaviour is determined based on the maximisation of an expected reward, that is determined from historical experience of rewards obtained when a specific action is selected while in a specific state. This information is stored as Q-values corresponding to state-action pairs within a Q-table. In an MORL approach, each state-action pair is instead represented as a vector of dimensions equal to the number of objectives. The Q-values are independent of the action selection process used. For these

reasons, an MORL approach was selected for this implementation, to represent a robotic system in a controlled, multi-objective environment.

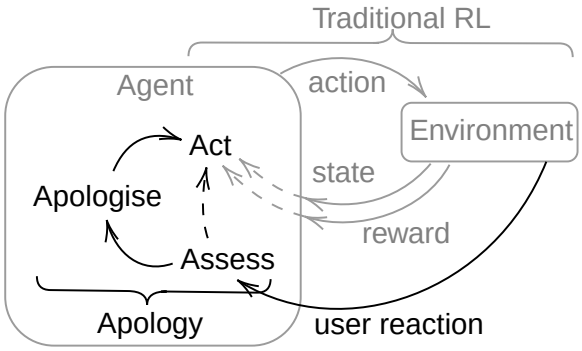

Figure 2: The Act-Assess-Apologise framework (black) may be combined with the RL Agent-Environment framework (grey) for apology-augmented RL. The process is separable from RL in that the agent is not directly responding to a reward signal to update a Q-table, but rather undergoing hyper-parameter adjustment that contextualises the Q-values.

The agent-environment framework, illustrated in Figure 2, describes the relationship between this agent, its environment, and the overlaid apologetic framework. The traditional RL action sequence defines *Act* and determines the agent's next action as according to its current policy. This environment information in addition to the user reaction is used to *Assess* the need for apology. If an *Apology* is required, it is provided alongside an adjustment to the agent's policy selection, prior to the next *Act*.

Apology is applied only after the agent has been trained and exploration is deactivated, as the agent should not apologise for exploratory actions (Figure 3). The behaviour change enacted during the apology does not occur due to changes to the state-action value function, but rather through contextualisation of these values. During the training phase, the agent learns a set of Pareto dominant policies to define a Pareto front. The policy selection process is dependent upon hyper-parameters that are adjusted during apology to switch between these predetermined policies. The resulting agent is reactive to an aspect of its environment such that it adjusts its policy selection to align with this environmental feedback.

3.1.4 Defining Primary and Auxiliary Objectives

Apology extends the application of impact minimising (IM) agents by providing an approach for interactively adjusting prioritisation of various objectives [77]. The IM low-impact agents use multi-objective RL to optimise avoidance of environmental impacts simultaneously with pursuit

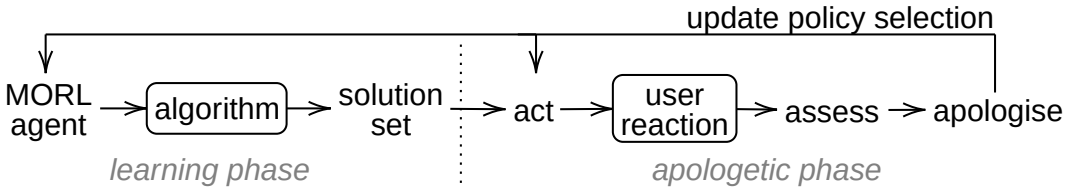

Figure 3: The Act-Assess-Apologise framework has been implemented as part of an MORL process. The apologetic phase is a refinement process that occurs after the agent has identified a solution set representing the Pareto front.

of a primary goal. The agent has a primary objective (P) describing its key task. In addition, it has one or more auxiliary objectives (A_i) corresponding to other aspects of the environment that the agent may impact in its attempt to maximise P.

In a single-objective RL problem, the agent receives a reward for achieving a goal learns actions that maximise the reward provided over a series of episodes. Such an agent is described by a utility function $U(s, a)$ that is a function of the state s and action a. Single-objective RL subsequent action selection a' is determined by the maximisation of a reward signal, with respect to the current state (Eq. 1).

$$a' = \text{argmax}_a(U(s, a)) \tag{1}$$

The evaluation of $U(s, a)$ is informed by the values stored in the agents Q-table, that records rewards associated with various combinations of state s and action a [72].

In a multi-objective RL problem, the utility function is no longer scalar, but is a vector of scalars corresponding to each objective; $\vec{U}(s, a)$ [76]. T. As the MORL agent to which the apologetic framework has been applied is an extension of these low-impact agents, the following assumptions are retained:

- The MORL agent is equipped with a true understanding of their environment and the states within, rather than producing this knowledge via camera imagery. This avoids the requirement to address the complicated possibilities of deliberate skewed observation to avoid penalty or obtain undeserved reward.
- State features may or may not be required to be interacted with as part of the primary objective. State features that are directly related to the primary objective; ie. the presence of rubbish on the floor, are not associated with a penalty for a state change. If

all features of the environment belong to a set *S*, then these features belong to a feature set S_P. All other features $S_A = S - S_P$ are associated with a penalty for a state change, to disincentivise the agent from unnecessary environmental impacts. All features not explicitly defined as within S_P are automatically included within S_A.

- All penalties enacted due to state changes upon features in S_A may be revoked if the state changes are undone. Penalties are derived from the original state of the feature not from any other previous state. Thus, the reward function for the impact minimisation objective can be described as $R_t^A = \phi(s_t^A) - \phi(s_{t-1}^A)$ where $\phi(s_t^A)$ is the potential function, given as $\phi(s_t^A) = -D(s_t^A, s_0^A)$ and D is some distance measure [77]. For this scenario, this distance measure will be Manhattan distance.
- The agent will use a lexicographical ordering selection process, subjected to minimum thresholds, to select a policy to optimise against the various objectives.

The following additional assumption has been made:

- State features included within S_A are not not represented as a singular auxiliary objective *A* but rather multiple auxiliary objectives corresponding to each mutable feature $\{A_1,...,A_i\}$ for the *i* mutable state features not included within S_P.

3.1.5 Determination of Fault

The process for the determination of fault includes recognition of an expression of discontentment from the user and discernment of potentially harmful recent actions. When a human experiences harm, they may express this through emotions such as anger or sadness [43] or through other modes of expression such as verbalisation or altered patterns of behaviour. This does not confirm that harm has occurred, but is an indication that it may have. A real-life apologetic system will involve the employment of sensory capabilities to detect this discontentment in the user. This was out of scope for this implementation and instead a simple simulated system has been used based upon assumptions discussed in the methodology (Section 3.2).

Determination of candidature for harm is a separate process to identifying the user's reaction. For each possible reason that the agent may become upset, an auxiliary objective should be

defined. For each objective, a candidature state must be defined or the agent may be given guidelines for it to be defined. This state would correspond to an undesirable outcome with respect to that objective. These two components may then be combined using an application of logic as required by the implementation; potentially though correlation, proximity or prior experience.

3.1.6 Policy Selection and Thresholding

To switch between prioritisation of separable objectives, a distinctly non-linear multi-objective action selection approach is required. The change in behaviour following an apology is defined by alterations to hyper-parameters used in this policy selection that determine the prioritisation of various objectives. Thresholded Lexicographical Ordering (TLO), introduced by Gabor et al. [31] for MORL and used by Vamplew et al. [77] for impact-minimising agents, is one such parameterised, non-linear multi-objective action selection approach. A TLO approach between two objectives will select an action that maximises the value of the second objective, subject to having reached the threshold specified for the first objective (Eq. 2) [31, 77]. Any policy selected via this operator will be pareto-optimal, subject to the agent's knowledge of the environment.

$$
\begin{aligned}
\forall s, a, a' \quad \vec{U}(s,a) >_{TLO^1} \vec{U}(s,a') \\
\Leftrightarrow \quad & \big(min(U_1(s,a), T_1) > min(U_1(s,a'), T_1)\big) \\
\vee & \Big(\big(min(U_1(s,a), T_1) = min(U_1(s,a'), T_1)\big) \wedge \big(U_2(s,a) > U_2(s,a')\big)\Big) \\
\vee & \Big(\big(min(U_1(s,a), T_1) = min(U_1(s,a'), T_1)\big) \wedge \big(U_2(s,a) = U_2(s,a')\big) \wedge \big(U_1(s,a) > U_1(s,a')\big)\Big)
\end{aligned}
\tag{2}
$$

In the impact minimisation experiments, two key agents were considered; the *SafetyFirst* agent used a TLO_A approach, and the *Satisficing* agent used TLO_P [77]. If this approach is extended to apply thresholding against both objectives, then a minimum performance as described by this threshold will be sought for each prior to maximising against either. Thus, the agent must prioritise any objective that has not yet satisfied its threshold before prioritising any

remaining objectives. This was also explored in the impact minimisation experiments, as a TLO_{PA} agent was introduced to address avoidance behaviours of the *SafetyFirst* agent where the agent ignores the primary objective if the auxiliary threshold cannot or is at threat of not being satisfied [77]. Dynamic alteration of these threshold values allows for manipulation of the prioritisation of objectives, such that an objective that penalises an undesirable outcome for the user can be given an increased priority so that behaviour is subsequently avoided.

The TLO process may also be extended to multiple auxiliary objectives, to require a specified minimum performance against all objectives prior to unbounded maximisation. Such an extension is proposed within the IM paper, but an approach was not proposed [77]. A default prioritisation order defines consistent preferential selection between equivalently thresholded objectives in exchange for a slight simplification. This resolves to TLO^{PMI}, as defined in Eq. 3. In this context, the superscript *PMI* is in reference to the prioritised multi-impact approach. The following shorthand notation has been introduced for readability: $U_i(s,a) \rightarrow U_i$, $U_i(s,a') \rightarrow U'_i$, and $min(U_i(s,a), T_i) \rightarrow \tau_i$.

$$
\begin{aligned}
\forall s, a, a' \vec{U}(s,a) >_{TLO^{PMI}} \vec{U}(s,a') \\
&\Leftrightarrow \left(\tau_1 > \tau'_1\right) \\
&\vee \left(\left(\tau_1 = \tau'_1\right) \wedge \left(\tau_2 > \tau'_2\right)\right) \\
&\vee \left(\left(\tau_1 = \tau'_1\right) \wedge \left(\tau_2 = \tau'_2\right) \wedge \left(\tau_3 > \tau'_3\right)\right) \\
&\vee \left(\left(\tau_1 = \tau'_1\right) \wedge \left(\tau_2 = \tau'_2\right) \wedge \left(\tau_3 = \tau'_3\right) \wedge \left(U_1 > U'_1\right)\right) \\
&\vee \left(\left(\tau_1 = \tau'_1\right) \wedge \left(\tau_2 = \tau'_2\right) \wedge \left(\tau_3 = \tau'_3\right) \wedge \left(U_1 = U'_1\right) \wedge \left(U_2 > U'_2\right)\right) \\
&\vee \left(\left(\tau_1 = \tau'_1\right) \wedge \left(\tau_2 = \tau'_2\right) \wedge \left(\tau_3 = \tau'_3\right) \wedge \left(U_1 = U'_1\right) \wedge \left(U_2 = U'_2\right) \wedge \left(U_3 > U'_3\right)\right)
\end{aligned}
\tag{3}
$$

In written terms, this equation seeks to maximise each objective until the threshold value is achieved, following a prioritisation order of U_1, U_2, then U_3. If each threshold is achieved, then there will be unbounded maximisation of the variables following this same prioritisation, with improvements against subsequent objectives as tie-breakers. Thresholding can be "switched off" for a specific objective with a threshold value below the minimum possible reward. This

causes the thresholding condition for that objective to be always satisfied and thus is silent in Eq. 3 above. Once the remaining threshold values are satisfied, the objective is revisited for maximisation, with respect to any higher priority objectives. This *PMI* approach provides the natural means to manage objective prioritisation via dynamically-specified thresholds, to facilitate constrained optimisation.

When using a threshold-adjustment approach to select for optimisation against conflicting objectives, inter-dependencies between the objectives also require consideration. For example, a primary objective that incurs a time-step penalty will exert a selection pressure between two auxiliary objectives if satisfaction of one requires a greater number of actions than satisfaction of the other. If this selection pressure is not intended to overwhelm thresholding prioritisation between these objectives, then the maximal threshold specified for the primary objective must be sufficiently lax as to be able to be met with satisfaction of either auxiliary. If the time sensitive objective is thresholded in this manner, it will only exert selection pressure between the auxiliary objectives post-thresholding if the auxiliary objectives remain otherwise equivalent.

3.1.7 Demonstrative Problems

A demonstration of apology in MORL requires a problem that an agent is unable to perfectly solve. Previous benchmark environments posed in AI Safety allow for a solution that is entirely 'safe'. If the agent were to learn this solution, it would have no candidate for which to apologise, or otherwise no alternative preferential behaviour to select. As such, we propose that this apology framework is best applied to a conflicted environment that cannot be fully solved. To complete its task in such an environment, the agent must learn policies that satisfy any combination of objectives to their fullest extent, as specified by the threshold values and prioritisation order. All objectives cannot be satisfied simultaneously. Thus, an apologetic approach is required to determine which, if any, objective can be ignored based on user preference, without causing harm or offense. One such environment has been used in this implementation.

3.2 Methodology

Implementation of the apologetic agent has been demonstrated through an extension of MORL impact minimisation. As no prior work exists demonstrating an apologetic approach, there exists no benchmark against which to compare. This agent's behaviour in context of improved user alignment, and thus the effectiveness of this approach, have been quantified by comparison to pre-apologetic behaviours.

The problem environment required multiple conflicting objectives to demonstrate selection between the based upon user preference (Section 3.1.7). The problem environment was modeled after a domestic living room: a discrete and otherwise static grid-world consisting of an assortment of obstacles. The agent is presented with a primary objective: collect the rubbish and return home (P). Impact of features of the environment not related to this collection of rubbish should be avoided, and this impact is quantified in auxiliary objective(s) (Section 3.1.4).

The initial proposed implementation of this environment (Section 4.1) considered a singular auxiliary objective, but this approach allowed for a 'safe' solution and thus was insufficient for the problem, according to the requirements discussed in Section 3.1.7. In the final implementation, the agent must manage multiple auxiliary objectives: avoid leaving the table displaced (A_1) and avoid running over the cat's tail (A_2). Penalties against these objectives are rewarded when the agent moves into the respective locations, however the table can be moved back into place to revoke the penalty.

Two environment configurations were proposed, described in Figure 4. These both represent non-trivial scenarios wherein the agent is unable to satisfy both auxiliary objectives at once, whilst still completing the primary objective of collecting the rubbish. Each environment maintains a discrepancy between the auxiliary objectives in problem complexity and time to complete, resulting in a selection bias. These environments are complementary in that the direction of this bias against the auxiliary objectives is exchanged between the two environment configurations.

