Neural Computing and Applications (2023) 35:16917–16930
https://doi.org/10.1007/s00521-023-08586-x

S.I. : HUMAN-ALIGNED REINFORCEMENT LEARNING FOR AUTONOMOUS AGENTS AND ROBOTS

AI apology: interactive multi-objective reinforcement learning for human-aligned AI

Hadassah Harland[1] · Richard Dazeley[1] · Bahareh Nakisa[1] · Francisco Cruz[2,3] · Peter Vamplew[4]

Received: 28 September 2022 / Accepted: 5 April 2023 / Published online: 22 April 2023

Abstract

For an Artificially Intelligent (AI) system to maintain alignment between human desires and its behaviour, it is important that the AI account for human preferences. This paper proposes and empirically evaluates the first approach to aligning agent behaviour to human preference via an *apologetic* framework. In practice, an apology may consist of an acknowledgement, an explanation and an intention for the improvement of future behaviour. We propose that such an apology, provided in response to recognition of undesirable behaviour, is one way in which an AI 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 Act-Assess-Apologise framework potentially could address both the practical and social needs of a human user, to recognise and make reparations against prior undesirable behaviour and adjust for the future. Applied to a dual-auxiliary impact minimisation problem, the apologetic agent had a near perfect determination and apology provision accuracy in several non-trivial configurations. The agent subsequently demonstrated behaviour alignment with success that included up to complete avoidance of the impacts described by these objectives in some scenarios.

Keywords AI apology · Multi-objective reinforcement learning · Human alignment · Impact minimisation · AI safety

1 Introduction

In human-to-human interactions, an apology is a powerful social tool to communicate awareness of harm caused and to repair trust [1, 2]. This trust and mutual understanding is a cornerstone of the functional society in which we work and live as humans. As Artificial Intelligence (AI) systems further increase their presence, this trust and mutual understanding needs to be similarly shared between humans and machines. Thus, in human/AI scenarios, it is reasonable to replicate apology in AI systems for use in human-agent interactions. This paper will introduce the first framework for an autonomously apologetic AI system, and provide the results of an implementation of one such system.

AI apology crosses between AI safety and explainability (XAI), as an approach for human-alignment and layperson-accessible explanations of decisions. An apology consists of three key components: acknowledgment that harm

✉ Hadassah Harland
h.harland@research.deakin.edu.au

Richard Dazeley
richard.dazeley@deakin.edu.au

Bahareh Nakisa
bahar.nakisa@deakin.edu.au

Francisco Cruz
f.cruz@unsw.edu.au

Peter Vamplew
p.vamplew@federation.edu.au

[1] School of Information Technology, Deakin University, Geelong, VIC, Australia

[2] School of Computer Science and Engineering, University of New South Wales, Sydney, NSW, Australia

[3] Escuela de Ingenieria, Universidad Central de Chile, Santiago, Chile

[4] Centre for Smart Analytics, Federation University, Ballarat, VIC, Australia

occurred, indication of understanding of responsibility for the harm, and commitment to avoiding causing this harm again in the future. Apology as a tool for reconciliation given broken trust in human–machine interactions has been studied [3], with findings that support the development of autonomous apologetic agents.

Beyond human behavioural perspectives, practical and communicative approaches for improving trust in AI have similarly been proposed. Explainable AI (XAI) focuses upon improving the transparency of AI decision-making processes, to provide clarity and justification to actions such as those that result in undesirable behaviour [4, 5]. Publications in AI Safety include pragmatic approaches for harm avoidance and self-supervisory wrapper systems [6, 7] as well as social approaches including exploration of legal regulation [8]. Recent work in Impact Minimisation (IM) seeks to generalise and penalise against any impactful behaviours that are not explicitly aligned with the agent's primary objective [9, 10]. The goals of each of these areas is to improve the practical and perceived standing of risks that improperly regulated AI systems pose. However, no prior work exists in which an agent may address both the practical and social consequences of undesirable behaviour through apology. This work presents the first framework and proof of concept for such an agent.

This paper presents the Act-Assess-Apologise framework for application to robotic and agent-based systems with a requirement for sensitivity to human needs and preferences. The framework proposes that following each action, the agent observes the user for a response that is assessed in context of its recent actions, and apologises where it recognises the need. In an AI context, behaviour improvement is demonstrated through prioritisation of objectives according to these user preferences. Human alignment in AI is best represented as a multi-objective (MO) problem [11, 12]. The framework proposed within this paper has been applied to one such problem, with a basis of reinforcement learning (RL) to represent a robotic system in a controlled environment.

This paper makes the following contributions:

- This is the first work to propose a framework for an AI agent to autonomously identify the need for and generate a formal apology.
- It proposes an approach for layperson-accessible interactive policy selection, to identify a desirable policy without explicit direction.
- It expands upon multi-objective impact minimisation by introducing contested additional auxiliary objectives.

