An AI Safety Threat from Learned Planning Models

Toryn Q. Klassen[1,2,3], **Sheila A. McIlraith**[1,2,3], **Christian Muise**[4]

[1]Department of Computer Science, University of Toronto, Toronto, Canada
[2]Vector Institute for Artificial Intelligence, Toronto, Canada
[3]Schwartz Reisman Institute for Technology and Society, Toronto, Canada
[4]School of Computing, Queen's University, Kingston, Canada
toryn@cs.toronto.edu, sheila@cs.toronto.edu, christian.muise@queensu.ca

Abstract

Historically, planning problems have often been constructed by hand, with the domain model and the goal developed together, leading to the model and goal being in harmony in the sense that the goal describes exactly which parts of the modelled state were desired to be changed (and not changed) as a consequence of the execution of the plan. With models learned from data, human goal specifiers may not know all the aspects of the model, nor have spent much time thinking about the real world situation that is being modelled. Also, naive users may expect the goals they specify to be interpreted in a commonsensical way by the automated planning system. These things may lead human goal specifiers to more often create incomplete goal specifications, failing to take into account all the different ways the environment can be changed – the potential side effects of plans. This could threaten safety. However, learned models may in some cases also have the feature of having detailed state representations, affording the opportunity for symbolic planning algorithms to recognize side effects that their human users did not think of, and to help avoid them. We propose in this position paper that researchers in symbolic planning should take up the challenge of developing planning algorithms that can safely deal with underspecified objectives – i.e., with problem goals that fail to specify everything that people want.

1 Introduction

Planning with a model relies on the model – the description of possible states and the transition system, as well as the initial state – being sufficiently faithful to the real world to ensure that a plan is valid in relation to the real world. That is, execution of the plan, starting in the real-world initial state, should lead to the achievement of the goal. Of course, this doesn't always happen, and *execution monitoring* techniques have long been developed to mitigate that (e.g., Fikes, Hart, and Nilsson 1972). However, there is another risk with automated planning that has been less well studied and that may increase with the use of learned planning artefacts – negative side effects arising from plans made for underspecified goals. In this position paper we discuss this risk and how it may be amplified by the use of learned planning models in symbolic planning. We further discuss how the learned models may also provide features that help ameliorate the risk.

The field of AI safety has considered how incomplete objective specifications may lead to undesirable *side effects*, for instance how a robot directed to move somewhere may break a valuable vase that's in the way (Amodei et al. 2016). In this context, a "side effect" is a change made by the AI system's actions that wasn't specified as part of the objective.[1] Amodei et al. wrote that

> [O]bjective functions that formalize "perform task X" may frequently give undesired results, because what the designer really should have formalized is closer to "perform task X subject to common-sense constraints on the environment," or perhaps "perform task X but avoid side effects to the extent possible."

The problem of side effects has attracted recent research interest (e.g., Zhang, Durfee, and Singh 2018; Krakovna et al. 2019, 2020; Turner, Hadfield-Menell, and Tadepalli 2020; Saisubramanian, Kamar, and Zilberstein 2020, 2022; Alizadeh Alamdari et al. 2021, 2022), though with few exceptions (Klassen and McIlraith 2021; Klassen et al. 2022) the problem has been considered in the context of Markov Decision Processes (MDPs) or similar formulations, and often with reinforcement learning (RL).

In this paper, we argue that objective underspecification may also become an important issue for future symbolic planning systems, and propose developing algorithms to deal with it as a challenge to the planning community. The symbolic planning community has devised various more restricted ways of symbolically modelling environments (e.g., STRIPS or FOND planning). Investigating side effects in such restricted settings may allow for finding different, more efficient algorithms. Furthermore, it may be easier to conceptually develop side-effect-related ideas in a simplified setting, which can later be generalized to more complex settings.

Some points that we raise are the following:

- Many plans generate side effects in the real world since actions can have effects beyond those necessary for goal achievement and/or exposed in a model.

[1]Note that this notion of "side effect" may not align with the everyday use of the phrase (Ashton 2022).

