

REAL ML: Recognizing, Exploring, and Articulating Limitations of Machine Learning Research

JESSIE J. SMITH, University of Colorado Boulder, USA

SALEEMA AMERSHI, Microsoft Research, USA

OLON BAROCAS, Microsoft Research, USA

HANNA WALLACH, Microsoft Research, USA

JENNIFER WORTMAN VAUGHAN, Microsoft Research, USA

Transparency around limitations can improve the scientific rigor of research, help ensure appropriate interpretation of research findings, and make research claims more credible. Despite these benefits, the machine learning (ML) research community lacks well-developed norms around disclosing and discussing limitations. To address this gap, we conduct an iterative design process with 30 ML and ML-adjacent researchers to develop and test REAL ML, a set of guided activities to help ML researchers recognize, explore, and articulate the limitations of their research. Using a three-stage interview and survey study, we identify ML researchers' perceptions of limitations, as well as the challenges they face when recognizing, exploring, and articulating limitations. We develop REAL ML to address some of these practical challenges, and highlight additional cultural challenges that will require broader shifts in community norms to address. We hope our study and REAL ML help move the ML research community toward more active and appropriate engagement with limitations.

CCS Concepts: • **Human-centered computing** → **User studies**; • **General and reference** → *Design*; *Validation*; *Reliability*; **Computing standards, RFCs and guidelines**; • **Computing methodologies** → **Machine learning**.

Additional Key Words and Phrases: Machine Learning, Research Practices, Limitations, Toolkit, Community Standards

ACM Reference Format:

Jessie J. Smith, Saleema Amershi, Solon Barocas, Hanna Wallach, and Jennifer Wortman Vaughan. 2022. REAL ML: Recognizing, Exploring, and Articulating Limitations of Machine Learning Research. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*, June 21–24, 2022, Seoul, Republic of Korea. ACM, New York, NY, USA, 31 pages. <https://doi.org/10.1145/3531146.3533122>

1 INTRODUCTION

Machine learning (ML) has emerged as one of the most active and impactful fields of research within computer science. The past decade has been marked by a steady stream of impressive technical achievements from the ML research community and many real-world applications of ML. These developments have fostered considerable hype about ML's potential, as well as growing concerns about the inflated and unsupported claims that are sometimes made about its true capabilities [22]. In this paper, we consider the role that ML researchers can play in ensuring appropriate interpretation of their research. In particular, we focus on the disclosure and discussion of limitations—and why such practices are important to the healthy functioning of a research community and the broader understanding of its achievements.

Limitations are drawbacks in the design or execution of research that may impact the resulting findings and claims. Limitations are different from “research ethics,” which focuses on the potential harms that may be caused to human

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

© 2022 Copyright held by the owner/author(s).

Manuscript submitted to ACM

subjects during the research process [29], and “broader impacts,” which focus on the potential downstream harms and consequences for society that may arise as a result of research findings and claims [21, 23]. In contrast, limitations concern aspects of the research process that may pose a threat to the validity of research findings and claims.

Many scientific fields have well-established norms around disclosing and discussing limitations, which are often recognized as necessary for improving scientific rigor and research integrity. These norms rest on a shared belief that recognizing, exploring, and articulating limitations can foster greater precision in the descriptions of research (making it easier to reproduce), help ensure appropriate interpretation of research findings, make research claims more credible, and highlight issues that would benefit from further research [5, 6, 18]. Although practices necessarily vary by field and by publication venue [12], the ML research community is notable for not having particularly well-developed norms around disclosing and discussing limitations. Although there have been recent efforts to encourage ML researchers to reflect on the potential impacts (intended or not) of their research on society—for example, the introduction of broader impacts statements at the Neural Information Processing Systems (NeurIPS) conference in 2020 [16] and the subsequent NeurIPS 2021 paper checklist, which encouraged authors to articulate the limitations of their research and even to create separate limitations sections in their papers [1]—disclosure and discussion of limitations remains uncommon in the ML research community.

In this paper, we present the Recognizing, Exploring, and Articulating Limitations in Machine Learning tool (REAL ML), a set of guided activities to help ML researchers recognize, explore, and articulate the limitations of their research. We developed and tested REAL ML via an iterative design process with 30 ML and ML-adjacent researchers. Specifically, using a three-stage interview and survey study, we 1) identified ML researchers’ perceptions of limitations, as well as the practical and cultural challenges they face when recognizing, exploring, and articulating limitations; and 2) iteratively developed REAL ML to address some of the practical challenges we identified. Additionally, we introduce a list of sources of limitations and a list of types of limitations that commonly occur in ML research, both of which were curated and refined over the course of our study. Our findings reveal many challenges faced by ML researchers, and show early evidence suggesting that REAL ML may help them feel more prepared and more willing to recognize, explore, and articulate the limitations of their research. Our study also exposes cultural challenges that go beyond the scope of REAL ML and will require broader shifts in community norms to address. We hope our study and REAL ML help move the ML research community toward more actively and appropriately engaging with limitations.

2 BACKGROUND AND RELATED WORK

Disclosing and discussing limitations can benefit both those doing research and those building on the research of others. Since limitations affect the validity of research findings and claims, there can be negative consequences when they are not recognized, explored, and articulated. Due to the severity of the potential harms that may arise if research findings are misinterpreted or taken out of context and the need to guard against both unintentional misreporting and deliberate spinning of research claims [3], some publication venues in fields like biomedicine require limitations to be stated in papers’ abstracts [11, 31]. The Journal of the American Medical Association advises authors to include “*a discussion section placing the results in context... and addressing study limitations*” [13], while the Journal of Neuroscience requires authors to include “*a discussion of the validity of the observations*” [14]. However, these practices are still relatively rare; one recent study reported that only one of the 25 top-cited scientific journals encourages disclosure and discussion of limitations [12], and, to the best of our knowledge, limitations sections are not yet required by any major ML publication venue.

Even when limitations are disclosed, researchers often fail to appropriately discuss how these limitations might impact their research findings and claims [24, 25]. Lingard and Watling [17] described three common approaches that researchers take to writing limitations sections: the confessional, where researchers beg forgiveness for their work’s flaws

from reviewers; the dismissal, where researchers acknowledge limitations only to diminish and dismiss their importance; and the reflection, where researchers acknowledge the uncertainty and assumptions that underlie their research, and reflect on their impacts. Although the latter approach is ideal, the first two are common in practice. There are a variety of reasons for this. Perhaps chief among them is the perceived stigma around disclosing limitations and the fear that doing so would increase the likelihood of paper rejection. Brutus et al. [4] observed that in the field of management, *“the pressure stemming from the increasingly low acceptance rates for peer reviewed journals and the emphasis on publications in academic reward structures represent clear motives for not acknowledging limitations and for offering only benign directions for future research.”* Similarly, Puhan et al. [24] noted that in biomedicine, researchers are reluctant to disclose limitations because *“they perceive a transparency threshold beyond which the probability of manuscript acceptance goes down (perhaps even to zero).”* In our study, we see evidence of a similar perceived stigma within the ML research community.

Another reason why researchers may fail to appropriately disclose or discuss limitations is a lack of guidance and appropriate training [12]. Although there are no widely agreed-upon approaches for recognizing, exploring, and articulating limitations, guidelines have begun to emerge in some fields. When developing REAL ML, we drew on a four-step process for articulating limitations in medical studies, proposed by Ross and Zaidi [25]. This process involves stating whether each limitation arises in the study design, data collection, data analysis, or study results; explaining the potential impacts of the limitation; providing potential alternative approaches that could have been taken and why they were not; and describing any steps that were taken to mitigate the limitation’s impacts. Within ML, researchers have begun to create tools and resources to help both researchers and practitioners think more critically about the impacts of their work, such as value cards [28], broader impacts statements [16], model cards [19], and datasheets for datasets [9]. REAL ML similarly encourages reflection, but we place less emphasis on ethical impacts and instead focus on the impacts of limitations on research findings and claims, looking across all aspects of ML research rather than just models and datasets. Importantly, REAL ML is intended for use *posthoc* via a process known as “reflection on action” [20], a contemplative practice that uses reflection on previous actions to gain a better understanding of their impacts. Given the cultural challenges faced by ML researchers, we believe this is a necessary step in moving the ML research community toward “reflection *in* action” [26], where informed trade-offs and just-in-time adjustments are made during the research process to reduce the likelihood of negative impacts and improve scientific rigor and research integrity; we return to this topic in Section 6.

Although encouraging reflection is an important step toward normalizing disclosure and discussion of limitations, additional obstacles remain. Community norms may hinder ML researchers’ abilities to recognize, explore, and articulate the limitations of their research. As discussed by Giraud-Carrier and Dunham [10], ML publication venues tend to favor positive results over negative results, with research failures rarely discussed openly. As noted by Liao et al. [15], who created a taxonomy of threats to internal and external validity from 107 papers surveying ML research, the use of benchmarks to assess progress in ML can place too much emphasis on outcomes at the expense of scientific inquiry. Finally, the ML review process—like those found in much of computer science—tends to emphasize abstraction and generalizability. Indeed, in an analysis of highly cited papers published at two major ML conferences, Birhane et al. [2] found that the third most common value expressed in ML papers was generalizability. This emphasis may encourage ML researchers to stretch their research claims in inappropriate ways, rather than explicitly describing their limited scope [8, 27]. Despite these community norms, disclosure and discussion of limitations is just as important in ML as in other scientific fields, particularly as ML research is often motivated by or meant to influence real-world applications [30].

Our findings provide insight into ML researchers’ perceptions of limitations, as well as the practical and cultural challenges they face when recognizing, exploring, and articulating limitations. Although we highlight some challenges that stem from community norms, our primary goal is to provide practical support for ML researchers. REAL ML therefore

contains structure to help ML researchers reflect on the ways that limitations arise from unavoidable constraints, unforeseen challenges, and decisions made during the ML research process, and on the impacts these limitations may have on the resulting findings and claims, with the aim of using this reflection to improve scientific rigor and research integrity.