2 Background: apology in human-aligned AI

This section presents the relevant literature from three key areas: apology, human-alignment in AI and multi-objective reinforcement learning. This describes the existing knowledge that has been consulted to inform the contributions of this paper, and provides support for the premise of apologetic AI.

2.1 What is an apology?

An apology is a social exchange, usually between two human parties, enacted in atonement for wrongful actions. It is a fundamental aspect of human communication [2]. When used appropriately, an apology communicates remorse and supports the repair of trust and relationship [3].

The definition of an apology is a well-traversed question within the fields of philosophy and psychology. Smith [2] deconstructed the complex social ritual to eleven components, practical aspects comprising; recognition of and responsibility for the harm, identification and endorsement of the moral underpinnings, regret and reform [2]. These are in addition to implicit considerations regarding performance, intention and reception. A culmination of other proposals on the matter [1, 13–15] converges with a central theme for a more concise working definition, consisting of affirmation, affect and action.

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 practice, these components are an acknowledgement, an explanation, and a promise to do better [1].

Previous works have addressed the use of apology to repair trust and build relationship in human-computer interactions. Considerations for the delivery of the apology in the establishment and repair of trust in these interactions is well researched, from the existence [16] and magnitude [17] of the apology, to the timing of the delivery [17, 18], to the language used for conversation elements [17, 19] and the attribution of blame [3, 20]. The research has found that apology may be beneficial to mitigating negative social effects of undesirable actions in human-robot interactions. The effectiveness of this approach depends on how the agent or robot is perceived [3, 21], as more social and anthropomorphic agents found greater success. However, these notable contributions are limited in scope to user studies with pseudo-AI agents and do not explore the implementation of such an agent or of AI-generated apology. Applications for AI-generated apology are not represented in the literature, and thus there is no benchmark established for this problem. Rather, this research provides significant insight to the motivations of this paper: that an approach for practical implementation is required.

2.2 Human alignment and safe AI

Human alignment in AI speaks to addressing the challenges and opportunities in the field to best suit the needs and requirements of the human user. This encompasses practical considerations of safe and beneficial practices, avoidance of harm to the agent or to the user, and social considerations of emotional well-being, comfort and understanding. Adherence to these requirements has been described as a social contract [22]. AI research without human alignment is inhibited by these issues [8, 23].

Humans and AI agents interpret and operate within the world in inherently different ways [24]. These differences need to be taken into consideration when proposing AI that operates within human spaces. Similarly, undesirable behaviours between humans and AI are born of different origins. For example, in agents driven by the maximisation of an expected utility (MEU), undesirable behaviour usually originates from inaccuracies in the agent's definition of this utility [9]. It is an inevitability that the agent might select an action misaligned with the desired outcome, improperly prioritised or otherwise unwanted: a mistake.

Many approaches to AI safety make use of restrictions, such as self-supervisory wrapper systems [6, 7]. The folly of these approaches is that it is not possible to infallibly define every undesirable outcome [25]. Other approaches address this issue through internal measures to disincentivise undesirable behaviours. Researchers have argued that a multi-objective approach is required to support human alignment in AI [11, 12]. Impact minimisation is one such approach that penalises against all environmental impacts outside a defined set of desirable changes [10].

2.3 Multi-objective reinforcement learning

Reinforcement Learning (RL) is an AI development approach that uses random exploration and prior experience to determine optimal patterns of behaviour. RL agents seek to maximise a reward that is provided upon the successful completion of a goal, through selection of a policy that corresponds to the greatest expected reward. Multi-objective RL (MORL) uses a vector of rewards corresponding to multiple objectives. MORL policy selection seeks to find and select from the set of Pareto dominant policies, one for which no improvement against a specific objective is possible without a loss against some other objective, as according to a set of priorities [26].

MORL-based Impact minimisation (IM) uses auxiliary objectives associated with undesirable environmental impacts, to produce low-impact agents [10]. This foundational algorithm is capable of recognising and considering prioritisation of impact management as an auxiliary objective, in opposition to a primary goal. The Act-Assess-Apology framework as proposed and implemented in this paper has been applied to an extension of MORL-based IM. In this application, it refines the agent's behaviour through the prioritisation of each auxiliary objective in accordance with the preferences of the user.

3 The Act-Assess-Apologise framework

An apologetic approach to AI, consisting of acknowledgement, explanation and change in behaviour, can be leveraged for both practical and social benefits in human-alignment. For the purpose of this paper, we propose an apologetic process that may be applied to an AI behavioural cycle for autonomously generated apology. This process is described as the Act-Assess-Apologise (AAA) framework, and is a novel contribution of this research.

This framework proposes a step-wise approach to generating an apology. It is an augmentation to an AI agent's action cycle, and relies upon observation of a human user. The framework consists of three stages, presented in a recursive cycle as shown in Fig. 1. In the *Act* stage, the agent undertakes an action as according to its underlying algorithm. In the *Assess* stage, the agent observes the human user for a reaction. The agent must also consider the impact of its prior actions as to whether there is reasonable correlation and potential for these actions to have been harmful. Implementation requires an assumption based upon specifics of the application for reasonable inference of causation to assign self-blame. If the user has reacted negatively and the agent has determined causation due to its action, the agent will proceed to the *apologise* stage.

