DOI: 10.1609/aaai.12028

ARTICLE

Avoiding Negative Side Effects due to Incomplete Knowledge of AI Systems

Sandhya Saisubramanian[1] | **Shlomo Zilberstein**[2] | **Ece Kamar**[3]

[1] Oregon State University

[2] University of Massachusetts

[3] Microsoft Research

Correspondence
Sandhya Saisubramanian, Oregon State University, OR, USA
Email: saisubrs@oregonstate.edu

Abstract

Autonomous agents acting in the real-world often operate based on models that ignore certain aspects of the environment. The incompleteness of any given model – handcrafted or machine acquired – is inevitable due to practical limitations of any modeling technique for complex real-world settings. Due to the limited fidelity of its model, an agent's actions may have unexpected, undesirable consequences during execution. Learning to recognize and avoid such *negative side effects* (NSEs) of an agent's actions is critical to improve the safety and reliability of autonomous systems. Mitigating NSEs is an emerging research topic that is attracting increased attention due to the rapid growth in the deployment of AI systems and their broad societal impacts. This article provides a comprehensive overview of different forms of NSEs and the recent research efforts to address them. We identify key characteristics of NSEs, highlight the challenges in avoiding NSEs, and discuss recently developed approaches, contrasting their benefits and limitations. The article concludes with a discussion of open questions and suggestions for future research directions.

INTRODUCTION

A world populated with intelligent and autonomous systems that simplify our lives is gradually becoming a reality. These systems are *autonomous* in the sense that they can devise a sequence of actions to achieve some given objectives or goals, without human intervention. Such systems are deeply integrated into our daily lives through various applications such as mobile health monitoring (Sim 2019), intelligent tutoring (Folsom-Kovarik, Sukthankar, and Schatz 2013), self-driving cars (Zilberstein 2015), and clinical decision making (Bennett and Hauser 2013). This broad deployment brings along new challenges and increased responsibility for designers of AI systems, particularly ensuring that these systems operate as expected when deployed in the real-world. Despite recent advances in artificial intelligence and machine learning, there are no ways to assure that systems will always "do the right thing" when operating in the open world (Lakkaraju et al. 2017).

For example, consider an autonomous vehicle (AV) that was carefully designed and tested for safety aspects such as yielding to pedestrians and conforming to traffic rules. When deployed, the AV may not slow down when driving through puddles and splash water on nearby pedestrians. Another documented example of undesirable behavior in AVs is the vehicle swerving left and right multiple times to localize itself for active lane keeping. During this process, the vehicle rarely prompted the driver to take control (Insurance Institute for Highway Safety 2018). This behavior, especially on curvy and hilly roads, can startle the driver or cause panic.

Undesirable behaviors may occur even when performing relatively simple tasks. For example, robot vacuum cleaners are becoming increasingly popular and they have a simple task – to remove dirt from the floor. A robot

 © 2021, Association for the Advancement of Artificial Intelligence. All rights reserved.

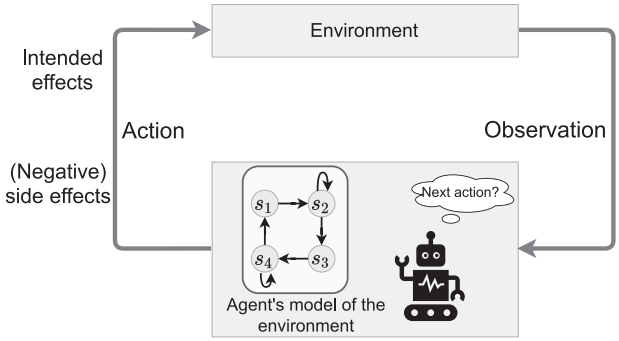

FIGURE 1 Negative side effects of an agent's behavior

vacuum cleaner in Florida ran over animal feces in the house and continued its cleaning cycle, smearing the mess around the house (Solon 2016). In an extreme case in South Korea, a robot vacuum cleaner locked into the hair of a woman who was sleeping on the floor, mistaking her hair for dust (McCurry 2015).

A key factor affecting an agent's performance is its knowledge of the environment in which it is situated. In these examples, the agent was performing its task, perhaps optimally with respect to the information provided to it, but there were serious negative side effects (NSEs) to the agent's actions. In the AV example, driving fast through puddles is optimal when optimizing travel time. The side effects are due to the limited scope of the agent's model, not accounting for the undesirability of splashing water on pedestrians. In practice, it is not feasible to anticipate all possible NSEs and accurately encode them in the model at design time. Due to the practical limitations in data collection and model specification, agents operating in the open world often rely on incomplete knowledge of their target environment which may lead to unexpected, undesirable consequences. Addressing the potential undesirable behaviors of autonomous systems is critical to support long-term autonomy and ensure that a deployed AI system is reliable.

There have been numerous recent studies focused on the broad challenge of building safe and reliable AI systems (Russell, Dewey, and Tegmark 2015; Amodei et al. 2016; Saria and Subbaswamy 2019; Thomas et al. 2019). Here, we examine the particular problem of identifying and mitigating the impacts of undesirable side effects of an agent's actions when operating in the open world. We do not consider system failure or NSEs that result from intentional adversarial attack on the system (Biggio and Roli 2018; Cao et al. 2019).

Negative side effects are undesired effects of an agent's actions that occur in addition to the agent's intended effects when operating in the open world (Figure 1).