3 METHODS

We conducted an iterative design process aimed at developing and testing REAL ML, a set of guided activities to help ML researchers recognize, explore, and articulate the limitations of their research. Specifically, we used a three-stage interview and survey study to answer the following four research questions:

- **RQ1:** What is a limitation of ML research? What types of limitations are there? How do they arise?
- **RQ2:** What challenges do ML researchers face when seeking to recognize limitations?
- **RQ3:** What challenges do ML researchers face when exploring and articulating limitations?
- **RQ4:** What practical support would help alleviate some of these challenges for ML researchers?

As we explain in detail below, our study included interviews with 20 ML researchers while using evolving versions of REAL ML (stage 1), interviews and reviews of REAL ML with six ML-adjacent researchers (stage 2), and a final stage with four additional ML researchers who provided feedback via an online survey using a near-final version of REAL ML (stage 3). Throughout the study, we iterated on REAL ML based on the feedback we received from participants. All interviews were conducted virtually on a video conferencing platform. Interviews were recorded and transcribed using third-party software. Participation was voluntary; each interview participant was compensated with a \$30 voucher, while each survey participant was compensated with a \$45 voucher. The study was approved by our institution’s IRB.

Initial prototype. Prior to beginning our study, we developed an initial prototype of REAL ML. When developing this prototype, we drew on our team’s interdisciplinary expertise in ML, HCI, and science and technology studies, as well as our decades of combined experience writing and reviewing ML papers and engaging with the ML research community. The prototype consisted of three lists intended to encourage reflection: a list of types of limitations that commonly occur in ML research (e.g., generalizability limitations, robustness limitations) and descriptions of each; a list of common decision-making points in the ML research process where limitations could arise (e.g., formalism of the problem, technical approach) and descriptions of each; and a list of probing questions to answer when preparing to articulate a limitation (e.g., questions about potential alternative approaches that could have been taken, questions about how the limitation’s impacts were mitigated). The list of commonly occurring types of limitations drew on the limitations uncovered in Nanayakkara et al.’s analysis of NeurIPS 2020 broader impacts statements [21] and on Birhane et al.’s analysis of the values expressed in ML papers [2]. The probing questions were adapted from Ross and Zaidi’s four-step process for articulating limitations in medical studies [25]. After constructing an initial version of each list, we piloted the lists using our own ML papers and updated them based on our experiences. We additionally piloted the initial prototype with and solicited informal feedback from colleagues within and outside the ML research community and further iterated on the prototype based on this feedback. The initial prototype is included in the appendix.

Stage 1. After creating our initial prototype, we conducted semi-structured interviews with 20 ML researchers. We recruited participants through social media, posting links to a recruitment form on our Twitter accounts. We specifically sought to recruit ML researchers who had previously published at least one peer-reviewed ML paper. Although we obtained a relatively diverse sample of 100 researchers, we acknowledge that social media recruitment can lead to selection bias, so the researchers who expressed interest in participating in our study may have already been more

Table 1. Information about participants.

ID	Country	Experience	Stage	Area(s) of expertise	Version of REAL ML
P1	USA	< 5 years	1	NLP, computational social science	V1 Prototype with guidance
P2	USA	< 5 years	1	model based relational learning, robotics	V2 Prototype with guidance
P3	USA	> 10 years	1	NLP, information retrieval, medical informatics	V2 Prototype with guidance
P4	USA	> 10 years	1	NLP, interpretable ML, reinforcement learning, AI ethics	V2 Prototype with guidance
P5	USA	5-10 years	1	NLP, computer vision	V2 Prototype with guidance
P6	China	< 5 years	1	medical informatics, healthcare	V2 Prototype with guidance
P7	India	< 5 years	1	data mining	V2 Prototype with guidance
P8	Spain	< 5 years	1	information retrieval, recommender systems	V2 Prototype with guidance
P9	India	< 5 years	1	NLP, deep learning	V2 Prototype with guidance
P10	USA	5-10 years	1	AI fairness, recommender systems	V3 Prototype with guidance
P11	Germany	5-10 years	1	adversarial attacks, trustworthy AI	V3 Prototype with guidance
P12	USA	5-10 years	1	NLP	V3 Prototype with guidance
P13	USA	5-10 years	1	NLP, reinforcement learning	V1 Tool with guidance
P14	USA	< 5 years	1	NLP, AI fairness, privacy in AI	V1 Tool with guidance
P15	Canada	5-10 years	1	explainability, human-centered AI, medical informatics	V2 Tool with guidance
P16	USA	5-10 years	1	NLP, structured prediction	V2 Tool with guidance
P17	USA	< 5 years	1	NLP	V2 Tool with guidance
P18	USA	5-10 years	1	AI ethics, fairness	V2 Tool no guidance
P19	USA	> 10 years	2	social science, HCI	V2 Tool
P20	USA	> 10 years	2	cognitive psychology, behavioral economics, actuarial decision making	V2 Tool
P21	USA	> 10 years	2	responsible AI, standardization, research ethics	V3 Tool
P22	USA	5-10 years	1	learning theory, foundations of ML	V3 Tool no guidance
P23	USA	> 10 years	2	HCI, social computing	V3 Tool
P24	Canada	< 5 years	1	NLP	V3 Tool no guidance
P25	USA	> 10 years	2	social science methods, ethics in AI/ML	V3 Tool
P26	USA	> 10 years	2	ethics in AI/ML, AI governance	V3 Tool
S1	USA	5-10 years	3	–	V4 Tool no guidance
S2	USA	< 5 years	3	–	V4 Tool no guidance
S3	Canada	> 10 years	3	–	V4 Tool no guidance
S4	USA	5-10 years	3	–	V4 Tool no guidance

likely to use a tool like REAL ML. Starting with the 100 researchers who completed our recruitment form, we filtered out researchers who had not both authored and reviewed at least one peer-reviewed ML paper (saving some for stage 2), binned the remaining researchers based on their years of experience with ML research and their geographic locations, and then randomly selected a fixed number of researchers from each bin, yielding 20 participants in total. We intentionally selected participants with different levels of experience and different geographic locations because we wanted to ensure REAL ML would be suitable for different research contexts. Table 1 contains more information about the participants.

Prior to each participant’s interview, we asked them to share with us a publicly available (e.g., published or available on arXiv) ML paper they had authored to use as a case study, with the guidance that the paper should fall into “the topic areas covered by ML venues such as NeurIPS, ICML, ICLR, COLT, and AISTATS or related venues like ACL, EMNLP, or

CVPR.” The interviewer read each participant’s paper before conducting their interview in order to provide personalized guidance on using REAL ML. We began each interview by asking the participant to reflect on their previous experiences recognizing, exploring, and articulating the limitations of their research, including any challenges they faced, and to provide their own definition of limitations of ML research. Next, we asked them to walk through the process of recognizing limitations in the paper they had shared with us, as if it had not yet been published, using the latest version of REAL ML. Several participants’ papers already mentioned limitations, so we encouraged these participants to focus on new limitations they had not previously recognized. With early versions of REAL ML (labeled as “Prototype” in Table 1, last column) that lacked any guiding prompts, the interviewer provided extensive verbal guidance, walking participants through its intended use. As REAL ML evolved to include more guiding prompts (labeled as “Tool” in Table 1, last column), the interviewer reduced the amount of guidance they provided to participants. In later interviews, when REAL ML was more robust, participants were encouraged to follow the guiding prompts on their own while “thinking aloud” and asking questions as they arose. In addition to observing participants’ use of REAL ML, we solicited explicit feedback on what was and was not helpful for participants and how they thought REAL ML could be improved. On average, each interview lasted 45–55 minutes. The full interview protocol is in the appendix.

Stage 2. After most of the stage 1 interviews were complete, we conducted interviews with six researchers who were knowledgeable about ML (e.g., use ML in their work or regularly read ML papers) and also experts in more sociotechnical fields, such as HCI, psychology, social science, responsible AI, and research ethics. Our goal in including these participants was to surface any community norms or assumptions that might have been taken for granted by ML researchers, but would stand out to ML-adjacent researchers, many of whom have subjected the field to more critical interrogation. Some of these participants had completed our recruitment form, while others were recruited through convenience sampling with the goal of ensuring coverage of different perspectives. Table 1 contains more information about these participants.

During the stage 2 interviews, we asked each participant to describe their impressions of community norms around limitations, both within and outside the ML research community, as well as any challenges or success stories. We then walked them through the tool and asked them for feedback on different sections, placing emphasis on helping ML researchers articulate limitations more effectively for different audiences, such as reviewers, researchers, and practitioners. Again, on average, each interview lasted 45–55 minutes. The full interview protocol is in the appendix.

Stage 3. After the stage 1 and stage 2 interviews were complete, we conducted an online survey with four additional ML researchers. We asked these participants to use REAL ML and complete the corresponding worksheet for an ML paper they were currently working on and planning to submit for publication. Because of the sensitivity of requesting access to information about others’ unpublished work, we limited participation to ML researchers from our own institutions and their immediate collaborators. Participants were recruited via internal mailing lists and direct emails. We sent a copy of REAL ML to each participant via email and asked them to use it on their own, without any additional guidance. After using REAL ML, each participant provided feedback via an online survey and shared their completed worksheet and limitations section with us. The feedback from participants in this stage led to very minimal requests for design changes. These requests influenced our last development iteration, resulting in the final version of REAL ML included in the appendix.

Thematic analysis and iterative design. Throughout our study, we conducted a version of thematic analysis [7] using open coding. We first coded participants’ responses and feedback into a few high level categories (e.g., general challenges, types of limitations, tool needs). We then iterated on REAL ML as new challenges and needs were identified by participants, always showing participants the latest version. Sufficient saturation was achieved on the themes before

the stage 1 interviews were complete, with relatively little new information collected during later interviews. After all interviews were complete, we did one last open-coding pass on all of the responses and feedback from participants and came up with a final set of themes (e.g., fear of paper rejection, the double-edged sword of limitations, and detrimental community norms). We discuss these final themes throughout the remaining sections of this paper.