The *apologise* stage involves the articulation of an apology that requires expression of this predetermined self-

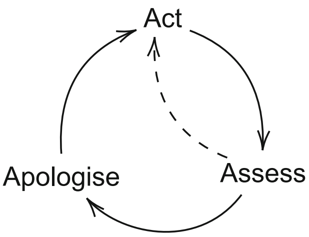

Fig. 1 The Act-Assess-Apologise framework presents a three-stage approach to implementing apology within an AI system

blame and the reason for the negative reaction. The agent may acknowledge that they have recognised the user is upset, and explain the impact of their prior actions that caused it. The agent completes the apology by describing and implementing the manner through which it will avoid this upsetting behaviour in the future. This concludes the apologetic cycle, and the agent will proceed to the next action.

This framework has been applied and empirically evaluated in the context of multi-objective reinforcement learning, but the same approach could be used in other agent-based systems.

3.1 Apology-augmented RL

The apologetic framework has been implemented using a MORL approach. The agent-environment framework, illustrated in Fig. 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, analogous to the arrow between *Act* and *Assess* in the AAA framework (Fig. 1). 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 extends the application of IM agents by providing an approach for interactively adjusting prioritisation of various objectives [10]. The RL agent will have a primary objective (P) describing its key task. In addition, it will have one or more auxiliary objectives $((A_i)_{i\in\{1..k\}})$ corresponding to other aspects of the environment that the agent may impact in its attempt to maximise P.

Apology is applied only after the agent has been trained and exploration is deactivated, as we have assumed that the agent is fully trained in its task prior to live operation. What is unknown during training is the preferences of the human user, and this post-training adjustment approach allows the agent to select for different preferential behaviours for different users. Figure 3 presents this process an adaptation of a 'review and adjust' scenario as discussed in the MORL literature [11]. 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.

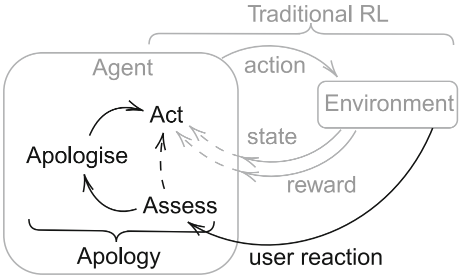

Fig. 2 The Act-Assess-Apologise framework may be combined with the RL Agent-Environment framework for apology-augmented RL. The RL action-reaction process is analogous to the arrow between *Act* and *Assess* in Fig. 1. The process is separable from RL in that the agent is not directly responding to a reward signal to update a Q-table, but rather undergoing hyper-parameter adjustment that contextualises the Q-values.

3.2 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 [27] 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 (Sect. 4).

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.

3.3 Policy selection and thresholding

Reinforcement learning agents seek to maximise expected utility by selecting an action a that, given the current system state s, maximises a utility function $U(s, a)$. In a multi-objective context, this utility function is a vector $\vec{U}(s,a)$ with elements corresponding to the utility contribution associated with each of the objectives. To switch

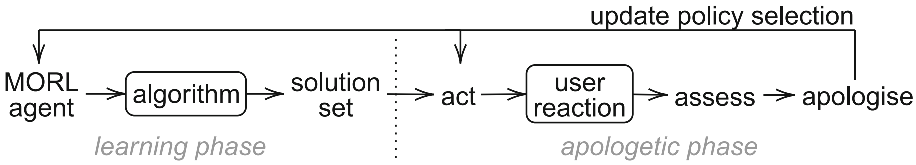

Fig. 3 The apologetic agent evolves through two phases of learning, applying a 'review and adjust' process based on the user's predicted needs

between prioritisation of separable objectives, a distinctly nonlinear multi-objective action selection approach is required. The change in behaviour following an apology occurs via alterations to hyper-parameters used in this policy selection. Thresholded Lexicographical Ordering (TLO), introduced by Gabor et al. [28] for MORL and used by Vamplew et al. [10] for impact-minimising agents, is one such parameterised, nonlinear multi-objective action selection approach. TLO selects an action that maximises the reward of the second objective, subject to having reached the threshold specified for the first objective [10, 28].

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. 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.