NSEs occur because the agent's model and objective function focus on some aspects of the environment but its operation could impact additional aspects of the environment. The value alignment problem studies the unsafe behavior of an agent when its objective does not align with human values (Hadfield-Menell et al. 2016; Russell 2017, 2019). Misaligned systems are more likely to produce NSEs. However, the occurrence of NSEs *does not* necessarily indicate that there is a value alignment problem. NSEs can occur even in settings where the agent optimizes legitimate objectives that align with the user's goals, due to incomplete knowledge and distributional shift. For example, while driving in Boston, AVs that are programmed to not run into obstacles were stopped by the local breed of unflappable seagulls standing on the street (Coren 2018). Not running into obstacles is well-aligned with the users' intentions and objectives, but there are side effects because the agent lacks knowledge that it can edge to startle the birds and then continue driving. In fact, such knowledge was later added to the system to resolve the problem. In addition, some systems may cause unavoidable NSEs that cannot be mitigated. While the side effects may be undesirable, the user may accept the system as is, once they learn about it and recognize that the side effects are unavoidable. In such cases, we cannot say that there is a value alignment problem, even though the NSEs may occur.

Certainly, some NSEs could be anticipated or detected during system development and appropriate mechanisms to mitigate their impacts could be implemented prior to deployment. This article focuses on NSEs that are discovered when the system is deployed, due to a variety of factors such as unanticipated domain characteristics, unanticipated consequences of system or software upgrade, or cultural differences among the target user and development team. Design decisions that may be innocuous during initial testing may have a significant impact when a system is widely deployed. For example, the issue of a Roomba locking into the hair of a person lying on the floor emerged only after the system was deployed in Asia. Overcoming NSEs is an emerging area that is attracting increased attention within the AI community (Hibbard 2012; Amodei et al. 2016; Hadfield-Menell et al. 2017; Russell 2017; Zhang, Durfee, and Singh 2018; Krakovna et al. 2019; Shah et al. 2019; Saisubramanian, Kamar, and Zilberstein 2020; Turner, Hadfield-Menell, and Tadepalli 2020).

The severity of NSEs may range from mild to severe impacts. Often, the discussions around the risk of encountering NSEs have highlighted catastrophic events. While these discussions are critical and essential, AI systems in general are carefully designed and tested for such failures before deployment. With the increasing growth in the capabilities and deployment of AI systems, it is equally important to address the NSEs that are not catastrophic but have significant impacts. Such side effects occur more frequently but are often overlooked, particularly when

TABLE 1 Taxonomy of negative side effects

Property	Property values
Severity	Ranges from mild to safety-critical
Reversibility	Reversible or irreversible
Avoidability	Avoidable or unavoidable
Frequency	Common or rare
Stochasticity	Deterministic or probabilistic
Observability	Full, partial, or unobserved
Exclusivity	Prevent task completion or not

the only remedy available is to remove the product from deployment and develop a new version that can avoid the undesired behavior. Hence, providing end users with the tools to identify and mitigate the impacts of NSEs is critical in shaping how users view, interact, collaborate, and trust AI systems (Saisubramanian, Roberts, and Zilberstein 2021).

The rest of this article identifies key characteristics of NSEs, highlights the challenges in overcoming NSEs, and discusses the recent research progress in this area. To promote a better understanding of the prevalence of NSEs and to provide common test cases for the research community, we have created a public repository that allows AI researchers to report new cases. We conclude the article with a discussion of open questions to encourage future research in this area.

TAXONOMY OF NEGATIVE SIDE EFFECTS

We introduce a taxonomy of NSEs, as outlined in Table 1. Understanding the characteristics of NSEs helps design better solution approaches to detect and mitigate their impacts in deployed systems.

Severity: The severity of NSEs ranges from mild side effects that can be largely ignored to safety-critical failures that require suspension of the system deployment. Safety-critical side effects are typically addressed by redesigning the model and hence require extensive evaluation before redeployment. An example of a safety-critical NSE is an AV failing to detect a construction worker's hand gestures (Crane, Logue, and Pilz 2017). We conjecture that many NSEs lie in the middle with significant impacts that require attention, but not sufficiently critical to suspend the service. An AV that does not slow down when going through puddles can cause significant impacts, but those are unlikely to be considered sufficiently critical to rollback its deployment, particularly if mechanisms are provided to mitigate the negative impacts. Addressing such NSE without suspension of service requires agent adaptation and online planning.

Reversibility: Side effects are reversible if the impact can be reversed or negated, either by the agent causing it or via external intervention. For example, breaking a vase is an irreversible side effect, regardless of the agent's skills (Amodei et al. 2016). Side effects such as leaving marks on a wall can be fixed by repainting it, but the agent may require external assistance to achieve that.

Avoidability: In some problems, it may be impossible to avoid the NSEs during the course of the agent's operation to complete its assigned task. This introduces a trade-off between performing agent's assigned task and avoiding the side effects. For example, the side effects of driving through puddles are unavoidable if all roads leading to the destination have puddles. Addressing unavoidable NSE requires a principled approach to balance the trade-off between avoiding side effects and optimizing the completion of the assigned task.

Frequency: The frequency of occurrence of NSEs depends on the environmental conditions and the action plan. Certain NSE may occur rarely, considering all use cases, but may occur frequently for a small subset of cases. A robot pushing a box over a rug may dirty it as an NSE. This is an example of a frequently occurring NSE when the domain of operation is largely covered with a rug. The frequency of occurrence could impact the approach to identify NSEs and the corresponding mitigation approach.