The agent was a low-impact MORL agent, that used the TLO_{PMI} (Eq. 3) action selection process. The primary objective, P, takes values in the interval [-999, 50], consisting of a $+50$ reward for completing the task and a -1 time-step penalty for each action required. The auxiliary objectives are each 0 unless a -50 penalty is evoked when the associated

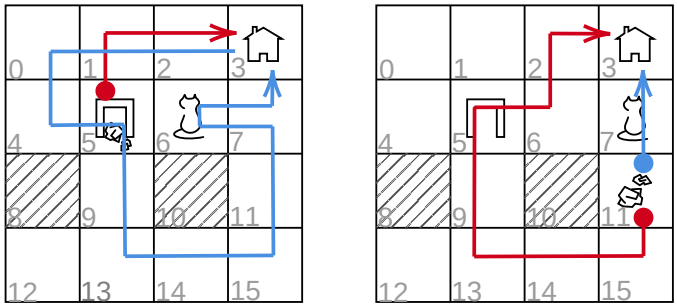

Figure 4: The 'Living Room with Table and Cat' gridworld environment in the positions A and B represent two non-trivial environment configurations against which the apologetic agent was tested. In both environments, the agent is unable to satisfy both auxiliary objectives. To complete the task, the agent must select between these objectives, resulting in two distinct solutions. The optimal policy for prioritisation of the table (A_1) and the cat (A_2) objectives are given by the blue and red paths, respectively. The two configurations are of interest as the task complexity and selection pressure exerted on the auxiliary objectives by the primary objective (P) are swapped.

Identity	Thresholds (P, A_1, A_2)	Identity	Thresholds (P, A_1, A_2)
Index 0	$\vec{T}_0$ = (35, 0, 0)	Index 4	$\vec{T}_4$ = (-1000, 0, 0)
Index 1	$\vec{T}_1$ = (35, 0, -50)	Index 5	$\vec{T}_5$ = (-1000, 0, -50)
Index 2	$\vec{T}_2$ = (35, -50, 0)	Index 6	$\vec{T}_6$ = (-1000, -50, 0)
Index 3	$\vec{T}_3$ = (35, -50, -50)	Index 7	$\vec{T}_7$ = (-1000, -50, -50)

Table 1: Eight threshold configurations representing each possible max-min combination of three objectives: P, A_1, A_2. These threshold sets are used to describe both learning scenarios and post-learning behaviour analysis in the subsequent experiments.

impact occurs. The representative threshold set, consisting of eight combinations of maximum/minimum threshold values is given in Table 1. The maximum threshold value for P was given as 35, to correspond with the 15 step minimum required to satisfy the more complex auxiliary objective in each scenario, as can be seen in Figure 4.

The experiment consisted of two phases. In the first phase, the agent was trained in a traditional, non-apologetic scenario according to a pre-determined threshold protocol to establish the set of Pareto-dominant policies. This corresponds to the *learning phase* of the apologetic MORL process as given in Figure 3. A representative set of thresholds consisting of all possible prioritisation configurations between the three objectives (Table 1) guided the agent to learn a best-effort representation of the Pareto front. In the second phase, the agent was tested in an apologetic scenario against four configurations of user. This corresponds to the *apologetic phase* also described in Figure 3. During this phase, exploration and Q-table updates were disabled and the agent's behaviour was altered exclusively by the changes made to the

thresholds following provision of an apology. These two phases are discussed in detail in the sections below.

3.2.1 Learning Phase

The learning phase of the experiment involved the preparation of the agent for sufficient knowledge to switch between pre-known policies in response to user preference. The agent was required to know an appropriate policy that best meet the needs of the user, for all possible configurations of user. For a binary specification of user preferences in a three objective environment, this is similarly represented by the eight combinations of min-max thresholds given by Table 1.

The agent was exposed to eighteen learning protocols based upon these threshold configurations and subsequently tested against them to provide a benchmark for the agent's unapologetic behaviour. Protocols S0-S7 describe a traditional agent learning process with a single specified threshold and 4000 episodes, corresponding to the threshold configuration indexes 0-7. Protocols A0-A9 describe aggregate approaches for learning, as follows:

- A0-A3 uses a S0 base with subsequent learning through exposure to each threshold configuration for 0, 10, 100 and 1000 episodes.
- A4-A7 uses the same subsequent learning through exposure approach with an aggregate base produced by averaging Q-values for each state-action pair across S0 to S7.
- Finally, A8 and A9 were created with episode-alternating thresholds for 4000 and 8000 episodes, respectively.

Each learning protocol was run for 10 fully independent trials and results are reported as an average between them.

The crucial outcome for this experiment was a trained MORL agent that conforms to the expectations of the threshold configurations. That is, if the user preferences corresponded to a given threshold configuration, the agent could be provided with this information and select a policy that satisfies these preferences. For this reason, the criteria for selection from among these learning protocols was minimisation of shortfall of the final reward obtained from

the reward threshold specified. It is important to note that the shortfall cannot be zero for threshold configuration $\vec{T_0}$ = (35, 0, 0) as no such solution exists, and will always be zero for $\vec{T_7}$ = (-1000, -50, -50), are these are the minimum achievable rewards for every objective. For the remaining threshold configurations $\vec{T_1}$ - $\vec{T_6}$, shortfall arises from an in-optimal policy. Final reward that is greater than the specified threshold values is not of interest, as this indicates good performance in an objective that the user does not care about. For this reason, the value function generation protocol was primarily evaluated with respect to the cumulative shortfall values.

3.2.2 Apologetic Phase

The apologetic phase utilised the best performing agent from the learning phase (Section 3.2.1) as the basis for the apologetic agent. The apologetic experiment involved introduction of a simulated user, and implementation of the apologetic cycle to *assess* and *apologise* to this user as needed.

The user was represented by a live updated *attitude* that described whether or not the user was upset, and a *sensitivity* corresponding to each objective. The user's attitude was updated in response to environmental changes, and could be retrieved by the agent when required. The response given by the user was determined by the sensitivity, which is a vector of boolean values. The agent has no visibility of the user's sensitivity.

Four configurations of user sensitivities were considered; each auxiliary objective alone (denoted as A_1, A_2), both auxiliaries (A_1+A_2), and none. A reactive state condition has been described for each objective: an episode length greater than 50, displaced table or disturbed cat. Algorithm 1 describes the heuristic approach through which the simulated user responds to any of these changes in the environment. In this implementation, the user's attitude is fully observable by the agent, thus demonstrating the capabilities of this framework given perfect predictive ability. The user is exclusively reactive to the presence of impacts compatible with its sensitivities, without interference with external stimuli. As a result, this system does not give opportunity for false-positive errors. The 'none' user mimics the pre-apologetic results, as this implementation involves no interpretation error and the apology sequence is only activated if the user has become upset, thus this configuration evokes no apology-driven behaviour changes.

Algorithm 1 Simulation of a user reacting to changes in the environment; an intermediary for the apologetic framework

require Environment State s_t
ensure $attitude$
Initialise $Sensitivity[]$, $StateConditions[]$ *conditions for undesirable state*
for each episode **do**
  Initialise $attitude \leftarrow 0$ *set as neutral*
  Initialise $justification \leftarrow -1$ *set as negative*
  repeat
    Given $a_t - 1$, receive current state s_t
    for each objective: i **do**
      if($Sensitivity_i =$ **true and** $s_{t,i}$ satisfies $StateConditions_i$ **and** $attitude = 0$) **then**
        $attitude \leftarrow -1$ *set as negative*
        $justification \leftarrow i$ *set as index*
      end if
    end for
  until s_t is terminal (goal or max t)
end for

Given the user has a negative attitude, the agent must consider its recent actions to identify any candidates for offence. In this implementation, the undesirable state for each objective is defined and candidature is determined if an objective is in that state and has recently transitioned to that state. Self-blame against this objective is assigned where this candidature corresponds with a step in which the user has become upset. Thus, this implementation represents a minimalist and somewhat under-nuanced approach to determination of blame, as a baseline for future enhancement. Algorithm 2 describes the agent's assessment and apologetic process. The justification refers to the knowledge or belief held by the user or the agent, regarding the reason the user is upset. Its value corresponds with the index for the blamed objective, or -1 if the user is not upset, or upset for a reason other than any of the agent's objectives.

Once the misaligned objective has been identified, the apology may be constructed. This implementation has focused on the behaviour correction rather than generation of the explanation, and so has used a templated approach. This is described in algorithm 3. The agent is restricted against apologising more than once per episode, as the apology does not remove the offensive state. That is, an apology does not undo a mistake, but rather promises not to repeat it in future.

The key components of this algorithm are aligned with the *affirmation*, *affect* and *action* components identified in Section 3.1.2. Previously established priorities consist of those for

Algorithm 2 Apologetic framework applied to Multi-Objective Reinforcement Learning for policy realignment

Initialise $\vec{T}$, $\vec{\delta}$
Load $Q(s, a)$
Given $T_{max,j}$ and $T_{min,j}$ as the maximum and minimum threshold values specified for objective j
For each episode **do**
  Initialise *apologised*← **false**, s_t *the agent has not yet apologised*
  Initialise *prioritised*← [**false**, **false**, **false**] *thus no priorities have been established*
  repeat
    Choose an action a_t according to $Q(s, a)$ w.r.t $\vec{T}$
      Take action a_t
    Observe reward $\vec{r}_{t+1}$ and next state s_{t+1}
    Update *attitude* via **Algorithm 1**
    Observe *attitude*
    if (*attitude*< 0 **and** *apologised* $=$ **false**) **then** *user is upset*
      if $(min(\vec{r}_{t+1}) < 0)$ **then** *determine candidature*
        justification ←Index$(min(R))$
        Agent 'apologises' w.r.t *justification*
        apologised ← **true**
        prioritised$_{justification}$ ← **true**
        for each Threshold: j in $\vec{T}$ **do** *set thresholds to max for all prioritised objectives*
          if *prioritised*$_j$= **true then**
            $T_j \leftarrow T_{max,j}$
          else
            $T_j \leftarrow T_{min,j}$
          end if
        end for
      end if
    end if
  until s_t is terminal (goal or max t)
end for

which the agent has previously apologised, and are not overwritten for subsequent apologies. If the agent is apologising for an objective that has already been prioritised, the agent articulates that they are unable to further improve that behaviour. This apology provides a concise but articulate overview of the recognised harm and the subsequent behaviour alteration.

The apologetic experiment consisted of 10 independent trials for each of eight initial threshold configurations corresponding to the eight representative configurations (Table 1). Each trial consisted of three stages of 10 episodes each. In all three stages, the agent's exploration and Q-table updates were disabled, and the agent referenced the same final Q-table described in the learning phase (Section 3.2.1). The first and final stages consisted of a traditional offline RL scenario that demonstrated the agent's behaviour before and after the apologetic framework

Algorithm 3 Template for generation of an apology, given the agent's knowledge of a derived justification and previously established priorities.

require *Justification*, *ExistingPriorities[]* *those previously established*
Affirmation ← "I recognise that you are upset. I believe that it is due to my recent behaviour, where I [failed *Justification*]."
Affect ← "I would like to apologise for how this behaviour has upset you."
if *Justification* $\notin$ *ExistingPriorities[]* **then**
    Action ← "To avoid this in future, I will now select a policy to prioritise [*Justification* + *ExistingPriorities[]*]."
else
    Action ← "Unfortunately, I have already maximised my prioritisation of [*Justification*] and it seems I am unable to avoid this behaviour with my existing knowledge and resources."
end if
Apology = *Affirmation* + *Affect* + *Action*

was applied. In the intermediate stage, the apologetic framework was enacted and the agent apologised as according to Algorithms 1, 2 and 3. This experiment was undertaken for each of four user configurations across the two environments, and results were averaged between the 10 trials.

The aim of this experiment was to determine apologetic accuracy and post-apology behavioural alignment successes. For each apology-enabled episode in each of these trials, the agent's and the user's justification were recorded. The apologetic accuracy was then determined as the percentage of cases where the agent's (predicted) justification is equal to the user's (known) justification.

Reporting the change in behaviour involved the distillation of final rewards to binary results for each objective, corresponding to a satisfied (non-minimum value) or unsatisfied (minimum value) objective results. Pre-apology and post-apology final rewards were recorded for every trial and presented as an average across each user and initial threshold configuration, as well as aggregate across all trials with that user, for each separate environment. The pre-apology behaviour for every user was exclusively described by the behaviour associated with the initial threshold configuration applied. As previously mentioned, the 'none' user preserves the initial behaviour as they are nonreactive and thus do not elicit a change in behaviour. The proportions and change in proportions of satisfied outcomes post-apology for each user quantifies the agent's success in behaviour realignment. These results were also assessed for statistical significance using a McNemar test for differences in proportionality [52].

3.2.3 Tools and Technologies

The algorithm described in the previous section was implemented using the MORL-Glue software package [78]. The program environment consists of a central 'glue' component, with separate environment, agent and experiment programs ported into it (Figure 5). The Glue itself is compiled in C++, however the environment, agent and experiment programs are each implemented as Java classes. The implementation of this simulation and subsequent testing is the full scope of the algorithm produced within this thesis.

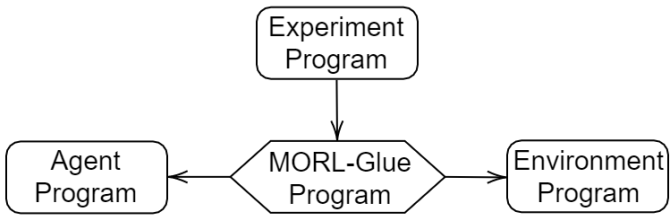

Figure 5: The MORL-Glue program environment consists of the central program and interfaced peripherals that are independently or dependently designed to suit the Multi-Objective problem posed. These peripherals include the agent, the environment and the experiment, which each communicate via the Glue.

The algorithm was developed using the Intellij IDK [39] and Java SE Development kit 8 [39].

3.3 Project Risks

Risk	Severity	Likelihood	Mitigation Strategy
Computation power on local PC may be insufficient to train image recognition models.*	Minor	Unlikely	Student will assess the performance limitations and either upgrade the relevant computer part or use a pre-trained model.
Computational power on local PC may be insufficient to process images for prediction, using a pre-trained model.*	Moderate	Rare	Student will assess the performance limitations and either upgrade the relevant computer part or simulate prediction using hard-coded probability model based on the known accuracy of existing models.
The agent simulated in the proposed environment may be unable to learn the expected behaviours.	Moderate	Unlikely	Likelihood of this risk is low as prior research has indicated that this is possible for simpler but similar environments. If agent is unable to learn the appropriate behaviours, the student may simplify the problem such that the agent is able to complete the task.
The agent simulated in the proposed environment may not sufficiently prioritize completion of the task or presentation of an apology and may instead develop a model of avoidance behaviours, which may undermine the premise of the project.	Moderate	Possible	If agent is unable to learn the appropriate behaviours, the student may simplify the problem until the point where the robot is able to complete the task.

Continues on next page

The agent may be unable to successfully distinguish at-fault behaviours sufficiently to provide an apology that would support the premise of this research.	Major	Possible	the student may reassess the cognitive processes leading up to the construction of the apology to determine at what point this breaks down, and thus may determine alternative approaches or write a thesis with negative result.
The limited time-frame of the thesis may be insufficient for the student to learn key multi-objective reinforcement learning and explainable reinforcement learning concepts and practically apply them to build a functional AI agent.	Major	Unlikely	the student may access external resources to build and rectify lacking knowledge, may reduce the scope of alternative aspects of the project to allow for greater time for upskilling, or else may reduce the scope of the problem to better reflect their capabilities.
The student's Java or broader programming knowledge may be insufficient and prohibitive to completing the required coding aspects and producing a functional system.	Major	Unlikely	The student may access external resources to build and rectify lacking knowledge, or else may reduce the scope of the problem to better reflect their capabilities.
The MORL benchmark suite created by Vamplew et al. may not be appropriate for the purpose described and may not supply the solid starting point required for this project, thus increasing the scope and complexity of the project if required to start from scratch.	Moderate	Unlikely	The student may request assistance from the creator to address these concerns or else the student may alter the scope and research question addressed in the thesis to create an applicable algorithm from scratch.

