An Introduction to Machine Learning (ML) Safety

Summary

Artificial Intelligence (AI), and in particular Machine Learning (ML), are having a great impact on our lives through vision systems, bioinformatics, recommendation systems, language models, and conversational AI. AI will become more important in the future as government entities rely on it to help achieve mission success in such areas as making decisions about peoples' immigration benefits, controlling autonomous vehicles, or serving as military decision aids. The Department of Defense and many federal civilian agencies are already investing in AI and ML. However, as some people, groups, and research have already noted, there are risks in using AI and ML, such as bias and unintended consequences that come with using models that learn from data and are difficult to inspect. Such risks, if left unaddressed, can result in errors that can negatively affect a government entity's mission performance.

The discipline of ML Safety exists to mitigate AI and ML risks, and includes such approaches as including Confidence, improving interpretability, ensuring Robustness to Distributional Shift, and approaches to formal verification. If not already a part of AI and ML programs or experimentation, we recommend ML Safety be incorporated into new or ongoing work to help mitigate AI and ML's potentially harmful effects.

In this paper, we describe the benefits of AI and, in particular, ML, before introducing the types of common ML risks that exist. Next, we describe the discipline of ML Safety as encompassing the study of mitigations for the unique risks posed by ML. Through mitigating the likelihood and impact of these risks, we seek to make the most of ML.

Promise of ML

Formally, AI consists of an agent that takes direct observation of its environment (input) and a separate fitness or error function grading its behavior (output). AI is sometimes easiest to understand by differentiating it from optimization, which takes its input from a fitness or error function. ML is a type of AI where model parameters are trained from data. A ML model is not an algorithm, but rather is trained by an algorithm and data. Representation learning is a set of techniques that allow a model to automatically learn the representations required for feature detection in raw data. Deep learning is another type of ML, consisting of many neural network architectures, use cases, and training methodologies, with the defining characteristic being that the network has more than one hidden layer. The relationships between AI, ML, and deep learning are shown in Fig. 1.

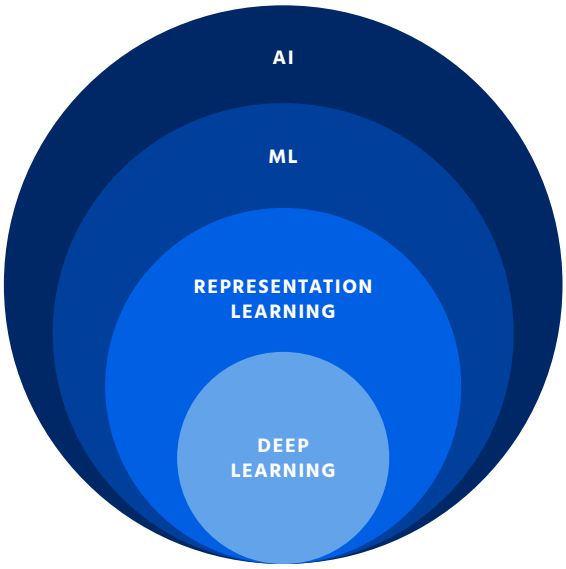

Figure 1: From Deep Learning by Ian Goodfellow, Yoshua Bengio, and Aaron Courville

AI, particularly ML, and more particularly deep learning, is having a great effect on our lives currently and will have an even greater effect in the near future. These systems automate or speed up many previously human tasks, and a key benefit of ML are methods that train models and learn rather than being explicitly programmed. They are used for pattern discovery in scientific and medical datasets, decision aids in real time management systems, perception of the environment for automated systems, prediction of difficult-to-model phenomena, games, planning, navigation, generative art, general computer vision, natural language processing, and autonomous vehicles. Now and in the next few years ML is expected to have a tremendous impact in medicine, autonomous vehicles, government or business processes, military decision aids and targeting, the financial sector, and robotics.

AI has received much hype and criticism, and this is particularly true for deep learning. There are those who promise great feats in order to secure funding and popularize their efforts. However, this technology is both practically useful and has its limitations. Like any useful tool, there are drawbacks and risks to its use. Responsible scientists and engineers seek to address the safety issues peculiar to ML so that we may gain the most benefit from it. Expertise in ML safety will allow the government to gain the most advantage from this technology while balancing risks to critical system and individuals.