TLO may 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 [10]. A default prioritisation order defines consistent preferential selection between equivalently thresholded objectives in exchange for a slight simplification. The three objective case used in this paper resolves to $\mathrm{TLO}^{\mathrm{PMI}_3}$, as defined in Eq. 1. In this context, the superscript PMI is in reference to the prioritised multi-impact approach, and the subscript 3 clarifies the three objective case. The following shorthand notation has been introduced for readability: $U_i(s,a) \rightarrow U_i$, $U_i(s,a') \rightarrow U'_i$, and $\min(U_i(s,a),T_i) \rightarrow \tau_i$, where i takes values of 1, 2, and 3 to reference the primary objective and two auxiliary objectives, respectively.

$$\begin{aligned}
&\forall s,a,a' \vec{U}(s,a) >_{TLO^{PMI_3}} \vec{U}(s,a') \\
&\quad\Leftrightarrow (\tau_1 > \tau'_1) \vee \big((\tau_1 = \tau'_1) \wedge (\tau_2 > \tau'_2)\big) \\
&\qquad \vee \big((\tau_1 = \tau'_1) \wedge (\tau_2 = \tau'_2) \wedge (\tau_3 > \tau'_3)\big) \\
&\qquad \vee \big((\tau_1 = \tau'_1) \wedge (\tau_2 = \tau'_2) \wedge (\tau_3 = \tau'_3) \wedge (U_1 > U'_1)\big) \\
&\qquad \vee \big((\tau_1 = \tau'_1) \wedge (\tau_2 = \tau'_2) \wedge (\tau_3 = \tau'_3) \\
&\qquad \wedge (U_1 = U'_1) \wedge (U_2 > U'_2)\big) \\
&\qquad \vee \big((\tau_1 = \tau'_1) \wedge (\tau_2 = \tau'_2) \wedge (\tau_3 = \tau'_3) \\
&\qquad \wedge (U_1 = U'_1) \wedge (U_2 = U'_2) \\
&\qquad \wedge (U_3 > U'_3)\big)
\end{aligned} \tag{1}$$

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. 1 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 a 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.

3.4 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 (Fig. 3). All objectives cannot be satisfied simultaneously. Thus, an apologetic approach is required to determine which objective, if any, can be ignored based on this user preference. Hence, allowing the agent to adapt its behaviour to the preferences of the individual user without any additional training as the underlying optimal policies are already known to the agent. One such environment has been used in this implementation.

4 Evaluation methodology

Implementation of the apologetic agent was demonstrated through an extension of Mmpact Minimisation [10]. As no prior work exists demonstrating an apologetic approach, there exists no benchmark against which to compare. The agent's behaviour was evaluated by comparison to pre-apologetic behaviours, in context of alignment to the user's preferences.

The problem environment was modeled after a domestic living room: a discrete and otherwise static grid-world consisting of an assortment of obstacles. This environment was selected as it is both reflective of a plausible real-life robotic service scenario and facilitates the flexibility and problem complexity required as discussed in Sect. 3.4. The agent is presented with a primary objective: collect the rubbish and return home (P). The agent must also 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 and described in Fig. 4. Environment A and Environment B 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.

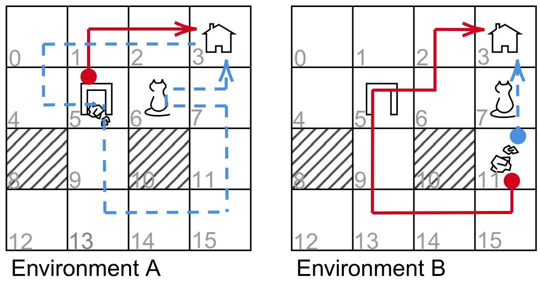

Fig. 4 In these two non-trivial environment configurations, the agent is unable to collect the rubbish and satisfy both auxiliary objectives. The optimal policy for prioritisation of the table (A_1) and the cat (A_2) are given by the blue (dashed) and red (solid) paths, respectively

The agent was a low-impact MORL agent, that used the $\mathrm{TLO}_{\mathrm{PMI}}$ (Eq. 1) 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 impact occurs. The actual values provided as rewards and penalties to the agent are not significant beyond their relations to each other, and thus these values have been adopted from literature [10]. The representative threshold set, consisting of eight combinations of maximum/minimum threshold values is given in Table 1. Using the maximum and minimum reward values obtainable for each objectives as threshold values acts as an on/off switch for prioritisation of these objectives. Any objectives with a minimum-reward threshold value will always meet the threshold value and thus will not be prioritised in the

Table 1 Eight threshold sets represent each possible on/off combination of the final reward values for the three objectives (P, A_1, A_2). For each objective, the maximum value approaches the highest achievable reward under reasonable execution of the task, thus representing a reasonable goal, while the minimum value is the lowest achievable result such that the reward always satisfies the threshold

Identity	Thresholds	Identity	Thresholds
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	$\vec{T}_7 = (-1000, -50, -50)$

reward optimisation process until optimum outcomes are obtained in the other objectives. Similarly, maximum-reward threshold values are not met unless the given objective is satisfied, and are prioritised as such. 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 represented by the blue (dashed) path in Environment A and the red (solid) path in Environment B in Fig. 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. 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. 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.

A simulated user reacts to changes in the environment. Responses were determined by the user's sensitivity, which is a vector of boolean values corresponding to each objective. Four configurations of user sensitivities were considered; each auxiliary 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.

