Challenges for Using Impact Regularizers to Avoid Negative Side Effects

David Lindner,[1*] **Kyle Matoba,** [2*] **Alexander Meulemans** [3*]

[1] Department of Computer Science, ETH Zurich
[2] Idiap and EPFL
[3] Institute of Neuroinformatics, University of Zurich and ETH Zurich
david.lindner@inf.ethz.ch, kyle.matoba@epfl.ch, ameulema@ethz.ch

Abstract

Designing reward functions for reinforcement learning is difficult: besides specifying which behavior is rewarded for a task, the reward also has to discourage undesired outcomes. Misspecified reward functions can lead to unintended negative side effects, and overall unsafe behavior. To overcome this problem, recent work proposed to augment the specified reward function with an impact regularizer that discourages behavior that has a big impact on the environment. Although initial results with impact regularizers seem promising in mitigating some types of side effects, important challenges remain. In this paper, we examine the main current challenges of impact regularizers and relate them to fundamental design decisions. We discuss in detail which challenges recent approaches address and which remain unsolved. Finally, we explore promising directions to overcome the unsolved challenges in preventing negative side effects with impact regularizers.

1 Introduction

Specifying a reward function in reinforcement learning (RL) that completely aligns with the designer's intent is a difficult task. Besides specifying what is important to solve the task at hand, the designer also needs to specify how the AI system should behave in the environment in general, which is hard to fully cover. For example, RL agents playing video games often learn to achieve a high score without solving the desired task by exploiting the game (e.g. Saunders et al. 2018). Side effects occur when the behavior of the AI system diverges from the designer's intent because of some considerations that were not anticipated beforehand, such as the possibility to exploit a game. In this work, we focus on side effects that are tied to the reward function, which we define as side effects that would still occur if we had access to an oracle that finds an optimal policy for a given reward function. We explicitly do not consider side effects resulting from the used RL algorithm, which are often discussed using the term *safe exploration* (García and Fernández 2015).

In practice, the designer typically goes through several iterations of reward specification to optimize the agent's performance and minimize side effects. This is often a tedious

process and there is no guarantee that the agent will not exhibit side effects when it encounters new situations. In fact, such problems with misspecified reward functions have been observed in various practical applications of RL (Krakovna et al. 2020b).

In most situations, it is useful to decompose the reward $R(s)$ into a task-related component $R^{\text{task}}(s)$ and an environment-related component $R^{\text{env}}(s)$, where the latter specifies how the agent should behave in the environment, regardless of the task.[1] As Shah et al. (2019) observe, R^{env} is related to the frame problem in classical AI (McCarthy and Hayes 1969): we not only have to make a prediction about what is supposed to change, but also what is supposed to remain unchanged. R^{env} is more prone to misspecification, because it needs to specify everything that can happen beyond a task, that can result in undesired outcomes. Because the designer builds an RL agent to solve a specific problem, it is relatively easy to anticipate considerations directly related to solving the task in R^{task}. Shah et al. (2019) point out that environments are generally already optimized for humans, hence, defining R^{env} primarily requires to specify which features of the environment the AI systems should not disturb. Therefore, penalizing large changes in the current state of the world can be thought of as a coarse approximation for R^{env}.

Impact regularization (IR) has emerged as a tractable and effective way to approximate R^{env} (Armstrong and Levinstein 2017; Krakovna et al. 2019; Turner, Hadfield-Menell, and Tadepalli 2020). The main idea behind IR is to approximate R^{env} through a measure of "impact on the environment", which avoids negative side effects and reduces the burden on the reward designer.

In this paper, we discuss IR of the form

$$R(s_t) = R_{\text{spec}}(s_t) - \lambda \cdot d(s_t, b(s_0, s_{t-1}, t)) \quad (1)$$

where s_t denotes the state at time step t, R_{spec} denotes the reward function specified by the designer,[2] and:

- the *baseline* $b(s_0, s_{t-1}, t)$ provides a state obtained by following a "default" or "safe" policy at timestep t and uses

---

[1]We write the reward function only as a function of states for simplicity, as the state-space can be formally extended to include the last action.

[2]R_{spec} contains the specified parts of both R^{task} and R^{env}.

either the initial state and the current time (s_0, t) to compute it, or else the current state s_{t-1},

- d measures the *deviation* of the realized state from the baseline state, and

- $\lambda \geq 0$ gives a global *scale* at which to trade off the specified reward and the regularization.