Stochasticity: The occurrence of NSEs may be deterministic or probabilistic. Deterministic NSEs always occur when some action preconditions arise in the open world. Side effects are probabilistic when their occurrence is not certain even when the right preconditions arise. For example, there may be a small probability that a robot may accidentally slide and scratch the wall while pushing a box, but that undesired effect may happen only 20 percent of the times the robot slips.

Observability: The agent's observability of the actual NSE or the conditions that trigger them are generally determined by the agent's state representation and sensory input. The side effects may be fully observable, partially observable, or even unobserved by the agent. Observing a side effect is *different* from identifying or recognizing the impact as a side effect. For example, the agent may observe the scratch it made on the wall but may not be aware that it is undesirable, and as a result may not try to avoid it. Observability is a critical factor when learning to avoid NSEs. When an external authority provides feedback to the agent, it may be sufficient for the agent to observe the conditions that trigger the NSE. However, when an agent may need to identify NSEs on its own, it needs more complex general knowledge about the open world.

Exclusivity: NSEs may prevent the agent from completing its assigned task. This category is relatively easier to identify. Often, however, the side effects negatively impact

the environment without preventing the agent from completing its assigned task. Such side effects are typically difficult to identify at design time. Much of the current research on avoiding NSEs focuses on side effects that do not prevent the agent from completing its current primary task.

CHALLENGES IN AVOIDING NEGATIVE SIDE EFFECTS

The challenges in avoiding NSEs broadly stem from the difficulty in obtaining knowledge about NSE a priori, gathering user preferences to understand their tolerance for side effects, and balancing the potential trade-off between completing the task and avoiding the side effects.

Model imprecision: Agents designed to operate in the open world are either trained in a simulator, or operate based on models created by a designer or generated automatically using data. Regardless of how much effort goes into the system design and how much data is available for training and testing, it is generally infeasible to obtain a perfect description of open-world environments. Practical challenges in model specification, such as the qualification and ramification problems, and computational complexity consideration often cause the agent to reason based on models that do not represent all the relevant details in the open world (Dietterich 2017). Simulators also suffer from this drawback, as they are also built by designers, resulting in mismatches between a simulator and the actual environment (Ramakrishnan et al. 2019). As a result of reasoning with incomplete information, agents may not consistently behave as intended, leading to unexpected and costly errors, or may completely fail in complex settings.

There are three key reasons why the agent may not have prior knowledge about the NSEs of its actions. First, identifying NSEs a priori is inherently challenging. As a result, this information is often lacking in the agent's model. Second, many AI systems are deployed in a variety of settings, which may be different from the environment used in training and testing of the agent. This distributional shift may cause NSE and is difficult to assess during the design process. Third, NSEs in many settings arise due to user preference violation. It is generally difficult to precisely learn or encode human preferences, and account for individual or cultural differences. Techniques such as online model update and policy repair to minimize side effects, and building more realistic simulators (Dosovitskiy et al. 2017) are some of the promising directions to handle NSEs due to model imprecision.

Feedback collection: An agent that is unaware of the side effects of its actions can gather this information through feedback from users or through autonomous exploration and model revisions. Although learning from feedback produces good results in many problems (Lakkaraju et al. 2017; Zhang, Durfee, and Singh 2018; Ramakrishnan et al. 2019; Basich et al. 2020; Saisubramanian, Kamar, and Zilberstein 2020; Zhang, Durfee, and Singh 2020), there are three main challenges in employing this approach in real-world systems. First, the learning process may not be sample efficient or may require feedback in a certain format to be sample efficient, such as correcting the agent policy by providing alternate actions for execution. Feedback collection in general is an expensive process, particularly when the feedback format requires constant human oversight or imposes significant cognitive overload on the user. Second, feedback may be biased or delayed or both, which in turn affects the agent's learning process. Finally, it is generally assumed that the agent uses human-interpretable representations for querying and feedback collection, but there may be mismatches between the models of the agent and human. There are some recent efforts toward addressing the problem of sample efficiency in learning (Wang et al. 2016; Buckman et al. 2018) and investigating the impact of bias in feedback for agent learning (Ramakrishnan et al. 2018; Saisubramanian, Kamar, and Zilberstein 2020). Identifying and evaluating human-interpretable state-action representations for querying humans is largely an open problem.

Managing tradeoffs: When NSEs are unavoidable and interfere with the performance of the agent's assigned task, there is a trade-off between completing the task efficiently and avoiding the NSE. In an extreme case, it may be impossible for the agent to achieve its goal without creating NSEs. How far should an agent deviate from its optimal plan in order to minimize the impacts of NSEs? Balancing this trade-off requires user feedback since it depends on their tolerance for NSEs. This can be challenging when the agent's objective and the side effects are measured in different units.

APPROACHES TO MITIGATE NEGATIVE SIDE EFFECTS

This section reviews the emerging approaches to mitigating the impacts of NSEs. Table 2 summarizes the characteristics of side effects handled by each one of the methods we mention.