4 PERCEPTIONS OF LIMITATIONS AND CHALLENGES FACED BY ML RESEARCHERS

In this section, we summarize ML researchers' perceptions of limitations, as well as the challenges they face when recognizing, exploring, and articulating limitations, thereby answering the first three research questions in Section 3.

4.1 Defining Limitations of ML Research

We begin with our first research question, which asks, "What is a limitation of ML research?" In general, participants did not agree upon a single definition of limitations of ML research and suggested many types of limitations.

4.1.1 Limitations as inherent vs. as indicative of bad research. Participants had different views on limitations. Many participants took the position that limitations occur in all ML research, no matter how scientifically rigorous. P3 said, "a limitation is... something in the [methodology] that might cause us to put an asterisk on our conclusion... Like if we are saying model A is better than model B or if we are saying that you know our automated thing is comparable in performance to a human at doing a certain thing... anything that would necessitate an asterisk on that conclusion." P25, an expert in social science research and ML, defined a limitation as "anything that isn't perfect." They called this an "extreme, but fitting" definition, as it (accurately, in their opinion) implies that limitations are unavoidable. Similarly, P22 indicated that in ML theory, limitations are always present, inherent in the assumptions made by researchers, stating, "from a theoretical perspective, a limitation is precisely defined as 'when the assumptions are removed, your stuff doesn't work anymore.'" This participant went on to describe the value of disclosing and discussing limitations as a reflection exercise:

"To me, limitations are an exercise of self reflection. It's a section of the research paper where authors should have a frank conversation and discussion with the readers and make them aware of some of the challenges they faced in the study. It should not be a list of things the researchers could have done or want to do in the future, it is not just admission of error... to me, it should be a 'zoom-in, zoom-out' process, where the zooming-in includes taking a deeper dive into the internal and external validity of the research, acknowledging some of the assumptions and how strong these assumptions are behind a particular model. And then by zooming out, I think the limitations section should include critical thinking on the research question and how does it fit into the broader academic context, look into its policy and real-world application, as well as the potential impact on individuals, organizations, and society as a whole." (P22)

In contrast, other participants suggested that limitations were indicative of bad research, rather than being an inherent part of the ML research process. As one example, P19 defined limitations as "vulnerabilities" in the ML research process, and indicated that stating limitations exposes aspects of one's research design or execution that could invite criticism. This was a common sentiment among participants, reflecting an understanding of limitations as "weaknesses" [25] or "flaws" [17] that may serve as reasons to dismiss research and the resulting findings and claims. Participants who took this position were generally unenthusiastic about highlighting these aspects of their research.

4.1.2 Types of limitations identified by participants. Rather than providing a definition of limitations of ML research, some participants narrowed in on a single type of limitation, often drawing on examples they had encountered previously. Most strikingly, before seeing REAL ML, many participants defined limitations as a lack of either generalizability or robustness. For example, P20 defined a limitation as an *“attempt to draw conclusions outside of the context in which they are merited.”* P26 instead narrowed in on technosolutionism, suggesting that ML research often involves building technologies that are not an appropriate way to address the motivating problem, a common practice that stems from the assumption that complex problems, including complex societal problems, are *“solvable or computable.”*

Some participants expressed a distinction between limitations that are somehow fundamental (e.g., the types of questions a model can be used to answer), limitations that arise from explicit decisions made by researchers (e.g., which hyperparameter values to use), and other limitations that arise from forces outside researchers’ control (e.g., constraints on time or resources, experimental failures, or negative results). P23 referred to some limitations as *“future work”* limitations—that is, things researchers did not scope or test for, but could be explored in the future.

Many participants used the terms “knowns” and “unknowns.” P21, an expert in research ethics, said they want to see *“what they know, what they don’t know, and what they’ve explored”* when reading ML papers. Several participants brought up limitations that arise from explicit decisions made by researchers as examples of “known knowns.” For example, P1 explained that after testing multiple stopping criteria, they chose one that lowered accuracy but maximized efficiency. They described this as a known known because it involved a tradeoff that existed because of an explicit decision they had made.

Other participants argued that limitations should cover things researchers know they haven’t explicitly tested for or measured—that is, “known unknowns.”¹ P4 gave an example in which they had chosen to evaluate their model using a music dataset that only included Western music, which they knew limited the generalizability of their research claims. Some participants drew analogies between this type of limitation and broader impacts, stating that it is important for ML researchers to think critically about the potential impacts (intended or not) of their research. P25 described known unknowns as a researcher’s *“I don’t knows”* and argued that it was important to appropriately articulate them: *“Let’s hear the ‘I don’t knows.’ Let’s make the ‘I don’t knows’ very explicit, and the conditions under which I do know and the conditions under which I don’t know.”* In contrast, several participants strongly opposed this idea. For example, P20 felt that speculative warnings could hinder scientific innovation because they might discourage future work or even cause harms: *“I think about [speculating about known unknowns] with product warnings and health warnings and things like that. There have been a lot of myths that have endured for like 75 years, like ‘pregnant women shouldn’t do this,’ and no one ever actually empirically figured that out. And then when they do they were like ‘oh yeah there was no risk all along.’ We were just assuming there might be and we were putting a lot of people under duress for no reason.”*

4.2 Challenges in recognizing limitations

One clear theme from our interviews was that junior researchers expressed difficulties recognizing the limitations of their research considerably more than senior researchers. This is partly due to a lack of transparency within the ML research community about common limitations. Several participants said it can take junior researchers years to gain the disciplinary knowledge necessary to fully recognize the limitations of their research, in part because there are few peer-reviewed ML papers that discuss limitations, meaning their only option is to learn from their advisors or mentors

¹Participants were not consistent in their terminology here. Some used the term “unknown unknowns” to describe things researchers haven’t explicitly tested for or measured and can therefore only speculate about. We avoid this usage since even speculation is not possible for true unknown unknowns—that is, things researchers do not even know are possibilities.

over time. Some junior researchers even described how the more they participate in the ML research process, the more they realize they don't know. As P3 explained, *"it's harder to identify limitations that you aren't aware of."*

Some junior researchers said they had a hard time articulating limitations without unnecessarily overemphasizing them. Both P1 and P6 said they had previously been told by their advisors that they had called out too many limitations of their research. Some junior researchers attributed their overemphasis of limitations to a lack of confidence in their research stemming from their underrepresented identities (e.g., being a woman in a field dominated by men or having English as a second language in a field in which English is the default language of publication). P24, for example, described their experience attempting to recognize limitations as a non-native English speaker as a *"constant struggle."*

Although most participants who had more than five years of experience with ML research said they were confident in their abilities to recognize limitations, this confidence did not always translate to a holistic view of possible limitations. As an example, when asked to define limitations, two participants who said they did not face challenges when recognizing the limitations of their research, each with more than five years of experience, gave only limited descriptions like *"lack of generalizability"* or *"lack of robustness."* It is therefore possible that senior researchers could still benefit from increased support recognizing limitations even if they feel confident doing so on their own.

4.3 Challenges in exploring and articulating limitations

Participants expressed a variety of challenges related to exploring and articulating limitations. Some challenges, like fear of paper rejection, echo those observed in other fields [4, 17, 24, 25], while others are more specific to community norms.

4.3.1 Limitations as grounds for rejection. One theme touched on by over half of the participants was the fear that disclosing limitations would increase the likelihood of paper rejection. As P2 explained, limitations are a double-edged sword, since disclosing them can make research appear less scientifically rigorous, but omitting them can suggest that the researchers do not understand the implications of their research design or execution: *"not mentioning enough limitations is grounds for rejection, but mentioning too many limitations is also grounds for rejection,"* (P2). This participant went on to say they tried to disclose only a few limitations in their papers, even if they knew of other limitations, because they did not want to give *"fuel"* to reviewers to *"trash their paper."* P15 said, on average, they would include only 3–4 limitations per paper, in part due to pragmatic reasons, but also *"to not admit to too many limitations for the reviewers."* Similarly, P14 mentioned that even though they thought it was important to disclose limitations, they were hesitant to do so because a long list of limitations might undermine the perceived importance or benefits of their research and *"you don't want to shoot yourself in the foot when you are writing a limitations section."*

Relatedly, some participants decided to omit limitations due to the perceived stigma around disclosing them, as discussed in Section 4.1.1. P11 admitted that some of the limitations of their research were important due to their potential negative impacts on society, but chose not to disclose them *"because societal impacts aren't mentioned in other ML research papers,"* going on to say they would not feel comfortable disclosing them because of community norms. In general, the choice to disclose less information about limitations appeared to be primarily motivated by a desire to appease reviewers.

Concerns about paper rejection may be compounded by perceived differences in reviewers' expectations of which limitations are important to disclose—and, more generally, which values are important to prioritize—which can also vary by subfield and by publication venue. P2 shared the example that reviewers for robotics publication venues tend to prioritize practicality over novelty or theoretical claims, while reviewers for other ML publication venues may not. As P5 put it, *"as you are going through this process, you understand what certain conferences are looking for."*

Multiple participants confessed that the emphasis on abstraction and generalizability in the ML review process had led them to underemphasize or omit some of the limitations of their research, since limitations often relate to the ways in which research cannot or should not be generalized. As P16 explained, *“work that is generalizable is more valued, so in research you want to overemphasize your generalizability—even if it isn’t necessarily true.”*

4.3.2 Struggles with prioritization and organization. Another commonly expressed challenge involved the page limits imposed by ML publication venues. P7 said, *“usually authors are constrained to be very precise about their limitations, and it can be hard to compress all of that information into the page limits that they have.”* Given limited space, participants felt the need to prioritize limitations, and several participants mentioned not knowing how to do this. For example, is it best to narrow in on the limitation the researchers personally find most important and discuss it in depth, or to cover more limitations at a higher level? P21 suggested that *“in the event of page limits... authors should prioritize the limitations with the highest known severity of impact,”* but severity is not always easy to judge. Indeed, P23, who had previously done work in research ethics, mentioned that the ML researchers they worked with consistently requested a “Richter scale” of some kind to measure the severity of the impacts of limitations so they would know which to articulate first.