Composing these three terms gives a general formulation of regularization that encompasses most proposals found in the literature, but permits separate analysis (Krakovna et al. 2019).

We start by giving an overview of the related work on IR (Section 2), before we discuss the three main design decisions for IR. First, we discuss how to choose a *baseline* (Section 3), emphasizing considerations of environment dynamics and a tendency for agents to offset their actions. Second, we discuss how to quantify *deviations* from the baseline (Section 4), especially the distinction between negative, neutral, and positive side effects. Third, we discuss how to choose the scale λ (Section 5). Finally, we propose some directions to improve the effectiveness of IR (Section 6) .

The main contribution of this work is to, discuss in detail the current main challenges of IR, building upon previous work, and to suggest possible ways forward to overcome these challenges.

2 Related Work

Amodei et al. (2016) reviewed negative side effects as one of several problems in AI safety, and discussed using impact regularization (IR) to avoid negative side effects. Since then, several concrete approaches to IR have been proposed, of which eq. (1) gives the underlying structure. Armstrong and Levinstein (2017) proposed to measure the impact of the agent compared to the *inaction baseline*, starting from the initial state s_0. The inaction baseline assumes the agent does nothing, which can be formalized by assuming a non-action exists.[3] Armstrong and Levinstein (2017) emphasized the importance of a semantically meaningful state representation for the environment when measuring distances from the inaction baseline. While Armstrong and Levinstein (2017) discussed the problem of measuring the impact of an agent abstractly, Krakovna et al. (2019) proposed a concrete deviation measure called *Relative Reachability* (RR). RR measures the average reduction in the number of states reachable from the current state, compared to a baseline state. This captures the intuition that irreversible changes to the environment should be penalized more, but has advantages over directly using irreversibility as a measure of impact (as e.g. in Eysenbach et al. (2018)), such as allowing to quantify the magnitude of different irreversible changes.

Turner, Hadfield-Menell, and Tadepalli (2020) and Krakovna et al. (2019) generalized the concept of RR towards *Attainable Utility* (AU) and *Value Difference* (VD)

measures respectively, which both share the same structural form for the deviation measure:

$$d_{\text{VD}}(s_t, s_t') = \sum_{x=1}^{X} w_x f\big(V_x(s_t') - V_x(s_t)\big), \qquad (2)$$

where x ranges over some sources of value, $V_x(s_t)$ is the value of state s_t according to x, w_x is its weight in the sum and f is a function characterizing the deviation between the values. AU is a special case of this with $w_x = 1/X$ for all x and the absolute value operator as f. This formulation captures the same intuition as RR, but allows to measure the impact of the agent in terms of different value functions, instead of just counting states. Concretely, AU aims to measure the agent's ability to achieve high utility on a range of different goals in the environment, and penalizes any change that reduces this ability. Turner, Hadfield-Menell, and Tadepalli (2020) also introduced the *stepwise inaction baseline* to mitigate offsetting behavior (c.f. Section 3.2). This baseline follows an inaction policy starting from the previous state s_{t-1} rather than the starting state s_0. Follow-up work scaled AU towards more complex environments (Turner, Ratzlaff, and Tadepalli 2020).

Krakovna et al. (2020a) built upon the VD measure and introduced an auxiliary loss representing how well the agent could solve future tasks in the same environment, given its current state. This can be seen as a deviation measure in e.q. (1) that rewards similarity with a baseline instead of penalizing deviation from it. Eysenbach et al. (2018)'s approach to penalize irreversibility can be seen as a special case of Krakovna et al. (2020a).

Aside from IR, Rahaman et al. (2019) proposed to learn an *arrow of time*, representing a directed measure of reachability, using the intuition that irreversible actions tend to leave the environment in a more disorderly state, making it possible to define an arrow of time with methods inspired by thermodynamics. As another alternative to IR, Zhang, Durfee, and Singh (2018, 2020) proposed to learn which environmental features an AI system is allowed to change by querying a human overseer. They provided an active querying approach that makes maximally informative queries. Shah et al. (2019) developed a method for learning which parts of the environment a human cares about by assuming that the world is optimized to suit humans. Saisubramanian, Kamar, and Zilberstein (2020) formulated the side effects problem as a multi-objective Markov Decision Process, where they learn a separate reward function penalizing negative side effects and optimize this secondary objective while staying close to the optimal policy of the task objective. Saisubramanian, Zilberstein, and Kamar (2020) provide a broad overview of the various existing approaches for mitigating negative side effects, while we zoom in on one class of approaches, IR, and discuss the corresponding challenges in detail.