*Risks associated with image recognition requirements were not relevant for the final outcomes of the project as this consideration was removed from scope.

4 Artefact Development Approach

This section discusses the approach used to address the aims of this project and produce the implementation discussed in Section 3.1. This includes the development of preliminary artefacts that drove discussion, and justification of the design decisions that preceded the final artefact design and evaluation. The final artefact designs were used as the basis for the experiments discussed in Section 3.2. These artefacts were produced within a series of three successive sprints. The development process was derived from agile principles of iterative implementation incorporating regularly revisiting project requirements, communicating these requirements through an explicit set of acceptance criteria. The cycle has been visualised in Figure 6.

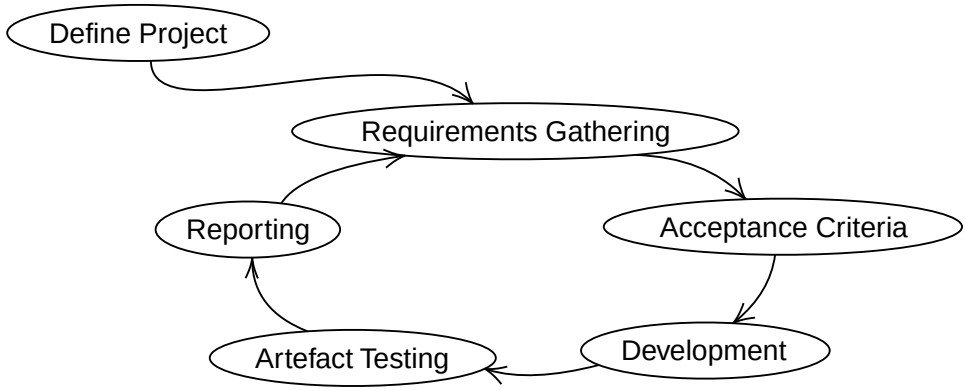

Figure 6: The artefact development cycle implemented an agile approach of iterative gathering and specification of requirements, development and testing. At the conclusion of each cycle, the knowledge and progress gained feeds into the following cycle for continuous improvement.

Section 3.1 has provided a project definition. The subsequent development stages are described as follows.

1. Requirements gathering: Define and propose an aspect of the project as a singular artefact. Undertake further research to assess the work required and refine scope.
2. Acceptance criteria: Using a Given/When/Then statement framework, produce a robust set of acceptance criteria from which the artefact will be built and assessed.
3. Development: Produce the artefact as according to the specifications given in the acceptance criteria.
4. Artefact Testing: Test the artefact produced against the acceptance criteria. Verify that the artefact is meeting these requirements. Ensure that newly introduced artefacts do not interfere with the existing system.

5. Reporting: Assess and reflect upon the outcomes of the development cycle and the system's ability to respond to the research questions. Declare any newly obtained knowledge and propose adjustments to the project requirements to ensure alignment with the project requirements.

Each cycle corresponds to a sprint, consisting of a 3-week block of time within which development was undertaken with respect to a sprint goal. A short overview of the goals of each sprint are as follows.

- **Sprint 1: Minimum Viable Artefact** (Section 4.1).

  A simple multi-objective environment was produced, and tested using the low-impact *SafetyFirst* and *Satisficing* agents.

- **Sprint 2: Multi-Impact Environment** (Section 4.2).

  This environment was expanded to include an additional auxiliary objective and a new agent was introduced to manage the prioritisation of these two auxiliaries with an impact minimisation approach.

- **Sprint 3: Apologetic Change in Behaviour** (Section 4.3).

  The Act-Assess-Apologise framework was implemented through inclusion of a user within the environment and an apologetic agent.

The outcomes of these sprints are discussed in the sections below.

4.1 Sprint 1: Minimum Viable Artefact

The minimum viable artefact (MVA) was the recreation of a simple impact minimisation reinforcement learning environment, wherein an existing agent may complete a required task. The environment consists of a primary objective and a single auxiliary impact minimisation objective. The environment was assessed using the low-impact SafetyFirst and Satisficing agents of the impact minimisation literature [77] as a benchmark.

The environment was produced in Java [61], using the MORL-Glue program [78] and uses the impact minimisation experiments based on this program [77] as a starting point.

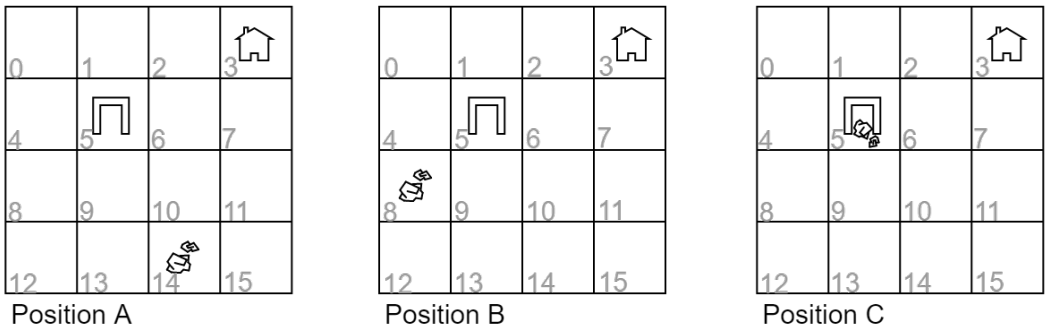

Figure 7: The Living-Room-With-Table environment represents a simple 4x4 grid-world, with a single table obstacle and a single piece of rubbish. Three configurations are posed, in which the agent must maneuver about the table or else learn to replace it once it has been moved, to avoid a negative impact reward.

The environment was a model of the living room scenario discussed in the motivating example, represented by an enclosed 4x4 grid (Figure 7). It contains only a single table and a single piece of rubbish. The environment defines how the agent is able to interact with it, through the communication of states and provision of rewards. The agent exists within one of the spaces in the grid, as does the table, rubbish, and a home location. The home location is where the agent begins and ends the episode, and the agent returns to this location to dispose of carried rubbish. The environment allows for four interactions associated with moving to an adjacent position in the grid, through one of four possible movement actions (up, down, left, right). The agent is able to move the table if the agent's movement ends in a location in which the table is currently, pushing the table into an adjacent space. The actions of picking up the piece of rubbish and putting the rubbish in the bin are automatic, and occur if the agent is in the correct location.

The environment is potential-based and multi-objective; with a primary objective P associated with putting the rubbish in the bin and an auxiliary objective A associated with moving the table. The agent receives a reward of $+50$ for completion of the primary goal, upon return to the home location. Furthermore, the primary objective receives a time-step penalty of -1 for each action undertaken during an episode. The auxiliary objective is potential-based such that the agent receives a reward of -50 corresponding to displacement of the table, that is revoked if the table is returned to its original location.

The motivation for creation of this artefact was to produce a unique environment for

exploration of impact management and thus implementation of subsequent project requirements. The environment is designed to suit scenario described in the motivating example, in addition to being sufficiently flexible to support multiple configurations with minor adjustments to allow for adjustment of the problem space. The environment was evaluated using the established low-impact agents to determine whether these agents are capable of learning an optimal solution to this environment. This result can be determined by the number of steps required to complete the task, and whether the table remains displaced at the end of the episode; both during the online learning process, and the offline final optimal behaviour solution. It was expected that the SafetyFirst agent will be impacted negatively with time to achieve the goal but will cause less impact to the environment, in comparison to the Satisficing agent.

The acceptance criteria can be found in Appendix B.

4.1.1 Report

Three configurations of the MVA environment were considered; a simple problem in which the rubbish is located in an open and easily accessible space in the room (Figure 7 and Figure 8, configuration A), a slightly more complex problem wherein the table lies in the direct path to the rubbish (Figure 7 and Figure 8, configuration B), and a most complex problem in which the rubbish is located beneath the table (Figure 7 and Figure 8, configuration C). In this latest scenario, the table must be moved to access the rubbish, but the table must be returned to its original location to remove the displacement penalty. The SafetyFirst agent prioritized impact-minimisation actions most highly, and only wishes to select the primary objective beneficial actions that are impact-neutral. The Satisficing agent prioritizes rubbish collection primarily, selecting for impact minimisation among the fastest rubbish collection solutions.

The two former problems were demonstratively simple to solve, and both agents appeared to find an optimal solution with relatively similar speed. Completing the episode provides the agent with a reward of 50. The agent is incentivised to complete the episode in as few actions as possible by a penalty of 1 for each action that does not result in completion of the episode. The average rewards of the online and offline learning cycles are shown in Figure 4.1.1. Across the four trials completed, this demonstrates that the SafetyFirst agent is slightly more effective at finding the optimal pathway, as it was able to find the optimal path each time. However,

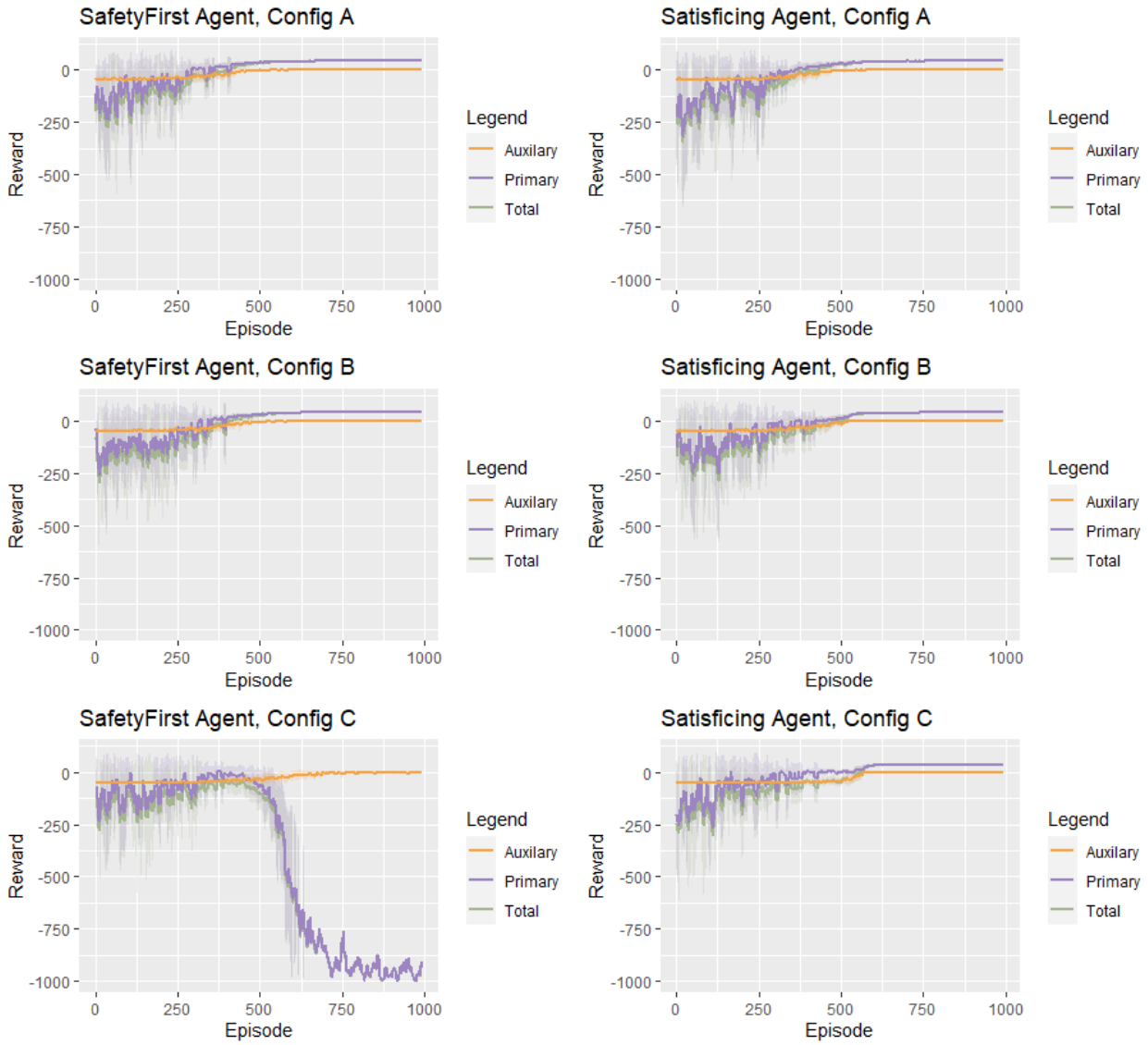

Figure 8: In each of three configurations of the Single-Impact Living-Room-With-Table environment, the Satisficing and SafetyFirst agents demonstrate learning performance behaviours, usually converging within 1000 episodes.

both agents are able to converge consistently on a close to optimal pathway.

Configuration C demonstrated the difference between the two agents more distinctly. The SafetyFirst agent learned the penalty associated with moving the table and sought to avoid it, as in the previous scenario. However, in the latter scenario, the only way to achieve the goal and complete the episode prior to reaching the 1000 action limit was to move the table. In this goal, the agent suffered and optimised to a solution that involved the table being in place at the conclusion of the episode, having never completing the task. Due to the limitations of the data collected, it is not possible to know whether the agent had interacted with the table during the episode, or the pathway that the agent took. This has identified an improvement in

Agent	Environment	Online			Offline		
		R^P	R^A	R^*	R^P	R^A	R^*
SafetyFirst	A	-8.64	-16.45	-23.79	43	0	43
	B	-11.80	-17.79	-28.15	42	0	42
	C	-414.39	-27.6	-440.51	-999	0	-999
Satisficing	A	-14.16	-17.075	-29.84	42.5	0	42.5
	B	-15.10	-19.39	-30.10	40.50	0	40.50
	C	-15.63	-26.36	-40.60	38.25	0	38.25

Table 2: Within the Single-Impact environment, this table lists the mean results for online (exploratory) and offline (greedy) episodes of the Satisficing and SafetyFirst Agents against the three room configurations.

the data collection wherein the final resting place of the table, and some manner of determining the pathway of the agent should be collected alongside the final results data. This information would be valuable for identifying the reasons for this strange behaviour of the SafetyFirst agent. It would also be useful for determining other particularities such as whether the agents in configuration B prefer a pathway that greater avoids the possibility of bumping the table through stochasticity.

Computation time was also of interest. Despite the earlier cutoff of each episode due to a substantially higher completion rate, the runtime for a 1000 episode trial averaged at 33 minutes for the Satisficing agent as opposed to 13 seconds for the SafetyFirst agent. Considerations may need to be made as the complexity of the problem increases to allow sufficient computational time for these agents. One potential solution to the blow-out of episode length may be to allow the agent to prematurely cease the episode, leaving the rubbish partially or entirely uncollected. However, this may cause the agent to optimise to immediate closure of the episode, thus minimising opportunity to acquire negative rewards. This would be problematic from the prospect of learning the apology aspect of this project.

