Human-in-the-loop Learning for Safe Exploration through Anomaly Prediction and Intervention

Prajit T Rajendran[1], **Huascar Espinoza**[3], **Agnes Delaborde**[2],
Chokri Mraidha[1]

[1]Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France
[2]Laboratoire national de métrologie et d'essais, Trappes, France
[3]ECSEL JU, Avenue de la Toison d'Or 56-60, 1060 Brussels, Belgium
prajit.thazhurazhikath@cea.fr, chokri.mraidha@cea.fr, agnes.delaborde@lne.fr, huascar.espinoza@ecsel.europa.eu

Abstract

Deep-learning based approaches for learning autonomous driving policies comes with a set of safety challenges. Human-in-the-loop (HITL) learning can be used to improve the safety and reliability of such systems by embedding the human understanding of the complex notion of safety. As AI systems are increasingly deployed in situations with real-world consequences for humans, it can be beneficial to involve humans in various stages of the life-cycle of AI systems to ensure safe and compliant behavior by the systems. In this position paper, we propose a new method to incorporate human-in-the-loop learning to facilitate safe exploration.

Introduction

Deep-learning based components are becoming a popular alternative in the field of autonomous driving, replacing hand-made rule-sets and formula-based pre-defined modules. End-to-end driving policy learning have been attempted through approaches such as reinforcement learning and imitation learning (Tampuu et al. 2020). The major challenge when it comes to a complex task such as autonomous driving is the high dimensional input space and combinatorially explosive number of plausible scenarios that comes with it. Approaches making use of pure reinforcement learning usually require a large amount of time and computational capacity to reach a significant level of driving performance. Moreover, due to issues such as reward hacking, safe learning can not be guaranteed (Amodei et al. 2016). Some constraints can be placed on learning, which impacts the performance and yet does not guarantee that the driving policy achieved would be preferable or comfortable for humans (Zhu et al. 2020). Thus full self-exploration ie. the agent acting in the environment on its own to learn the policy, is infeasible.

Imitation learning on the other hand is an approach wherein demonstration samples or historical data from human experts are used to train the driving policy. Imitation learning suffers from several data related issues: train-test distribution shift, anomalies in data and bias can affect the run-time driving policy in such approaches (Hussein et al. 2017). A hybrid approach wherein the demonstration samples are used to generate the initial policy with further exploration using reinforcement learning, usually performs better (Ross, Gordon, and Bagnell 2011). It is incorrect to assume that all the historical data we have amounts to safer behaviors: there could have been various anomalies due to factors such as driver inattention, out-of-distribution samples and so on. The presence of erroneous or biased samples could have an adverse effect on the safety of the learnt policy. Safe exploration needs to be a priority even after the initial policy is mimicked from humans. Here, we propose an approach to embed the complex human understanding of safety into the learning process to facilitate safe learning. The proposed approach makes use of human-in-the-loop in three ways to facilitate learning a safer exploration policy- providing demonstrations for the initial policy (learning by demonstration), as an oracle for intervention (learning by intervention) and to categorize unsafe samples to train the anomaly predictor (learning by evaluation) thereby covering all of the stages of human-in-the-loop as mentioned in (Goecks 2020).

Background and prior work

The use of deep learning components is increasingly explored in autonomous systems due to the immense potential of modern learning algorithms. However, adoption of fully autonomous systems are challenging due to various factors like vulnerability to out of distribution data, adversarial inputs, anomalies, lack of transparency in black box deep learning components, stochastic nature of training in deep learning, uncertainty in model predictions and unknown unknowns (high confidence wrong predictions). Traditional approaches do not facilitate safe learning, but adding a human expert in the loop can guide the system to safe behavior making use of their knowledge and experience. Humans are necessary in safety critical systems because of their flexibility and capability to adapt to changing conditions and to the incorrect assumptions made at the design phase. (Leveson 2011)

There are various works dealing with safe exploration in reinforcement learning and human-in-the-loop approaches. Some of the works such as (Wang et al. 2020) focus on designing specific loss functions to ensure safe behavior. Here the focus is on the environmental uncertainty. The