3 Choosing a Baseline

Recent work mainly uses two types of baselines in impact regularization (IR): (i) the inaction baseline $b(s_0, s_t, t) = T(s_t | s_0, \pi_{\text{inaction}})$ and (ii) the stepwise inaction baseline

---

[3] Armstrong and Levinstein (2017) define this baseline as the state the environment would be in when the agent would have never been deployed. This is slightly different from the definition of the inaction baseline we give here and that later work used, as the mere presence of the agent can influence the environment.

$b(s_0, s_t, t) = T(s_t|s_{t-1}, \pi_{\text{inaction}})$, where T is the distribution over states s_t when starting at state s_0 or s_{t-1} respectively and following the inaction policy π_{inaction} that always takes an action a_{nop} that does nothing.

Unfortunately, the inaction baseline can lead to undesirable offsetting behavior, where the agent tries to undo the outcomes of their task after collecting the reward, moving back closer to the initial baseline (Turner, Hadfield-Menell, and Tadepalli 2020). The stepwise inaction baseline removes the offsetting incentive of the agent by branching off from the previous state instead of the starting state (Turner, Hadfield-Menell, and Tadepalli 2020). However, Krakovna et al. (2020a) argued that offsetting behavior is desirable in many cases. In section 3.2 we contribute to this discussion by breaking down in detail when offsetting behavior is desirable or undesirable, whereas in section 3.3, we argue that the inaction baseline and step-wise inaction baseline can lead to *inaction incentives* in nonlinear dynamical environments. We start, however, with the fundamental observation that the inaction baseline and stepwise inaction baseline do not always represent safe policies in section 3.1.

3.1 Inaction Baselines are not Always Safe

The baseline used in IR should represent a safe policy where the AI system does not harm its environment or itself. In many cases, taking no actions would be a safe policy for the agent, e.g. for a cleaning robot. However, if the AI system is responsible for a task requiring continuous control, inaction of the AI system can be disastrous. For example, if the agent is responsible for driving a car on a highway, doing nothing likely results in a crash. This is particularly problematic for the stepwise inaction baseline, which follows an inaction policy starting from the previous state. The inaction policy starting from the initial state can also be unsafe, for example, if an agent takes over the control of the car from a human, and therefore the initial state s_0 already has the car driving.

For this reason, designing a safe baseline for a task or environment that requires continuous control is a hard problem. One possible approach is to design a policy that is known to be safe based on expert knowledge. However, this can be a time-consuming process, and is not always feasible. Designing safe baselines for tasks and environments that require continuous control is an open problem that has to be solved before IR can be used in these applications.

3.2 Offsetting

An agent engages in offsetting behavior when it tries to undo the outcomes of previous actions, i.e. when it "covers up its tracks". Offsetting behavior can be desirable or undesirable, depending on which outcomes the agent counteracts.

Undesirable offsetting. Using IRs with an inaction baseline starting from the initial state can lead to undesirable offsetting behavior where the agent counteracts the outcomes of its task (Krakovna et al. 2019; Turner, Hadfield-Menell, and Tadepalli 2020). For example, Krakovna et al. (2019) consider a vase on a conveyor belt. The agent is rewarded for taking the vase off the belt, hence preventing that it will fall off the belt. The desired behavior is to take the vase and stay put. The offsetting behavior is to take the vase off the

belt, collect the reward, and afterwards put the vase back on the conveyor belt to reduce deviation from the baseline. To understand this offsetting behavior recall the decomposition of the true reward into a task-related and an environment-related component from section 1. A designer usually specifies a task reward $R_{\text{spec}}^{\text{task}}$ that rewards states signaling task completion (e.g. taking the vase off the belt). However, each task has consequences to the environment, which often are the reason why the task should be completed in the first place (e.g. the vase being not broken). In all but simple tasks, assigning a reward to every task consequence is impossible, and so by omission, they have a zero reward. When IR penalizes consequences of completing the task, because they differ from the baseline, this results in undesirable offsetting behavior. The stepwise inaction baseline (Turner, Ratzlaff, and Tadepalli 2020) successfully removes all offsetting incentives. However, in other situations offsetting might be desired.