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 $Sensitivity_i \wedge s_{t,i}$ satisfies $StateConditions_i$
                    $\wedge\ attitude = 0$ 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.

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 $apologised \leftarrow$ false, $s_t$
    Initialise $prioritised \leftarrow$ [false, false, false]
    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 $(attitude < 0 \ \wedge \ apologised =$ false$)$ then
            if $(min(\vec{r}_{t+1}) < 0)$ then
                $justification \leftarrow$Index$(min(R))$
                Agent ‘apologises’ w.r.t justification
                $apologised \leftarrow$ true
                $prioritised_{justification} \leftarrow$ true
                for each Threshold: $j$ in $\vec{T}$ do
                    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
```

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

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
$Apology = Affirmation + Affect + Action$
```

The key components of this algorithm align with the *affirmation*, *affect* and *action* components of apology. Previously established priorities consist of those 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.

5 Results

5.1 Pre-apologetic behaviour

The agent was pre-trained by alternating through the threshold configurations specified in Table 1, for 4000 online training episodes. This demonstrated the closest adherence to the expected policies and the fewest infringements against max-thresholded objectives, of 18 considered approaches. 10 independent trials were performed, with the final trained agent from each trial retained for use in the apologetic agent trials.

The agent demonstrated behaviour that was aligned with expectation of existing literature for such a problem [10]. The agent found and converged to the simple solution, the red path in Environment A and the blue path in Environment B, with consistent success. Environment A allows for two alternative paths, equivalent in P and A_1 but impactful against A_2, that the agent learns to avoid. For thresholds prioritising this simple solution, the agents demonstrate success in managing both P and this easier objective simultaneously (see Environment A, $\vec{T}_2$ and Environment B, $\vec{T}_1$ in Table 2).

In no trials in either environment does the agent successfully learn the complex pathway required to promptly complete the task and simultaneously satisfy the condition of the more complex auxiliary objective. This aligns with known limitations of MO exploration wherein the agent's path does not vary far from an easily identified 'good enough' solution and thus struggles to find the true optimal policy [29]. The agent does, however, learn a policy that avoids impact against this objective by sacrificing P and not completing the primary task at all. The agent stumbles upon a time-inefficient solution that it exploits, in some cases, or else repetitively selects a redundant action for 1000 timesteps as to satisfy the Environment A, A_1 or Environment B, A_2 objectives (Table 2). This behaviour arises when the threshold configuration prioritises this objective but not the primary objective, as thresholding against the primary objective takes priority over either auxiliaries.

5.2 Apologetic scenarios

The apologetic experiment consisted of 10 independent trials, each consisting 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 pre-apology results (Table 2). The first and final stages consisted of a traditional offline RL scenario that demonstrated the agent's behaviour before and after the apologetic framework was applied. In the

Table 2 Final reward outcomes for pre-apologetic agent for each threshold configuration

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*
$\vec{T}_3 = (\mathbf{35}, \mathbf{-50}, \mathbf{-50})$	**45**	**− 50**	**0**	**47**	**0**	**− 50**
$\vec{T}_4 = (-1000, 0, 0)$	− 718.5	0	*− 15*	47	0	*− 50*
$\vec{T}_5 = (\mathbf{-1000}, \mathbf{0}, \mathbf{-50})$	**− 718.5**	**0**	**− 15**	**47**	**0**	**− 50**
$\vec{T}_6 = (\mathbf{-1000}, \mathbf{-50}, \mathbf{0})$	**− 75.2**	**− 50**	**0**	**− 183.9**	**− 50**	**0**
$\vec{T}_7 = (\mathbf{-1000}, \mathbf{-50}, \mathbf{-50})$	**30**	**− 50**	**0**	**47**	**0**	**− 50**

Rewards that do not meet the threshold are highlighted in italic. For single- or no-priority thresholds (in bold), all thresholds are satisfied

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 “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.

Rewards for the auxiliary objectives A_1 and A_2 took binary values corresponding to whether or not the objective was satisfied. For the primary objective P, the distribution of data clustered heavily around the maximum threshold value, with the exception of results of -999, corresponding with an incomplete task. This data was coerced into a binary result of satisfied or unsatisfied, aligned with the auxiliary objectives. Figure 5 demonstrates that the proportions of satisfied and unsatisfied results differ with the user type.

In most cases, sensitivity to a given objective by a user results in greater proportions of satisfaction in that objective, post-apology. This is true in all cases of single-sensitivity scenarios, and for one of the objectives in each of the dual-sensitivity scenarios. Statistical analysis of the prior and subsequent proportions of satisfaction of objectives corresponding to the user's sensitivity found this proportionality difference to be significant given $\alpha = 0.05$ for each of these cases. This statistical analysis is further detailed in Appendix A.

In Environment B, the bias towards satisfaction of objective A_1 over A_2 is stronger than for Environment A (Table 2). 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 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 3).

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 3, Fig. 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 2). 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.