- People may create underspecified objectives for planning models, leading to plans with (negative) side effects.
- Furthermore, using learned models may make such underspecified objectives more likely, by changing how much the typical objective (goal) designer knows (about the model, the part of the world being modelled, and how goals are interpreted by the planning system).
- Some learned models may provide large vocabularies (of fluents) that at least allow for representing side effects, which affords opportunities for algorithmically dealing with them, even though the human goal designer didn't refer to them (e.g., in the simplest case, by trying to minimize how many fluents are changed).
- In order for the last point to apply, model-learning algorithms have to find models with sufficiently rich vocabularies – in the case where the vocabularies are not given but are a learned abstraction of lower-level states like images – that is, sufficiently rich to describe not just the sorts of goals people may explicitly state but also possible side effects. So development of such algorithms is also an important component of dealing with side effects.
- What characterizes *negative* side effects? How should they be identified (so they can be avoided)?

We expand on the risk of objective underspecification with learned models in Section 2, reflect on some existing approaches that relate or could be related to side effects in Section 3, and conclude with some research suggestions in Section 4.

2 Side Effects with Learned Models

In this section we consider in more detail why underspecified objectives may be created, how using learned models may increase the risk of that, and also how learned models may allow for creating safer systems. To start, it will be useful to be explicit about the definition of a side effect we have in mind:

Definition 1 (Side effect (informal definition))**.** A side effect of a plan is any change *in the real world* caused by the execution of the plan, that was not prescribed explicitly as part of the goal.

This is similar to Klassen et al.'s Definition 5 (2022), but here we are *not* assuming that the planning model agrees with the real world, so we have emphasized that the definition of a side effect is in terms of change in the real world (similarly to Saisubramanian, Zilberstein, and Kamar (2021)). A plan might have side effects in the real world which an inaccurate model fails to predict.

Note that this definition does not distinguish whether side effects are negative or not. Krakovna et al. (2020) suggested "side effects matter because we may want the agent to perform other tasks after the current task in the same environment," and so proposed an approach for an (RL) agent to avoid interfering with its own ability to act in the future. In more recent work, we have observed that whether a side effect is negative is sometimes difficult to measure objectively and that we should consider the impact of side effects on other agents' abilities and well-being (Klassen and McIlraith 2021; Klassen et al. 2022; Alizadeh Alamdari et al. 2021, 2022). We will discuss that more later.

Undesirable side effects may result from having underspecified goals that fail to include everything that the designer actually wanted. Historically, planning problems have often been constructed by hand, with the domain model and the goal developed together. So the model and goal are typically in harmony in the sense that the goal describes exactly which parts of the modelled state are desired to be changed and the dynamics may only include those action effects that are relevant to achievement of the specified goal. (The model might not agree with the world, though, so there still was some risk.) The use of learned models, which may increase in the future, may lead to changes in how much goal designers typically know. Below we suggest a number of situations that might cause people to produce underspecified objectives, and consider how using a learned model may make some of these situations more likely.

1. The model and real world may agree, but the human goal designer lacks knowledge or awareness that some side effect is possible in the real world, so doesn't design the goal to preclude it. For example, when designing a goal requiring moving a robot, the human may not think about how the robot would break a vase if it drove into it, and so will make a goal that doesn't refer to vases. (If the human had designed the model by hand, they may have been forced to think more carefully about the relevant part of the real world and may have averted this problem.)
2. The human knows how the real world works, and also how the model works, but doesn't fully understand the relation between them – what the fluents in the learned model refer to (for example, how small an object has to be for a predicate called `small` to apply to it). So the human misspecifies the goal (this case might lead to errors other than just underspecification). If the human had also designed the vocabulary, this case would seem much less likely.
3. The human lacks knowledge of the model – they don't know all the fluents that exist in the vocabulary – and so don't know how to encode their complete goal. Again, if the human had designed the vocabulary, this would seem less likely to happen.
4. The model doesn't even have the vocabulary to represent some *side effect*, so the human (whether or not they can anticipate the side effect in the real world) can't make avoiding it part of the goal, at least not in a direct way.
5. The model's dynamics are inaccurate and show some side effect as being impossible, and the human relies on that and doesn't design the goal to preclude it. (In this case, even if the human had made avoiding the side effect part of the goal – perhaps thanks to having real-world knowledge – a plan found for the model might still cause the side effect in the real world, so goal underspecification may not be the biggest problem.)
6. The untrained human may expect the AI system to be able to fill in underspecified objectives in a commonsensical way, taking into account things like social norms

