EECS 4080: Investigating the Safety of Agents with Machine Learning Components

Part 1: Survey on the safety of systems built using machine learning techniques

Jonathan Azpur
Supervisor: Dr. Yves Lespérance

ABSTRACT

This paper consists of a survey of the literature related to the safety of agents with machine learning components. We showcase the definitions of safety from multiple perspectives, provide formal definitions on safety techniques (such as risk minimization) in machine learning, illustrate the current literature on their effectiveness in the field of machine learning, and describe some examples of current research on safety techniques in specific machine learning applications.

1. Introduction

Over the last few years we have seen an increasing use of machine learning algorithms to build autonomous agents. Nowadays these machine learning based agents are involved in every aspect of our lives, including; health, finance, transportation, communications, entertainment and law. However, this trend has also given rise to a widespread of concerns over the safety of such systems [7][9]. In this paper we are going to survey the current literature on safety of agents built using machine learning algorithms. Safety is a broad term, so depending on how you choose to define it, the safety of machine learning algorithms can be studied from different perspectives. We review the literature on the current status of machine learning agent's safety from a variety of perspectives.

2. Safety and Machine Learning

N. Möller's definition of safety is used to study the safeness of machine learning agents from a statistical perspective [1]. Möller defines safety as a minimization of the risk and uncertainty that is associated with harmful outcomes [2]. Based on the input received and its state an agent comes up with an outcome. This outcome can be either wanted or unwanted. An unwanted outcome is said to harm if and only its costs are greater than a certain threshold. Then, the risk is the expected value of the cost of harm, and by knowing its distribution we can calculate its expectation. On the other hand, uncertainty is the lack of knowledge or ability to predict what the outcome will be. Going back to Möller's definition, safety involves reducing the probability of expected harm (risk) and the possibility of unexpected (uncertainty) harms. To further look into these statistical methods in the context of machine learning we will need to provide some formal definitions [3][4]. Given a set of joint variables $\boldsymbol{x} \in X = (\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n})$ and its target $\boldsymbol{t} \in T = (t_1, t_2, \ldots, t_n)$ that has been sampled from a distribution D, a mapper $h \in H: X \to T$ and a loss function $\mathcal{L}: Y \times Y \to \mathbb{R}$, the risk $R(h)$ is the expectation:

$$\mathrm{E}[\mathcal{L}(h(\boldsymbol{x_i}), \boldsymbol{t_i})] = \int \int \mathcal{L}(h(\boldsymbol{x_i}), \boldsymbol{t_i}) D(\boldsymbol{x_i}, \boldsymbol{t_i}) dx\, dt$$

Research has been done on the effectiveness of empirical and structural risk minimization for safety machine learning [5]. The literature points out that empirical is not capable of dealing with uncertainty for machine learning agents. In the context of machine learning we do not know the distribution D, therefore we can only calculate the empirical risk through our input data $\boldsymbol{x}$ and $\boldsymbol{t}$:

$$R_n^{emp}(h) = \frac{1}{n}\sum \mathcal{L}(h(\boldsymbol{x_i}), \boldsymbol{t_i})$$

The risk $R_n^{emp}(h)$ will eventually converge to $R(h)$ when n approximates to infinity. But, when considering safety, a machine can only process a finite number of inputs and therefore the risk calculated is empirical, not the true risk. Therefore, we have proven that machine learning agents are not capable of dealing with uncertainties.

Furthermore, empirical risk minimization relies on the loss functions $\mathcal{L}(h(\boldsymbol{x_i}), \boldsymbol{t_i})$, and these loss functions (square loss, binary loss, hinge loss etc.) cannot be implemented into application-specific values that measure real-life safety aspects (i.e. value of life, quality of life, loss of life, etc.), which limits the ability of machine learning agents in terms of safety. As of now, according to the literature, there is no existing work on analyzing machine learning's probability foundations to include the minimization of uncertainty when developing risk minimization [1] in AI and it needs to be included in order to be able to ensure certain standards of safety on these agents.

Moving away from the probability and technical perspective of safety we find literature that focuses on safety as the possibility of machine learning agents having accidents. In this case accidents are seen as unintended harmful behavior that emerges from poorly designed AI adaptations of the real world. The focus on this paper is to present five research problems related to accident risk; the problem with negative side effects, the problem with reward hacking, the problem with scalable oversight, the problem with self exploration, and finally the problem with adaptability. After exploring all five problems there are suggestions on how to approach them in order to ensure the safety of machine learning agents facing these types of issues. In conclusion, the research finds that so far we have been able to handle safety issues with case-by-case rules or specific fixes, but there is still a need for a unified approach to ensure a more generalized definition of safety [6].

