OWASP Top Ten Feedback

**General feedback - the list and the home page**

I think this is great! Some feedback below:

* I would also add something about these kinds of mitigations being risk-based, and that each organisation should make a risk-based assessment of which specific attacks are more likely to be employed against their models, how impactful the ramifications would be (ie. would they lose massive amounts of PII or would someone just get a bad customer experience) and then decide on their ML security posture from there.
* I would also highlight some of the reasons why ML systems are different to traditional cyber systems (ie. ML systems are probabilistic while cyber systems are rules based, ML systems learn and evolve while cyber systems are more static). From an organisational perspective this means that it is unfair to expect cyber security professionals to automatically know how best to secure ML systems, and that there should be investment in training and/or the creation of new roles
* I would also add a comment about the terminology here being ML security vs AI security. Clarify the difference between ML and AI (ie. AI is a broader set of technologies while ML refers specifically to the AI sub-field of ML models). For example, I usually refer to my work as AI Security to be more encompassing of an audience who is used to hearing about AI, but most of what I talk about is actually ML security. Adding something here about these terms would be useful for non-ML folk.

**Some questions before I give feedback on these things**

* Is the intention of the top 10 list that it is also ordered so that #1 is most threatening, or is it unordered?
* What is the methodology/reference for the risk factor values under each threat? (I know it comes from the specific scenarios, but is it meant to also be representative of ‘common’ ways of implementing the attack? Because it could be very different and might be interpreted as generic)

**ML01:2023 Input Manipulation Attack**

* Where Adversarial Attacks are referenced in the intro, I usually see this referred to as an Adversarial Example instead and this language would be more consistent. I’m also not sure what it means by saying it’s an umbrella term, what are the other attacks underneath this umbrella?
* The list Adversarial Training, Robust Models, Input Validation are all good to include but could be rephrased to be mutually exclusive (ie. the Robust Models section references both Adversarial Training and Input Validation as the way to make models robust). You could start with Robust Models first and still mention that those other two techniques are some ways to make model robust, but is also dependent on good data quality (clean data, good feature engineering, complete data), model selection (having to choose an appropriate balance between accuracy and interpretability, which are usually trade-offs), considering ensemble methods instead of a single model (ie. you get results from multiple models as a way of checking and validation output). You can then into system level security measures (input validation) and process-based security measures (incorporating adversarial training).
* Love the scenarios, that really helps to put it in perspective.

**ML02:2023 Data Poisoning Attack**

* Would be worth including more here in the ‘about’ section around how this normally happens (an organisation’s data repository could have new data added, or existing data altered), why it works (an ML model’s heavy reliance on training data and the fact that it needs LOTS of it, which means it’s also very hard to spot if it has been tampered with), who is likely to do this (someone either needs access to the data repository if the model is trained on internal data only in which case an insider threat scenario is more likely, or if the model learns from ‘new’ data fed from the outside like a chatbot that learns off inputs without validation (ie. Tay) then this is a much easier attack) etc. Also clarify how this is different to the cyber security challenge of just keeping data secure (ie. it’s more about the statistical distribution of the data than the likelihood of data breach).
* The ‘how to prevent this section’ is great, and adding more info beforehand might help make it clearer as to why these measures are important, and why the measure to secure data is for different reasons than traditional cyber security threat models.
* Again, great examples. You could add something like a chatbot or social media model here to underscore how publicly injected data is also a risk here.

**ML03:2023 Model Inversion Attack & ML04:2023 Membership Inference Attack**

* These two attacks could be clarified as to exactly how they’re different, and depending on how you expect someone might use the Top 10, you could consider combining them. For example, they both leak information about data used to train the model but model inversion is about leaking any data, and membership inference is more about seeing if a specific piece of data was used. Or you could clarify that they are separate based on the different controls used to prevent it (ie. is the focus on differentiation here the goal of attack, or the means of securing it)
* Model inversion - you could explain what it means to ‘reverse-engineer’ in this context. Ie. it is more about targeted prompting than reverse engineering in a cyber context. This would also explain why the suggested controls (ie. input verification) work.

**ML05:2023 Model Stealing**

* This is good! Again, I think a longer description would be helpful.

**ML06:2023 AI Supply Chain Attacks**

* Yes, love that this is included!
* Why are we switching from ML to AI here?

**ML07:2023 Transfer Learning Attack**

* This is interesting.. In my mind this is more an AI misuse issue than an AI security issue, but it really depends on the scenario I think. Could be worth clarifying here.

 **ML08:2023 Model Skewing**

* This also seems the same as the data poisoning attack.. It’s not clear how they’re different. Usually when I saw model skewing the attack is based on being able to change the model outcome without actually touching the training data.. Would be worth clarifying here.

**ML09:2023 Output Integrity Attack**

* Again, this attack overlaps a lot with other attacks already listed that aim to impact the output, it’s almost more of an umbrella term in my mind. Or it could be different but it depends on how an organisation creates its threat model. The controls mentioned here are good though, so it’s worth clarifying why this is listed as a different attack and if it’s based on the controls suggested (which would also apply to the other attacks listed).

**ML10:2023 Model Poisoning**

* Yes, this is good, and worth adding a bit more about how this can happen (ie. direct alteration of the model through injection, the interaction with hardware here and how this can be done in memory etc)