Model and policy update

The occurrence of NSEs in a system depends on the agent's trajectory, which is determined by its policy derived using its reasoning model. Hence, a natural approach to mitigate

TABLE 2 Summary of the characteristics of the surveyed approaches to mitigate negative side effects

	Severity	Reversibility	Avoidability	Frequency	Stochasticity	Observability	Exclusivity
(Hadfield-Menell et al. 2017)	–	Irreversible	–	Frequent	Deterministic	–	–
(Zhang, Durfee, and Singh 2018)	–	Irreversible	Avoidable	–	Deterministic	Observable	Non-interfering
(Krakovna et al. 2019)	–	–	–	–	–	Observable	Non-interfering
(Shah et al. 2019)	–	Irreversible	–	Frequent	Deterministic	Observable	Non-interfering
(Zhang, Durfee, and Singh 2020)	–	Irreversible	–	–	Deterministic	Observable	–
(Turner, Hadfield-Menell, and Tadepalli 2020)	–	Irreversible	Avoidable	Frequent	Deterministic	Observable	Non-interfering
(Saisubramanian, Kamar, and Zilberstein 2020)	Not safety-critical	Irreversible	–	Frequent	Deterministic	–	Non-interfering
(Turner, Ratzlaff, and Tadepalli 2020)	–	–	–	Frequent	Deterministic	–	–
(Krakovna et al. 2020)	Not safety-critical	–	–	–	–	Observable	–
(Saisubramanian, Roberts, and Zilberstein 2021)	–	–	–	Frequent	Deterministic	–	Non-interfering

"–" indicates the approach is indifferent to the values of that property. Although some existing works do not explicitly refer to the severity of the side effects they can effectively handle, in general these approaches target side effects that are undesirable and significant, but not safety-critical

NSE is to update the model such that the agent's policy avoids NSE as much as possible. When the side effects are safety-critical, the model update may include significant changes such as redesign of the reward function. Hadfield-Menell et al. (2017) address such a setting where the NSEs occur due to unintentional misspecification of rewards by the designer. It is assumed that the designer prescribes a proxy reward function and the agent is assumed to be *aware* of a possible reward misspecification. The proxy reward function is treated as a set of demonstrations, and the agent learns the intended reward function using approximate solutions for inference. As acknowledged by the authors, this approach is not scalable to large, complex settings.

Redesigning the reward function may degrade the agent's performance with respect to its assigned task or introduce new risks, and hence requires exhaustive evaluation before redeployment. This could be very expensive and likely require suspension of operation until the newly derived policies could be deemed safe for autonomous operation. In problem domains where the side effects are undesirable but not safety-critical, the impact can be minimized by augmenting the agent's model with a penalty function corresponding to NSE. This exploits the reliability of the existing model with respect to the agent's assigned task, while allowing a deployed agent to adjust its behavior to minimize the side effects.

In related work (Saisubramanian, Kamar, and Zilberstein 2020), we describe a multi-objective formulation of this problem with a lexicographic ordering of objectives that prioritizes optimizing the agent's assigned task (primary objective) over minimizing NSE (secondary objective). A slack value to the primary objective determines the maximum allowed deviation from the optimal expected reward of the primary objective so as to minimize side effects. This work considers a setting in which the agent has *no prior knowledge* about the side effects of its actions. Information about NSE is gathered using feedback, which is then encoded by a reward function. The agent may not be able to observe the NSE except for the penalty, which is proportional to the severity of the NSE provided by the feedback mechanism. The model is updated with this learned reward function and an updated policy that avoids NSEs as much as possible, within the allowed slack, is computed. This formulation can hence handle both avoidable and unavoidable NSE. However, this approach is not suitable

for safety-critical consequences since it prioritizes optimizing the completion of the agent's assigned task.

Both these approaches address the side effects associated with the execution of an action, independent of its outcome.

Constrained optimization

Negative side effects occur when an agent alters features in the environment that the user does not expect or desire to be changed. Such side effects can be addressed by constraining the features that can be altered by the agent during its operation. In (Zhang, Durfee, and Singh 2018), the authors consider a setting in which the uncertainty over the desirability of altering a feature is included in the agent's model and considers deterministic side effects that are irreversible, but avoidable. The agent first computes a policy assuming all the uncertain features are "locked" for alteration. If a policy exists, then the agent executes it. If no policy exists, the agent queries the human to determine which features can be altered and recomputes a policy. A regret minimization approach is used to select the top-k features for querying. Recently, the authors extended this approach to identify if NSEs are unavoidable by casting it as a set-cover problem (Zhang, Durfee, and Singh 2020). If the side effects are unavoidable, the agent ceases operation. Therefore, these approaches are not suitable for settings where the agent is expected to alleviate (unavoidable) NSEs to the extent possible, while completing its assigned task.

Minimizing deviations from a baseline

Another class of solution methods defines a penalty function for NSEs as a measure of deviation from a baseline state, based on the features altered. The deviation measure reflects the degree of disruption to the environment caused by the agent's actions. The agent is expected to minimize the disruption while pursuing its goal, thereby mitigating NSEs. In (Krakovna et al. 2019), the authors present a multi-objective formulation with scalarization, with the deviation from baseline state measured using reachability-based metrics. The agent's sensitivity to NSEs can be adjusted by tuning the scalarization parameters. The relative reachability approach (Krakovna et al. 2019) is not straightforward to apply in settings more complex than grid-worlds, as acknowledged by the authors. Furthermore, the resulting performance is sensitive to the metric used to calculate deviations, particularly the choice of baseline state. Different candidates for baseline states have been proposed, such as start state and inaction in a state (Krakovna et al. 2019). These baselines do not consider human preferences and may penalize all side effects. To overcome this, Shah et al. (2019) present a Maximum Causal Entropy approach to infer human preferences from the start state. They assume that an environment is typically optimized for human preferences and the agent can mitigate NSEs by inferring human preferences before it starts acting. This approach, however, requires knowledge about the dynamics of the environment to determine if the environment has been optimized for human preferences or not.