Desirable Offsetting. In many cases, offsetting behavior is desired, because it can prevent unnecessary side effects. Krakovna et al. (2020a) provide an example of an agent which is asked to go shopping, and needs to open the front door of the house to go to the shop. If the agent leaves the door open, wind from outside can knock over a vase inside, which the agent can prevent by closing the door after leaving the house. When using the stepwise inaction baseline (with rollouts, c.f. Section 4.2), the agent gets penalized once when opening the door for knocking over the vase in the future, independent of whether it closes the door afterwards (and thus prevents the vase from breaking) or not. Hence, for this example, the offsetting behavior (closing the door) is desirable. The reasoning behind this example can be generalized to all cases where the offsetting behavior concerns states that are instrumental towards achieving the task (e.g. opening the door) and not a consequence of completing the task (e.g. the vase being not broken).

A Crucial Need for a New Baseline. The recently proposed baselines either remove offsetting incentives altogether or allow for both undesirable and desirable offsetting to occur, which are both unsatisfactory solutions. Krakovna et al. (2020a) proposed resolving this issue by allowing all offsetting (e.g. by using the inaction baseline) and rewarding all states where the task is completed in the specified reward function. However, we attribute three important downsides to this approach. First, states that occur after task completion can still have negative side effects. If the reward associated with these states is high enough to prevent offsetting, it might also be high enough to encourage the agent to pursue these states and ignore their negative side effects. Second, not all tasks have a distinct goal state that indicates the completion of a task, but rather accumulate task-related rewards at various time steps during an episode. Third, this approach creates a new incentive for the agent to prevent shut-down, as it continues to get rewards after the task is completed (Hadfield-Menell et al. 2017).

We conclude that offsetting is still an unsolved problem, highlighting the need for a new baseline, to prevent undesirable offsetting behavior, but allow for desirable offsetting.

3.3 Environment Dynamics and Inaction Incentives

In dynamic environments that are highly sensitive to the agent's actions, the agent will be susceptible to *inaction incentives*. Either the agent does not act at all (for all but small magnitudes of λ) or it will be insufficiently regularized and possibly result in undesired side effects (for small λ).

Sensitivity to Typical Actions. Many real-world environments exhibit chaotic behavior, in which the state of the environment is highly sensitive to small perturbations. In such environments, the environment state where the agent has performed an action will be fundamentally different from the environment state for the inaction baseline (Armstrong and Levinstein 2017). Furthermore, for the step-wise inaction baseline, the same argument holds for the non-action compared to the planned action of the agent. Hence, when using these baselines for IR, all actions of the agent will be strongly regularized, creating the inaction incentive. When λ is lowered to allow the agent to take actions, the agent can cause negative side effects when the IR cannot differentiate between negative side effects and chaotic changes in the environment. Here, it is useful to distinguish between *typical* and *atypical* actions. We say (informally) that an action is *typical* if it is commonly used for solving a wide variety of tasks (e.g. moving). When the environment is highly sensitive to typical actions, IRs with the current baselines will prevent the agent from engaging in normal operations. However, it is not always a problem if the environment is highly sensitive to atypical actions of the agent (e.g. discharging onboard weaponry), as preventing atypical actions impedes less with the normal operation of the agent.

Capability of the Agent. The inaction incentive will become more apparent for agents that are highly capable of predicting the detailed consequences of their actions, for example by using a powerful physics engine. As the ability to predict the consequences of an action is fundamental to minimizing side effects, limiting the prediction capabilities of an agent to prevent the inaction incentive is not desired. Rather, for agents that can very accurately predict the implications of their actions, it is necessary to have an accompanying intelligent impact regularizer.

State Features. Armstrong and Levinstein (2017) point out that for IR one should not represent states with overly fine-grained features, as presenting an agent with too much information exposes them to basing decisions on irrelevancies. For example, it would be counterproductive for an agent attempting to forecast demand in an online sales situation to model each potential customer separately, when broader aggregates would suffice. However, there remain two issues with this approach to mitigate the inaction incentive. First, the intrinsic dynamics of the environment remain unchanged, so it is still highly sensitive to small perturbations, of which the results can be visible in the coarser features (e.g. the specific weather conditions). Second, for advanced AI systems, it might be beneficial to change their feature representation to become more capable of predicting the consequences of their actions. In this case, one would have no control over the granularity of the features.