This exploration demonstrated some limitations of this environment in context of intentions to implement the apologetic framework. Given the correct combination of opportunity to learn and an appropriate objective prioritisation approach, as demonstrated by the Satisficing agent, the agent is able to achieve both the primary goal and avoid causing impact to the environment. In such a scenario, the agent will have no misaligned behaviours for which to apologise. In the opposite scenario, if the agent does not learn either of these objectives, it would be unable to demonstrate satisfaction of this objective if it recognised that failing to achieve the objective was harmful.

This exploration has also demonstrated that it is important that the primary objective be prioritised so that the agent is motivated to explore the environment to complete the task. As demonstrated by the SafetyFirst agent, when the primary objective is of a lower priority than the auxiliary objective, the agent may demonstrate avoidance behaviour to the extend of never pursuing the primary objective. If the agent's first priority is to ensure that impact is minimised, the only way to ensure that this is the case is not to attempt the task at all.

The behaviour of these benchmark agents in the MVA environment provided the baseline of considerations and possible improvements for the project. These considerations were taken into account at later stages of development, as the problem space becomes more complex.

4.2 Sprint 2: Multi-Impact Environment

The second sprint goal was to improve the environment and create an agent that addressed the limitations of the approach demonstrated in the MVA. The premise of the apologetic agent requires that the problem facilitates a clearly visible change in priority to effectively demonstrate a behavioural adjustment. That is, the environment needs to present a solution that aligns with a policy defined by one set of thresholds, that does not align with a policy defined by a different set of thresholds. The agent must be required to make a decision, weighing up the significance of two different objectives, so that it is possible to assess whether the agent has been receptive to the user's reactions and appropriately adjusted its behaviour.

The agent changes its behaviour as part of an apology by making an adjustment to the thresholds that guide action selection. This approach relies upon a state-action value function that is threshold-independent, such that any set of thresholds can be held against it to obtain an appropriately optimised policy. The MVA environment posed a difficulty here, as it presented a problem with an outright solution. Independent of any thresholds, the best policy in the MVA environment is explicitly defined.

Thus, a multi-impact (MI) environment was introduced that presented two obstacles: the existing table, and an additional cat. The cat obstacle applies a singular irreversible penalty in the first instance where an agent or object enters its location – "running over the cat's tail". The agent now seeks to maximise three objectives, as follows: the primary rubbish collection objective, the table impact objective, and the cat impact objective. These are identified as P,

A_1 and A_2, respectively. In addition to this, there was a further introduction of two unenterable, unmovable grid locations representing navigation obstacles. Thus, the environment poses a configuration of these components as an unsolvable problem, in which the agent is not capable of completing the task while avoiding both impacts. This requires the agent to decide as to which objective to prioritise and presents easily defined prioritisation misalignments for which the agent can apologise.

Two MI environment configurations were considered (Figure 4) and are discussed in detail in the methodology (Section 3.2). Within this sprint, further reporting detail including step-wise agent action reporting was introduced to provide visibility to unexpected agent behaviours. An experiment regarding the state-space and runtime considerations was also undertaken.

The acceptance criteria can be found in Appendix B.

4.2.1 Report

The Satisficing agent that applies a threshold to the primary objective requires a state-space augmentation that is multiplicative with respect to the number of bins [77]. This is required to provide the agent with time-binned flags so that it is aware. In absence of these bins, the agent is not aware of how much time has passed. The number of bins used is a configurable parameter, dependent upon a balance between the influences of too many or too few bins.
If the agent's state-space is augmented with too many bins, the states associated with time steps less frequented may be underpopulated and so the agent will require many more training episodes to generate experience to inform action selection in these states. If the agent has too few bins, the final reward for many time-step episodes will be convoluted with the final reward for fewer time-step episodes. If the agent were to wander for 1000 steps in refusal of finishing the episode, this reward will be reflected in the Q-values of the state-action selections made and thus penalising them with respect to P, when there is no explicit reason that these selections should result in such a low reward.

In the original IM agents, 10 evenly-spaced bins were used to provide this state-augmentation. However, in initial explorations of the agent's learning behaviours it was observed that these additional states are rarely utilised, as most episodes were completed fairly quickly. Within a 1000-step episode, the primary reward can take any value within a range of $+45$ and -999.

The -999 result correlates to 1000 steps without completion of the task while all others indicate completion and thus inclusion of a +50 reward. However, frequency analysis of among some of those demonstrates that even among those that select a policy with a primary reward of -999, few episodes achieve a final reward in the interval (-150, -950) (Figure 9).

Figure 9: Histogram of compiled learning outcomes for the primary objective across a selection of the threshold sets

This finding is impactful regarding the time-augmented state space used against this objective, as the agent is not utilising this space meaningfully to produce different behaviours. After 200 time-steps, most cases demonstrated the agent waiting out the episode timer. Furthermore, most final rewards indicate completion within 100 time-steps. Given a linear 10-bin discretion, these all fall within the first time-bracket. Overall, this suggests that the state-space augmentation is being underutilised within this configuration and that it may be beneficial to alter the distribution to better suit the problem space. Reducing the size of the state-space in this manner may improve the runtime without impacting the agent's performance.

A small experiment was undertaken to confirm this hypothesis, in which linear discretisation will be replaced with a tailored one that uses the following discretisations: $R_P > 0$, $0 \geq R_P >$ -150, and $R_P \leq$ -150. The viability of the time-augmentation heuristic was assessed in terms of its ability to produce similar training results and reduction of run-time. The results of this exploration supported this conclusion, as the agent optimised to the same final solution and demonstrated a similar standard deviation between trials, in 10% of the time. These results suggested that for this problem there was negligible quality compromise to be made if making use of the more resource-efficient approach. This approach is represented in later experiments.

The representative threshold configurations (Table 1) were coined during this exploration of the MI environments, for demonstration of the breadth of possible outcomes. Figures 10 and 11 visualise the agent's learning results as an average of 10 trials. For ease of interpretation, the reward values were re-scaled to lie within the same interval and smoothing was applied. This graph demonstrates the patterns of convergence behaviour towards the final policies.

In Environment A (Figure 10), the problem is posed such that the agent has greater ease completing the task by leaving the table displaced, thus causing an impact against objective A_1. In this environment, the primary task cannot be completed without displacing the table, so completion of the primary task in conjunction with satisfaction of A_1 requires finding the solution that returns the table to its original location. The data suggests that when the threshold for P is 35, the agent's prioritisation of this objective is so absolute that it does not attempt exploration sufficiently to discover the solution that satisfies maximised thresholds for P and A_1 simultaneously, despite this solution being achievable. This is demonstrated by the poor optimisation of this objective for all threshold configurations except $\vec{T_4}$ = (-1000, 0, 0) and $\vec{T_5}$ = (-1000, 0, -50), in which the threshold for P is minimised. Unlike A_1, objective A_2 does not involve have implications for P and can simply be avoided by an alternative path of equal length. For this reason, it is interesting to note that the agent does not care to do this for any threshold configuration that prioritizes A_1, whether or not it finds a solution that satisfies this objective.

In Environment B (Figure 11), the problem difficulty introduces an experimental bias towards impacting A_2. This environment does not require the table to be displaced to access the rubbish except where the agent must avoid disturbing the cat. In a contrasting manner to Environment A, the agent converges towards a policy that disregards the A_2 objective for all threshold configurations except $\vec{T_6}$ = (-1000, -50, 0), in which only the A_2 objective is prioritised. It is interesting to note that on average it appears that the final policy for all environment and threshold configurations in these trials converges towards satisfaction of P.

4.3 Sprint 3: Apologetic change in Behaviour

The goals of the third sprint were to complete a functional prototype of an apologetic system. This involved generation and assessment of the viability of a singular value function to support

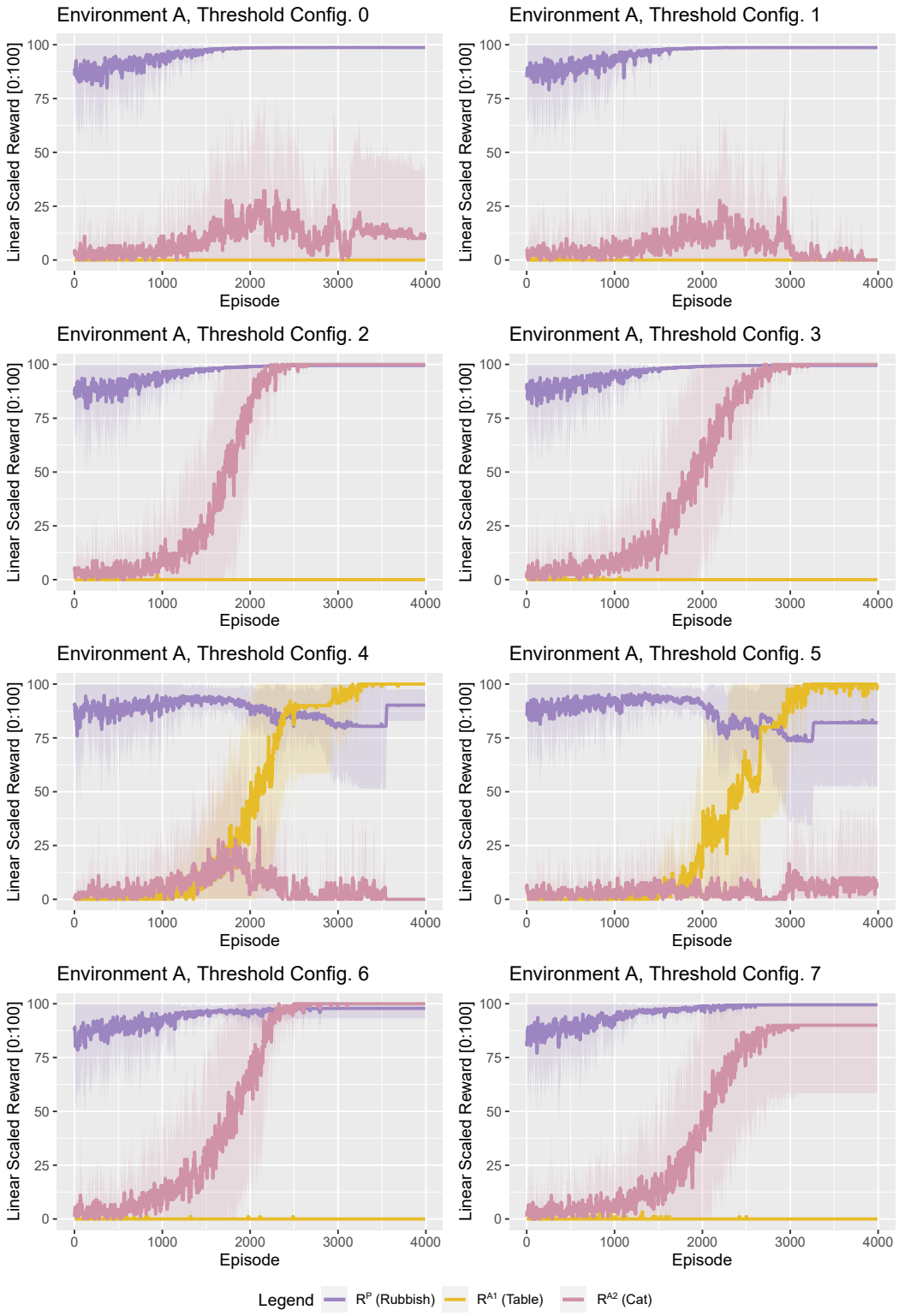

Figure 10: The agent was trained using each of the eight representative threshold configurations in the multi-impact Environment A, for 4000 episodes. In this graph, the rewards are re-scaled according to their maximum and minimum possible values to lie in the interval [0, 100]. The sum of rewards R* is not described in this graph, and a single standard deviation error margin is shown. Smoothing using a moving average with k = 9 was applied. This environment configuration is biased towards satisfaction of auxiliary objective A_2.

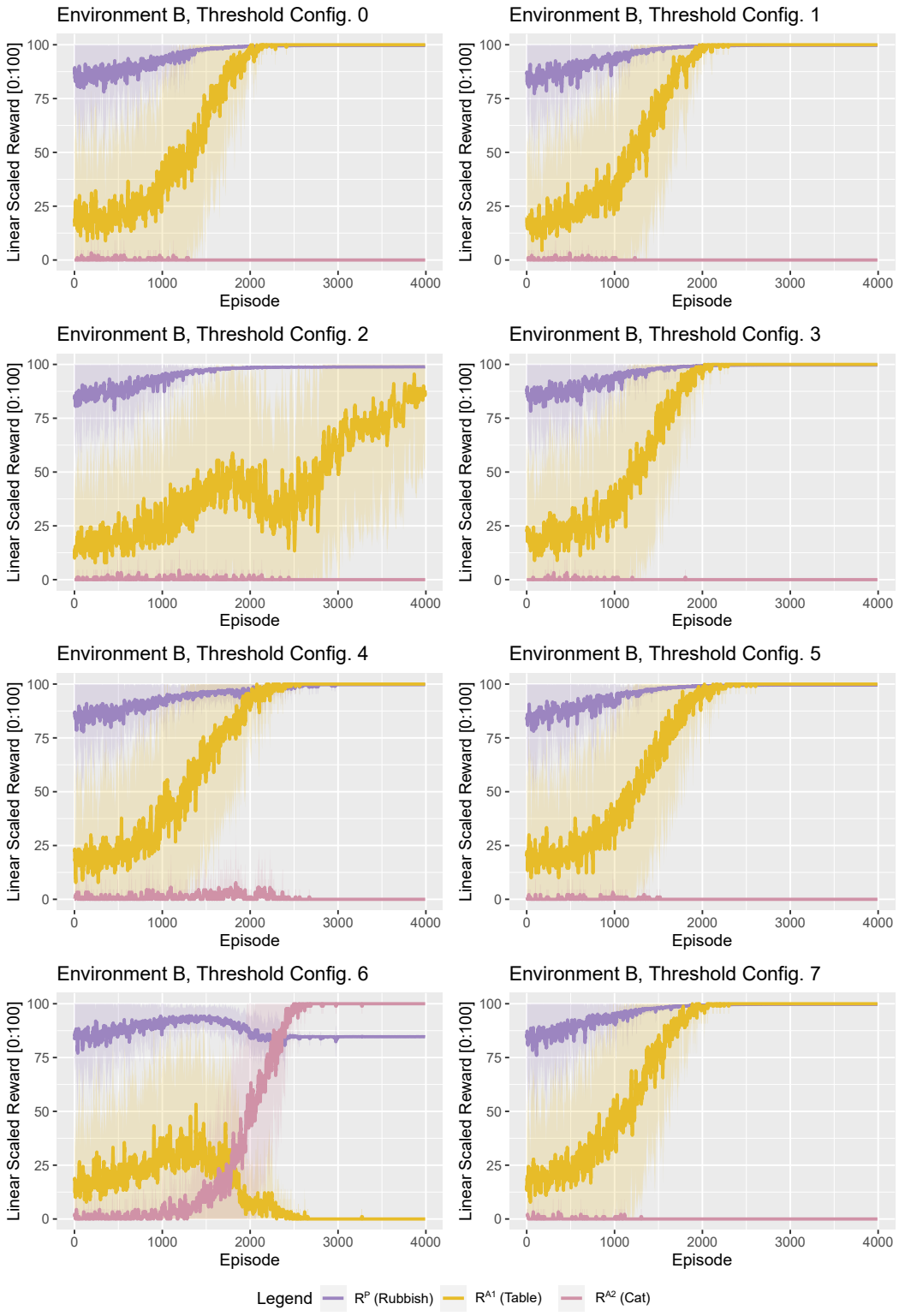

Figure 11: Similarly to Environment A, the agent was trained using each of the eight representative threshold configurations in the multi-impact environment B, for 4000 episodes. The rewards are re-scaled according to their maximum and minimum possible values to lie in the interval [0, 100]. The sum of rewards R* is not described in this graph, and a single standard deviation error margin is shown. Smoothing using a moving average with k = 9 was applied. This environment configuration is biased towards satisfaction of auxiliary objective A_1.

agent function in all representative thresholds, followed by implementation of the apologetic framework.