Human-agent collaboration

Approaches such as policy update, constrained optimization, and minimizing deviations from a baseline rely heavily on the fidelity of agent's state representation. In many cases, however, the agent's state representation may only include the features relevant to its assigned task. This limited representation can impact the agent's ability to learn and mitigate NSEs. In recent work (Saisubramanian and Zilberstein 2021), we describe a human-agent team approach that mitigates NSEs via environment shaping. Environment shaping is the process of applying simple modifications to the current environment to make it more agent-friendly and minimize the occurrence of side effects. The agent optimizes its assigned task, unaware of the side effects of its actions. The human mitigates the side effects of the agent through simple reconfigurations of the environment. This approach is applicable to settings where the user can assist the agent actively, beyond providing feedback, and there are one or more agents with limited state representation. This approach is not suitable for environments that are not configurable by the user or when the agent's model and policy are frequently updated.

Accounting for auxiliary objectives and future tasks

Attainable utility (Turner, Hadfield-Menell, and Tadepalli 2020; Turner, Ratzlaff, and Tadepalli 2020) measures the impact of side effects as the shifts in the agent's ability to optimize for auxiliary objectives, generalizing the relative reachability measure. Often, the occurrence of NSEs may not impact the agent's ability to complete its current assigned task but may affect future task completion. To minimize the interference with future tasks, Krakovna et al. (2020) present an approach that provides the agent an auxiliary reward for preserving agent ability to perform future tasks in the environment. These approaches assume

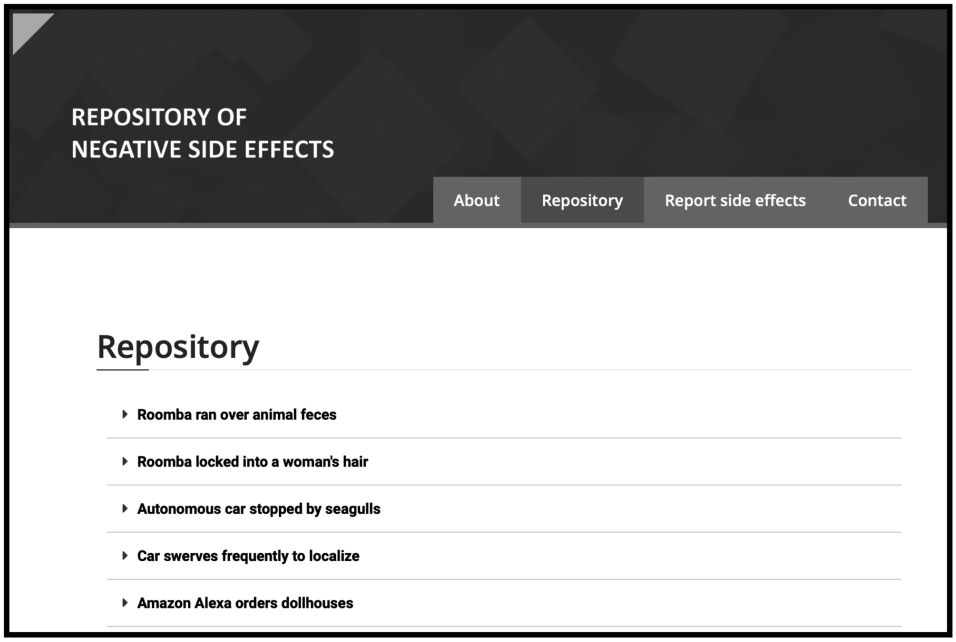

FIGURE 2 A public repository of negative side effects

that the agent's state representation is sufficient to calculate the deviations and are therefore not directly applicable to settings with mismatches between the agent's state representation and the environment.

A REPOSITORY OF NEGATIVE SIDE EFFECTS

Since the problem of NSEs is an emerging topic, current research relies on proof-of-concept toy domains for performance evaluation. Moving forward, understanding the occurrence of NSEs in deployed AI systems is necessary for a realistic formulation of the problem and to design effective solution approaches to address it. To that end, we have created a repository of NSEs (Saisubramanian 2020b). This publicly available repository is shown in Figure 2.

This repository contains real-world instances from scientific reports or news articles, identified by us. For each instance, details such as the problem setting in which NSEs were observed, a description of the side effects, location, and date of incident, are provided. We believe this repository will promote a deeper understanding of the problem, provide insights about which assumptions are valid, and facilitate moving beyond simple grid-world type domains as common test cases to evaluate techniques. We invite the readers to contribute to this repository by reporting cases of NSEs of deployed AI systems or laboratory prototypes, based on user experiences, published papers, or media reports, using an online form we provide (Saisubramanian 2020a). Each submission will be reviewed by our team before adding it to the repository.

OPEN QUESTIONS AND FUTURE WORK

Some key open questions and research directions that can further the understanding of NSEs and development of strategies to mitigate their impacts are discussed below.