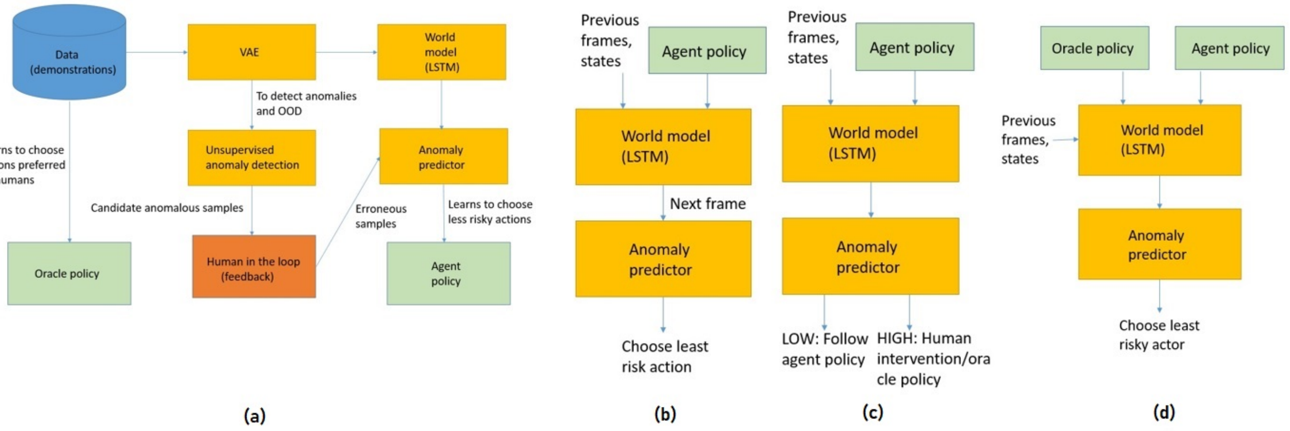

Figure 1: Block diagrams of proposed approach: (a) shows the training phase, (b)-(d) are the variants that can be employed in run-time

work (Lütjens, Everett, and How 2019) conversely makes use of model uncertainty as a proxy for potentially unsafe actions. Uncertainty is a very important measure of model confidence, and reducing model uncertainty could lead to safer policies. However, we should note that all data points wherein the model is uncertain are not unsafe. Conversely, all the data points wherein the model is highly confident are not safe. Therefore a human in the loop approach could be used to identify the unsafe scenarios and prescribe corrective actions. SafeDAgger (Zhang and Cho 2016) describes an approach wherein safety thresholds are used in the training stage to switch action control from an AI agent policy to an expert driving policy. Defining the thresholds which work for all scenarios could be challenging in this scenario.

In the work Crash Prediction Network (CPN) (Nair et al. 2019), the action decision obtained as output from the driving module is fed to a specific module to determine whether it is likely to lead to a crash given the sensory information about the state. In this approach, the training phase consists of the agent interacting with the environment and the trajectories leading up to a crash event are labelled as unsafe data points and the others are safe data points. CPN makes use of self-exploration to generate crash scenarios which are used to train the network. One potential issue is that the crashes generated from self-exploration may not be similar to crashes generated due to environmental anomalies or human deficiencies. Moreover, crash events are just a quantitative proxy for safety, but safety could have a more complex definition and even near misses or deviations or sudden lane changes which did not lead to a crash in training time could be incorrectly labelled as safe data points.

The authors of Trial without error (Saunders et al. 2017), propose a method wherein the agent learns via human intervention. In the training phase, the agent interacts with the environment and when it is about to reach an unsafe state, a human present in the loop blocks the unsafe action. A blocker module learns to predict when humans block the unsafe action, and eventually after the blocker module reaches a certain level of performance it can replace the human to perform the blocking operation. The issue with this approach is that the human needs to be in the loop for a long time during the exploration phase, which is costly. Moreover, there could be a delayed response from the human which could affect the feedback. The learning process is also slow because the agent starts with random exploration and the human only intervenes on unsafe actions.

In the work task-aware generative uncertainty (McAllister et al. 2019), the condition for intervention are based on the satisfaction of two conditions simultaneously: a high collision probability and novelty, where novelty is defined as a significant deviation from in-distribution samples. Similar to crash prediction, this approach makes use of collision as a representative of unsafe states. However, the presence of a human in the loop could aid in capturing a more complex understanding of safety.