ML Risks

All manmade systems can include mistakes or errors in design that lead to harmful consequences. It is hard to predict those negative consequences or detect errors in systems that learn from data and interact with a complex world. ML systems, and their safety concerns, touch all aspects of our lives including admissions and candidate adjudication, threat detection, targeting, medical diagnoses, and driverless cars, among many others. Here we will divide ML risks into three broad categories: General Issues and Classifiers, Reinforcement Learning, and Social Concerns.

General Issues and Classifiers

Traditional statistical confidence is not a part of the current and best-performing ML models. This makes the reliability of a model difficult to judge. The risk is that a model will misclassify an input in the same manner as a correct classification with no indication that the misclassified input is somehow unusual or not well-modeled.

Interpretability can be defined as the ability to explain or to present in understandable terms to a human. This is a major challenge within deep learning because of the size of the models, scope of explanation, and human understandable meaning of a given calculation. There are approaches to interpretability or ability to explain, but none of them are complete and they can lead to problems of overconfidence, sometimes known as Fairwashing.

While neural networks are known to generalize well to new input that is similar to their training data, they can perform poorly and unexpectedly on input outside of their training. This demonstrates a lack of robustness to different kinds of input and potential reliability problems with deep neural networks. It is difficult to know what input is inside the bounds of a model's trained classes, and models can fail in unexpected and surprising ways. Such as the recent news that hackers at McAfee Advance Threat Research were able to trick Tesla's first generation autopilot into classifying 35 MPH signs as 85 MPH signs and accelerating the car. The modification made to the signs was done with stickers and quite subtle; a human being may not even notice that something is wrong with some of the examples.

Formal verification of some ML models, such as deep neural networks, is difficult if not impossible to achieve. This is due to the (currently) unfeasible number of input configuration that would need to be tested. There are some methods that can verify small networks. Validation is even more difficult since it would require human understanding as well.

In data poisoning the attacker injects data into a systems training dataset in order to influence a misclassification. This is possible with traditional offline learning because the datasets can be extremely large and only partially validated. Online, constantly learning, systems can be taught to learn malicious behavior as normal through repeated false positives, leaving them open to attack using that behavior.

Famously, deep neural networks can misclassify input with just a little change to that input. Examples have included adding noise to images, carefully selected single pixel attacks, and slight modifications to real world objects, such as road signs, which leads to misclassification. This is particularly concerning because the modifications can go unnoticed by people, or at least not anticipated as being negative. Camouflage patterns specifically designed to hide people from face detection has been around for a number of years.

Reinforcement Learning (RL)

In RL, the objective function does not evaluate an agent's output directly but rather the effect that the agent has on its environment. This makes reinforcement learning a more challenging task but the training becomes more similar to how a human might train a child or dog. Reinforcement learning involves an agent interacting with its environment.[1]

Negative side effects can result when a designer does not (or is not able to) address all aspects of the environment, and thus implicitly tells the agent that the designer does not care about that aspect of the environment. For example, a robot may knock something over on its way to complete its task, or a swarm may destroy some of its own units in order to simplify the solution. The problem here is that in the real world, and in sufficiently complex virtual environments, it would be impossible (or effectively impossible) for the designer to account for all the potentially bad things an agent could get into as a result of optimizing its objective.

Another ML risk, Reward Hacking, is when an agent finds an unexpected and undesirable method for satisfying its objective. RL agents have found ways of cheating games just as humans have. Goodhart's law, "When a measure becomes a target, it ceases to be a good measure" is used when discussing the problem of predicting the effects of policy.

Some objective functions are too expensive to evaluate frequently enough for traditional reinforcement learning. An important example of oversight that scales poorly is human feedback. If the amount of user effort required to access an AI agent's work is great enough, it may negate the value of the automation in the first place.

An important part of robustness and thorough optimization is safe exploration of an agents' environment. In order to ensure than an AI doesn't act unexpectedly, it needs as much experience with different input as possible, and as much of the environment as possible needs to be explored. However, some parts of the environment, or configurations of the environment and the agent, are harmful. For example, we don't want a flying robot to perform extreme maneuvers when near the ground.

[1] *This section uses the 2016 seminal paper "Concrete Problems in AI Safety", by Dario Amodei et al as its primary reference.*

Social Concerns

Recommendation systems have a major impact on our lives currently. They are the technology behind the advertisement, news selection, and recommended reading and viewing systems employed by Google, Facebook, Amazon, and almost every other major technology company with a user-facing system. These are the systems that encourage innate human tribalism and stove-piping by feeding them news, media, and group recommendations that they are likely to already agree with enforcing confirmation bias, and encouraging grouping around .