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

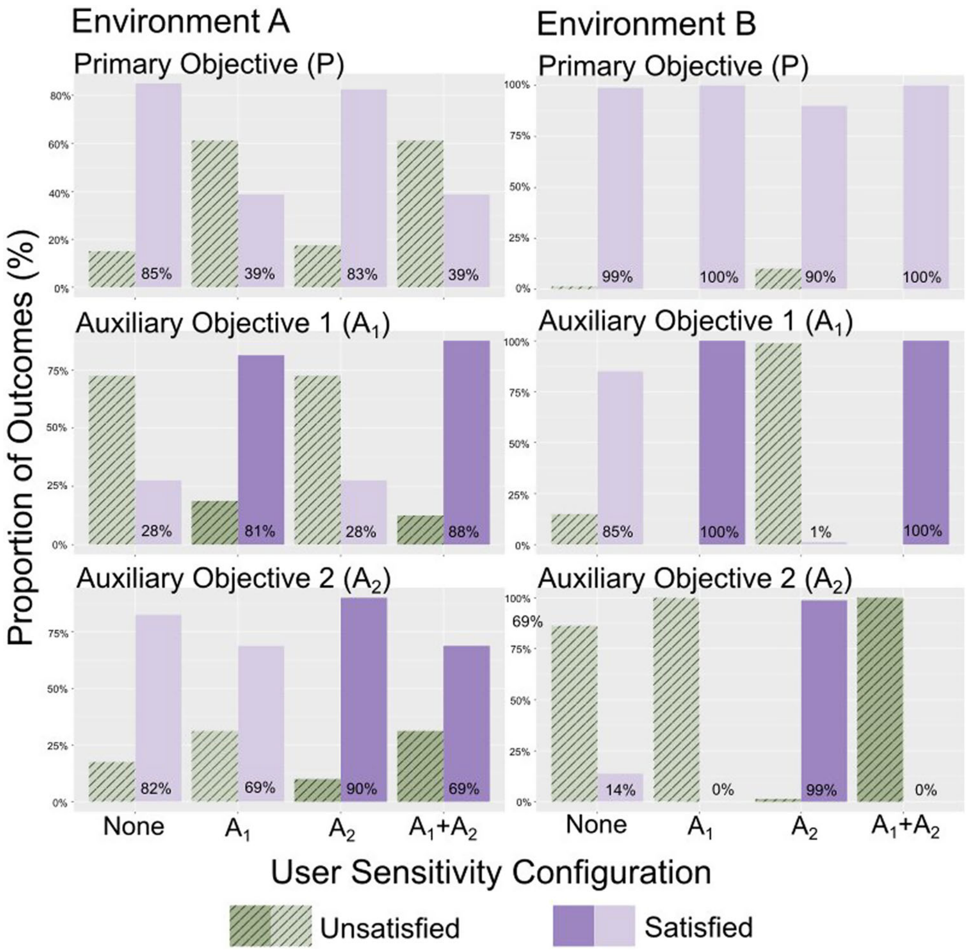

Fig. 5 The proportion of post-apology episodes for which each objective is satisfied and unsatisfied is demonstrated by the three parallel plots. Darkened emphasis is applied to user-prioritised objectives. In general, greater proportions of satisfied objectives correlate with the user prioritisation

Table 3 Accuracy and apologies provisioned (Accuracy% / total apologies) for each user, behavioural index and environment

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									
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									
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)**

Bold highlight has been applied where the accuracy is greater than 90%

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 [30].

6 Conclusion

This paper has made three main contributiosn to knowledge regarding the establishment of AI apology. It has proposed a framework for an AI agent to autonomously identify the need for and generate a formal apology. In so doing, it has also presented an approach for interactive policy selection for a layperson. Finally, it has explored an expansion of Impact Minimisation to learn two auxiliary objectives, where these objectives are in conflict.

This paper 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.

Future work will consider prospects for improvement to the determination of blame and apology provision processes, in addition to investigations of impacts of more realistic user scenarios. It will also seek to validate the intrinsic assumptions made within this paper in respect to the value of AI apology used in this manner, for establishing trust and relationship with a user and for its capability to understand the needs of a user. In a real-life application, AI apology may be applied as a tool for improved user experience, to facilitate real-time acknowledgement and realignment of behaviours in accordance with the preferences of a human user.