Proposed approach

Pure reinforcement learning usually requires a lot of training time, especially on tasks with high-dimensional input space, such as self-driving based on camera and sensor inputs. Moreover, pure reinforcement learning can not ensure safe behavior because the learning process can be susceptible to reward hacking. Due to these drawbacks, prior historical data or human demonstrations are often used as a starting point to ensure faster convergence and safer behavior (Kelly et al. 2019). The assumption is that imitating human experts ensures that the autonomous agent learns the preference of the humans, thereby resulting in safe behavior (Christiano et al. 2017). However, human demonstrations might not be able to help the agent learn the dynamics of the environment because the agent would just copy the behavior of the expert in a supervised manner. In the event it encounters an unknown scenario or anomalous situation during run-time, this would prove to be inadequate. Thus, an approach which uses demonstrations as a starting point to facilitate further exploration is appropriate for complex tasks such as navigation. A module which can predict anomalous behavior, powered with the knowledge of the environment dynamics can help in ensuring that the exploration of the agent remains safe. The environment dynamics are predicted by a module called the world model, which could be based on physical equations for features like velocity or an LSTM when it comes to images and other sensor inputs (Ha and Schmidhuber 2018). This module can be trained on the basis of historical data and updated periodically in run-time. The anomaly prediction module can be trained with the help of the human expert, thereby acting as an embedding of the human notion of safety. In our paper, we introduce such an approach which incorporates human-in-the-loop learning for safety guidance. This is performed by using a human in the loop to identify unsafe or anomalous samples and training a module to predict the probability of risky behaviour for each possible action.

Training phase

The training phase can be divided into two parts: The non-exploratory training phase and the exploratory training phase.

Non-exploratory training phase: In this phase, the AI system does not perform active self-exploration. Instead it makes use of a pre-determined policy, either by imitation of an oracle or from available historical data or pre-trained policies. However, we do not trust this policy completely, and keep a human in the loop to monitor for anomalous data points. An unsupervised anomaly detector is used to make the job of the human easier, as it recommends data points which deviate from the normal by a significant degree. The candidate data points, which are pointed out by the unsupervised anomaly detector are observed by the human, who classifies them as either a ”good” (normal or safe) or ”bad” (erroneous or unsafe) sample. The ”bad” samples are treated as anomalies that we wish to teach the agent to avoid, so we use these samples to train the anomaly predictor module. Additionally, the human could provide an explanation in terms of a label with the reason why he or she believes that the sample should be classified as erroneous. The explanation could take the form of a label specifying the anomaly type, or any information relative to the conditions during which this anomaly occurred

The anomaly predictor module is trained with the environment dynamics as the input and the outcome of ”good” or ”bad” sample as the output. This way, the module is able to predict anomalous behavior before it happens. This module could be placed right before the policy learning module so that we can incorporate the knowledge of future anomalous behavior into our actions. If the human in the loop provides explanations regarding the reason for assigning erroneous labels to certain data points, this information could be used by the module to provide a reasoning of its decision process. This could be important, especially in the context of a human operator being present in the loop in run-time.

Exploratory training phase: In this phase, the AI system interacts with the environment and actively fine-tunes its policy. However, this would be different from pure reinforcement learning in the sense that we facilitate safe exploration by taking previous human feedback into consideration. This is implemented by choosing actions based on the knowledge of the predicted anomaly score from the module that was trained in the non-exploratory training phase. Thus, the AI system does not explore potentially unsafe regions because the exploration space would be constrained by the embedding of the concept of anomalous behavior in the anomaly predictor module as learnt from humans. This could be important in safety critical tasks and situations where we we want to converge to a safe policy quickly without too many mistakes or damage in the training phase. The predicted anomaly score could be made use of in multiple ways in the policy learning module as the following section demonstrates.

Algorithm 1: Algorithm of the safe self-exploration variant

Input: Previous and current states of environment
Parameter: Agent policy
Output: Selected action

```
while driving do
    for all agent actions do
        Predict next frame using world model
        Predict anomaly score using anomaly predictor
    end for
    return Agent action with minimum risk
end while
end
```

Run-time variants in the exploratory phase

There are three approaches by which the proposed model could be used in run-time:

Safe self-exploration: The safe self-exploration variant is one where the human is no longer present in the loop in the exploratory phase, as in (Nair et al. 2019). Here, the AI system explores the environment on its own, subject to the predicted anomaly score of the anomaly prediction module. The next expected frame is predicted using the known environment dynamics world model, and the potential anomaly score for each possible action is checked. The AI system then selects the least risky action. In this manner, we can ensure safe exploration of the environment.