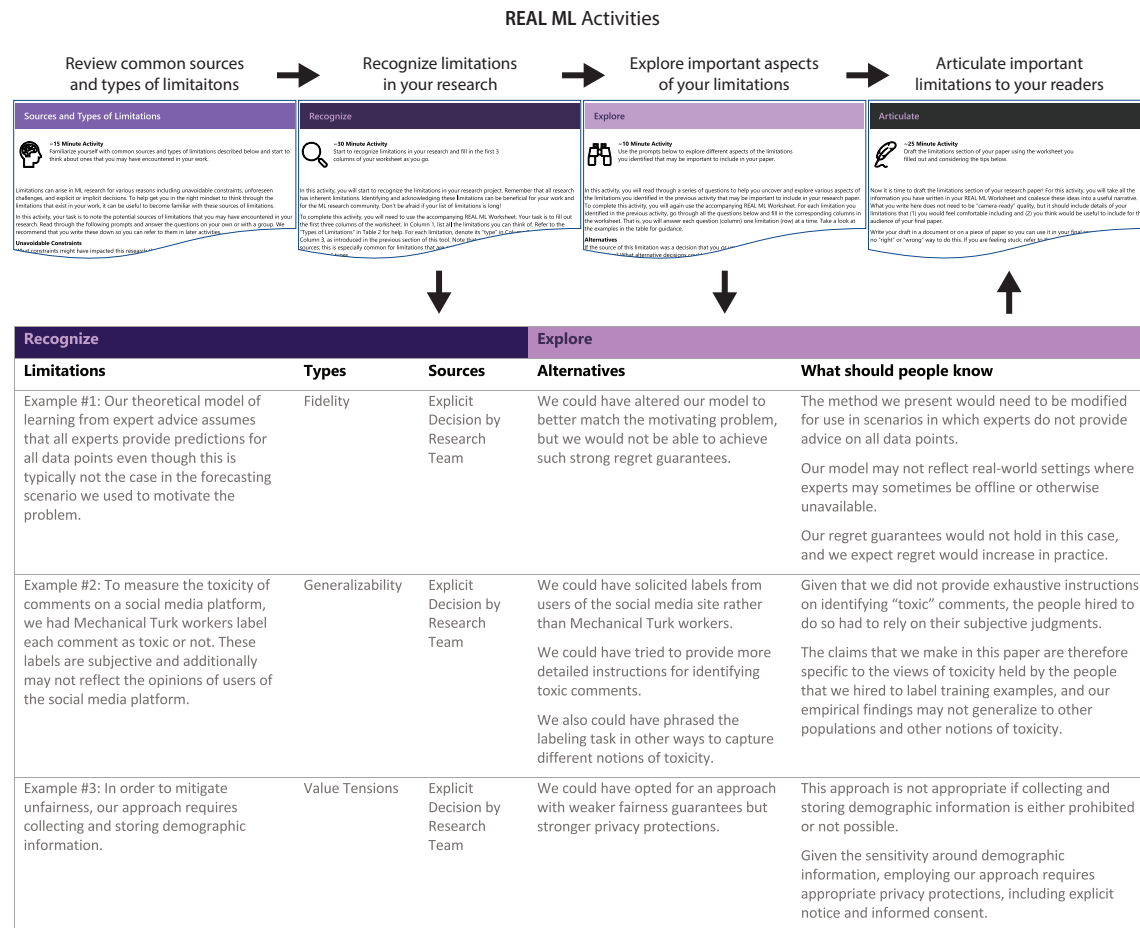
Participants also struggled with where and how to articulate limitations. Some participants had difficulty knowing how to develop a narrative around limitations that would be valuable to different audiences. As P6 said, *“the most difficult thing for me is how to write a [...] coherent story.”* Several other participants said they were not sure if this “story” should appear in a dedicated limitations section or if limitations should be introduced throughout the paper. Finally, several participants said they were unsure how much information was too much. They noted that some types of limitations occur so commonly in ML research that discussing such limitations would risk wasting readers’ time. P12 said they struggled with knowing where to draw the line when articulating limitations, noting that *“it’s hard to [strike] the balance between simply listing some caveats and it turning into a total philosophical paper.”* Similarly, P4 said they could have included an extra two pages in their paper on their dataset’s lack of generalizability across cultures and the societal impacts of this limitation, but thought it *“would probably be a waste of time for the ML research audience.”*

5 REAL ML

Our findings reveal many challenges faced by ML researchers. Some of these challenges stem from community norms that disincentivize disclosure and discussion of limitations, but other challenges can be attributed to a lack of guidance and appropriate training on recognizing, exploring, and articulating limitations. When developing REAL ML, we therefore took into account participants’ expressed needs for guidance, iteratively updating the tool based on their feedback and on our observations of their attempts to use it. In this section, we describe REAL ML and the reasoning behind our main design decisions. Figure 1 provides an overview of REAL ML, which is broken down into an introduction plus four content sections: 1) sources and types of limitations, 2) recognizing limitations, 3) exploring limitations, and 4) articulating limitations. Each section includes guided activities and resources for ML researchers to use when writing limitations sections. We describe each section below; the full tool is in the appendix and available at <https://github.com/jesmith14/REAL-ML>.

5.1 Sources and Types of Limitations

The first section asks ML researchers to familiarize themselves with sources of limitations and types of limitations that occur commonly in ML research and to start thinking about how these might relate to their research. Sources of limitations are broken into three broad categories: unavoidable constraints (e.g., constraints on time or resources), unforeseen challenges (e.g., experimental failures or negative results), and implicit and explicit decisions made during



REAL ML Worksheet

Fig. 1. An overview of the instructions, activities, and worksheet included in REAL ML.

the ML research process. For this last category, REAL ML directs ML researchers to a list of common decision-making points in the ML research process where limitations could arise, as shown in Table 2. They are prompted to use this guidance to reflect on possible sources of limitations of their research. Finally, they are presented with a list of types of limitations that occur commonly in ML research, as shown in Table 3, to reflect on and return to later.

All versions of REAL ML, starting with the initial prototype, included something akin to Tables 2 and 3, but the content and placement evolved over the course of our study. Some decision-making points (e.g., the composition of the research team and ablation studies) were suggested by participants. Participants who saw early versions of REAL ML also pointed out that limitations can arise from sources other than decisions made by researchers. For example, P2 noted that one of their biggest sources of limitations was “time constraints.” P5 mentioned that limitations were sometimes caused by experimental failures. They said, as a reader of ML papers, they often wished they knew about limitations arising from other ML researchers’ failed experiments so they would know what not to do in the future. P7 suggested that limitations fall into

Table 2. Decision-making points presented in the “sources and types of limitations” section of REAL ML.

Decision-making points	Examples
Composition of the research team	Demographic features (e.g., race, gender), disciplinary training (e.g., computer science, medicine), epistemological perspectives (e.g., Bayesian vs. Frequentist), or other researcher characteristics that can influence the approach to research and interpretation of the findings
Related work	The specific fields with which your current study is engaging and which may shape your research; prior work to which your current work is responding; prior work upon which your current work is building
Problem formulation	The general problem that motivates the research; the specific research questions developed to get at that problem
Formalism of the problem	Mathematical statement of the problem that your study is trying to address; technical assumptions (e.g., i.i.d. data points)
Technical approach	Learning algorithm; statistical model; hyperparameter choices
Theoretical claims	Theoretical guarantees such as error bounds; analyses of computational complexity; mathematical derivations
Datasets	The collection, curation, and selection of datasets; the use of particular datasets for training or evaluating
Empirical evaluation setup and metrics	Experimental setup including approaches to be compared, metrics, parameter settings; research subjects
Ablation studies	Setup for ablation studies, including components removed and metrics

two categories: those that researchers have control over and those that they do not. They went on to explain that “[lack of access to compute] is a type of limitation that I have no control over, I can do my best... but I can [only] do whatever I can with the small amount of compute that I have, so this is a limitation that is out of my control.” Our decision to include unavoidable constraints and unforeseen challenges as sources of limitations was based on this early feedback from participants.

Initially there was disagreement among our team as to whether it would be more effective to ask ML researchers to begin by reflecting on possible sources of limitations of their research before considering types of limitations or to instead begin with types of limitations before considering sources of limitations. After trying both approaches with different participants, we found that beginning with sources of limitations was a more successful way to spark open-ended brainstorming. This approach gave participants an opportunity to reflect on why they had made various decisions, which may have led to limitations, during the ML research process. In contrast, jumping straight to types of limitations without this initial reflection caused participants to anchor too much on the examples we provided.

5.2 Recognizing Limitations

The next section prompts ML researchers to build on the brainstorming activity in the first section by filling in a worksheet with a list of the limitations of their research, along with their sources and types. The worksheet is in the appendix.

This section evolved conjointly with the previous section. We found that participants were able to more easily recognize the limitations of their research after brainstorming about sources and types of limitations. For example, P22 was pleasantly surprised by their ability to recognize new limitations of their research after being exposed to Table 3, and later said, “I can honestly see all of these limitation types applying to my work—and honestly to all ML work.” As discussed in Section 4.2, junior researchers were particularly eager for guidance about recognizing limitations, and of the four survey participants, we found that the one junior researcher rated this section as more useful than the three senior researchers.

Table 3. Types of limitations presented in REAL ML. (The tool includes additional examples of each type of limitation.)