Appendix A: Statistical analysis

The difference between the proportion of satisfied outcomes in the prior state and the subsequent state have been recorded in Table 4. A McNemar test for the statistical significance of the proportionality differences between the

Table 4 Change in proportion of satisfied outcomes, given initial threshold configuration

User + Obj		Initial threshold configuration								p-value (All)
		$\vec{T}_0$	$\vec{T}_1$	$\vec{T}_2$	$\vec{T}_3$	$\vec{T}_4$	$\vec{T}_5$	$\vec{T}_6$	$\vec{T}_7$	
Environment A										
A$_1$ (table)	P	− 0.7	− 0.7	− 0.7	− 0.7	0	0	− 0.1	− 0.6	**3.3E−09**
	$\boldsymbol{A_1}$	*0.7*	*0.7*	*0.9*	*0.9*	*− 0.1*	*− 0.1*	*0.3*	*0.8*	***9.8E−10***
	A$_2$	0.1	0.1	− 0.3	− 0.3	0	0	− 0.4	− 0.3	**0.0045**
A$_2$ (cat)	P	− 0.1	− 0.1	0	0	0	0	0	0	0.41
	A$_1$	0.2	0.2	0	0	− 0.2	− 0.2	0	0	1
	$\boldsymbol{A_2}$	*0.1*	*0.1*	*0*	*0*	*0.2*	*0.2*	*0*	*0*	***0.014***
A$_1$+A$_2$ (both)	P	− 0.7	− 0.7	− 0.7	− 0.7	− 0.1	− 0.1	− 0.1	− 0.6	**1.2E−09**
	$\boldsymbol{A_1}$	*0.8*	*0.8*	*1*	*1*	*0*	*0*	*0.3*	*0.9*	***2.8E−11***
	$\boldsymbol{A_2}$	*0.1*	*0.1*	*− 0.3*	*− 0.3*	*0*	*0*	*− 0.4*	*− 0.3*	***0.0045***
Environment B										
A$_1$ (table)	P	0	0	0	0	0	0	0.1	0	0.32
	$\boldsymbol{A_1}$	*0*	*0*	*0.2*	*0*	*0*	*0*	*1*	*0*	***5.3E−4***
	A$_2$	0	0	− 0.1	0	0	0	− 1	0	**9.1E−4**
A$_2$ (cat)	P	− 0.1	− 0.1	− 0.1	− 0.1	− 0.1	− 0.1	0	− 0.1	**0.0082**
	A$_1$	− 1	− 1	− 0.7	− 1	− 1	− 1	0	− 1	**7.3E−16**
	$\boldsymbol{A_2}$	*1*	*1*	*0.8*	*1*	*1*	*1*	*0*	*1*	***1.6E−16***
A$_1$+A$_2$ (both)	P	0	0	0	0	0	0	0.1	0	0.32
	$\boldsymbol{A_1}$	*0*	*0*	*0.2*	*0*	*0*	*0*	*1*	*0*	***5.3E−4***
	$\boldsymbol{A_2}$	*0*	*0*	*− 0.1*	*0*	*0*	*0*	*− 1*	*0*	***9.1E−4***

Statistically significant changes in proportion (given by McNemar differences in proportionality test, $\alpha = 0.05$) are in bold. The prioritised objectives for each scenario are highlighted in italic.

prior and subsequent state was undertaken. Given a threshold of $\alpha = 0.05$, the statistical analysis suggests that the behaviour change is significant in all of the examined scenarios. The exception to this is the dual prioritisation case in both environments, in which proportion of satisfied outcomes decreases in the subsequent state for one of the objectives. This is aligned with the outcomes discussed in the main body of the paper, in that the agent does not demonstrate capability to consistently find a policy that allows for satisfaction of both auxiliary objectives.

Author contributions All authors contributed to the study conception and design, and provided theoretical, intellectual, conceptual, technical, and analytical input. Material preparation, data collection and analysis were performed by Hadassah Harland. The first draft of the manuscript was written by Hadassah Harland and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions.