Learning from intervention: In the learning from intervention variant the oracle (human proxy) continues to be present in the loop in the exploratory phase. Here the AI system explores the environment on its own, with the oracle present to propose an alternate action if necessary, similar to the mechanism proposed in (Menda, Driggs-Campbell, and Kochenderfer 2019). The next expected frame is predicted using the known environment dynamics world model, and the potential anomaly score for the predicted agent action from the policy module is checked. If the predicted anomaly score is lower than a threshold value, the action is considered unsafe and the oracle would take over control and the subsequent action would be taken according to the oracle policy. If the predicted anomaly score is higher than a threshold value, the action is deemed to be safe and the subsequent action would be taken according to the agent policy. The threshold

Algorithm 2: Algorithm of the learning from intervention variant

Input: Previous and current states of environment
Parameter: Agent policy and oracle policy
Output: Selected action

```
Define THRESHOLD
while driving do
    Predict next frame using world model
    Predict anomaly score of agent action using anomaly predictor
    if anomaly score lesser than THRESHOLD then
        return Agent action
    else
        return Oracle action/Human intervention
    end if
end while
end
```

Algorithm 3: Algorithm of the joint execution variant

Input: Previous and current states of environment
Parameter: Agent policy and oracle policy
Output: Selected action

```
Define THRESHOLD
while driving do
    Predict next frame using world model
    Predict anomaly score of agent action using anomaly predictor
    Predict anomaly score of oracle action using anomaly predictor
    if agent anomaly score lesser than oracle anomaly score then
        return Agent action
    else
        return Oracle action
    end if
end while
end
```

value could be tuned to ensure that there are as few false negatives in terms of the classification of an action as safe or unsafe

Joint execution: In the joint execution variant inspired by (Ramakrishnan et al. 2019), the oracle (human proxy) continues to be present in the loop in the exploratory phase. Here, both the AI system and the oracle propose actions, and the safer action, as determined by the anomaly predictor, is selected. The next expected frame is predicted using the known environment dynamics world model, and the potential anomaly scores for the predicted AI system's action and the oracle action are checked. If the predicted anomaly score of the AI system is lower than that of the oracle action, the action is taken as per the oracle's policy.

Generating explanations

A major advantage of having a human-in-the-loop in the learning phase is that we can use the human to create explanations of the decisions made. If the AI system also learns to generate explanations of its decisions, the advantages are two-fold: Humans can determine when to intervene during run-time and post-hoc analysis of system failures becomes easier. In the context of the proposed idea, the anomaly predictor block could be more powerful if it could explain why it thinks a particular action is risky, or why it prefers one action to another. In the non-exploratory training phase, along with identifying the bad samples, the human can record an explanation of why the sample is erroneous and should not be used to train the agent. This explanation could be in the form of a classification between different categories such as collision risk, out of lane, environmental anomaly and so on.

In the exploratory training phase, the anomaly predictor could thereby provide a probability score for a potentially anomalous event based on the current state and resulting from the proposed action. Additionally it could contain an explanation of why it determines that the sample is an anomaly, as a label in the same format provided by the human during training. This could facilitate easier take-over and intervention by the oracle or human expert when extended to run-time situations.

Evaluation metrics

The metrics to evaluate the proposed system would be as follows:

Data quality: Data quality could be defined in terms of completeness of the data, or in terms of its accuracy for use in the policy module. Completeness is related to the proportion of state transitions existing in the data to the total number of possible transitions. This is easy to measure in simple grid world tasks but extremely hard in complex tasks like autonomous driving. In the latter case, we can use the ratio of erroneous samples as categorized by the human to the total number of original samples as a proxy for data quality.

Data quantity: Data quantity could be measured by number of samples, type and amount of human involvement needed and query budget ie. the number of times the agent is allowed to query the oracle.

Performance: We could evaluate the performance improvement of the system over baseline methods in terms of task completion rate, average reward and speed of completion.

Safety: Estimating the safety of a device is complex, and litterature often relies on proxy measures such as frequency of catastrophic and risky states, rate of catastrophic/anomalous events or number of ODD (Operational Design Domain) infractions (Weng et al. 2021).

User trust: User trust is a subjective metric that is linked, among other properties, to the estimated level of risks in the system. For example, this could be performed using Likert scale from surveys or questionnaires. Additionally, number of human interventions undertaken in test-time could be an indirect way to measure user trust.