Negative side effects in multi-agent settings: The existing works have studied the NSEs of a single agent's actions on the environment. In collaborative multi-agent systems, agents work together to optimize performance and may have complementary skills. For example, the NSEs produced by an agent may be reversible by another agent. How can we leverage collaborative multi-agent settings to effectively mitigate NSEs? One solution approach is to devise a joint policy to mitigate the NSEs, in addition to optimizing the utility of the assigned task. The existing rich body of work on cooperative multi-agent systems examines how the intended effects of each agent's actions may affect the other agents when devising a joint policy that maximizes the performance (Pynadath and Tambe 2002; Goldman and Zilberstein 2003; Zhang and Lesser 2007; Ramakrishnan et al. 2019). Extending such frameworks to handle the side effects problem requires knowledge about the NSEs of each agent's actions and how it affects the behavior and rewards of other agents in the environment. External feedback may indicate the occurrence of NSEs as a result of a joint action of the agents. Effectively mitigating the

side effects requires mechanism design for precise identification of the agent whose actions produce these undesirable effects, based on the feedback provided for joint actions.

Addressing side effects in partially observable settings: In partially observable settings, an agent operates based on a belief distribution over the states. The problem is further complicated when the agent has no prior knowledge of the side effects, which may be partially observable or unobserved. How can an agent effectively learn to avoid NSEs in partially observable settings? Due to partial observability, the agent maps the external feedback indicating the occurrence of NSEs to a belief distribution and not an exact state. As a result, a belief distribution may be associated with multiple conflicting feedback. Depending on how the feedback signals are aggregated, different types of agent behavior emerge with varying sensitivity to NSEs.

Combination of side effects: Many AI systems, such as AVs, are comprised of multiple entities that function together to achieve a goal. Each of these entities may contribute to different forms of NSEs. It is likely that multiple forms of NSEs, with varying impacts and severity, co-exist and require different solution techniques to mitigate the overall impact. Reasoning about multiple forms of risks together is a cornerstone in achieving safe AI systems. How to ensure that approaches designed to eliminate one form of side effect do not introduce new risks? One approach is to evaluate the effects of an impact regularizer on other modules in the system that interact with the module of interest. This requires broad background knowledge about the architecture and functionality of each component, which may not be available in systems with black-box components.

Skill discovery to mitigate NSEs: Skill discovery (Konidaris and Barto 2009; Eysenbach et al. 2018) in reinforcement learning allows an agent to discover useful new skills autonomously. High-level skills or *options* are temporally extended courses of actions that generalize primitive actions of an agent. These closed-loop policies speed up planning and learning in complex environments and are generally used in hierarchical methods for reasoning. Exploring the feasibility of *skill discovery for avoiding NSEs* is an interesting direction that could accelerate agent behavior adaptation, especially to avoid side effects during agent exploration. For example, if the agent learns to push a box without scratching the walls or dirtying the rug, this skill is useful in a variety of related settings and enables faster behavior adaptation.

Beyond safety and control: This article has discussed the undesirable side effects in the context of safety and control in embodied autonomous systems. Investigating NSEs of decision-support systems and recommender systems is an important direction for the future. NSEs in these contexts may not be immediate, such as the effect on climate change, human health, or cognitive ability caused by the system's decisions.

AI systems may also suffer from other factors that affect their reliability, such as biases and privacy concerns. Amplifying underlying biases in a system or increased vulnerability to attacks may occur when the system optimizes incorrect or incompletely specified objectives, which can be treated as serious side effects that require entire model redesign. There are growing efforts in the machine learning community to address many forms of biases and to improve the security for safeguarding against adversarial attacks (Kurakin, Goodfellow, and Bengio 2016; Barocas et al. 2017; Gleave et al. 2019; Peng et al. 2019; Galhotra, Saisubramanian, and Zilberstein 2021).

CONCLUSION

This article examines the concept of NSEs of AI systems and offers a comprehensive overview of recent research efforts to address the challenges presented by side effects. In doing so, we aim to advance the general understanding of this nascent but rapidly evolving area. We present a taxonomy of NSEs, discuss the key challenges in avoiding side effects, and summarize the current literature on this topic. This article also presents potential future research directions that are aimed at deepening the understanding of the problem. While some of these issues can be addressed using problem-specific or ad hoc solutions, developing general techniques to identify and mitigate NSEs will facilitate the design and deployment of more robust and trustworthy AI systems.

ACKNOWLEDGEMENTS

This work was supported in part by the Semiconductor Research Corporation under grant #2906.001.

REFERENCES

Amodei, D., C. Olah, J. Steinhardt, P. Christiano, J. Schulman & D. Mane 2016. Concrete problems in AI safety. arXiv: 1606.06565 [cs.AI]. Ithaca, NY: Cornell University Library.

Barocas, S., K. Crawford, A. Shapiro & H. Wallach 2017. "The problem with bias: Allocative versus representational harms in machine learning." In Proceedings of the 9th Annual Conference of the Special Interest Group for Computing, Information and Society. October 2017. Philadelphia, Pennsylvania, USA.

Basich, C., J. Svegliato, K. H. Wray, S. J. Witwicki, J. Biswas & S. Zilberstein 2020. "Learning to optimize autonomy in competence-aware systems." In Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems. May 2020. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.

Bennett, C. C., and K. Hauser. 2013. "Artificial intelligence framework for simulating clinical decision-making: A markov decision process approach." *Artificial Intelligence in Medicine* 57(1): 9–19.

Biggio, B., and F. Roli. 2018. "Wild patterns: Ten years after the rise of adversarial machine learning." *Pattern Recognition* 84: 317–31.