Types of Limitations	Probes to Uncover Limitation	Examples
Fidelity	How faithfully do the formalism of the problem, the technical approach, and the results map onto the motivating problem that drives the work?	The training data was labeled even though similar real-world data is not usually labeled.
Generalizability	To what extent do the results hold in different contexts? How broadly or narrowly should the claims in the paper be interpreted? How broadly can the technical approach be applied across domains?	Model was developed for a particular scenario and does not apply to other scenarios or contexts.
Robustness	How sensitive are the results to minor violations of assumptions (e.g., small tweaks to mathematical model, metrics, hyperparameters)?	Adding a small amount of noise in the data dramatically reduces accuracy.
Reproducibility	To what extent could other researchers reproduce the study?	Researchers provide details on parameter settings used but cannot share code or data because they are proprietary.
Resource Requirements	Is the technical approach computationally efficient? Does it scale? What other resources does the technical approach require?	Technical approach requires specialized hardware.
Value Tensions	Are some values (e.g., novelty, simplicity, high accuracy, low false positive rate, ease of implementation, interpretability, efficiency) sacrificed in pursuit of others?	The model has high accuracy on a test dataset but is a black box and hard to interpret.
Vulnerability to Mistakes and Misuse	How sensitive are the results to human errors, unintended uses, or malicious uses?	System operators are liable to misinterpret results without sufficient training.

5.3 Exploring Limitations

This section asks ML researchers to answer a series of questions designed to help them explore the limitations they recognized in the previous section, with the goal of uncovering information that may be important to articulate. They are asked to record their answers in the worksheet. For each limitation, they are first asked to think through potential alternative approaches that could have been taken—that is, alternative decisions that could have been made or alternative research designs that could have been explored—as well as pros and cons of each. Next, they are prompted to reflect on what different audiences might need to know about this limitation, considering the distinct needs of reviewers, researchers, practitioners, and people who use or are affected by ML systems that build on their research.

This section was designed to address challenges around determining which information to focus on. We took the approach of encouraging ML researchers to think through the information that would be most valuable to different audiences, rather than dictating what we felt was most important. In early versions of REAL ML, we included probing questions drawn directly from the four-step process of Ross and Zaidi [25], asking about potential alternative approaches, the potential impacts of each limitation, and how these impacts were mitigated. Based on participant feedback, we later deemphasized mitigating impacts so that researchers who had not already taken explicit steps to mitigate impacts would still have an opportunity to reflect on how these impacts could be mitigated in the future. Finally, three of the four survey participants indicated that listing impacts of limitations in the worksheet felt redundant with other questions, leading us to merge the consideration of impacts into the reflection about what different audiences might need to know.

5.4 Articulating Limitations

In the final section, ML researchers are asked to build on the information they recorded in the worksheet to draft a limitations section. The goal is to develop a narrative around the limitations of their research that is valuable to different audiences. The narrative does not need to be of “camera-ready” quality and, indeed, they may choose to introduce information about limitations throughout their paper rather than including it in a dedicated limitations section. To help with narrative development, REAL ML includes a set of “tips and tricks” for articulating limitations, including guidance about speculation, prioritization, broader impacts of limitations, and concerns about paper rejection.

The “tips and tricks” were targeted at addressing needs repeatedly expressed by our participants, like struggles with prioritization and organization, as discussed in Section 4.3. The specific guidance on prioritization—that is, focusing on limitations that might have the most severe impacts, in addition to those that would be most valuable for different audiences to know about—was inspired, in part, by remarks from P23, who mentioned the need for a “Richter scale” to measure the severity of the impacts of limitations. However, in order to avoid embedding our own biases about impacts into REAL ML, we opted to leave the determination of severity to ML researchers using the tool.

All four survey participants said they “agreed” or “strongly agreed” with the statement “I felt better prepared to write about these limitations than I would have been without the tool,” an indication that—at least for this very small set of participants—even senior researchers who felt comfortable recognizing the limitations of their research found the tool useful for articulating those limitations, although a larger evaluation study would be needed to fully measure the benefits.

6 LIMITATIONS, DISCUSSION, AND FUTURE WORK

Our study itself has limitations that could impact both our research findings and REAL ML. First, we conducted our qualitative research through an interpretivist lens. As such, our research findings reflect our own biases and subjectivities. Second, because we largely recruited participants via our personal networks, selection bias was unavoidable. Notably, we ended up with an imbalanced sample in terms of research backgrounds. For example, many interview participants had expertise in natural language processing, but only one had expertise in learning theory. It is also possible that participants were more open to recognizing, exploring, and articulating the limitations of their research than is typical of ML researchers. These limitations may have impacted both the perspectives reflected in our research findings and the evolution of REAL ML. ML researchers who use REAL ML should recognize they may have different needs when it comes to articulating the limitations of their research (e.g., preferring to introduce information about limitations throughout the paper rather than including it in a dedicated limitations section, recognizing types of limitations other than the ones listed in REAL ML). REAL ML is meant to act as a guide for ML researchers. It is not all-encompassing nor prescriptive, and we recommend that ML researchers adapt both the suggested activities and the outputs to meet their needs. (Although this paper is not a typical ML paper and therefore not the type of paper that REAL ML was intended for, we adapted the suggested activities to guide us when writing this limitations section.) Additionally, since our study focused on iteratively developing and testing REAL ML rather than empirically evaluating it in action, we have not yet conducted a full evaluation of the final version with ML researchers beyond the four survey participants. Additional research is needed to measure the effectiveness of REAL ML in practice.

As discussed in Section 1, REAL ML is intended for use *posthoc* via a process known as “reflection on action” [20], helping ML researchers look back on their research during the paper-writing stage of the ML research process. Several interview participants said they would find it valuable to have a similar tool targeted at earlier stages of the ML research process, noting that if they had thought through the limitations of their research earlier, they would have pivoted their

research direction or made different decisions. A natural next step would therefore be to adapt REAL ML to explicitly target earlier stages of the ML research process in order to encourage “reflection in action” [26].

REAL ML was developed to address the challenges faced by ML researchers when recognizing, exploring, and articulating limitations. However, we note that some of these challenges go beyond the scope of what can be accomplished with a tool, requiring broader shifts in community norms to address. For example, many participants expressed concerns about the emphasis on generalizability in the ML review process, mentioning that this had led them to underemphasize or omit some of the limitations of their research. As others have noted [2, 8], generalizability is highly valued in the ML research community. Deemphasizing generalizability would require a major shift in community norms. Along similar lines, as discussed in Section 4.3.1, many participants expressed concerns about whether disclosing limitations would increase the likelihood of paper rejection, in part because of ML publication venues’ reviewing norms and in part because of perceived differences in reviewers’ opinions about disclosing limitations. Participants suggested that a version of REAL ML be provided to reviewers in order to standardize the way limitations are critiqued during the ML review process. Standardized guidance for reviewers could serve as a complement to emerging efforts to promote transparency on the part of authors, such as the introduction of broader impacts statements at NeurIPS 2020 [16] and the subsequent NeurIPS 2021 paper checklist [1], which encouraged authors to disclose the limitations of their research. Of course, more research is needed to explore how reviewers would respond to heightened transparency around limitations and whether standardizing the way limitations are critiqued would improve the ML review process.

7 CONCLUSION

Despite previous work calling out the importance of transparency around limitations, the ML research community lacks well-developed norms around disclosing and discussing limitations. In this paper, we uncovered the practical and cultural challenges faced by ML researchers when recognizing, exploring, and articulating limitations. Specifically, we found that the ML research community does not have a single, agreed-upon definition of limitations of ML research, nor does it have a standardized process for disclosing and discussing limitations. We discovered that junior researchers were particularly eager for guidance about recognizing limitations, and that both junior and senior researchers would benefit from guidance about articulating limitations. Using a three-stage interview and survey study, we conducted an iterative design process to develop and test REAL ML, a set of guided activities to help ML researchers recognize, explore, and articulate the limitations of their research. REAL ML was intended to address some of the practical challenges faced by ML researchers. However, our study also exposes cultural challenges that go beyond the scope of REAL ML and will require broader shifts in community norms to address. We hope our study and REAL ML help move the ML research community toward more actively and appropriately engaging with limitations.

ACKNOWLEDGMENTS

We are grateful to all of our study participants for their time. We also thank members of Microsoft’s FATE research group, who provided valuable feedback throughout our study and the development of REAL ML. Special thanks goes to David Alvarez-Melis, Zana Bućinca, Mahsan Nourani, Divya Shanmugam, and Angelina Wang, who volunteered their time to help us pilot our interview protocol.

FUNDING/SUPPORT

Most of this research was conducted while the first author was an intern at Microsoft. All other authors are full-time employees of Microsoft.