3. Safety techniques in ML applications

Machine learning is involved in almost every aspect of our lives, and for each different field there is extensive literature on why there is a need to work on safety for machine learning agents and how to achieve such standards of safety. For example, if we look at transportation and autonomous driving we find literature that focuses on addressing the functional insufficiencies in the machine learning techniques used for automated driving and looks to make the case for safety of machine learning agents in autonomous driving [8].

We can also find research that dives deeper into specific techniques used in order to obtain safety standards. There is research on the use of reinforcement learning in order for autonomous agents to form long term driving strategies for ensuring functional safety [10]. We can dive even deeper and look into research on the use of reinforcement learning to ensure safety for autonomous machines in agriculture [11]. There is also research that presents new techniques for vehicle-to-vehicle networks with the objective of providing road safety to connected autonomous drivers [12] or research on the use of tree algorithms to plan safe trajectories in high traffic situations [13].

Another stream of interest being surveyed for safety in AI agents is interruption. The more complex AI systems get; the more focus we have to put into making sure that these systems are not capable of adopting policies that inhibit humans to shut them down. Maybe a machine learning agent wants to maximize its safety, or maybe a rational agent determines that it cannot achieve its goals if it dies, and therefore they need to make sure they can't get turned off. In any case the possibility of these agents choosing to prevent humans from switching them off raises serious safety concerns. It is found that implementing certain levels of uncertainty in terms of their goals leads to safer designs of intelligent systems [14]. For a majority of fields in which we have introduced machine learning agents there is extensive literature in regards to safety of these agents.

4. Conclusions

In conclusion, there seems to be a widespread concern about the safety of machine learning agents, and a rise in research for techniques to meet certain standards of safety in specific fields where machine learning is applied. However, there seems to be a void in terms of research for unified approaches to ensure the safety of these intelligent agents, and a lack of inclusion of certain aspects of safety (i.e. uncertainty) when developing machine learning algorithms.

5. References

[1] Kush R. Varshney: Engineering safety in machine learning. ITA 2016: 1-5

[2] N. Möller, "The concepts of risk and safety," in *Handbook of Risk Theory*, S. Roeser, R. Hillerbrand, P. Sandin, and M. Peterson, Eds. Dordrecht, Netherlands: Springer, 2012, pp. 55–85.

[3] Christopher M. Bishop: *Pattern Recognition and Machine Learning*. Secaucus: Springer, 2006, p. 148

[4] S. Shai, B. Shai, *Understanding Machine Learning: From Theory to Algorithms*. New York: Cambridge University Press, 2014, p. 35

[5] V. Vapnik, "Principles of risk minimization for learning theory," in *Adv. Neur. Inf. Process. Syst.* 4, 1992, pp. 831–838.

[6] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul F. Christiano, John Schulman, Dan Mané: Concrete Problems in AI Safety. *CoRR* abs/1606.06565 (2016)

[7] A. Conn, "The AI wars: The battle of the human minds to keep artificial intelligence safe," http://futureoflife.org/2015/12/17/the-ai-warsthe-battle-of-the-human-minds-to-keep-artificial-intelligence-safe, Dec. 2015.

[8] Simon Burton, Lydia Gauerhof, Christian Heinzemann: Making the Case for Safety of Machine Learning in Highly Automated Driving. *SAFECOMP Workshops* 2017: 5-16

[9] Katharina Holzinger, Klaus Mak, Peter Kieseberg, Andreas Holzinger: Can we Trust Machine Learning Results? Artificial Intelligence in Safety-Critical Decision Support. *ERCIM News* 2018(112) (2018)

[10] Shai Shalev-Shwartz, Shaked Shammah, Amnon Shashua: Safe, Multi-Agent, Reinforcement Learning for Autonomous Driving. CoRR abs/1610.03295 (2016)

[11] Kim Arild Steen, Peter Christiansen, Henrik Karstoft, Rasmus N. Jørgensen: Using Deep Learning to Challenge Safety Standard for Highly Autonomous Machines in Agriculture. *J. Imaging* 2(1): 6 (2016)

[12] Billy Kihei, John A. Copeland, Yusun Chang: Automotive Doppler sensing: The Doppler profile with machine learning in vehicle-to-vehicle networks for road safety. SPAWC 2017: 1-5

[13] Amit Chaulwar, Michael Botsch, Wolfgang Utschick: A machine learning based biased sampling approach for planning safe trajectories in complex, dynamic traffic-scenarios. *Intelligent Vehicles Symposium* 2017: 297-303

[14] Dylan Hadfield-Menell, Anca D. Dragan, Pieter Abbeel, Stuart J. Russell: The Off-Switch Game. *AAAI Workshop: AI, Ethics, and Society* 2017