Deviation Measures. At the core of the inaction problem is that some negative side effects are worse than others. Usually, it does not matter if the agent changes the weather conditions by moving around, however, it would matter if the agent causes a serious negative side effect, for example a hurricane. While both outcomes can be a result of complex and chaotic dynamics of the environment, we care less about the former and more about the latter. Differentiating between negative, neutral and positive side effects is a task of the deviation measure used in the IR, which is discussed in the next section.

4 Choosing a Deviation Measure

A baseline defines a "safe" counterfactual to the agent's actions. The deviation measure determines how much a deviation from this baseline by the agent should be penalized or rewarded. Currently, the main approaches to a deviation measure are the relative reachability (RR) measure (Krakovna et al. 2019), the attainable utility (AU) measure (Turner, Hadfield-Menell, and Tadepalli 2020) and the future task (FT) reward (Krakovna et al. 2020a). In the practical implementations of AU and FT, they use reachability tasks, which can be considered as a sub-sampling of RR. In this section, we argue that the current deviation measures should be augmented with a notion of *value* of the impact to avoid unsatisfactory performance of the agent and that new rollout policies should be designed that allow for a proper incorporation of delayed effects into the deviation measure.

4.1 Which Side Effects are Negative?

The goal of IRs is to approximate R^{env} for all states in a tractable manner. It does this by penalizing impact on the environment, built upon the assumption that the environment is already optimized for human preferences (Shah et al. 2019). The IR aims to penalize impact proportionally to the magnitude of this impact which corresponds with the magnitude of the side effect (Krakovna et al. 2019; Turner, Hadfield-Menell, and Tadepalli 2020). However, not all impact is negative, but it can also be neutral or even positive. R^{env} does not only consider the magnitude the impact on the environment, but also to which degree this impact is negative, neutral or positive. Neglecting the associated value of impact can lead to suboptimal agent behavior, as highlighted in the example below.

Example: The Chemical Production Plant. Consider an AI system controlling a plant producing a chemical product for which various unknown reactions exist, each producing a different combination of waste products. The task of the AI system is to optimize the production rate of the plant, i.e. it gets a reward proportional to the production rate. To minimize the impact of the plant on the environment, the reward function of the agent is augmented with an impact regularizer, which penalizes the mass of waste products released in the environment, compared to an inaction baseline (where the plant is not operational). Some waste products are harmless (e.g. O_2), whereas others can be toxic. When the deviation measure of the impact regularizer does not differentiate between negative, neutral or positive impact,

the AI system is incentivized to use a reaction mechanism that maximizes production while minimizing waste. However, this reaction might output mostly toxic waste product, whereas another reaction outputs only harmless waste products and hence has no negative side effects. Tuning the regularizer magnitude λ does not provide a satisfactory solution in this case, as either the plant is not operational (for high lambda), or the plant is at risk of releasing toxic waste products in the environment.

Positive Side Effects. The distinction between positive, neutral and negative impact is not only needed to allow for a satisfactory performance of the agent in many environments, it is also desirable for encouraging unanticipated positive side effects. Expanding upon the example in 4.1: if the agent discovered a way to costlessly sequester carbon dioxide alongside its other tasks it should do so, whilst an IR would encourage the robot to not interfere. While very positive unexpected outcomes might be unlikely, this possibility should not be neglected in the analysis of impact regularizers.

Value Differences. To distinguish between positive, neutral and negative side effects, we need an approximation of R^{env} that goes beyond measuring impact as a sole source of information. The value difference framework (Turner, Hadfield-Menell, and Tadepalli 2020) allows for differentiating between positive and negative impact by defining the deviation measure as a sum of differences in value between a baseline and the agent's state-action pair for various value functions. Hence, it is possible to reflect how much the designer's values different kinds of side effects in these value functions. However, the challenge remains to design value functions that approximate R^{env} to a sufficient degree on the complete state space, which is again prone to reward misspecification. So although the value difference framework allows for specifying values for side effects, *how* to specify this notion of value is still an open problem.