(such as not injuring people) the way another human would. There may be more untrained people designing goals in the future if learned models lead to broader use of automated planning in society. Furthermore, planning goals might be automatically generated from natural language instructions, which might increase the expectation for human-like understanding.

We will not say much more about possible model inaccuracy (involved in point 5 above) in this paper, since we are focusing on underspecified objectives. See the work of Saisubramanian, Zilberstein, and Kamar (2021) for a conception of how inaccurate models relate to negative side effects.

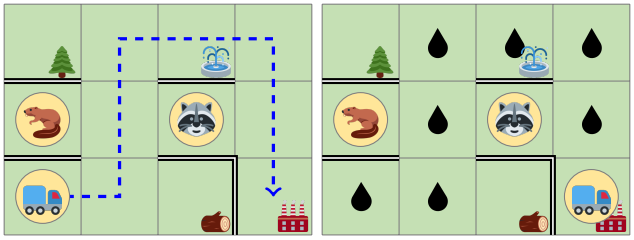
Let's return to the problem of an incomplete vocabulary (point 4 above). Saisubramanian, Zilberstein, and Kamar (2021) also note that "the agent's state representation may only include the features relevant to its assigned task. This limited representation can impact the agent's ability to learn and mitigate [negative side effects]." However, some future learned models may be intended to be general-purpose, and have very large vocabularies. If the learned model is complete enough, such negative side effects (e.g., a vase being broken) may be expressible in the language of the model. So a *feature* of a learned model is that it can expose side effects. Some of these side effects may be irrelevant in relation to the intended objective, and that's why they don't appear in the human goal specification, but others may not be irrelevant and they necessitate consideration. That may help to allow symbolic methods to potentially find ways to avoid negative side effects. Some of the approaches to dealing with side effects discussed in the next section take advantage of this.

3 Approaches to Dealing with Side Effects

In this section we discuss existing work that deals with, or could be related to dealing with, side effects in automated planning.

The problem of avoiding side effects has recently been considered in the context of STRIPS planning by Klassen et al. (2022). In that work, we identify a class of *negative* side effects: effects of the agent's plan that compromise the agency and well being of other agents in the environment. In the example of the Canadian wildlife domain (Figure 1), the robot truck leaves behind a trail of oil which blocks the movement of wildlife (the beaver and raccoon). This creates the possibility of the robot causing negative side effects (from the point of the view of the wildlife). In the domain, there are fluents indicating whether grid cells are contaminated with oil, and the cause and effects of oil contamination are encoded in the transition dynamics. This allows for algorithms operating on the planning model to try to avoid interfering with the other agents, even though the given goal is only for the robot to reach the factory.

Note that considering the ability of other agents to reach goals or follow plans gives a way to identify which side effects are negative (we have also described a version of this in the context of MDPs that considers possible value functions of other agents (Alizadeh Alamdari et al. 2022)). Klassen et al. considered a number of symbolic planning algorithms,



(a) (b)

Figure 1: The Canadian wildlife domain from Klassen et al. (2022). The robot truck's movement leaves behind a trail of leaked oil, so if the robot just goes directly to the factory, it obstructs subsequent movement by the beaver and raccoon. By cleaning a few cells of oil, the robot can do better.

which optimize for different things:

- minimizing how many possible goals are made unreachable for other agents (given a set of possible goal-agent pairs),
- minimizing how many goals are made unreachable for agents following particular policies (given a set of possible goal-policy pairs),
- or minimizing how many fluents are changed (this does not try to identify in any sense which side effects are negative).

