# Steps To Reduce Risk In Artificial Intelligence And Machine Learning

As the amount of available data and the power of machine learning continues to increase exponentially there is also an exponential increase in the potential for harm to individuals. The potential for harm can be intentional or unintentional. Harm can arise from bias within the model being built or the data employed in the model. The potential negative effect of artificial intelligence often comes as a shock to people who generally think of machine learning and artificial intelligence as tools for the common good. While these tools can have a positive societal effect, we can ill afford naïve assumptions that this is a universal truth. Recognition of the potential for harm and then knowing how to address it will go a long way in mitigating risk.

Learning models can fall prey to several traps which include framing, portability, formatting, ripple effect, and solutionism. A full listing of examples is beyond the scope of this paper but an example of the singular trap of portability follows. In Medicine, lesion and anomaly scanning of skin has effectively been used to identify various dermal cancers. To make a leap that the same scanning algorithm used for diagnosis would also be successful for prevention could lead to fatal consequence. Understanding potential model traps such as this can lead to identification which is the starting point for addressing bias and risk.

Even when used for its intended purpose, a model can still have risk of harm due to the inherent bias of the data it utilizes. In prior example, the algorithm may have learned through being fed a collection of images of cancerous lesions. Those images were likely collected from geographic areas that have the tools to image the lesion, catalog it, and then pass that information along to a repository where it is properly classified. Aside from the risks such as improper transference from one step to the next, there is already an inherent selection bias because individuals with access to treatment and technology options may not represent the population that is most likely to get a particular type of cancer. The algorithm variance potentially also increases because the learning data may mainly represent a particular race or ethnicity. There is the additional risk regarding consent for the collection of the data. While not necessarily something that would affect bias or variance, it does pose a risk. Fortunately, this particular risk can be addressed by getting proper consent.

Once potential risk or bias has been identified, it should be disclosed. Before disclosure the issues with a dataset can be teased out further by simply taking a "Devils Advocacy" position in a series of questions about the data. These could include identifying what classifications exist, how they are defined, whether they represent the entire universe that could be impacted, and many others. For example, limitations regarding the learning or test data used can be identified before the data is utilized. The primary problem with disclosure is that it really does nothing with respect to removing the bias or harm. The challenge of addressing the bias is primarily handled by the user once disclosure has been made by cleaning the data and using alternative testing models.

The process identified above can be employed to address limitations of the data after generation but many times simply asking some preliminary questions at the beginning of the work can go a long way to avoiding or minimizing bias and harm. Questions such as;

- How will the model affect, people communities, and society?
- Is technology the correct solution?
- How do we expect the system to be used?

There are no guarantees that bias and harm can be eliminated but following a thoughtful plan and asking the right questions will at a minimum allow the developer to identify potential risks, mitigate them and disclose thought process as well as potential risks that remain to future users.