REFERENCES

- [1] Alina Beygelzimer, Yann Dauphin, Percy Liang, and Jennifer Wortman Vaughan. 2021. Introducing the NeurIPS 2021 Paper Checklist. (2021). <https://neuripsconf.medium.com/introducing-the-neurips-2021-paper-checklist-3220d6df500b> Medium Article.
- [2] Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2021. The values encoded in machine learning research. *arXiv preprint arXiv:2106.15590* (2021).
- [3] Isabelle Boutron and Philippe Ravaud. 2018. Misrepresentation and distortion of research in biomedical literature. *Proceedings of the National Academy of Sciences* 115, 11 (2018), 2613–2619.
- [4] Stéphane Brutus, Herman Aguinis, and Ulrich Wassmer. 2013. Self-reported limitations and future directions in scholarly reports: Analysis and recommendations. *Journal of Management* 39, 1 (2013), 48–75.
- [5] Fiona Burlig. 2018. Improving transparency in observational social science research: A pre-analysis plan approach. *Economics Letters* 168 (2018), 56–60.
- [6] Garret Christensen and Edward Miguel. 2018. Transparency, reproducibility, and the credibility of economics research. *Journal of Economic Literature* 56, 3 (2018), 920–80.
- [7] Victoria Clarke and Virginia Braun. 2014. Thematic analysis. In *Encyclopedia of critical psychology*. Springer, 1947–1952.
- [8] Catherine D’ignazio and Lauren F Klein. 2020. *Data Feminism*. MIT press.
- [9] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (2021), 86–92.
- [10] Christophe Giraud-Carrier and Margaret H Dunham. 2011. On the importance of sharing negative results. *ACM SIGKDD Explorations Newsletter* 12, 2 (2011), 3–4.
- [11] Carlos Alberto Guimarães. 2006. Structured abstracts. Narrative review. *Acta Cirúrgica Brasileira* 21 (2006), 263–268.
- [12] John PA Ioannidis. 2007. Limitations are not properly acknowledged in the scientific literature. *Journal of clinical epidemiology* 60, 4 (2007), 324–329.
- [13] JAMA. 2021. Instructions for Authors. (2021). <https://jamanetwork.com/journals/jama/pages/instructions-for-authors> Publishing Instructions.
- [14] JNeuroSci. 2021. Information for Authors. (2021). <https://www.jneurosci.org/content/information-authors> Publishing Instructions.
- [15] Thomas Liao, Rohan Taori, Inioluwa Deborah Raji, and Ludwig Schmidt. 2021. Are We Learning Yet? A Meta Review of Evaluation Failures Across Machine Learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.
- [16] Hsuan-Tien Lin, Maria-Florina Balcan, Raia Hadsell, and Marc’Aurelio Ranzato. 2020. Getting Started with NeurIPS 2020. (2020). <https://neuripsconf.medium.com/getting-started-with-neurips-2020-e350f9b39c28> Medium Article.
- [17] Lorelei Lingard and Christopher Watling. 2021. The art of limitations. In *Story, Not Study: 30 Brief Lessons to Inspire Health Researchers as Writers*. Springer, 53–59.
- [18] Matthew C Makel and Jonathan A Plucker. 2017. *Toward a more perfect psychology: Improving trust, accuracy, and transparency in research*. American Psychological Association.
- [19] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*. 220–229.
- [20] Hugh Munby. 1989. Reflection-in-action and reflection-on-action. *Current issues in education* 9, 1 (1989), 31–42.
- [21] Priyanka Nanayakkara, Jessica Hullman, and Nicholas Diakopoulos. 2021. Unpacking the expressed consequences of AI research in broader impact statements. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 795–806.
- [22] Arvind Narayanan. 2019. How to recognize AI snake oil. *Arthur Miller Lecture on Science and Ethics* (2019).
- [23] Partnership on AI. 2021. Managing the Risks of AI Research: Six Recommendations for Responsible Publication. *white paper* (2021).
- [24] Milo A Puhon, Elie A Akl, Dianne Bryant, Feng Xie, Giovanni Apolone, and Gerben ter Riet. 2012. Discussing study limitations in reports of biomedical studies-the need for more transparency. *Health and quality of life outcomes* 10, 1 (2012), 1–4.
- [25] Paula T Ross and Nikki L Bibler Zaidi. 2019. Limited by our limitations. *Perspectives on medical education* 8, 4 (2019), 261–264.
- [26] Donald A Schön. 1984. The architectural studio as an exemplar of education for reflection-in-action. *Journal of Architectural Education* 38, 1 (1984), 2–9.
- [27] Andrew Selbst, danah boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet A. Vertesi. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*.
- [28] Hong Shen, Wesley H Deng, Aditi Chattopadhyay, Zhiwei Steven Wu, Xu Wang, and Haiyi Zhu. 2021. Value Cards: An Educational Toolkit for Teaching Social Impacts of Machine Learning through Deliberation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 850–861.
- [29] Kate Sim, Andrew Brown, and Amelia Hassoun. 2021. Thinking Through and Writing About Research Ethics Beyond “Broader Impact”. *arXiv preprint arXiv:2104.08205* (2021).
- [30] Kiri L Wagstaff. 2012. Machine learning that matters. In *Proceedings of the 29th International Conference on Machine Learning*. 1851–1856.
- [31] Amélie Yavchitz, Philippe Ravaud, Sally Hopewell, Gabriel Baron, and Isabelle Boutron. 2014. Impact of adding a limitations section to abstracts of systematic reviews on readers’ interpretation: A randomized controlled trial. *BMC Medical Research Methodology* 14, 123 (2014).

A REAL ML

This section includes screenshots of REAL ML. The full tool is available at <https://github.com/jesmith14/REAL-ML>.

Fig. 2. REAL ML, Page 1.

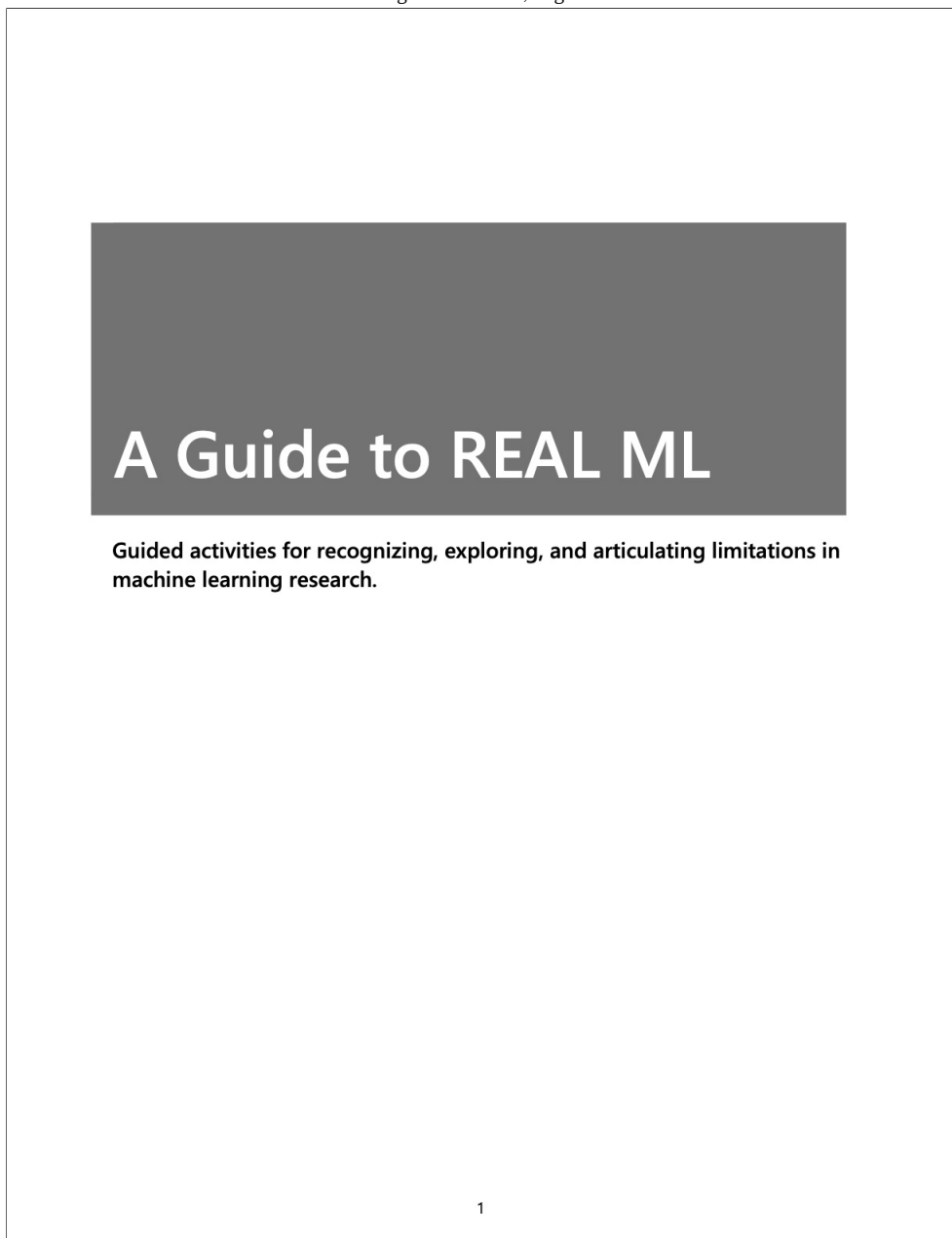


Fig. 3. Real ML, Page 2.

Introduction

What is this tool?

The Recognizing, Exploring, and Articulating Limitations in Machine Learning research tool (REAL ML) is a set of guided activities to help ML researchers recognize, explore, and articulate the limitations that arise in their research. This document acts as an instructional guide and is meant to be completed with the corresponding REAL ML worksheet that you will fill out as you go through the activities below.

What is a limitation?

The process of conducting research is never without some limitations—*drawbacks in the design or execution of the research that present a possible threat to the validity of your findings and claims*. Research rarely takes place under ideal conditions; instead, researchers must frequently navigate practical constraints and unexpected challenges in seeking to answer their research questions. Researchers also make all sorts of explicit and implicit choices in the design and execution of their studies. All of these may affect the conclusions that researchers are able to draw from their work. In many fields, limitations are understood to be an inherent part of research and researchers are expected and encouraged to disclose and discuss them. Doing so is viewed as critical to maintaining the integrity of a field of research and to helping advance collective knowledge.

How do limitations relate to broader impact statements?

While limitations and broader impacts share some characteristics, the two are not the same. Limitations focus on parts of the research process that might threaten the validity of your conclusions; broader impacts focus on the downstream implications of your research when put to some practical use. For guidance on writing a Broader Impact statement in your paper, we suggest you look at the NeurIPS Broader Impact statement or the NeurIPS Paper Checklist Guidelines.

Who should use this tool?

We have developed this tool for use by ML researchers specifically.

Why should I write about limitations?