Sufficient exploration of the agent to learn other pareto-optimal policies, despite a different set of thresholds, was an item of concern for the project. The agent's change in behaviour relies upon prior exploration sufficient that the agent discovers all other potential optimal policies. In the literature, this is often reasonably achieved within a normal training sequence [34]. The additional concern in this context is that the complexity of the MI environment is sufficient that the agent often does not learn the true Pareto-optimal policies for more complex objectives. This is often observed in scenarios where a simple solution can be easily obtained and so the agent seeks to exploit this solution rather than explore to obtain other, potentially better, solutions [75].

As this was the final sprint of the project, the final approach and implementation is that which was presented in the design and methodology (Section 3). Considerations of this development was discussed in the report, however the final outcomes are presented with the final project discussion (Section 5), limitations (Section 6) and conclusion (Section 7).

The acceptance criteria can be found in Appendix B.

4.3.1 Report

In attempting the implementation of the apologetic framework, one considered approach involved a step-wise alteration to the thresholds. This approach allowed the thresholds to take values other than those that the agent was trained on, including values that did not correspond to meaningful thresholds for the auxiliary objectives. This approach was discarded in preference of that described in the Section 3.2 as the agent would become confused when presented with these unfamiliar thresholds and perform poorly in all objectives. In some initial threshold configurations, this demonstrated a cyclic behaviour with the agent as the user would become upset in each episode, and the erratic behaviour would result in alternating between the objective blamed. This behaviour was avoided by moving to the 'switch' scenario, in which the agent's thresholds are only ever selected from. It may be worthwhile undertaking further investigation of this behaviour for implications in research regarding thresholded-MORL approaches.

The initial apologetic framework intended for implementation involved a series of checks occurring between each action sequence within an episode. Once the user decides upon an action, the environment is updated in accordance with this action. As part of this update, the user reacts to the new state of the environment and updates an internally maintained variable *attitude*.

When the agent has confirmation that all required updates have been completed, which occurs just prior to determining the next action, the agent completes its assess and apologise sequence. The agent obtains knowledge of the user's attitude through observation, and then uses this attitude and its own knowledge of accumulated rewards to determine if harm has occurred and own fault. The agent is restricted to apologising once per episode, as neither subsequent actions nor the provision of the apology is able to immediately influence the user's attitude. In the preferential scenario, the agent will supply the apology to the user and the user would update their attitude reflect the outcome of the apology. The agent would then have their course of action confirmed before updating the thresholds, or would have their assumption contradicted with the opportunity to correct this in future. However, two limitations to this implementation prevented this from occurring. The first limitation, in that the agent is unable to remove the impact within the episode, is discussed in Section 5.3.3. The second limitation was due to the direction of communication in the MORL-glue software (Figure 5). The initial design relied upon using the inbuilt *RL_Agent_message* and *RL_env_message* functions, which use the glue to pass and return strings between these areas [78]. However, an unforeseen complication was that these functions are only configured to respond when called by the experiment program. Calls by and between the agent and environment themselves were unresponsive. Using the experiment program as an intermediary between these two programs was also an approach that was considered however the experiment program only has authority over the RL-Glue to the discretisation of a whole episode, and was not able to interject during an episode to provide additional direction. This path of exploration thoroughly debunked, an alternative solution was proposed.

The alternative solution involved sharing the variable between the functions via a utility class, storing the regularly updated variable of the agent's attitude. Due to complexities with threading, the calls required the structure of the existing RL glue framework to ensure appropriate synchronisation, which forced the problem into a simpler form. This required that the structure of the apology as it was implemented was necessarily simplified to include only the

assessment and apologetic aspects, with no latter confirmation.

In initial trials, the format of the impact minimisation approach of not penalising for the table outright introduced a difficulty in the step-wise apology as the agent was unable to see the negative table reward at the time that the actor became upset. A further issue that was encountered was using rewards for the actor's reactions. The table impact penalty was not felt until the end of the episode, at which point the table has been distinctly left out of place. The reactivity of the actor, then, was linked more decisively with the environment. Rather than passing a reward vector, the actor received knowledge on three states with regard to the environment, to answer three questions: is the table displaced? Has the cat been bothered? Has the agent taken more than 50 steps to complete the task? The actor is then classed in terms of sensitivity to these concerns, as discussed in the methodology (Section 3.2).

5 Results & Discussion

5.1 Results

5.1.1 Learning Phase

The extent to which the agent is capable of recognising how its actions may have caused harm to a user and subsequently correcting these behaviours, is intrinsically linked to the agent's capability to achieve its objectives overall. In the learning phase of the experiment, an agent was sought that demonstrated sufficient capability to achieve the minimum thresholds for each objective for a representative set of objectives. The set of objectives considered are those presented in the design, that represents each possible max-min configuration of thresholds across the three objectives (Table 1). Section 3.2.1 describes how these results were obtained.

Figure 12 demonstrates the results across eight threshold configurations, and Table 3 provides the sum of negative results, cumulative across each threshold. The points lying below zero represent a result where the agent failed to meet the specified threshold against that objective. For example, in Environment A, the S0 protocol successfully satisfies all thresholds for P, but has shortfall of approximately -50 for $\vec{T_0}$, $\vec{T_1}$, $\vec{T_4}$ and $\vec{T_5}$ for A_1 and $\vec{T_0}$, $\vec{T_2}$, $\vec{T_4}$ and $\vec{T_6}$ for A_2. These four threshold configurations for each objective correspond to a maximised threshold, implying that this S0 protocol performed poorly for both A_1 and A_2. This result was expected, as the $\vec{T_0}$ trained agent in Environment A performed poorly. Similar results are observed for the remaining S1-S7 protocols, demonstrating that the agent struggled to learn any other policies when presented with a constant threshold configuration.

Value Function Generation methods A0-A9 represent attempts to improve upon the results of S0-S7 through exposure to multiple threshold configurations during training. In contrast to the S0 protocol, the A8 protocol demonstrates improved performance across all objectives in both environments. The A8 protocol corresponds with an alternating approach between threshold configurations, with no initial training sequence, for 4000 episodes. In both Environment A and Environment B, the A8 protocol satisfies the threshold for P for every threshold configuration,

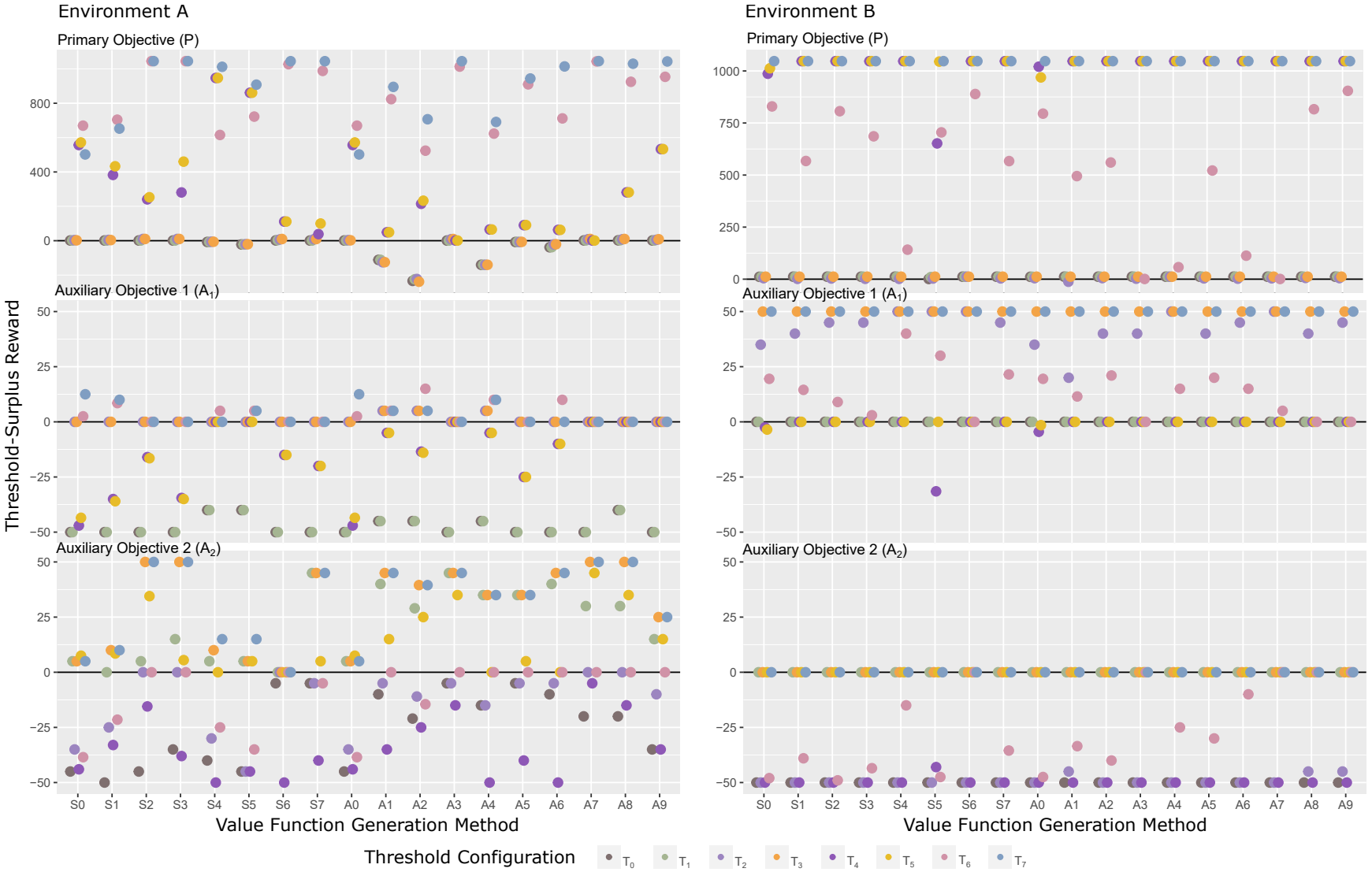

Figure 12: Value-function generation approaches, consisting of eight single-threshold protocols and ten aggregate protocols. Numeric results represent surplus reward beyond the threshold value specified. Negative values indicate the agent failed to meet the threshold.

Obj	Value Function Generation Approach																	
	S0	S1	S2	S3	S4	S5	S6	S7	A0	A1	A2	A3	A4	A5	A6	A7	*A8*	A9
Environment A																		
P	0	0	0	0	-26	-85	0	0	0	-473	-920	0	-561	-29	-118	0	*0*	0
A1	-191	-171	-133	-170	-80	-80	-130	-140	-191	-100	-118	-100	-100	-150	-120	-100	*-80*	-100
A2	-163	-130	-61	-73	-145	-170	-55	-55	-163	-50	-72	-25	-80	-50	-65	-25	*-35*	-80
total	**-353**	**-301**	**-193**	**-243**	**-251**	**-335**	**-185**	**-195**	**-353**	**-623**	**-1109**	**-125**	**-741**	**-229**	**-303**	**-125**	***-115***	**-180**
Environment B																		
P	0	0	0	0	0	0	0	0	0	-12	0	0	0	0	0	0	*0*	0
A1	-6	0	0	0	0	-31	0	0	-6	0	0	0	0	0	0	0	*0*	0
A2	-198	-189	-199	-194	-165	-191	-150	-186	-198	-179	-190	-150	-175	-180	-160	-150	*-145*	-145
total	**-204**	**-189**	**-199**	**-194**	**-165**	**-222**	**-150**	**-186**	**-204**	**-191**	**-190**	**-150**	**-175**	**-180**	**-160**	**-150**	***-145***	**-145**

Table 3: Cumulative shortfall results of value-function generation approaches, consisting of eight single-threshold protocols and ten aggregate protocols. Bold text highlights the column sum across the three objectives for each environment, and italicised text highlights the best performing protocol, A8, in both environments.

as well as A_1 in Environment B. For objective A_1 in Environment A, this protocol meets the threshold for $\vec{T}_4$ and $\vec{T}_5$, unlike most other protocols, and demonstrates slight improvement in $\vec{T}_0$ and $\vec{T}_1$. Similarly for A_2, the A8 protocol performs equally or better than other approaches considered. These results are reflected numerically in the Table 3, that gives A8 as the minimum or tied-minimum cumulative shortfall. The A8 protocol has been used as the basis for the subsequent apologetic experiments.

Interestingly, the A9 protocol demonstrated a decreased performance from A8 despite utilising the same approach for an increased episode count. This outcome may indicate over-training causing the agent's memory of complex policies to decay, especially in the context of hyper-parameters optimised for a 4000-episode training sequence limiting exploration in later episodes. Overall, these results demonstrate an avenue for further exploration in addition to providing a basis agent for the subsequent apologetic experiments.

5.2 Apologetic Phase

The apologetic experiment involved the exposure of four user configurations (Section 3.2.2) to the behaviour of an MORL agent with one of eight initial threshold configurations describing its objective prioritisation (Table 1) across two environments (Figure 4) and 10 trials. Figure 13 demonstrates that the proportions of satisfied and unsatisfied results differ with the user type, and Table 5 quantifies the differences in proportions of satisfied and unsatisfied results.

The experimental results demonstrate that in most cases, sensitivity to a given objective by a user results in greater proportions of satisfaction in that objective, post-apology. In Table 5, statistically significant changes in proportion of satisfied outcomes are denoted with bold text. The results report that in all single-sensitivity scenarios and dual-sensitivity scenarios with caveat, a statistically significant result was reported in the objective corresponding to the user's sensitivity.