Recent developments in language models and the success of Transformers, a type of modern language model, have raised concerns over using them to generate fake text for misinformation campaigns. Transformer-based language models can generate convincing text that is difficult to detect as fake. To help mitigate concerns, OpenAI delayed the release of its GPT-2 model in order to prepare the AI community with test cases of a full staged release. Additionally, they published a release strategy document describing its intent for risk and benefit analyses to be conducted on increased model size releases.

As discussed above, it can be difficult-to-impossible to fully explain or properly interpret some models. This "right to explanation" is an individual right to have the output of an algorithm (model based or not) explained when that output affects a decision regarding an individual's health, finances, or legal status. For example, a Government agent may require information on why a system flagged an individual as a potential threat. In individual private lives, people they may wish to know why a model and/or algorithm denied them medical treatment or coverage, denied them a loan, or flagged for additional screening. In light of these problems, the Government may require explanations from itself and from private businesses. Part of the art of ML model-building is generalization and preventing memorization. Memorization is when a model stores information about a specific example (such as a person) instead of learning general concepts. Memorization can be caused by too many parameters, and data leakage from a training set to a testing dataset, and is not uncommon among non-expert users. Memorization can also lead to privacy concerns in data that contains Personally Identifiable Information (PII).

ML models learn the biases expressed in their datasets. Examples in professional literature include: many top facial recognition models currently perform poorly on nonwhite males; Amazon's facial recognition system matched 28 Members of Congress to arrest mugshots; a US hospital decision aid was less likely to recommend additional healthcare to black people; the Correctional Offender Management Profiling for Alternative Sanctions software was nearly twice as likely to label black people as potential reoffenders than whites; and Amazon's recruiting tool was biased against women. Not only is this concerning to our social health, but it also means that mission-critical recognition systems may operate sub-optimally or make mistakes. This is because optimization is often greedy (i.e., it looks for simple answers in the same way that people develop prejudices), general concepts are not learned, but short cuts are. Even though a lot of effort goes into making these models generalizable in research and development, that is not necessarily true of those that deploy them in the real world. One of the most problematic facts about ML bias is that it can only be detected if you know what to look and test for.

ML Safety as a Mitigation

AI safety encompasses the line of thought, research, and practice concerning how to prevent negative consequences to the human species caused by AI and ML. AL safety topics address safeguards on intelligences greater than our own such as: artificial general intelligence; the law of unintended consequences and poor design choices; the right to explanation and difficult-to-interpret models; and impacts to individual rights and human social systems. Specific to ML are social responsibility concerns regarding privacy, security, and bias, systems safety and validation of learning models, interpretability, and fundamental problems in specifying behavior in a complex environment. ML safety focuses on problems and mitigations concerning current learning models.

Not all ML risks issues have solutions. In fact, the methods listed below are mitigations. Currently the best approach to benefit from the utility of modern AI and ensuring safe, responsible, and robust performance is to have team members well versed in the benefits and risks of AI. More success can be achieved with teammates that further understand the theoretical and working details of ML systems, and the importance and implementation of ML safety methods. Key ML mitigations include confidence; enabling interpretability and explainability; formal verification; logical scaffolding; and improving robustness to distributional shift.

Confidence

Including measures of confidence in ML systems is likely to be an important research topic in 2020 and beyond. This is because more safety-critical systems are adopting ML. For example, perception systems are being used for driverless cars, medical diagnosis, and target recognition both in and out of the military. A necessary improvement to these systems is that they be able to convey how confident they are, and that this measure have a traditional Bayesian interpretation. For example: “I am 95% confident that this is the right target given this video stream and environmental conditions.”

Interpretability and Explainability

Methods currently exist to improve both the interpretation of models and their ability to explain their results. Something can be said to be interpretable if it is presented in understandable terms to a human being. Explainability is a characteristic usually associated with post hoc analyses and techniques used to explain the behavior of a previously trained model. Currently, techniques exists to explain local behavior of models and to visualize dependence between effects either locally or, in a limited manner, globally. However, no explanations currently exists for the entirety of a large neural network. This is due to the immense dimensionality of deep neural networks.