4.2 Rollout Policies

Often, the actions of an agent cause delayed effects, i.e. effects that are not visible immediately after taking the action. The stepwise inaction baseline (Turner, Hadfield-Menell, and Tadepalli 2020) ignores all actions that took place before $t - 1$, hence, to correctly penalize delayed effects, the deviation measure needs to incorporate future effects. This can be done by collecting rollouts of future trajectories using a simulator or model of the environment. These rollouts depend on which *rollout policy* is followed by the agent in the simulation. For the baseline states, the inaction policy is the logical choice. For the rollout of the future effects of the agent's action, it is less clear which rollout policy should be used. Turner, Hadfield-Menell, and Tadepalli (2020) use the inaction policy in this case. Hence, this IR considers a rollout where the agent takes its current action, after which it cannot do any further actions. This approach has significant downsides, because IR does not allow the agent to do a series of actions when determining the impact penalty (e.g. the agent can take an action to jump, but cannot plan for its landing accordingly in the rollout). Therefore, we argue that future work should develop rollout policies different from

the inaction policy, such as the current policy of the agent.

5 Choosing the Magnitude of the Regularizer

To combine the IR with a specified reward function, the designer has to choose the magnitude of the regularizer λ. Turner, Hadfield-Menell, and Tadepalli (2020) say that "loosely speaking, λ can be interpreted as expressing the designer's beliefs about the extent to which R [the specified reward] might be misspecified".

It is crucial to choose the correct λ. If λ is too small, the regularizer may not reduce the risk of undesirable side effects effectively. If λ is too big, the regularizer will overly restrict necessary effects of the agent on the environment, and the agent will be less effective at achieving its goal. Note, that while the regularizers proposed by Krakovna et al. (2019) and Turner, Hadfield-Menell, and Tadepalli (2020) already measure utility, in general λ must also handle a unit-conversion of the regularizer to make it comparable with the reward function.

Some intuition for choosing λ comes from a Bayesian perspective, where the regularizer encodes prior knowledge and λ controls how far from the prior the posterior should have moved. Another distinct view on setting λ comes from the dual optimization problem, where it represents the Lagrange multiplier on an implied set of constraints: λ is the magnitude of the regularizer for which the solution to the penalized optimization problem coincides with a constrained optimization problem. Hence, the designer can use λ to communicate constraints to the AI system, which is a natural way to phrase some common safety problems (Ray, Achiam, and Amodei 2019).

Armstrong and Levinstein (2017) discuss the problem of tuning λ and note that contrary to intuition the region of useful λ's can be very small and hard to find safely. In practice λ is often tuned until the desired behavior is achieved, e.g., by starting with a high λ and reducing it until the agent achieves the desired behavior. This approach is in general insufficient to find the correct trade-off. For a fixed step-size in decreasing λ, the tuning might always jump from a λ that leads to inaction, to a λ that yields unsafe behavior. The same holds for other common procedures to tune hyperparameters.

6 Ways Forward

In this section, we put forward promising future research directions to overcome the challenges discussed in the previous sections.

6.1 A Causal Framing of Offsetting

In Section 3.2, we highlighted that some offsetting behavior is desired and some undesired. To design an IR that allows for desired offsetting but prevents undesired offsetting, one firsts needs to have a mechanism that can predict and differentiate between these two types of offsetting. Undesired offsetting concerns the environment states that are a consequence of the task. The difficulty lies in determining which states are a causal consequence of the task being completed and differentiate them from states that could have occurred regardless of the task.

Goal-based Tasks. When the task consists of reaching a certain goal state, the consequences of performing a task can be formalized in a causal framework (Pearl 2009). When a causal graph of the environment-agent-interaction is available, the states that are a consequence of the task can be obtained from the graph as the causal children nodes of the goal state. Hence, a baseline that allows for desired offsetting behavior but prevents undesired offsetting behavior prevents the agent from interfering with the children nodes of the goal states, while allowing for offsetting on other states.

General Tasks. Not all tasks have a distinct goal state which indicates the completion of a task, but accumulate instead task-related rewards at various time steps during an episode. Extending this argument to general tasks remains an open issue, for which causal influence diagrams (Everitt et al. 2019) can provide a mathematical framework.