Buckman, J., D. Hafner, G. Tucker, E. Brevdo & H. Lee 2018. "Sample-efficient reinforcement learning with stochastic ensemble value expansion." In Advances in Neural Information Processing Systems, 8224–34. December 2018. Red Hook, NY: Curran Associates, Inc.

Cao, Y., C. Xiao, B. Cyr, Y. Zhou, W. Park, S. Rampazzi & Q. A. Chen et al. 2019. "Adversarial sensor attack on lidar-based perception in autonomous driving." In Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, 2267–8. November 2019. London, UK.

Coren, M. J. 2018. "All the things that still baffle self-driving cars, starting with seagulls." Quartz, September 23.

Crane, D. A., K. D. Logue, and B. C. Pilz. 2017. "A survey of legal issues arising from the deployment of autonomous and connected vehicles." *Michigan Telecommunications and Technology Law Review* 23: 191–320.

Dietterich, T. G. 2017. "Steps toward robust artificial intelligence." AI Magazine. 38: 3–24.

Dosovitskiy, A., G. Ros, F. Codevilla, A. Lopez & V. Koltun 2017. "CARLA: An open urban driving simulator." In Conference on Robot Learning, 1–16. November 2017. Mountain View, California: Proceedings of Machine Learning Research

Eysenbach, B., A. Gupta, J. Ibarz & S. Levine 2018. "Diversity is all you need: Learning skills without a reward function." In International Conference on Learning Representations. May 8, 2018. Vancouver, Canada.

Folsom-Kovarik, J. T., G. Sukthankar, and S. Schatz. 2013. "Tractable POMDP representations for intelligent tutoring systems." *ACM Transactions on Intelligent Systems and Technology* 4(2): 1–22.

Galhotra, S., S. Saisubramanian & S. Zilberstein 2021. "Learning to generate fair clusters from demonstrations." In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. May 2021. New York, NY: Association for Computing Machinery.

Gleave, A., M. Dennis, C. Wild, N. Kant, S. Levine & S. Russell 2019. "Adversarial policies: Attacking deep reinforcement learning." In Proceedings of the 7th International Conference on Learning Representations. May 2019. New Orleans, Louisiana, USA.

Goldman, C. V. & S. Zilberstein 2003. "Optimizing information exchange in cooperative multi-agent systems." In Proceedings of the 2nd International Conference on Autonomous Agents and Multi Agent Systems. July 2003. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.

Hadfield-Menell, D., S. J. Russell, P. Abbeel & A. Dragan 2016. "Cooperative inverse reinforcement learning." In Advances in Neural Information Processing Systems, 3909–17. December 2016. Barcelona, Spain: Curran Associates, Inc.

Hadfield-Menell, D., S. Milli, P. Abbeel, S. J. Russell & A. Dragan 2017. "Inverse reward design." In Advances in Neural Information Processing Systems. December 2017. Long Beach, California, USA: Curran Associates, Inc.

Hibbard, B. 2012. "Avoiding unintended AI behaviors." In International Conference on Artificial General Intelligence, 107–16. December 2012. Oxford, UK: Springer.

Insurance Institute for Highway Safety. 2018. "Reality check: Research, deadly crashes show need for caution on road to full autonomy." *Status Report Newsletter* 53(4): 1–12.

Konidaris, G. & A. G. Barto 2009. "Skill discovery in continuous reinforcement learning domains using skill chaining." In Advances in Neural Information Processing Systems, 1015–23. December 2009. Vancouver, Canada: Curran Associates, Inc.

Krakovna, V., L. Orseau, M. Martic & S. Legg 2019. "Penalizing side effects using stepwise relative reachability." In AI Safety Workshop, IJCAI. August 10-12, 2019. Macao, China.

Krakovna, V., L. Orseau, R. Ngo, M. Martic & S. Legg 2020. "Avoiding side effects by considering future tasks." In Advances in Neural Information Processing Systems. December 2020. Red Hook, NY: Curran Associates, Inc.

Kurakin, A., I. Goodfellow & S. Bengio 2016. "Adversarial examples in the physical world." arXiv: 1607.02533 [cs.CV]. Ithaca, NY: Cornell University Library.

Lakkaraju, H., E. Kamar, R. Caruana & E. Horvitz 2017. "Identifying unknown unknowns in the open world: Representations and policies for guided exploration." In Proceedings of the 31st AAAI Conference on Artificial Intelligence. February 2017. Palo Alto, CA: AAAI Press.

McCurry, J. 2015. "South korean woman's hair 'eaten' by robot vacuum cleaner as she slept." The Guardian, February 8.

Peng, A., B. Nushi, E. Kıcıman, K. Inkpen, S. Suri, and E. Kamar. 2019. "What you see is what you get? the impact of representation criteria on human bias in hiring." In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7(1): 125–34.

Pynadath, D. V. & M. Tambe 2002. "Multiagent teamwork: Analyzing the optimality and complexity of key theories and models." In Proceedings of the 1st International Conference on Autonomous Agents and Multi Agent Systems. July 2002. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.

Ramakrishnan, R., E. Kamar, D. Dey, J. Shah & E. Horvitz 2018. "Discovering blind spots in reinforcement learning." In Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems. July 2018. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.

Ramakrishnan, R., E. Kamar, B. Nushi, D. Dey, J. Shah & E. Horvitz 2019. "Overcoming blind spots in the real world: Leveraging complementary abilities for joint execution." In Proceedings of the 33rd AAAI Conference on Artificial Intelligence. January 2019. Palo Alto, CA: AAAI Press.