Formal Verification

Xiaowei Huang et al. (2017) introduced an approach to formal verification of neural networks that has attracted much attention and follow-on research and development. The methods, based on the Satisfiability Modulo Theory, takes advantage of the hierarchical design of feedforward neural networks, and defines safety as invariance to classification given small perturbations to the input (in this case, images). The authors state that this method guarantees finding adversarial examples, if they exist, for feedforward neural networks. Human beings can be used for certain difficult examples, and this method can be used to estimate robustness. It works layer-by-layer, calculating the distance between activations known to be in the class and changes to the input. Small changes, in a given layer, are assumed to be required to be the same class. Similarly, points interior to the bounds of the classification as calculated per-layer are forced to be within the class. Determining the exterior points of classification, per layer, is done using the help of human beings. This method can produce a neural network with fewer manipulations than the starting example, and could theoretically find the minimal set of manipulations.

Katz (2019) summarizes formal verification methods as belonging to complete and incomplete methods. Complete methods always succeed, but incomplete methods are more scalable. However, at this time these techniques do not scale well to the size of current, useful neural networks. Osbert Bastani et al. (2017) construct the formal verification problem as a linear programming problem. They are able to do this because the piecewise rectified linear unit (ReLU) activation function can be replaced with a linear function if its specific phase (y=x or y=0) is known. A linear program solver can be used to find adversarial inputs. *AI*^{2} by Timon Gehr et al. (2018) overapproximates the input of the neural network layer by layer. If properties of this abstraction hold, they hold for the actual representation itself. They introduce abstract transformers for fully connected, convolutional, and pooling for ReLU-based neural networks with perception tasks.

Logical Scaffolding

A logical scaffold, as introduced by Nikow Aréchiga et al. (2019), is an observable consequence of appropriate behavior of a learning model and not a formal specification of behavior. This is necessary because these function approximators are trained and the behavior is not imperatively specified. The authors state that logical scaffolds can take the form of any implicit behavior that a human being can reason to be true about an application, its data, or a learning model, such as label consistency, class specific information, physical knowledge, and behavior modeling. This technique can help in mitigating problems related to implicit specifications, distributional shift between the training dataset and real world application, and robustness to adversarial attacks for applications in automobiles, airplanes, legal decision aids, and military decision aids and targeting systems.

Robustness to Distributional Shift

An important trait we seek in human professionals is their knowledge of their own limitations, and for them to admit their inexperience or ignorance when faced with situations they are not prepared for. With this information, their conclusion can be weighed appropriately or further expertise may be sought. However, this is not a feature of traditional ML classifiers. They will "confidently" misclassify unexpected input, fail unexpectedly, and fail silently without warning. This is because the model is unaware of its own experience and limitations, and traditional statistics and rigorous confidence measures are not included as part of these models. This is the most researched problem area, and the one that is closely associated with explainability, interpretability, and human and intelligent system interaction. If a machine could access the true confidence of its performance, it could more effectively team with a human than a model that just produces a classification (decision). For example, a vision system in a vehicle may say, "my classification performance is reduced by 95% because it is raining." Or, "I am only 60% sure that is the target, based on our currently available data."

Reinforcement Learning Specific Mitigations

Dario Amodei et al. (2016) suggest several methods that minimize changes in the environment, under the assumption that humans like the status quo. In this way the agent will seek to achieve its task while not disturbing the environment more than necessary. Some of the proposed solutions and experiments, for solving or mitigating reward hacking, involve an adversarial system to restrict the greedy optimization of the primary agent. Other methods restrict how much reward can be accumulated, or which variable the agent can see.

Most of their proposed solutions and mitigations for expensive reward functions involve reward modeling, where learning an approximation of the real objective function is a separate task from the agent learning the task to be completed. In this way, the reward model can train the task agent sufficiently. Whether or not the reward model can be trained sufficiently is a separate issue. However, since the reward model is smaller than it otherwise would be if it were part of the task learning agent, it should be easier to train than it would be as part of the tasking agent.

A traditional solution to this problem is hard-coding constraints on the AI. However, if the task is complex enough to warrant a learning AI, then it is presumed that human beings can't think of every conceivable dangerous condition (otherwise we wouldn't use a learning agent, and instead would have hard coded the solution). Some of their proposed solutions involve making reward risk sensitive, reverting to safe (but not optimal) policies in certain situations, and using simulation and pre-training.