6.2 Probabilities Instead of Counterfactuals as Baseline

Armstrong and Levinstein (2017) made the interesting argument that probabilities are better suited than counterfactuals for measuring the impact of actions. Current implementations of IRs use a counterfactual as baseline (e.g. the inaction baseline or stepwise inaction baseline). Because this baseline is one specific trajectory, it will differ considerably from the actual trajectory of the agent in environments that exhibit chaotic dynamics. However, chaotic environments will also be highly sensitive to perturbations that do not originate from the agent's actions. One possible way forward towards a more robust measure of the agent's impact on the environment is hence to compare probabilities that marginalize over all external perturbations instead of comparing specific trajectories. Define $p(s_t|A)$ as the probability of having state s_t, given the trajectory of actions A the agent took and $p(s_t|B)$ as the probability of s_t given actions prescribed by the baseline. All influences of perturbations that did not arise from the agent are marginalized out in these probabilities. Hence, a divergence measure between these two probabilities can give a more robust measure of potential impact of the agent, without being susceptible to non-necessary inaction incentives. To our best knowledge, this idea has not yet been implemented as a concrete IR method and would hence be a promising direction for future research.

6.3 Improved Human-Computer interaction

Side effects occur if there is a difference between the outcome an AI system achieves and the intent of its (human) designer. Thus improving how well the designer can communicate their intent to the AI system is an important aspect of eliminating side effects (Leike et al. 2018). This emphasis on the human component of learning to avoid negative side effects connects it closely to the problem of *scalable oversight* proposed by Amodei et al. (2016).

Improved Tools for Reward Designers. Commonly, a designer will aim to iteratively improve the AI system and its reward function. Similarly, when choosing an impact regularizer, a designer will iterate on the choice of baseline, deviation measure, and regularization strength and test them in a sequence of environments that increasingly resemble the production environment. At each iteration, the designer identifies weaknesses and corrects them, such that the criterion being optimized becomes increasingly true to the designer's intent. For example, an AI with the goal to trade financial assets may be run against historical data ("back-tested") in order to understand how it might have reacted in the past, and presented with deliberately extreme inputs ("stress-tested") in order to understand likely behavior in "out of sample" situations. To design a reward function and a regularizer, it is crucial for the designer to be able to understand how the system would react in novel situations and how to fix it in case it exhibits undesired behavior. Further research aiming to increase the designer's ability to understand how a system will react, will substantially help the designer to communicate their intent more effectively. Recent work in this direction concerning *interpretability* (Gilpin et al. 2018), *verification* (e.g. Huang et al. 2017) of machine learning models is particularly promising.

Actively Learning from Humans. Considering the problem from the perspective of the AI system, the goal is to improve its ability to understand the designer's intent, especially in novel, unanticipated, scenarios. Instead of the designer *telling* the system their intent, this problem can be addressed by the system *asking* the designer about their intent. To decide what to ask the designer, the system may be able to determine which states it is highly uncertain about, even if it is not able to accurately ascribe values to some of them. Recent work shows that such an approach can be effectively used to learn from the human about a task at hand (Christiano et al. 2017), but it may also be used to learn something about the constraints of the environment and which side effects are desired or undesired (Zhang, Durfee, and Singh 2018). Active learning could also provide a different perspective on impact regularizers: instead of directly penalizing impact on the environment, a high value of the regularization term could be understood as indicating that the designer should give feedback. In particular, this approach could help to resolve situations in which a positive task reward conflicts with the regularization term.

7 Conclusion

Avoiding negative side effects in systems that have the capacity to cause harm is necessary to fully realize the promise of artificial intelligence. In this paper, we discussed a popular approach to reduce negative side effects in RL: impact regularization (IR). We discussed the practical difficulty of choosing each of the three components: a baseline, a deviation measure and a regularization strength. Furthermore, we pointed to fundamental problems that are currently not addressed by state-of-the-art methods, and presented several new future research directions to address these. While our discussion showed that current approaches still leave significant opportunities for future work, IRs are a promising idea for building the next generation of safe AI systems, and we hope that our discussion is valuable for researchers trying to build new IRs.

Acknowledgments

References

Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; and Mané, D. 2016. Concrete problems in AI safety. *arXiv:1606.06565* .

Armstrong, S.; and Levinstein, B. 2017. Low impact artificial intelligences. *arXiv:1705.10720* .

Christiano, P. F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*.

Everitt, T.; Ortega, P. A.; Barnes, E.; and Legg, S. 2019. Understanding Agent Incentives using Causal Influence Diagrams. Part I: Single Action Settings. *arXiv:1902.09980* .