In Environment B, the bias towards satisfaction of objective A_1 over A_2 is stronger than for Environment A (Table 4). The agent only differed from its A_1-preferential policy for threshold configurations where A_1 is minimal and A_2 is maximal. This is likely expounded by the deviance in the paths and physical distance between the associated impacts, limiting exploration of A_2-preferential policies as the agent seeks to exploit that which it has already learned. Unlike

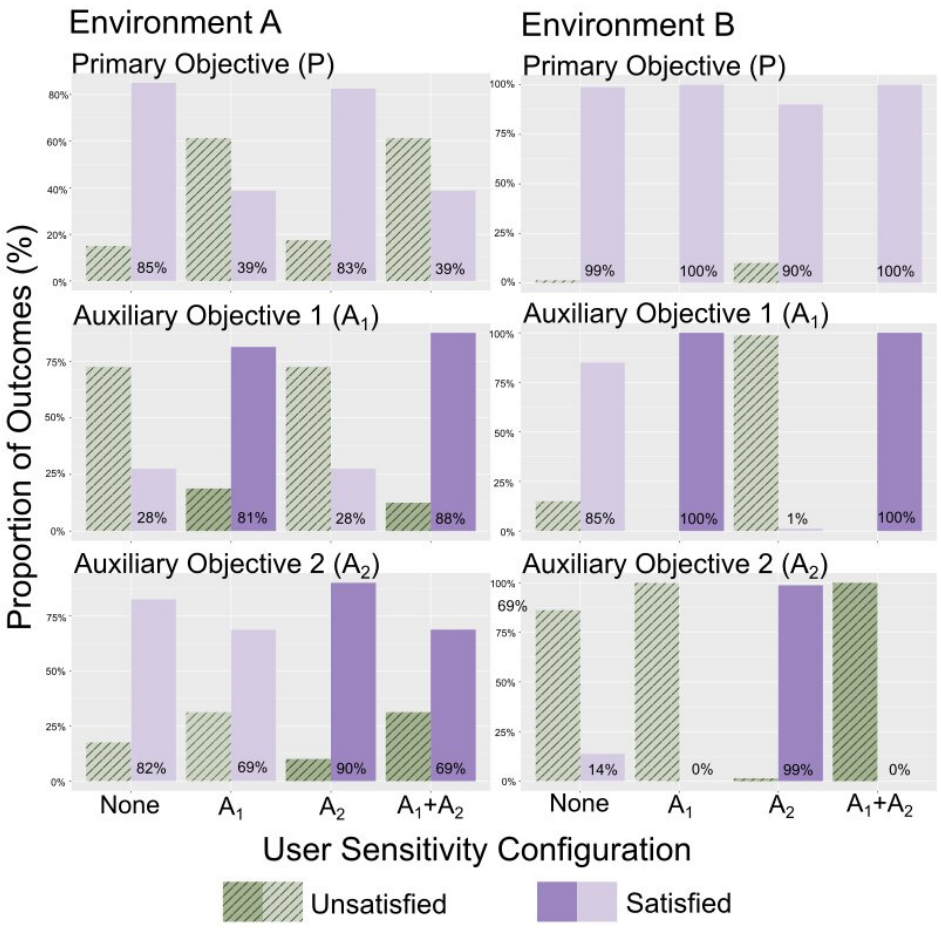

Figure 13: For each user configuration, the proportion of post-apology episodes for which the objective is satisfied and unsatisfied is demonstrated by the three parallel plots. For each objective, the darkened emphasis is applied where the user prioritises that objective. In general, greater proportions of satisfied objectives correlate with the user prioritisation.

Thresholds (P, A_1, A_2)	Final Rewards					
	Environment A			Environment B		
	R_P	R_{A_1}	R_{A_2}	R_P	R_{A_1}	R_{A_2}
$\vec{T_0}$ = (35, 0, 0)	37.2	*-40*	*-20*	47	0	*-50*
$\vec{T_1}$ = (35, 0, -50)	37.8	*-40*	-20	47	0	-50
$\vec{T_2}$ = (35, -50, 0)	43.2	-50	0	40.9	-10	*-45*
$\vec{T_3}$ = (35, -50, -50)	**45**	**-50**	**0**	**47**	**0**	**-50**
$\vec{T_4}$ = (-1000, 0, 0)	-718.5	0	*-15*	47	0	*-50*
$\vec{T_5}$ = (-1000, 0, -50)	**-718.5**	**0**	**-15**	**47**	**0**	**-50**
$\vec{T_6}$ = (-1000, -50, 0)	**-75.2**	**-50**	**0**	**-183.9**	**-50**	**0**
$\vec{T_7}$ = (-1000, -50, -50)	**30**	**-50**	**0**	**47**	**0**	**-50**

Table 4: Final reward outcomes for pre-apologetic agent for each threshold configuration. Rewards that do not meet the threshold are highlighted with italics. For single- or no-priority thresholds (in **bold**), all thresholds are satisfied.

User + Obj		Initial Threshold Configuration								p-value (All)	Initial Threshold Configuration								p-value (All)
		$\vec{T}_0$	$\vec{T}_1$	$\vec{T}_2$	$\vec{T}_3$	$\vec{T}_4$	$\vec{T}_5$	$\vec{T}_6$	$\vec{T}_7$		$\vec{T}_0$	$\vec{T}_1$	$\vec{T}_2$	$\vec{T}_3$	$\vec{T}_4$	$\vec{T}_5$	$\vec{T}_6$	$\vec{T}_7$	
		Environment A									Environment B								
None	P	0	0	0	0	0	0	0	0	NAN	0	0	0	0	0	0	0	0	NAN
	A_1	0	0	0	0	0	0	0	0	NAN	0	0	0	0	0	0	0	0	NAN
	A_2	0	0	0	0	0	0	0	0	NAN	0	0	0	0	0	0	0	0	NAN
A_1	P	-0.7	-0.7	-0.7	-0.7	0	0	-0.1	-0.6	**3.3E-09**	0	0	0	0	0	0	0.1	0	0.32
(table)	**A_1**	0.7	0.7	0.9	0.9	-0.1	-0.1	0.3	0.8	**9.8E-10**	0	0	0.2	0	0	0	1	0	**5.3E-4**
	A_2	0.1	0.1	-0.3	-0.3	0	0	-0.4	-0.3	**0.0045**	0	0	-0.1	0	0	0	-1	0	**9.1E-4**
A_2	P	-0.1	-0.1	0	0	0	0	0	0	0.41	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	0	-0.1	**0.0082**
(cat)	A_1	0.2	0.2	0	0	-0.2	-0.2	0	0	1	-1	-1	-0.7	-1	-1	-1	0	-1	**7.3E-16**
	A_2	0.1	0.1	0	0	0.2	0.2	0	0	**0.014**	1	1	0.8	1	1	1	0	1	**1.6E-16**
A_1+A_2	P	-0.7	-0.7	-0.7	-0.7	-0.1	-0.1	-0.1	-0.6	**1.2E-09**	0	0	0	0	0	0	0.1	0	0.32
(both)	**A_1**	0.8	0.8	1	1	0	0	0.3	0.9	**2.8E-11**	0	0	0.2	0	0	0	1	0	**5.3E-4**
	A_2	0.1	0.1	-0.3	-0.3	0	0	-0.4	-0.3	**0.0045**	0	0	-0.1	0	0	0	-1	0	**9.1E-4**

Table 5: Change in proportion of satisfied outcomes, given initial behaviour index, for each objective, user setting and environment. Also shown is the p-value of the McNemar proportionality test across the all initial behaviours, values of $p < 0.05$ suggesting a statistically significant alteration in the agent's behaviour against that objective.

User	Initial Threshold Configuration (Accuracy% (Total Apologies))								Average (all)
	$\vec{T_0}$	$\vec{T_1}$	$\vec{T_2}$	$\vec{T_3}$	$\vec{T_4}$	$\vec{T_5}$	$\vec{T_6}$	$\vec{T_7}$	
Environment A									
None	na (0)	na (0)	na (0)	na (0)	na (0)	na (0)	na (0)	na (0)	na (0)
A_1	**95% (37)**	**95% (37)**	**95% (37)**	**95% (37)**	**93% (30)**	**93% (30)**	**90% (91)**	**93% (46)**	**94% (345)**
A_2	29% (31)	29% (31)	na (0)	na (0)	8% (12)	8% (12)	na (0)	na (0)	9% (86)
A_1+A_2	41% (37)	41% (37)	51% (37)	51% (37)	10% (30)	10% (30)	55% (91)	59% (46)	40% (345)
Environment B									
None	na (0)	na (0)	na (0)	na (0)	na (0)	na (0)	na (0)	na (0)	na (0)
A_1	na (0)	na (0)	**100% (2)**	na (0)	na (0)	na (0)	**90% (10)**	na (0)	24% (12)
A_2	**100% (10)**	**100% (10)**	**94% (18)**	**100% (10)**	**100% (10)**	**100% (10)**	na (0)	**100% (10)**	**99% (78)**
A_1+A_2	**99% (100)**	**99% (100)**	**100% (100)**	**99% (100)**	**99% (100)**	**99% (100)**	**99% (100)**	**99% (100)**	**99% (800)**

Table 6: Accuracy and apologies provisioned (Accuracy% / Total Apologies) for each user, behavioural index and environment. Bold highlight has been applied where the accuracy is greater than 90%.

in Environment A, the agent does not tend towards a holding pattern, demonstrated by an unsatisfied P, to wait the episode out. Thus, when asked to prioritise both A_1 and A_2 objectives by the A_1+A_2 user, the agent eagerly selected this A_1-preferential policy and neglected A_2 entirely. In following this policy, the agent continues to disturb the cat, apologise and repeat the behaviour. It is aware of the reason the user is upset as beyond the first episode, A_2 is the only possible candidate, thus demonstrating a 99% apology provision accuracy. However it cannot avoid this impact, thus it continues to upset the user and subsequently needs to apologise during every episode, resulting in the maximum possible total apology count of 800 (10 episodes in 10 trials of 8 configurations, Table 4).

It can be inferred that a similar result occurred for A_1 in Environment A. The fewer apologies reported in this environment is likely due to a greater, yet imperfect, rate of success in realigning behaviour post-apology. Further evidence of this can be observed in the inverted pattern of change in proportion of satisfied outcomes and number of apologies given (Table 5). This suggests that the agent continues to apologise when it fails to correct the behaviour, which is the expected and desired result.

The agent is less accurate in the A_2 and A_1+A_2 scenarios in Environment A, and behavioural alignment is less pronounced. This is likely due to the increase in noise associated with the Environment A threshold results (Table 4). The bias in Environment A towards its simpler objective is less pronounced than in Environment B, likely due to the physical closeness of the two impact triggers.

5.3 Discussion

To restate the research question:

To what extent is an AI agent capable of learning to produce the components of a formal apology?

And sub-questions:

- *To what extent is an AI agent capable of learning to identify self-blame associated with a prior action, when recognising the presence of harm following an interaction?*

- *Given the identified prior action, to what extent is the agent capable of adjusting its policy selection, such that the agent demonstrates a reduction in likelihood of reproducing said harm?*

In pursuit of an answer to this question, this thesis has proposed a framework to support the generation of a formal apology by an autonomous agent. This framework provides a basis for the implementation of apology within an AI system though the interpretation of a reaction from a human user and self-reflection. Furthermore, this thesis has implemented this framework in an MORL context, to demonstrate identification of self-blame associated with a prior action, and the subsequent change in behaviour that results in a reduction in likelihood of reproducing undesirable behaviours. This thesis also presents and discusses an expansion of MORL impact minimisation to a dual-auxiliary scenario. The following discussion addresses the results of this research.

5.3.1 An Apologetic Agent

The prototype apologetic agent presented in this thesis provided a promising proof of concept for further development of apologetic AI. In some scenarios, such as those relating to difficult objectives in the environment, the agent is highly successful in detecting undesirable behaviour based on a user response, and demonstrates a significant subsequent behaviour improvement.

The agent did demonstrate weaknesses in accuracy in problems with less complex solutions, indicating that some improvements to the agent's assessment process may be of benefit. Approaches for improving the determination of fault, such as those that appeal to human behaviours surrounding apology, may improve the agent's accuracy in these circumstances. These behaviours include verifying the justification of an apology through conversation with the user, and using this knowledge to improve future behaviours by avoiding repeating mistakes.

The agent learns the user's preferences through an association between negative user feedback and the presence of stimulus by way of a candidate objective. Available knowledge that this approach does not utilise is the presence of this stimulus in absence of the negative feedback. This information could be leveraged to decrease the likelihood of selecting a particular objective for apology, if this objective has also been present while the user was not upset. In human learning this is referred to as stimulus discrimination [41].

5.3.2 Dual-Auxiliary Impact Minimisation

The TLO_{PMI} agent, in a dual-auxiliary impact minimisation environment, is an expansion and exploration of ideas discussed in the existing literature [77]. The approach considers whether multiple impact objectives can be learned when they are in conflict, and an appropriate policy aligned with the needs of the human user selected after learning is complete.

In general, the TLO_{PMI} agent demonstrated some success in its multi-objective optimisation. the agent was able to learn avoidance of both auxiliary impacts in most cases where P was not thresholded; a desirable outcome in which the agent avoids both grid locations during its 1000-step refusal. Code logs suggest that the agent often decides to walk into a wall space repeatedly in these cases, thus not changing state and avoiding the need to learn a safe 'holding pattern' of movement between different locations. The agent was able to optimise effectively against less complex problems, and was able to optimise exclusively against a singularly-prioritised objective when directed. However, the agent was limited in its success for some combination of objectives, converging towards a policy that only optimises a single objective when the ignored objective is particularly complex. This difficulty in balancing multiple objectives is likely due to the complexity of the problem, as those objectives required up to 15 specific steps to achieve success where similar success could be obtained in much fewer elsewhere. Thus, the agent never identified the policy that allowed satisfaction of this complex objective and a further objective, such as completion of the task.

5.3.3 Considerations of the Problem Environment

In this environment, the two impacts that the agent may cause are difficult or impossible to reverse. In either environment, the impact of disturbing the cat cannot be corrected. The impact of displacing the table is reversible but presents a significant challenge. In the Environment A, the agent must learn a policy consisting of 15 steps with minimal opportunity for any variation if it wishes to replace the table. In Environment B, the objective associated with the table impact can be satisfied with greater ease as the task can be completed without ever moving the table. However, if the agent were ever to displace it, the pathway to replace the table would be similarly complex.

In a reinforcement learning approach, the agent learns through experience that is driven by

random exploration. As the agent begins to develop a more refined understanding of the environment, it will converge towards a policy that non-exploratory steps will seek to exploit for reward. This is the origin of the learning difficulties demonstrated in problems where multiple solutions of differing complexity exist, as the agent has no reason to believe and thus motivation to explore in search of solutions far displaced from the 'good enough' solution that it has identified.

The implication of this is that the agent is unable to correct its behaviour during the episode. Despite the alteration of the thresholds occurring prior to the next action, the policy does not change as there is no alternative policy given the initial course of action the agent had already taken. A different problem space and a different type of impact would potentially be able to overcome this limitation, and might also address the problem complexity bias in the two dual-impact environments. If the agent has knowledge and opportunity to correct and impact during an episode after it has been identified and apologised for, it would be interesting to observe whether the agent is able to demonstrate this exchange of policy mid-episode.

Examples include tasks for which all pareto-optimal policies require a similar number of steps to obtain. Similarly, navigation complexity could be exchanged for problem complexity by introducing multiple pieces of rubbish. This would require the agent to explore the space more thoroughly as the primary objective could not easily be obtained, and so could introduce alternative pathways between obstacles associated with different impacts.