- Addressing limitations in your research can improve the scientific rigor of your work by ensuring that you are more precise in describing what the research entailed and what claims your research supports.
- Making the limitations of your research explicit can help readers develop a more accurate understanding of the research project, its findings, and the conclusions that you've drawn from these findings.
- Papers that include critical reflection on limitations can improve readers' justified trust in the research findings; papers that lack such reflection can make readers overly skeptical of research claims.
- Imparting a clearer understanding of the limitations of your research can help to avoid inappropriate applications of your research in practice.
- Pro-actively reflecting on limitations in research papers may help to address possible concerns on the part of reviewers, improving the chances of having a paper evaluated correctly and thus being accepted.
- Describing the limitations of your research—and the sources of these limitations—can foster collective scientific progress by highlighting opportunities for future research; disclosing limitations can also mean that others are more likely to build on your work.

When should I use this tool?

This tool is intended to be used during the paper-writing phase of research. You may also find some of the activities useful earlier in your research, including during the development of your research design.

Fig. 4. Real ML, Page 3.

How do I use this tool?

We recommend that you use this tool for one ML research project/paper at a time. You can complete the activities in this tool on your own, with a partner, or with a group of researchers. When working in a group, you can discuss the answers to questions out loud and have someone write down your thoughts and takeaways. It is recommended that you use the REAL ML Worksheet to help structure your note-taking.

How long should I expect to spend on this?

It will take about 45 minutes for you to get through the first two activities in this tool—learning about “Sources and Types of Limitations” and then “Recognizing” some of them in your own work. If you continue onto the “Exploring” and “Articulating” activities, you should expect to spend at least 1 more hour to complete the full set of activities.

How was this tool created?

This tool was created by researchers at Microsoft Research who interviewed researchers from the ML community about their experience with limitations and with various iterations of this tool. Participants’ input and feedback shaped the ultimate design of this tool. More information can be found in the associated publication:

Jessie J. Smith, Saleema Amershi, Solon Barocas, Hanna Wallach, and Jennifer Wortman Vaughan. REAL ML: Recognizing, Exploring, and Articulating Limitations of Machine Learning Research. In *Proceedings of the 5th ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2022.

Where can I find more information?

The latest version of REAL ML and the accompanying REAL ML Worksheet can be found at <https://github.com/jesmith14/REAL-ML>. You can contact the team at realml@microsoft.com.

Fig. 5. Real ML, Page 4.

Sources and Types of Limitations



~15 Minute Activity

Familiarize yourself with common sources and types of limitations described below and start to think about ones that you may have encountered in your work.

Limitations can arise in ML research for various reasons including unavoidable constraints, unforeseen challenges, and explicit or implicit decisions. To help get you in the right mindset to think through the limitations that exist in your work, it can be useful to become familiar with these sources of limitations.

In this activity, your task is to note the potential sources of limitations that you may have encountered in your research. Read through the following prompts and answer the questions on your own or with a group. We recommend that you write these down so you can refer to them in later activities.

Unavoidable Constraints

What constraints might have impacted this research that were out of your control? For example:

- Time constraints (e.g., not enough time to run experiments on the entire dataset, had to complete the research before funding was terminated)
- Resource constraints (e.g., lack the desired technical infrastructure to perform the study; not enough money to pay for desired compute power)
- Lack of access (e.g., unable to obtain access to a dataset that would have been more relevant to the research question, not able to secure necessary license to use certain code)

Unforeseen Challenges

What unforeseen challenges did you encounter in this research that might have resulted in limitations? For example:

- Experimental failures (e.g., your dataset was too sparse to create a meaningful model, your approach didn't perform as well as previously set benchmarks)
- Negative results (e.g., the results of your experimental study do not confirm your hypothesis, a theorem you initially set out to prove turned out to be false)

Implicit and Explicit Decisions

What decisions did you make in your research? Which decisions were more likely to lead to limitations than others?

Fig. 6. Real ML, Page 5.

Decision-making points

The following table gives examples of common decision-making points that you may have encountered during your research:

Decision-making points	Examples
Composition of the research team	Demographic features (e.g., race, gender), disciplinary training (e.g., computer science, medicine), epistemological perspectives (e.g., Bayesian vs. frequentist), or other researcher characteristics that can influence the approach to research and interpretation of the findings
Related work	The specific fields with which your current study is engaging and which may shape your research; prior work to which your current work is responding; prior work upon which your current work is building
Problem formulation	The general problem that motivates the research; the specific research questions developed to get at that problem
Formalism of the problem	Mathematical statement of the problem that your study is trying to address; technical assumptions (e.g., i.i.d. data points)
Technical approach	Learning algorithm; statistical model; hyperparameter choices
Theoretical claims	Theoretical guarantees such as error bounds; analyses of computational complexity; mathematical derivations
Datasets	The collection, curation, and selection of datasets; the use of particular datasets for training or evaluating
Empirical evaluation setup and metrics	Experimental setup including approaches to be compared, metrics, parameter settings; research subjects
Ablation studies	Setup for ablation studies, including components removed and metrics

Types of limitations

These unavoidable constraints, unforeseen challenges, and explicit or implicit decisions can result in a range of limitations, a list of which we provide in the following table:


Types of Limitations	Probes to Uncover Limitation	Examples
Fidelity	How faithfully do the formalism of the problem, the technical approach, and the results map onto the motivating real-world problem that drives the work?	<p>The formalism of the problem includes so many assumptions that results are not clearly applicable to the motivating problem in the real world.</p> <p>There are large gaps between the data/model/metrics and the motivation.</p> <p>The training data was labeled even though similar real-world data is not usually labeled.</p> <p>The distribution of the data is different than what you would encounter in the real world.</p>

Fig. 7. Real ML, Page 6.

Generalizability	To what extent do the results hold in different contexts? How broadly or narrowly should the claims in the paper be interpreted? How broadly can the technical approach be applied across domains?	Model was developed for a particular scenario and does not apply to other scenarios or contexts. Only a single dataset, small dataset, or datasets from a specific domain were used. If taken out of context, the results of this study could be misleading.
Robustness	How sensitive are the results to minor violations of assumptions (e.g., small tweaks to mathematical model, metrics, hyperparameters)?	Two equally reasonable ways of formalizing the problem would lead to dramatically different takeaways. Adding a small amount of noise in the data dramatically reduces accuracy. Selecting different reasonable metrics would lead to different outcomes or tradeoffs for the experiments.
Reproducibility	To what extent could other researchers reproduce the study?	Researchers provide details on parameter settings used but cannot share code or data because they are proprietary. The code and data are shared, but no guidance about running the experiments is given.
Resource Requirements	Is the technical approach computationally efficient? Does it scale? What other resources does the technical approach require?	Approach works well with a couple thousand training examples but cannot handle a couple million. Technical approach requires specialized hardware.
Value Tensions	Are some values (e.g., novelty, simplicity, high accuracy, low false positive rate, ease of implementation, interpretability, efficiency) sacrificed in pursuit of others?	The model has high accuracy on a test dataset but is a black box and hard to interpret. The model is optimized to favor certain kinds of errors over others (e.g., false positives are favored over false negatives). Seeking voluntary and informed consent from human subjects leads to selection bias in the dataset.
Vulnerability to Mistakes and Misuse	How sensitive are the results to human errors, unintended uses, or malicious uses?	System operators are liable to misinterpret results without sufficient training. There is a serious risk if the technique, which was developed in one context, is applied in others for which it is not suited.

Fig. 8. Real ML, Page 7.

Recognize



~30 Minute Activity
 Start to recognize limitations in your research and fill in the first 3 columns of your worksheet as you go.

In this activity, you will start to recognize the limitations in your research project. Remember that all research has inherent limitations. Identifying and acknowledging these limitations can be beneficial for your work and for the ML research community. Don't be afraid if your list of limitations is long!

To complete this activity, you will need to use the accompanying REAL ML Worksheet. Your task is to fill out the first three columns of the worksheet. In Column 1, list all the limitations you can think of. Refer to the "Types of Limitations" in Table 2 for help. For each limitation, denote its "type" in Column 2 and its "source" in Column 3, as introduced in the previous section of this tool. Note that one limitation might have multiple sources; this is especially common for limitations that are categorized under the "Generalizability" and "Robustness" types.

Still unable to come up with a limitation?
 If you are having difficulty thinking of a limitation, recall all of the different sources that you previously identified. There are always limitations to uncover, whether or not you decide to put them in your final paper.

7

Fig. 9. Real ML, Page 8.

Explore



~10 Minutes Per Limitation

Use the prompts below to explore different aspects of the limitations you identified that may be important to include in your paper.

In this activity, you will read through a series of questions to help you uncover and explore various aspects of the limitations you identified in the previous activity that may be important to include in your research paper. To complete this activity, you will again use the accompanying REAL ML Worksheet. For each limitation you identified in the previous activity, go through all the questions below and fill in the corresponding columns in the worksheet. That is, you will answer each question (column) one limitation (row) at a time. Take a look at the examples in the table for guidance.

Alternatives

If the source of this limitation was a decision that you or your research team made, what are the pros/cons of this decision? What alternative decisions could you and/or the research team have made instead?

If the source of this limitation was practical constraints or unforeseen challenges, what alternative research designs could you have explored if those could have been avoided? What are the pros/cons of those alternatives?

Write down your biggest takeaways in Column 4 of the worksheet.

What Should People Know?

What are the impacts of this limitation on your study's results and claims? What do different readers of your paper need to know about this limitation and its impacts?


- What might reviewers need to know (e.g., what did you do to minimize the possible impacts of this limitation)? This category is a great opportunity to prepare for reviewer feedback!
- What does the ML community need to know to understand and trust the claims of your work? What do future researchers building on this work need to know?
- What do practitioners who might build on this research need to know?
- What do people who are using or are affected by the systems these practitioners build need to know?

Write down your biggest takeaways in Column 5 of the worksheet.

Repeat this activity for each limitation you identified in the worksheet. It is recommended that you spend at least 10 minutes on this activity for each individual limitation you would like to explore. Once you have completed each row, move on to the next section.