Russell, S., D. Dewey, and M. Tegmark. 2015. "Research priorities for robust and beneficial artificial intelligence." *AI Magazine* 36(4): 105–14.

Russell, S. 2017. "Provably beneficial artificial intelligence." In Exponential Life, The Next Step. New York, NY.

Russell, S. 2019. *Human compatible: Artificial Intelligence and the Problem of Control*. London, UK: Penguin.

Saisubramanian, S. & S. Zilberstein 2021. "Mitigating negative side effects via environment shaping." In Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems, 1640–2. May 2021. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.

Saisubramanian, S., E. Kamar & S. Zilberstein 2020. "A multi-objective approach to mitigate negative side effects." In Proceedings of the 29th International Joint Conference on Artificial Intelligence. January 2021. California: IJCAI.

Saisubramanian, S., S. C. Roberts & S. Zilberstein 2021. "Understanding user attitudes towards negative side effects of AI systems." In Extended Abstracts of the CHI Conference on Human Factors in Computing Systems. May 2021. New York, NY: Association for Computing Machinery.

Saisubramanian, S. 2020a. Negative Side Effects Form. https://forms.gle/5MLZ7XMc9FzbDaoW7. Accessed: December 08, 2020.

Saisubramanian, S. 2020b. Negative Side Effects Repository. http://groups.cs.umass.edu/nse/. Accessed: December 08, 2020.

Saria, S. & A. Subbaswamy 2019. "Tutorial: Safe and reliable machine learning." arXiv:1904.07204 [cs.LG]. Ithaca, NY: Cornell University Library.

Shah, R., D. Krasheninnikov, J. Alexander, P. Abbeel & A. Dragan 2019. "Preferences implicit in the state of the world." In Proceedings of the 7th International Conference on Learning Representations. May 2019. New Orleans, Louisiana, USA.

Sim, I. 2019. "Mobile devices and health." *New England Journal of Medicine* 381(10): 956–68.

Solon, O. 2016. "Roomba creator responds to reports of 'poopocalypse': 'we see this a lot'." *The Guardian*, August 15.

Thomas, P. S., B. C. da Silva, A. G. Barto, S. Giguere, Y. Brun, and E. Brunskill. 2019. "Preventing undesirable behavior of intelligent machines." *Science* 366(6468): 999–1004.

Turner, A. M., D. Hadfield-Menell & P. Tadepalli 2020. "Conservative agency via attainable utility preservation." In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York, NY: Association for Computing Machinery.

Turner, A., N. Ratzlaff & P. Tadepalli 2020. "Avoiding side effects in complex environments." In Advances in Neural Information Processing Systems. February 2020. Red Hook, NY: Curran Associates, Inc.

Wang, Z., V. Bapst, N. Heess, V. Mnih, R. Munos, K. Kavukcuoglu & N. de Freitas 2016. "Sample efficient actor-critic with experience replay." arXiv: 1611.01224 [cs.LG]. Ithaca, NY: Cornell University Library.

Zhang, X. & V. Lesser 2007. "Meta-level coordination for solving negotiation chains in semi-cooperative multi-agent systems." In Proceedings of the 6th International Conference on Autonomous Agents and Multi Agent Systems. May 2007. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.

Zhang, S., E. H. Durfee & S. P. Singh 2018. "Minimax-regret querying on side effects for safe optimality in factored markov decision processes." In Proceedings of the 27th International Joint Conference on Artificial Intelligence, 4867–73. July 2018. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.

Zhang, S., E. H. Durfee & S. Singh 2020. "Querying to find a safe policy under uncertain safety constraints in markov decision processes." In Proceedings of the 34th AAAI Conference on Artificial Intelligence. February 2020. Palo Alto, CA: AAAI Press.

Zilberstein, S. 2015. "Building strong semi-autonomous systems." In Proceedings of the 29th AAAI Conference on Artificial Intelligence. January 2015. Palo Alto, CA: AAAI Press.

AUTHOR BIOGRAPHIES

Sandhya Saisubramanian is an Assistant Professor in the School of Electrical Engineering and Computer Science at Oregon State University. She received her B.Tech. in Computer Science from the Pondicherry Engineering College, India, her Master of Computing degree from the National University of Singapore, and her Ph.D. from the College of Information and Computer Sciences at the University of Massachusetts, Amherst. Her research focuses on safe and reliable decision-making in autonomous systems.

Shlomo Zilbertsein is a Professor and Associate Dean for Research and Engagement in the College of Information and Computer Sciences at the University of Massachusetts, Amherst. He received his B.A. in Computer Science from the Technion–Israel Institute of Technology, and his Ph.D. in Computer Science from the University of California, Berkeley. His research in artificial intelligence focuses on the foundations and applications of automated planning, particularly enabling autonomous systems to make decisions while coping with uncertainty, missing information, and limited computational resources.

Ece Kamar is a Senior Principal Research Manager at Microsoft Research, Redmond. She received her B.S. in Computer Science from the Sabanci University, Turkey, and her Ph.D. in Computer Science from Harvard University. Her research focuses on developing AI systems that can function reliably in the open world in collaboration with people.

How to cite this article: Saisubramanian, S., Zilberstein, S., and Kamar, E. 2021. "Avoiding Negative Side Effects due to Incomplete Knowledge of AI Systems." *AI Magazine*, 42: 62–71. https://doi.org/10.1609/aaai.12028