6 Threats to Validity

Apology in AI is a threat to the validity of this thesis is the accuracy of the definition of apology used and its application in this context. As discussed in Section 2.1, the literature debates whether the generation of such an apology via an unemotive artificially intelligent agent undermines the intention of the apology. Smith [70] proposed that an apology is insincere unless motivated by a moral compass. If this position was to be sufficiently established, this would invalidate the application discussed here. However, this assertion does not hold up to human use of apology, which is often driven by motivations other than moral ones.

This experiment has demonstrated significant results in a controlled simulated environment, with a simulated user. Many aspects of the user experience and expression have been simplified to facilitate a complete prototype of the apologetic system. However, these simplifications do not reflect real-life scenarios. A real human user will likely be reactive to, and reactions are unlikely to be as clearly identifiable as implied by this implementation. As such, the accuracy demonstrated in this prototype will likely be impacted and may be impacted to the extent that these findings are no longer significant.

Further to the limitations of this implementation, a discrete grid-word as such as the environments considered within this thesis are not a realistic environment. The agent demonstrated difficulty in sufficiently exploring this environment to learn policies for more complex problems. This may become more problematic if the state-space were to be expanded to represent a continuous problem.

As this implementation represented a prototype of a multi-faceted system, the breadth and depth of the investigation was limited. This investigation did not consider a user with sensitivity to the primary objective or different approaches to threshold updates, and only considered two similar configurations of a single environment type. These limitations inhibit a claim to the robustness of the proposed framework. Furthermore, each configuration was demonstrated with a small sample of only 10 trials. This sample size limited the capacity to identify potential outliers and . In order to have a sufficient sample size to compute statistical significance test scores, the results between the eight initial threshold configurations were combined. This reduced the resolution of the statistical analysis possible and would be avoidable with an

substantially increased sample size.

Finally, the implementation of this prototype did not involve an investigation as to alternative approaches than MORL for execution. Alternative approaches that were not considered may be better suited to this problem.

7 Conclusion & Future Work

This thesis presents the first framework and subsequent implementation of apology in an automated AI system, through an application of MORL. It investigated and demonstrated the extent to which an AI agent is capable of learning to produce the components of a formal apology; that which consists of acknowledgement of harm, explanations of the actions that caused the harm, and description of the changes to be made to the system to avoid this harm in the future. In so doing, this thesis has also presented an approach for interactive policy selection for a layperson, and explored an expansion of Impact Minimisation to learn two auxiliary objectives, where these objectives are in conflict.

This thesis has introduced and successfully demonstrated the Act-Assess-Apologise framework for AI apology. This framework has demonstrated success in specific environments to recognise undesirable behaviours and adjust behaviour in accordance, while also providing a templated articulation of apology. Variation in behaviour and apology provision accuracy was observed between configurations of problem complexity and prioritisation, demonstrating accuracy of up to 99% in some non-trivial scenarios. High accuracy was associated with complex problems and those with a distinct solution complexity bias. Post-apologetic behaviours demonstrated statistically significant improvements in user-sensitive objectives for all single-sensitivity scenarios, and in one of the objectives for multi-sensitivity scenarios. The behaviour improvement was resilient against configurations that resulted in lower apology accuracy. This agent also demonstrates selection of a policy that rejects the primary objective entirely to avoid causing harm to a user that is sensitive to both auxiliary objectives, where satisfaction of both objectives is incompatible with the primary goal. This is desirable for an impact minimisation problem where the consequences of a breach is high, in that the agent is able to recognise this requirement and cease pursuit of its primary goal.

7.1 Future Work

To further this research, a more complete investigation is required. There are three main areas of consideration that would benefit from further investigation, to establish AI Apology as an

approach for human-alignment in AI systems.

The first area of expansion regards exploration for more intelligent and in-depth apology and communication processes. In this implementation, the agent has used a simplified model of correlation to imply causation. This approach might be expanded with more intelligent approaches, such as through proximity or application of prior knowledge. The apology might also make use of conversational explanations to provide context and justification to actions that resulted in harm [22]. This could include providing an explanation of intentions describing why the action was selected, as well as confirmation of the user having received an apology that was given.

The second area of expansion regards addressing the weaknesses and assumptions demonstrated in the mode of implementation demonstrated in this thesis. As was explored in the discussion, the environment configuration heavily influenced the behaviours demonstrated by the agent. Both multi-impact environments were vulnerable to a weakness in the exploration processes of a MORL agent, as discussed in the literature [75]. These environmental influences highlight considerations of the MORL approach that were not thoroughly investigated in this implementation. Given that accuracy of apology provision demonstrates a negative correlation with the agent's success of completing a particular task, the viability of this approach given a problem that the agent does not struggle to solve requires consideration.

The final area regards assumptions of user needs and preferences, and user interactions with AI. This implementation has discussed the concept of a user reaction in response to actions undertaken by the agent, but has not proposed how this reaction may be expressed. Furthermore, this implementation has assumed that a human user would be comforted by an apology provisioned by an AI. Future work should consider undertaking a user study to understand the user's needs, in the context of these assumptions. This study should include considerations regarding how a user may express discontentment with the actions of a perceived autonomous robot, and, how might such a user perceive and respond to an autonomously generated apology provided by this robot in response. This knowledge will provide a firm basis to challenge or validate intrinsic assumptions about human perception and acceptance of AI apology.

References

[1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng, *TensorFlow: A System for Large-Scale Machine Learning*, in 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16), Savannah, GA, 11 2016, {USENIX} Association, pp. 265–283.

[2] R. Abduljabbar, H. Dia, S. Liyanage, and S. A. Bagloee, *Applications of Artificial Intelligence in Transport: An Overview*, Sustainability, 11 (2019).

[3] A. Adadi and M. Berrada, *Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)*, IEEE Access, 6 (2018), pp. 52138–52160.

[4] A. Allan, M. M. Allan, D. Kaminer, and D. J. Stein, *Exploration of the association between apology and forgiveness amongst victims of human rights violations*, Behavioral Sciences and the Law, 24 (2006), pp. 87–102.

[5] D. Amodei, C. Olah, G. Brain, J. Steinhardt, P. Christiano, J. Schulman, O. Dan, and M. Google Brain, *Concrete Problems in AI Safety*, tech. rep., arXiv, 2016.

[6] P. Andras, L. Esterle, M. Guckert, T. A. Han, P. R. Lewis, K. Milanovic, T. Payne, C. Perret, J. Pitt, S. T. Powers, N. Urquhart, and S. Wells, *Trusting Intelligent Machines: Deepening Trust Within Socio-Technical Systems*, IEEE Technology and Society Magazine, 37 (2018), pp. 76–83.

[7] T. B. Asafa, T. M. Afonja, E. A. Olaniyan, and H. O. Alade, *Development of a vacuum cleaner robot*, Alexandria Engineering Journal, 57 (2018), pp. 2911–2920.

[8] J. Asmuth, M. L. Littman, and R. Zinkov, *Potential-based shaping in model-based reinforcement learning*, Proceedings of the National Conference on Artificial Intelligence, 2 (2008), pp. 604–609.

[9] A. Bahrammirzaee, *A comparative survey of artificial intelligence applications in finance: Artificial neural networks, expert system and hybrid intelligent systems*, Neural Computing and Applications, 19 (2010), pp. 1165–1195.

[10] H. Baomar and P. J. Bentley, *An Intelligent Autopilot System that learns flight emergency procedures by imitating human pilots*, in 2016 IEEE Symposium Series on Computational Intelligence, SSCI 2016, Institute of Electrical and Electronics Engineers Inc., 2 2017.

[11] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi, *Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots*, 11 2009.

[12] J. M. Bernardo, *Expected Information as Expected Utility*, The Annals of Statistics, 7 (1979), pp. 686–690.

[13] A. Bhatt, *Safe AI Systems*, Tech. Rep. 31, 2018.

[14] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Z. Openai, *OpenAI Gym*, tech. rep.

[15] D. Brougham and J. Haar, *Smart Technology, Artificial Intelligence, Robotics, and Algorithms (STARA): Employees' perceptions of our future workplace*, 3 2018.

[16] J. Bullock, A. Luccioni, K. H. Pham, C. S. N. Lam, and M. Luengo-Oroz, *Mapping the landscape of artificial intelligence applications against COVID-19*, Journal of Artificial Intelligence Research, 69 (2020), pp. 807–845.

[17] P. S. Castro, S. Moitra, C. Gelada, S. Kumar, M. G. Bellemare, and G. Brain, *DOPAMINE: A RESEARCH FRAMEWORK FOR DEEP REINFORCEMENT LEARNING*, tech. rep.

[18] S. Cave and S. S. ÓhÉigeartaigh, *An AI Race for Strategic Advantage: Rhetoric and Risks*, New Orleans, LA, USA, 2018, In Proceedings of 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18).

[19] A. D. Cohen and E. Olshtain, *Developing a Measure of Sociocultural Competence: the Case of Apology*, Language Learning, 31 (1981), pp. 113–134.

[20] R. Dasoriya, J. Rajpopat, R. Jamar, and M. Maurya, *The Uncertain Future of Artificial Intelligence*, in Proceedings of the 8th International Conference Confluence 2018 on Cloud Computing, Data Science and Engineering, Confluence 2018, Institute of Electrical and Electronics Engineers Inc., 8 2018, pp. 458–461.

[21] R. Dazeley, P. Vamplew, and F. Cruz, *Explainable Reinforcement Learning for Broad-XAI : A Conceptual Framework and Survey*, arXiv, (2021).

[22] R. Dazeley, P. Vamplew, C. Foale, C. Young, S. Aryal, and F. Cruz, *Levels of Explainable Artificial Intelligence for Human-Aligned Conversational Explanations*, Artificial Intelligence, 299 (2021).

[23] S. Devlin, D. Kudenko, and M. Grze, *An empirical study of potential-based reward shaping and advice in complex, multi-agent systems*, Advances in Complex Systems, 14 (2011), pp. 251–278.

[24] F. K. Dosilovic, M. Brcic, and N. Hlupic, *Explainable artificial intelligence: A survey*, in 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2018 - Proceedings, Institute of Electrical and Electronics Engineers Inc., 6 2018, pp. 210–215.

[25] B. R. Duffy, *Anthropomorphism and the social robot*, in Robotics and Autonomous Systems, vol. 42, North-Holland, 3 2003, pp. 177–190.

[26] E. Fast and E. Horvitz, *Long-Term Trends in the Public Perception of Artificial Intelligence*, Tech. Rep. 1, 2 2017.

[27] M. D. Fethi and F. Pasiouras, *Assessing bank efficiency and performance with operational research and artificial intelligence techniques: A survey*, 7 2010.

[28] Frank Kaptein, Joost Broekens, Koen Hindriks, and Mark Neerincx, *The Role of Emotion in Self-Explanations by Cognitive Agents*, Institute of Electrical and Electronics Engineers, (2017).

[29] B. Fraser, *On Apologizing*, in Rasmus Rask Studies in Practicing Linguistics, F. Coulmas, ed., De Gruyter Mouton, 2011, pp. 259–272.

[30] N. Fulton and A. Platzer, *Safe AI for CPS (Invited Paper)*, in Proceedings - International Test Conference, vol. 2018-Octob, Institute of Electrical and Electronics Engineers Inc., 1 2019.

[31] Z. Gabor, K. Zsolt, and C. Szepesvari, *Multi-criteria Reinforcement Learning*, ICML, 98 (1998), pp. 197–205.

[32] S. E. Guthrie, *Anthropomorphism: A definition and a theory.*, in Anthropomorphism, anecdotes, and animals., SUNY series in philosophy and biology., State University of New York Press, Albany, NY, US, 1997, pp. 50–58.

[33] T. A. Han, L. Moniz Pereira, T. Lenaerts, and F. C. SantosID, *Mediating artificial intelligence developments through negative and positive incentives*, PLoS ONE, 16 (2021).

[34] C. F. Hayes, R. Rădulescu, E. Bargiacchi, J. Källström, M. Macfarlane, M. Reymond, T. Verstraeten, L. M. Zintgraf, R. Dazeley, F. Heintz, E. Howley, A. A. Irissappane, P. Mannion, A. Nowé, G. Ramos, M. Restelli, P. Vamplew, and D. M. Roijers, *A Practical Guide to Multi-Objective Reinforcement Learning and Planning*, Unpublished Manuscript, (2021).

[35] B. Hibbard, *Decision Support for Safe AI Design*, LNAI, 7716 (2012), pp. 117–125.

[36] M. Hutter and S. Legg, *A Formal Measure of Machine Intelligence*, arXiv, (2006).

[37] S. Ingle and M. Phute, *Tesla Autopilot : Semi Autonomous Driving, an Uptick for Future Autonomy*, International Research Journal of Engineering and Technology, (2016).

[38] R. Issabekov and P. Vamplew, *An Empirical Comparison of Two Common Multiobjective Reinforcement Learning Algorithms*, in AI 2012: Advances in Artificial Intelligence, M. Thielscher and D. Zhang, eds., Berlin, Heidelberg, 2012, Springer Berlin Heidelberg, pp. 626–636.

[39] JetBrains, *IntelliJ IDEA*, 2021.

[40] R. Kaushik, K. Chatzilygeroudis, and J. B. Mouret, *Multi-objective model-based policy search for data-efficient learning with sparse rewards*, arXiv, (2018).

[41] F. S. Keller and W. N. Schoenfeld, *Principles of psychology: A systematic text in the science of behavior*, 1950.

[42] T. Kim and H. Song, *How should intelligent agents apologize to restore trust? Interaction effect between anthropomorphism and apology attribution on trust repair*, Preprint submitted to Telematics and Informatics, (2021).

[43] Y. Y. Lee, C. C. S. Kam, and M. H. Bond, *Predicting emotional reactions after being harmed by another*, Asian Journal of Social Psychology, 10 (2007), pp. 85–92.

[44] E. Liang, R. Liaw, R. Nishihara, P. Moritz, R. Fox, J. Gonzalez, K. Goldberg, and I. Stoica, *Ray RLlib: A Composable and Scalable Reinforcement Learning Library*, in Proc. of the 35th International Conference on Machine Learning, no. Nips, 2018, pp. 3059–3068.

[45] P. H. Lin, A. Wooders, J. T. Y. Wang, and W. M. Yuan, *Artificial Intelligence, the Missing Piece of Online Education?*, IEEE Engineering Management Review, 46 (2018), pp. 25–28.

[46] M. Lippi, G. Contissa, A. J. Lonowska, F. Lagioia, H. W. Micklitz, P. Palka, G. Sartor, and P. Torroni, *The Force Awakens: Artificial intelligence for consumer law*, Journal of Artificial Intelligence Research, 67 (2020), pp. 169–190.

[47] C. Liu, X. Liu, F. Wu, M. Xie, Y. Feng, and C. Hu, *Using artificial intelligence (watson for oncology) for treatment recommendations amongst Chinese patients with lung cancer: Feasibility study*, Journal of Medical Internet Research, 20 (2018), p. e11087.

[48] S. Loren, *What are the implications of the virtual for the human? An analytical ethics of identity in pop culture narratives*, European Journal of American Culture, 23 (2004), pp. 173–185.

