|  |
| --- |
| REAL ML Worksheet |

This is the worksheet that accompanies the REAL ML tool. As you use the tool, refer to the columns in this table during activities that ask you to write down various details of your limitations. This table includes 3 examples that you can refer to when beginning to brainstorm your own limitations. These examples are meant to give you a sense of how to fill out the table, but are not intended to be prescriptive. Use the table in the way that you find most helpful. The three examples we have provided showcase three different limitations that could arise in three different ML projects—they are not related to one another.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Recognize | | | Explore | |
| Limitations | **Types** | **Sources** | **Alternatives** | **What should people know** |
| Example #1: Our theoretical model of learning from expert advice assumes that all experts provide predictions for all data points even though this is typically not the case in the forecasting scenario we used to motivate the problem. | Fidelity | Explicit Decision by Research Team | We could have altered our model to better match the motivating problem, but we would not be able to achieve such strong regret guarantees. | The method we present would need to be modified for use in scenarios in which experts do not provide advice on all data points.  Our model may not reflect real-world settings where experts may sometimes be offline or otherwise unavailable.  Our regret guarantees would not hold in this case, and we expect regret would increase in practice. |
| Example #2: To measure the toxicity of comments on a social media platform, we had Mechanical Turk workers label each comment as toxic or not. These labels are subjective and additionally may not reflect the opinions of users of the social media platform. | Generalizability | Explicit Decision by Research Team | We could have solicited labels from users of the social media site rather than Mechanical Turk workers.  We could have tried to provide more detailed instructions for identifying toxic comments.  We also could have phrased the labeling task in other ways to capture different notions of toxicity. | Given that we did not provide exhaustive instructions on identifying “toxic” comments, the people hired to do so had to rely on their subjective judgments.  The claims that we make in this paper are therefore specific to the views of toxicity held by the people that we hired to label training examples, and our empirical findings may not generalize to other populations and other notions of toxicity. |
| Example #3: In order to mitigate unfairness, our approach requires collecting and storing demographic information. | Value Tensions | Explicit Decision by Research Team | We could have opted for an approach with weaker fairness guarantees but stronger privacy protections. | This approach is not appropriate if collecting and storing demographic information is either prohibited or not possible.  Given the sensitivity around demographic information, employing our approach requires appropriate privacy protections, including explicit notice and informed consent. |
| *Your Limitations Here…* |  |  |  |  |