Conclusion and future work

In this paper, we proposed a human-in-the-loop learning approach to improve safety by actively identifying samples which could lead to anomalies and predicting future unsafe states for safer exploration. The extent to which the various metrics such as data quality, safety and user trust can be verified in our model will be further explored. We will also develop an experimental procedure for the design and test of such a model, in particular in contexts which are subjective in nature or when human contextual knowledge plays a major role. The work is still in an early stage and future steps include development of the experimental procedure for design and test of proposed model and evaluation of the system on pre-decided metrics on the target domain of autonomous systems.

Acknowledgments

This work is partially funded by TAILOR, an ICT-48 Network of AI Research Excellence Centers funded by EU Horizon 2020 research and innovation programme under grant agreement No 952215.

References

Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P. F.; Schulman, J.; and Mané, D. 2016. Concrete Problems in AI Safety. *CoRR*, abs/1606.06565.

Christiano, P.; Leike, J.; Brown, T. B.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep reinforcement learning from human preferences. *arXiv preprint arXiv:1706.03741*.

Goecks, V. G. 2020. Human-in-the-Loop Methods for Data-Driven and Reinforcement Learning Systems. *arXiv preprint arXiv:2008.13221*.

Ha, D.; and Schmidhuber, J. 2018. World models. *arXiv preprint arXiv:1803.10122*.

Hussein, A.; Gaber, M. M.; Elyan, E.; and Jayne, C. 2017. Imitation Learning: A Survey of Learning Methods. *ACM Comput. Surv.*, 50(2).

Kelly, M.; Sidrane, C.; Driggs-Campbell, K.; and Kochenderfer, M. J. 2019. Hg-dagger: Interactive imitation learning with human experts. In *2019 International Conference on Robotics and Automation (ICRA)*, 8077–8083. IEEE.

Leveson, N. G., ed. 2011. *Engineering a Safer World: Systems Thinking Applied to Safety*. Cambridge, Mass.: The MIT Press.

Lütjens, B.; Everett, M.; and How, J. P. 2019. Safe reinforcement learning with model uncertainty estimates. In *2019 International Conference on Robotics and Automation (ICRA)*, 8662–8668. IEEE.

McAllister, R.; Kahn, G.; Clune, J.; and Levine, S. 2019. Robustness to out-of-distribution inputs via task-aware generative uncertainty. In *2019 International Conference on Robotics and Automation (ICRA)*, 2083–2089. IEEE.

Menda, K.; Driggs-Campbell, K.; and Kochenderfer, M. J. 2019. Ensembledagger: A bayesian approach to safe imitation learning. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 5041–5048. IEEE.

Nair, S.; Shafaei, S.; Kugele, S.; Osman, M. H.; and Knoll, A. 2019. Monitoring safety of autonomous vehicles with crash prediction networks. In *SafeAI@ AAAI*.

Ramakrishnan, R.; Kamar, E.; Nushi, B.; Dey, D.; Shah, J.; and Horvitz, E. 2019. Overcoming blind spots in the real world: Leveraging complementary abilities for joint execution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 6137–6145.

Ross, S.; Gordon, G.; and Bagnell, D. 2011. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, 627–635. JMLR Workshop and Conference Proceedings.

Saunders, W.; Sastry, G.; Stuhlmueller, A.; and Evans, O. 2017. Trial without error: Towards safe reinforcement learning via human intervention. *arXiv preprint arXiv:1707.05173*.

Tampuu, A.; Matiisen, T.; Semikin, M.; Fishman, D.; and Muhammad, N. 2020. A Survey of End-to-End Driving: Architectures and Training Methods. *IEEE Transactions on Neural Networks and Learning Systems*, 1–21.

Wang, D.; Fan, T.; Han, T.; and Pan, J. 2020. A Two-Stage Reinforcement Learning Approach for Multi-UAV Collision Avoidance Under Imperfect Sensing. *IEEE Robotics and Automation Letters*, 5(2): 3098–3105.

Weng, B.; Capito, L.; Ozguner, U.; and Redmill, K. 2021. A Finite-Sampling, Operational Domain Specific, and Provably Unbiased Connected and Automated Vehicle Safety Metric. *arXiv preprint arXiv:2111.07769*.

Zhang, J.; and Cho, K. 2016. Query-efficient imitation learning for end-to-end autonomous driving. *arXiv preprint arXiv:1605.06450*.

Zhu, M.; Wang, Y.; Pu, Z.; Hu, J.; Wang, X.; and Ke, R. 2020. Safe, efficient, and comfortable velocity control based on reinforcement learning for autonomous driving. *Transportation Research Part C: Emerging Technologies*, 117: 102662.