[49] Y. Lu, *Artificial intelligence: a survey on evolution, models, applications and future trends*, Journal of Management Analytics, 6 (2019), pp. 1–29.

[50] X. Luo and L. Xie, *Research on artificial intelligence-based sharing education in the era of internet*, in Proceedings - 3rd International Conference on Intelligent Transportation, Big Data and Smart City, ICITBS 2018, vol. 2018-Janua, Institute of Electrical and Electronics Engineers Inc., 4 2018, pp. 335–338.

[51] D. McArthur, D. McArthur, M. Lewis, and M. Bishary, *The Roles of Artificial Intelligence in Education: Current Progress and...*, Journal of Educational Technology, 1 (2005), pp. 42–80.

[52] Q. McNemar, *Note on the sampling error of the difference between correlated proportions or percentages*, Psychometrika, 12 (1947), pp. 153–157.

[53] T. Miller, P. Howe, and L. Sonenberg, *Explainable AI: Beware of Inmates Running the Asylum*, arXiv, (2017).

[54] Mirka Snyder Caron and Abhishek Gupta, *The Social Contract for AI*, Cornell University, (2020).

[55] S. M. Muddamsetty, M. N. S. Jahromi, A. E. Ciontos, L. M. Fenoy, and T. B. Moeslund, *Introducing and assessing the explainable AI (XAI)method: SIDU*, arXiv, (2021).

[56] L. Muehlhauser and N. Bostrom, *WHY WE NEED FRIENDLY AI*, Think, Spring2014 (2013).

[57] M. Muraven, *Designing a Safe Autonomous Artificial Intelligence Agent based on Human Self-Regulation*, arXiv, 01 (2017).

[58] M. R. Napolitano and M. Kincheloe, *On-line learning neural-network controllers for autopilot systems*, Journal of Guidance, Control, and Dynamics, 18 (1995), pp. 1008–1015.

[59] T. T. Nguyen, N. D. Nguyen, P. Vamplew, S. Nahavandi, R. Dazeley, and C. P. Lim, *A multi-objective deep reinforcement learning framework*, Engineering Applications of Artificial Intelligence, 96 (2020), p. 103915.

[60] S. Omohundro, *Autonomous technology and the greater human good*, Journal of Experimental & Theoretical Artificial Intelligence, 26 (2014), pp. 303–315.

[61] Oracle, *Java SE Development Kit 8*, 2021.

[62] L. Pang, Y. Zhang, S. Coleman, and H. Cao, *Efficient Hybrid-Supervised Deep Reinforcement Learning for Person Following Robot*, Journal of Intelligent and Robotic Systems: Theory and Applications, 97 (2020), pp. 299–312.

[63] J. M. Pinder, *Multi-Objective Reinforcement Learning Framework for Unknown Stochastic & Uncertain Environments*, (2017).

[64] D. M. Roijers, P. Vamplew, S. Whiteson, and R. Dazeley, *A survey of multi-objective sequential decision-making*, Journal of Artificial Intelligence Research, 48 (2013), pp. 67–113.

[65] A. Santara, S. Rudra, S. A. Buridi, M. Kaushik, A. Naik, B. Kaul, and B. Ravindran, *MADRaS: Multi agent driving simulator*, Journal of Artificial Intelligence Research, 70 (2021), pp. 1517–1555.

[66] J. M. Schoenborn and K.-D. Althoff, *Recent Trends in XAI: A Broad Overview on current Approaches, Methodologies and Interactions*, in ICCBR Workshops, 2019.

[67] J. Searle, *Speech Acts: An Essay in the Philosophy of Language*, the Syndics of the Cambridge University Press, 1969.

[68] D. Shin, *The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI*, International Journal of Human Computer Studies, 146 (2021).

[69] D. Slocum, A. Allan, and M. M. Allan, *An emerging theory of apology*, Australian Journal of Psychology, 63 (2011), pp. 83–92.

[70] N. Smith, *I Was Wrong: The Meanings of Apologies*, Cambridge University Press, 2008.

[71] K. Sokol and P. Flach, *Counterfactual Explanations of Machine Learning Predictions: Opportunities and Challenges for AI Safety*, SafeAI@ AAAI, (2019).

[72] R. S. Sutton, A. G. Barto, and A. B. Book, *Reinforcement Learning : An Introduction*, MIT Press, 1998.

[73] B. Tanner and A. White, *RL-Glue: Language-Independent Software for Reinforcement-Learning Experiments*, Journal of Machine Learning Research, 10 (2009), pp. 2133–2136.

[74] I. Ulrich, F. Mondada, and J. D. Nicoud, *Autonomous vacuum cleaner*, Robotics and Autonomous Systems, 19 (1997), pp. 233–245.

[75] P. Vamplew, R. Dazeley, and C. Foale, *Softmax exploration strategies for multiobjective reinforcement learning*, Neurocomputing, 263 (2017), pp. 74–86.

[76] P. Vamplew, R. Dazeley, C. Foale, S. Firmin, and J. Mummery, *Human-aligned artificial intelligence is a multiobjective problem*, Ethics and Information Technology, 20 (2018), pp. 27–40.

[77] P. Vamplew, C. Foale, R. Dazeley, and A. Bignold, *Potential-based multiobjective reinforcement learning approaches to low-impact agents for AI safety*, Engineering Applications of Artificial Intelligence, 100 (2021), p. 104186.

[78] P. Vamplew, D. Webb, L. M. Zintgraf, D. M. Roijers, R. Dazeley, R. Issabekov, and E. Dekker, *MORL-Glue: A benchmark suite for multi-objective reinforcement learning*, 29th Benelux Conference on Artificial Intelligence November 8–9, 2017, Groningen, (2017), p. 389.

[79] P. Vamplew, J. Yearwood, R. Dazeley, and A. Berry, *On the Limitations of Scalarisation for Multi-objective Reinforcement Learning of Pareto Fronts*, in AI 2008: Advances in Artificial Intelligence, W. Wobcke and M. Zhang, eds., Berlin, Heidelberg, 2008, Springer Berlin Heidelberg, pp. 372–378.

[80] K. Van Moffaert, M. M. Drugan, and A. Nowe, *Scalarized multi-objective reinforcement learning: Novel design techniques*, in IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning, ADPRL, 2013, pp. 191–199.

[81] R. Vashistha, A. K. Dangi, A. Kumar, D. Chhabra, and P. Shukla, *Futuristic biosensors for cardiac health care: an artificial intelligence approach*, 8 2018.

[82] R. V. Yampolskiy, *Unpredictability of AI: On the Impossibility of Accurately Predicting All Actions of a Smarter Agent*, Journal of Artificial Intelligence and Consciousness, 07 (2020), pp. 109–118.

[83] B. Zhong and M. Zamani, *Towards Safe AI: Safe-visor Architecture for Sandboxing AI-based Controllers in Stochastic Cyber-Physical Systems*, Journal of the ACM, 37 (2020).

[84] D. Zhou, J. Chen, J. P. Morgan, and Q. Gu, *Provable Multi-Objective Reinforcement Learning with Generative Models*, tech. rep., 2021.

A Search Terms

Objective	Search Terms	Sources Consulted	Further Notes
Background information and literature on apology in human communication	"purpose" AND "apology"	Google Scholar	Selected literature presenting a broad overview of the topic
Construction and components of an effective apology	"components" AND "apology"	Google Scholar	Selected literature directly representing the topic. Excluded items where apology was a secondary topic
Links between eXplainable AI and applications in apology	"Explainable Artificial Intelligence"	Google Scholar + Deakin University Library	No date restrictions applied. 11 articles relevant to the subject were identified, from 19 results. Manually pruned by title
Instances or expansions on Broad-XAI	"Broad XAI"	Google Scholar + Deakin University Library	Only unpublished manuscripts found
Discussion of emotion in XAI applications	"Emotion-aware XAI"	Google Scholar + Deakin University Library	No further restrictions were applied. After manual pruning via title, a single relevant result was identified
Multi-Objective Reinforcement learning	"multi" AND "objective" AND "reinforcement" AND "learning" AND "AI" NOT ("Game" OR "Deep")	Google Scholar via Publish or Perish + Deakin University Library	Restricted to papers published from 2017 and beyond. Extraction of first 360 results and manually pruned by title. Additional papers identified through Deakin Library. Also avoided "fuzzy", "multi-agent", "cooperative" or "evolution" keywords. 55 relevant articles identified.
Multi-Objective Reinforcement learning with explainable systems	"multi objective" AND "reinforcement learning" AND "explainable"	Google Scholar via Publish or Perish	Restricted to papers published from 2017 and beyond. Extraction of first 100 results and manually pruned by title

Continues on next page

Background information on classification of human emotions	"psychological model of emotion"	Google Scholar	Selected a single textbook to cover some base concepts
Overview of current emotion recognition research using machine learning via secondary studies	Title search: survey AND content search: ("expression prediction" OR "expression recognition" OR "emotion prediction" OR "emotion recognition") AND "machine learning" AND "survey"	Deakin University Library	Filtered by date to 2017-current, and by subject tags: recognition, state, and measurement of emotions and the perception or recognition of faces and facial expressions. Excluded items where emotion recognition was not a primary topic.
Identification of current applications of AI	artificial intelligence applications	Google Scholar + Deakin University Library + JAIR	Publications since 2017, seeking examples (not a complete representation of work) This search revealed a number of surveys that provided further linkages and expansions to relevant papers.
Identification of public perception and social risk of AI applications	public perception of artificial intelligence	Google Scholar	Since 2017
Justification of assumption: humans express emotions in reaction to experiencing harm	reactions to being harmed	Google Scholar	none

B Acceptance Criteria

B.1 Sprint 1

	S1-AC01
Given	There is a piece of rubbish in a given location within the room the agent is carrying no pieces of rubbish
When	the agent enters that location
Then	the rubbish will no longer be present in that location in the room the agent will be carrying one piece of rubbish.

	S1-AC02
Given	There is a piece of rubbish in a given location within the room the agent is carrying one or more pieces of rubbish
When	the agent enters that location
Then	the rubbish will no longer be present in that location in the room the agent will be carrying one additional piece of rubbish.

	S1-AC03
Given	There is a 'home' in a given location within the room The agent is carrying no pieces of rubbish
When	the agent enters that location
Then	no changes will occur.

	S1-AC04
Given	There is a 'home' in a given location within the room The agent is carrying one or more pieces of rubbish
When	the agent enters that location
Then	the agent will be given a reward of 50 against the tidy objective for each item of rubbish the agent will no longer be carrying any pieces of rubbish.

	S1-AC05
Given	There is no pieces of rubbish in the room The agent is carrying one or more pieces of rubbish
When	the agent disposes of the rubbish
Then	the episode will end.

	S1-AC06
Given	There was a piece of rubbish in a given location within the room The agent has collected this piece of rubbish
When	the agent enters that location a second or subsequent time
Then	the will be no rubbish present in that location in the room The agent will not be carrying any additional pieces of rubbish.

B.2 Sprint 2

	S2-AC01
Given	An apologetic agent has determined that their behaviour is misaligned with the actor's preferences
When	the agent sufficiently alters its thresholds to reflect these preferences
Then	the agent's behaviour must alter to reflect these thresholds.

	S2-AC02
Given	An apologetic agent is required to assess its behaviour for fault
When	the agent considers its current reward signal
Then	the state of possible at-fault actions that the agent may have taken must be observable from this reward signal and align to an independent impact objective.

	S2-AC03
Given	The Satisficing Agent is presented with the dual impact environment
When	the agent interacts with the environment
Then	the agent refines a mapping of the environment state-action expected rewards via a value function these values are not dependent upon the prioritization thresholds.

	S2-AC04
Given	The Satisficing Agent is presented with the dual impact environment
When	the agent determines which policy to follow
Then	there should be no universally optimal policy the decision should be based upon thresholded lexicographical ordering based upon a provided set of thresholds.

	S2-AC05
Given	The Satisficing Agent is presented with the dual impact environment
When	the agent determines which policy to follow
Then	the decision should be based upon thresholded lexicographical ordering based upon a provided set of thresholds.

	S2-AC06
Given	The Satisficing Agent is presented with an environment containing the cat obstacle
When	the agent attempts to move into the space occupied by the cat in the first instance
Then	the agent successfully moves to that location, and a penalty is applied for having "run over the cat's tail".

	S2-AC07
Given	The Satisficing Agent is presented with an environment containing the cat obstacle
When	the agent attempts to move into the space occupied by the cat in the second or subsequent instances
Then	the agent successfully moves to that location, and no additional penalty is applied.

	S2-AC08
Given	The Satisficing Agent is presented with an environment containing the cat obstacle
When	the agent has received a penalty for "running over the cat's tail"
Then	no subsequent actions of the agent can revoke the penalty.

	S2-AC09
Given	The Satisficing Agent is presented with the dual impact environment
When	the agent moves the table to access the rubbish
Then	the agent is only able to return the table to its original location by moving into the same location as the cat and thus receiving a penalty.

	S2-AC09
Given	The Satisficing Agent is presented with an environment containing the Chair or TV obstacles
When	the agent attempts to move into the space occupied by the Chair or TV obstacles
Then	the agent receives an invalid location update, and thus does not complete the action but does consume a time-step.

B.3 Sprint 3

	S3-AC01
Given	An experiment involves a watched scenario with an Apologetic Agent
When	the Agent is initialised
Then	the Agent is initialised an Apologetic agent.

	S3-AC02
Given	An experiment involves a watched scenario without an Apologetic Agent
When	the Agent is initialised
Then	the Agent is initialised as a Considerate or Traditional agent.

	S3-AC03
Given	An experiment involves a watched scenario with any Agent
When	the Environment is initialised
Then	an Actor is initialised to observe the Agent.

	S3-AC04
Given	An experiment involves a watched scenario with any Agent
When	the Actor is initialised
Then	A type is assigned to the actor in accordance with the scenario requirements this type dictates manner in which the environment determines the Actor's Attitude.

	S3-AC05
Given	The Agent is completing its required task
When	the Agent takes an action
Then	the Environment state is updated to reflect this action.

	S3-AC06
Given	There is an Actor observing the Agent
When	the Environment state is updated
Then	the Environment parses the updated reward to the Actor the Actor updates its Attitude in accordance with its type.

	S3-AC07
Given	There is an Actor observing the Agent
When	the Actor updates its Attitude to happy or neutral
Then	there is a specified random chance that the actor will instead update its attitude to upset.

	S3-AC08
Given	The Agent is completing a watched scenario with an Apologetic Agent
When	the Agent takes an Action
Then	the Agent retrieves the Actor's Attitude with a specified probability of successful interpretation representing realistic expression recognition success rates.

	S3-AC09
Given	The Agent is completing a watched scenario with an Apologetic Agent
When	the Agent has determined that the Actor is upset
Then	the Agent examines its accumulated reward against each objective, and determines whether any are net negative if one or more is net negative, the agent determines fault the agent returns the most negative as the justification for fault.

	S3-AC10
Given	The Agent is completing a watched scenario with an Apologetic Agent
When	the Agent has determined that the Actor is upset and that they are at fault
Then	the Agent constructs and delivers an Apology to the Actor the apology consists of a justification that is the most negative of the accumulated reward. the Agent updates its thresholds to prioritise this objective as described in the apology.