These optimization problems are compiled into planning problems with costs. These approaches illustrate how, given an accurate model with a suitable fluent vocabulary (for this example, including fluents relating to oil contamination), it is possible to try to use planning techniques to avoid some side effects. While the Canadian wildlife domain and the other domains Klassen et al. experimented with were not learned from data but handcrafted to include the relevant fluents, similar techniques could find application in future learned models. There still is need for more efficient algorithms, though (and more features, like dealing with action costs).

The above algorithms are non-interactive, in that they do not involve any attempt to interact with humans to gather information about what side effects should be avoided. An alternative approach would be to find plans that involve querying humans about side effects. That sort of approach was used for finding plans for *factored MDPs* (Markov Decision Processes) by Zhang, Durfee, and Singh (2018). In a factored MDP, a state is described in terms of the values of features. Zhang, Durfee, and Singh's approach, which involves asking users about which features are safe to change, also relies on having the relevant features included in the state representation.

Another work that could be viewed as a sort of interactive approach to avoiding side effects is by Nguyen et al. (2012), who considered preference-based planning with incomplete preferences. They proposed having the planning system deal with its lack of knowledge about user preferences by generating a diverse set of plans and having the user pick among

them which plan should be executed. If the user has knowledge about action effects that are not encoded in the model they might, in choosing a plan, even be able to avoid side effects that the model does not predict and can not represent. However, picking a plan may require more human effort than is available in all circumstances.

While not specifically aimed at discovering side effects, the area of model reconciliation aims to convey the relevant aspects of a model to a human user in a variety of ways, all based on a mental model of what the human user understands about the domain (Sreedharan, Kulkarni, and Kambhampati 2022). By viewing the interaction with human users and contingent outcomes, Sreedharan, Chakraborti, and Kambhampati propose a method for gradually reconciling information on a partially understood model of the environment (2018). Zahedi et al. build on this further to focus on the user preferences and how they intersect with the explanation process itself (2019). This line of model reconciliation work pre-supposes that an agent has a complete model of the environment and is conveying this model to a human user. One could imagine this process taking side effects into account for the selection of explanations, but these techniques do not inherently consider this aspect.

4 Conclusion

We have discussed how the use of learned planning models may raise the risks of incomplete goal specifications being used, resulting in plans being found whose executions would have undesirable consequences. We advance that more work is needed on planning algorithms that try to reduce such side effects. The existing literature on avoiding side effects in planning has limitations, in algorithmic scalability or in requiring a lot of human effort (e.g., for evaluating proposed plans). We conclude with some suggestions for future work.

- To minimize human effort it may be useful to incorporate additional information into the planning process, like possible goals of other agents that shouldn't be interfered with (as Klassen et al. (2022) explored). While Klassen et al. manually constructed possible goals in their examples, for real-life problems that sort of information could potentially be learned from data. More generally, construction of general-purpose knowledge bases about typical human preferences and social norms could aid in avoiding side effects, by being used by planning algorithms to augment explicitly given goals.
- Another research idea that might be worth considering would be execution monitoring that kept track not just of whether the goal is still achievable but of what side effects might occur or had occurred.
- Standard planning benchmarks like those in the International Planning Competition (IPC) are not designed to expose safety issues – they assume the state-space representations are complete, and the goals are complete specifications of what is desired with respect to them. We propose that, similarly to how safety-related benchmarks have been developed for reinforcement learning (e.g., Leike et al. 2017), they should be developed for symbolic planning.
- Finally, as we increasingly rely on models that are learned in a data-driven fashion, we feel that increased effort should be spent on learning more effects of an action than just those relevant to achieving a particular objective. In the absence of this, methods that address safety will be inadequate on learned representations.

Acknowledgments

We gratefully acknowledge funding from the Natural Sciences and Engineering Research Council of Canada (NSERC), the Canada CIFAR AI Chairs Program, and Microsoft Research. Finally, we thank the Schwartz Reisman Institute for Technology and Society for providing a rich multi-disciplinary research environment.