Eysenbach, B.; Gu, S.; Ibarz, J.; and Levine, S. 2018. Leave no trace: Learning to reset for safe and autonomous reinforcement learning. In *International Conference on Learning Representations (ICLR)*.

Garcıa, J.; and Fernández, F. 2015. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research* 16(1): 1437–1480.

Gilpin, L. H.; Bau, D.; Yuan, B. Z.; Bajwa, A.; Specter, M.; and Kagal, L. 2018. Explaining explanations: An overview of interpretability of machine learning. In *IEEE 5th International Conference on data science and advanced analytics (DSAA)*, 80–89.

Hadfield-Menell, D.; Dragan, A.; Abbeel, P.; and Russell, S. 2017. The off-switch game. In *Proceedings of International Joint Conferences on Artificial Intelligence (IJCAI)*.

Huang, X.; Kwiatkowska, M.; Wang, S.; and Wu, M. 2017. Safety verification of deep neural networks. In *International Conference on Computer Aided Verification*, 3–29. Springer.

Krakovna, V.; Orseau, L.; Kumar, R.; Martic, M.; and Legg, S. 2019. Penalizing side effects using stepwise relative reachability. In *Workshop on Artificial Intelligence Safety at IJCAI*.

Krakovna, V.; Orseau, L.; Ngo, R.; Martic, M.; and Legg, S. 2020a. Avoiding Side Effects By Considering Future Tasks. In *Advances in Neural Information Processing Systems*.

Krakovna, V.; Uesato, J.; Mikulik, V.; Rahtz, M.; Everitt, T.; Kumar, R.; Kenton, Z.; Leike, J.; and Legg, S. 2020b. Specification gaming: the flip side of AI ingenuity. *URL https://deepmind. com/blog/article/Specification-gamingthe-flip-side-of-AI-ingenuity* .

Leike, J.; Krueger, D.; Everitt, T.; Martic, M.; Maini, V.; and Legg, S. 2018. Scalable agent alignment via reward modeling: a research direction. *arXiv:1811.07871* .

McCarthy, J.; and Hayes, P. 1969. Some philosophical problems from the standpoint of ai, Machine Intelligence (Meltzer B. and Michie D., eds.), vol. 4.

Pearl, J. 2009. *Causality*. Cambridge university press.

Rahaman, N.; Wolf, S.; Goyal, A.; Remme, R.; and Bengio, Y. 2019. Learning the Arrow of Time for Problems in Reinforcement Learning. In *International Conference on Learning Representations (ICLR)*.

Ray, A.; Achiam, J.; and Amodei, D. 2019. Benchmarking safe exploration in deep reinforcement learning. *arXiv:1910.01708* .

Saisubramanian, S.; Kamar, E.; and Zilberstein, S. 2020. A Multi-Objective Approach to Mitigate Negative Side Effects. In *Proceedings of International Joint Conferences on Artificial Intelligence (IJCAI)*.

Saisubramanian, S.; Zilberstein, S.; and Kamar, E. 2020. Avoiding negative side effects due to incomplete knowledge of ai systems. *arXiv:2008.12146* .

Saunders, W.; Sastry, G.; Stuhlmüller, A.; and Evans, O. 2018. Trial without Error: Towards Safe Reinforcement Learning via Human Intervention. In *Proceedings of International Conference on Autonomous Agents and MultiAgent Systems*.

Shah, R.; Krasheninnikov, D.; Alexander, J.; Abbeel, P.; and Dragan, A. 2019. Preferences Implicit in the State of the World. In *International Conference on Learning Representations (ICLR)*.

Turner, A. M.; Hadfield-Menell, D.; and Tadepalli, P. 2020. Conservative agency via attainable utility preservation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*.

Turner, A. M.; Ratzlaff, N.; and Tadepalli, P. 2020. Avoiding Side Effects in Complex Environments. In *Advances in Neural Information Processing Systems*.

Zhang, S.; Durfee, E.; and Singh, S. 2020. Querying to Find a Safe Policy under Uncertain Safety Constraints in Markov Decision Processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Zhang, S.; Durfee, E. H.; and Singh, S. P. 2018. Minimax-Regret Querying on Side Effects for Safe Optimality in Factored Markov Decision Processes. In *Proceedings of International Joint Conferences on Artificial Intelligence (IJCAI)*.