Current Movement within the Department of Defense (DoD)

The DoD's Joint Artificial Intelligence Center (JAIC) has already defined AI Ethics for DoD AI technologies. AI Ethics serve as a driving force for defining and solving ML Safety problems:

Responsible. DoD personnel will exercise appropriate levels of judgment and care, while remaining responsible for the development, deployment, and use of AI capabilities.

Equitable. The Department will take deliberate steps to minimize unintended bias in AI capabilities.

Traceable. The Department's AI capabilities will be developed and deployed such that relevant personnel possess an appropriate understanding of the technology, development processes, and operational methods applicable to AI capabilities, including with transparent and auditable methodologies, data sources, and design procedure and documentation.

Reliable. The Department's AI capabilities will have explicit, well-defined uses, and the safety, security, and effectiveness of such capabilities will be subject to testing and assurance within those defined uses across their entire life-cycles.

Governable. The Department will design and engineer AI capabilities to fulfill their intended functions while possessing the ability to detect and avoid unintended consequences, and the ability to disengage or deactivate deployed systems that demonstrate unintended behavior.

References & Recommended Readings

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, Dan Mané "Concrete Problems in AI Safety" In: *arXiv:1606.06565v2* (25 Jul 2016)

Department of Defense, Small Business Innovation Research (SBIR) Program "SBIR 20.1 Program Broad Agency Announcement" pulled from *www.dodsbirsttr.mil*

Doshi-Velez and Been Kim "Towards a Rigorous Science of Interpretable ML" In *arXiv:1702.08608v2 [stat.ML]* (02 Mar 2017)

Guy Katz "Verification of Deep Neural Networks" *The Hebrew University of Jerusalem, ForMal Spring School* (05 June 2019)

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, Miles McCain, Alex Newhouse, Jason Blazakis, Kris McGuffie, Jasmine Wang "Release Strategies and the Social Impacts of Language Models" OpenAI Report (November, 2019)

Jiawei Su, Danilo Vasconcellos Vargas and Kouichi Sakurai "One Pixel Attack for Fooling Deep Neural Networks" In: *arXiv:1710.08864v7 [cs.LG]* (17 Oct 2019)

José M. Faria "ML Safety: An Overview" Safe Perspective Ltd., UK (27 October 2017)

Osbert Bastani, Yani Ioannou, Leonidas Lampropoulos, Mimitri Vytiniotis, Aditya V. Nori, Antonio Criminisi "Measuring Neural Network Robustness with Constraints" In: *arXiv:1605.07262v2 [cs.LG]* (16 Jun 2017)

Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, Dario Amodei "Deep reinforcement learning from human preferences" In: *arXiv:1706.03741v3* (13 July 2017)

Patrick Hall, Navdeep Gill "An Introduction to ML Interpretability, Second Edition" O'Reilly Media, Inc. © 2019 (2nd Edition: August 2019)

Patrick Hall, Navdeep Gill, Nicholas Schmidt "Proposed Guidelines for the Resposible Use of Explainable ML" In: *arXiv:1906.03533v3 [stat.ML]* (29 Nov 2019)

Sandy Huang, Nicolas Papernot, Ian Goodfellow, Yan Duan, Pieter Abbeel "Adversarial Attacks on Neural Network Policies" In: *arXiv:1702.02284v1 [cs.LG]* (08 Feb 2017)

Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, Ananthram Swami "Practical Back-Box Attacks against ML" *In: arXiv:1602.02697v4 [cs.CR]* (19 March 2017)

Nikow Aréchiga, Jonathan DeCastro, Soonho Kong, Karen Leung "Better AI through Logical Scaffolding" In: *arXiv:1909.06965v1 [cs.LG]* (12 Sep 2019)

Timon Gehr, Matthew Mirman, Dana Drachsler-Cohen, Petar Tsankow, Swarat Chaudhuri, Martin Vechev "*AI*^{2}: Safety and Robustness Certification of Neural Networks with Abstract Interpretation" retrieved from: *www.cs.rice.edu/pubs* (2018)

Ulrich Aïvodji, Hiromi Arai, Olivier Fortneau, Sébastien Gambs, Satoshi Hra, Alain Tapp "Fairwashing: the risk of rationalization" In: *arXiv:1901.09749v3 [cs.LG]* (15 May 2019)

Xiaowei Huang, Marta Kwiatkowska, Sen Wang and Min Wu "Safety Verification of Deep Neural Networks" In *arXiv:1610.06940v3 [cx.AI]* (05 May 2017)