References

Alizadeh Alamdari, P.; Klassen, T. Q.; Toro Icarte, R.; and McIlraith, S. A. 2021. Avoiding Negative Side Effects by Considering Others. In *NeurIPS 2021 Workshop on Safe and Robust Control of Uncertain Systems*.

Alizadeh Alamdari, P.; Klassen, T. Q.; Toro Icarte, R.; and McIlraith, S. A. 2022. Be Considerate: Avoiding Negative Side Effects in Reinforcement Learning. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022)*, 18–26.

Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P. F.; Schulman, J.; and Mané, D. 2016. Concrete Problems in AI Safety. *arXiv preprint arXiv:1606.06565*.

Ashton, H. 2022. Defining and Identifying the Legal Culpability of Side Effects using Causal Graphs. In *Proceedings of the Workshop on Artificial Intelligence Safety 2022 (SafeAI 2022) co-located with the Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI2022)*, volume 3087 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Fikes, R.; Hart, P. E.; and Nilsson, N. J. 1972. Learning and Executing Generalized Robot Plans. *Artificial Intelligence*, 3(1-3): 251–288.

Klassen, T. Q.; and McIlraith, S. A. 2021. Planning to Avoid Side Effects (Preliminary Report). In *IJCAI Workshop on Robust and Reliable Autonomy in the Wild (R2AW)*.

Klassen, T. Q.; McIlraith, S. A.; Muise, C.; and Xu, J. 2022. Planning to Avoid Side Effects. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI 2022)*.

Krakovna, V.; Orseau, L.; Martic, M.; and Legg, S. 2019. Penalizing Side Effects using Stepwise Relative Reachability. In *Proceedings of the Workshop on Artificial Intelligence Safety 2019 co-located with the 28th International Joint Conference on Artificial Intelligence, AISafety@IJCAI 2019*, volume 2419 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Krakovna, V.; Orseau, L.; Ngo, R.; Martic, M.; and Legg, S. 2020. Avoiding Side Effects By Considering Future Tasks. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*.

Leike, J.; Martic, M.; Krakovna, V.; Ortega, P. A.; Everitt, T.; Lefrancq, A.; Orseau, L.; and Legg, S. 2017. AI Safety Gridworlds. *arXiv preprint arXiv:1711.09883*.

Nguyen, T. A.; Do, M. B.; Gerevini, A.; Serina, I.; Srivastava, B.; and Kambhampati, S. 2012. Generating diverse plans to handle unknown and partially known user preferences. *Artificial Intelligence*, 190: 1–31.

Saisubramanian, S.; Kamar, E.; and Zilberstein, S. 2020. A Multi-Objective Approach to Mitigate Negative Side Effects. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, 354–361.

Saisubramanian, S.; Kamar, E.; and Zilberstein, S. 2022. Avoiding Negative Side Effects of Autonomous Systems in the Open World. *Journal of Artificial Intelligence Research*, 74: 143–177.

Saisubramanian, S.; Zilberstein, S.; and Kamar, E. 2021. Avoiding Negative Side Effects Due to Incomplete Knowledge of AI Systems. *AI Magazine*, 42(4): 62–71.

Sreedharan, S.; Chakraborti, T.; and Kambhampati, S. 2018. Handling Model Uncertainty and Multiplicity in Explanations via Model Reconciliation. In *Proceedings of the Twenty-Eighth International Conference on Automated Planning and Scheduling, ICAPS 2018*, 518–526. AAAI Press.

Sreedharan, S.; Kulkarni, A.; and Kambhampati, S. 2022. *Explainable Human–AI Interaction: A Planning Perspective*, volume 16 of *Synthesis Lectures on Artificial Intelligence and Machine Learning*. Morgan & Claypool Publishers.

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*, AIES '20, 385–391.

Zahedi, Z.; Olmo, A.; Chakraborti, T.; Sreedharan, S.; and Kambhampati, S. 2019. Towards understanding user preferences for explanation types in model reconciliation. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 648–649. IEEE.

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 the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018*, 4867–4873.