Code availability Code for agent, environment and user simulation is available at https://github.com/hadassahharland/IM_MORL.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Slocum D, Allan A, Allan MM (2011) An emerging theory of apology. Aust J Psychol 63(2):83–92. https://doi.org/10.1111/j.1742-9536.2011.00013.x
2. Smith N (2008) I was wrong: the meanings of apologies. Cambridge University Press, Cambridge, pp 28–131
3. Kim T, Song H (2021) How should intelligent agents apologize to restore trust? Interaction effect between anthropomorphism and apology attribution on trust repair. Telematics Inform 61:101595. https://doi.org/10.1016/j.tele.2021.101595
4. Cruz F, Dazeley R, Vamplew P, Moreira I (2021) Explainable robotic systems: understanding goal-driven actions in a reinforcement learning scenario. Neural Computing and Applications 1–18. https://doi.org/10.1007/s00521-021-06425-5. arXiv:2006.13615
5. Dazeley R, Vamplew P, Foale C, Young C, Aryal S, Cruz F (2021) Levels of explainable artificial intelligence for human-aligned conversational explanations. Artif Intell 299:103525. https://doi.org/10.1016/j.artint.2021.103525
6. Omohundro S (2014) Autonomous technology and the greater human good. J Exp Theor Artif Intell 26(3):303–315. https://doi.org/10.1080/0952813X.2014.895111
7. Zhong B, Zamani M (2020) Towards safe AI: safe-visor architecture for sandboxing AI-based controllers in stochastic cyber-physical systems. J ACM 10(1145/1122445):1122456. https://doi.org/10.1145/3457335.3461705
8. Han TA, Moniz Pereira L, Lenaerts T, SantosID FC (2021) Mediating artificial intelligence developments through negative and positive incentives. PLoS One. https://doi.org/10.1371/journal.pone.0244592
9. Amodei D, Olah C, Brain G, Steinhardt J, Christiano P, Schulman J, Dan O, Google Brain M (2016) Concrete Problems in AI Safety. Unpublished Manuscript. arXiv:1606.06565
10. Vamplew P, Foale C, Dazeley R, Bignold A (2021) Potential-based multiobjective reinforcement learning approaches to low-impact agents for AI safety. Eng Appl Artif Intell 100:104186. https://doi.org/10.1016/j.engappai.2021.104186
11. Hayes CF, Rdulescu R, Bargiacchi E, et al (2022) A practical guide to multi-objective reinforcement learning and planning. Auton Agent Multi-Agent Syst 36:26. https://doi.org/10.1007/s10458-022-09552-y
12. Vamplew P, Dazeley R, Foale C, Firmin S, Mummery J (2018) Human-aligned artificial intelligence is a multiobjective problem. Ethics Inf Technol 20(1):27–40. https://doi.org/10.1007/s10676-017-9440-6
13. Allan A, Allan MM, Kaminer D, Stein DJ (2006) Exploration of the association between apology and forgiveness amongst victims of human rights violations. Behav Sci Law 24(1):87–102. https://doi.org/10.1002/bsl.689
14. Cohen AD, Olshtain E (1981) Developing a measure of socio-cultural competence: the case of apology. Lang Learn 31(1):113–134. https://doi.org/10.1111/j.1467-1770.1981.tb01375.x
15. Fraser B (2011) On Apologizing. In: Coulmas F (ed) Rasmus rask studies in practicing linguistics. De Gruyter Mouton, Berlin, pp 259–272. https://doi.org/10.1515/9783110809145.259
16. Fratczak P, Goh YM, Kinnell P, Justham L, Soltoggio A (2021) Robot apology as a post-accident trust-recovery control strategy in industrial human-robot interaction. Int J Ind Ergon 82:103078. https://doi.org/10.1016/J.ERGON.2020.103078
17. Galdon F, Wang SJ (2020). From apology to compensation: a multi-level taxonomy of trust reparation for highly automated virtual assistants. https://doi.org/10.1007/978-3-030-25629-6_7
18. Nayyar M, Wagner AR (2018) When should a robot apologize? Understanding how timing affects human-robot trust repair. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 11357 LNAI, 265–274. https://doi.org/10.1007/978-3-030-05204-1_26
19. Svenningsson N, Faraon M (2019) Artificial intelligence in conversational agents: a study of factors related to perceived humanness in chatbots. In: Proceedings of the 2019 2nd artificial intelligence and cloud computing conference. https://doi.org/10.1145/3375959
20. Buchholz V, Kulms P, Kopp S, (2017) It's (Not) your fault! Blame and trust repair in human-agent cooperation. https://doi.org/10.17185/duepublico/44538
21. Lee Y, Bae J-E, Kwak SS, Kim M-S (2011) The effect of politeness strategy on human - robot collaborative interaction on

malfunction of robot vacuum cleaner. RSS'11 (Robotics Science and Systems) Workshop on Human-Robot Interaction (October 2017)

22. Mirka Snyder Caron (2020) Abhishek Gupta: the social contract for AI. Cornell University
23. Cave S, ÓhÉigeartaigh SS (2018) An AI Race for strategic advantage: rhetoric and risks. In Proceedings of 2018 AAAI/ ACM conference on AI, ethics, and society (AIES '18), New Orleans. https://doi.org/10.1145/3278721.3278780
24. Dazeley R, Vamplew P, Cruz F (2021) Explainable reinforcement learning for broad-XAI: a conceptual framework and survey. Unpublished Manuscript. arXiv:2108.09003
25. Yampolskiy RV (2020) Unpredictability of AI: on the impossibility of accurately predicting all actions of a smarter agent. J Artif Intell Conscious 07(01):109–118
26. Vamplew P, Dazeley R, Berry A, Issabekov R, Dekker E (2011) Empirical evaluation methods for multiobjective reinforcement learning algorithms. Mach Learn 84(1–2):51–80. https://doi.org/10.1007/s10994-010-5232-5
27. Lee YY, Kam CCS, Bond MH (2007) Predicting emotional reactions after being harmed by another. Asian J Soc Psychol 10(2):85–92. https://doi.org/10.1111/j.1467-839X.2007.00215.x
28. Gabor Z, Zsolt K, Szepesvari C (1998) Multi-criteria reinforcement learning. ICML 98:197–205
29. Vamplew P, Dazeley R, Foale C (2017) Softmax exploration strategies for multiobjective reinforcement learning. Neurocomputing 263:74–86. https://doi.org/10.1016/j.neucom.2016.09.141
30. Keller FS, Schoenfeld WN (1950) Principles of psychology: a systematic text in the science of behavior, pp 115–163