Fig. 10. Real ML, Page 9.

Articulate



~25 Minute Activity

Draft the limitations section of your paper using the worksheet you filled out and considering the tips below.

Now it is time to draft the limitations section of your research paper! For this activity, you will take all the information you have written in your REAL ML Worksheet and coalesce these ideas into a useful narrative. What you write here does not need to be “camera-ready” quality, but it should include details of your limitations that (1) you would feel comfortable including and (2) you think would be useful to include for the audience of your final paper.

Write your draft in a document or on a piece of paper so you can use it in your final research paper. There is no “right” or “wrong” way to do this. If you are feeling stuck, refer to the tips and tricks below.

Tips and Tricks for Framing Your Limitations

Speculating about your claims?
If you encounter a limitation where you aren’t sure of the impact on your findings, begin by reporting what you do know. Then, if you choose to speculate about something unknown, be explicit about your speculation. For example: “we hypothesize that our solution could be used in these other ways, but future work should provide verifiable evidence for these.”

Running into page limits?
Consider focusing on the limitations, or details of your limitations, that you think may have the biggest impact on the validity of your findings. Additionally, think about your primary audience(s) and what is most important for them to know to accurately interpret your research claims. In many research venues, you can include further details or even your completed REAL ML Worksheet in an appendix to your research paper.

Worried that a limitation is grounds for rejection?
We understand that authors might fear that complete honesty about limitations could lead to their papers being rejected. As noted in the NeurIPS paper checklist, you should keep in mind that your paper might be more likely to be rejected if reviewers discover limitations in your paper that you didn’t acknowledge ahead of time and explain why your research still has merit despite these limitations. Recognize that transparency around limitations also helps to preserve the integrity of the ML research community, foster collective scientific progress, and ensure appropriate applications of research findings in practice.

Want to discuss a limitation’s broader impact?
You might have noticed that certain limitations have the potential to lead to serious downstream harms, especially if your research is misapplied or applied maliciously. We recommend that you separate this discussion of broader impacts from your limitations section. For guidance on writing a Broader Impact statement in your paper, we suggest you look at the NeurIPS Broader Impact statement or the NeurIPS Paper Checklist Guidelines.

9

REAL ML Worksheet				
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.	<p>The method we present would need to be modified for use in scenarios in which experts do not provide advice on all data points.</p> <p>Our model may not reflect real-world settings where experts may sometimes be offline or otherwise unavailable.</p> <p>Our regret guarantees would not hold in this case, and we expect regret would increase in practice.</p>
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	<p>We could have solicited labels from users of the social media site rather than Mechanical Turk workers.</p> <p>We could have tried to provide more detailed instructions for identifying toxic comments.</p> <p>We also could have phrased the labeling task in other ways to capture different notions of toxicity.</p>	<p>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.</p> <p>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.</p>
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.	<p>This approach is not appropriate if collecting and storing demographic information is either prohibited or not possible.</p> <p>Given the sensitivity around demographic information, employing our approach requires appropriate privacy protections, including explicit notice and informed consent.</p>
Your Limitations Here...				

Fig. 11. REAL ML, Page 10. The REAL ML Worksheet.

B TOOL PROTOTYPE

The initial prototype of REAL ML consisted of three lists intended to encourage reflection: a list of types of limitations that commonly occur in ML research (Table 4), a list of common decision-making points in the ML research process where limitations could arise (Table 5), and a list of probing questions to answer when preparing to articulate a limitation (Table 6). The versions included here correspond to the V1 prototype referenced in Table 1, the earliest version used in our study.

Table 4. The types of limitations included in the V1 prototype of REAL ML. Through our iterative design process, this table evolved into the one included in the “sources and types of limitations” section of REAL ML, shown in Table 3 (with pared-down examples).

Types of Limitations	Definition of Limitation	Examples & Impacts
Fidelity (or Applicability)	How faithfully do the formalism of the problem, the technical approach, and the results map onto the motivating real-world problem that drives the work?	The formalism of the problem includes so many assumptions that results are not clearly applicable to the motivating problem in the real world.
Robustness	How sensitive are the results to (minor) violations of assumptions?	Two equally reasonable ways of formalizing the problem lead to dramatically different takeaways. Adding a small amount of noise in the data dramatically reduces accuracy. If a system operator misinterprets results, bad things could happen.
Generalizability	To what extent do the results hold in different contexts? How broadly can the technical approach be applied?	Model is developed for a particular scenario and does not apply to other scenarios or contexts, e.g., face recognition system requires good lighting.
Replicability	To what extent could other researchers replicate the results?	Authors provide details on parameter settings used but cannot share code or data because they are proprietary.
Scope of Claims	How broadly or narrowly should the claims in the paper be interpreted? How do the implicit assumptions made in the evaluation impact interpretation of results?	Only a single random seed was tried. Only a single dataset was used. Only one part of a more complex system was modeled.
Validity/Rigor	Were there better research practices that could have been used?	Because access to data was limited, multiple experiments were run iteratively using the same test data.
Computational Efficiency	Is the technical approach computationally efficient? Does it scale?	Approach works well with a couple thousand training examples, cannot handle a couple million. Approach requires a lot of computational resources and therefore has bad carbon footprint.
Other Costs	What other resources does the technical approach require?	Technical approach requires specialized hardware.
Other Tradeoffs	What tradeoffs must be made to achieve the results?	Technical approach has good accuracy, but sacrifices privacy.

Table 5. The decision-making points included in the V1 prototype of REAL ML. Through our iterative design process, this table evolved into the one included in the “sources and types of limitations” section of REAL ML, shown in Table 2.

Decision making points	Definition
High level problem formulation	High level description of the problem this paper trying to solve; motivation for the problem; research questions
Formalism of the problem	Mathematical statement of the problem this paper is trying to solve; technical assumptions (e.g., i.i.d. data points, adversarial noise)
Technical approach	Learning algorithm; statistical model
Mathematical / theoretical claims	Theoretical guarantees such as error bounds; analyses of computational complexity
Dataset(s)	Any dataset(s) that are generated or used for any purpose, such as training or evaluating ML models
Empirical evaluation setup and metrics	Experimental setup including approaches to be compared, metrics, parameter settings

Table 6. The probing questions to answer included in the V1 prototype of REAL ML. Via our iterative design process, this table evolved into the activities included in the section on exploring limitations, as discussed in Section 5.3.

Step of Limitation Writing	Prompt (answer these questions to write your limitation)
Type of limitation (from areas list)	What type of limitation is this from the areas of limitations? Can you explicitly state the source of this limitation?
Broader impact of limitation	What is the broader (possibly negative if applicable) impact of this limitation on (1) your work? (2) the ML research community? (3) society?
Alternatives	What are potential alternatives you could have taken instead of the choice that led to this limitation?
How limitation was minimized	How did you minimize the impact of this limitation in your research? How do you plan to if you haven't yet?

C STAGE 1 INTERVIEW PROTOCOL

As described in Section 3, stage 1 of our study involved conducting semi-structured interviews with 20 ML researchers. The following interview protocol is the final version that was used to conduct these interviews. Small changes were made throughout stage 1 to ensure that the interview protocol matched the latest version of REAL ML. Specifically, Section #2 of the interview protocol originally prompted participants to begin with types of limitations before considering sources of limitations, but later evolved to begin by asking participants to reflect on possible sources of limitations of their research before considering types of limitations, as described in Section 5. When using the interview protocol, the interviewer personalized the follow-up questions and guidance on using REAL ML for each participant.

Section #1

Goal: Get background information about the participant and their views of limitations

- (1) What areas of machine learning do you work on? Do you typically write theoretical or applied ML papers?
- (2) Have you ever written a limitations section in a paper before? Why or why not? What was the experience like?
- (3) What challenges (if any) have you faced before when it comes to identifying your assumptions and/or limitations in a research paper? When it comes to writing these things?
- (4) How do you define a limitation? What do you think indicates a limitation has occurred in an ML research paper?

Section #2

Goal: Understand how ML researchers might use the tool, and ways that it can be improved

Interviewer indicates that it is time to start reviewing the paper that the interviewee has brought to the interview. They explain that the next set of questions relates specifically to the research conducted in that paper. They mention “Our intention is to reflect on the research process and paper writing process. There is no judgement here; this is a safe space.” Begin with the first high level question to see what they say without showing them the tool.

- (1) What kinds of assumptions did you make in your paper? What kinds of limitations did this research have?
Interviewer now shows the taxonomy “questionnaire” to the participant as a reference and lets them use it as a guide to refer to when thinking through the kinds of limitations of their ML research (making it very clear that this is a preliminary mockup and a very early version of the tool, and by no means will be the end tool). The participant selects one of their own research papers or a research paper that they are familiar with. They will be guided to the “limitations” of a ML research paper. The interviewer asks the following questions for each limitation type or for each decision type (depending on the interview). If each type is taking long, have the participant choose two or three types to focus on.
- (2) What is coming up for you in this limitation/decision category? What possible limitations can you think of that you might have encountered in this work?
If starting with limitations and participant is struggling to identify limitations, guide them to the “Decisions” sheet and prompt them there to see if that helps them remember the stages of research. If starting with decisions and participant is struggling to identify decisions, guide them to the “Limitations” sheet and prompt them there to see if that helps them remember what they may have encountered. When a new limitation from participant is encountered, interviewer guides participants to go through the “Writing a Limitation” prompts:
 - (a) What was the source of this limitation (e.g., an explicit decision, unforeseen consequences, lack of resources)?
 - (b) What were alternatives? Tradeoffs?

- (c) What are the potential broader impacts of this limitation?
- (d) What (if anything) did you do to mitigate this limitation? What could you have done?
- (3) Was this limitation included or excluded from the original paper?
 - If limitation was not included in participants' original paper:*
 - (a) Were you aware of this limitation before publishing your paper? If not, if you had been made aware of this limitation while writing the paper, would you have included it?
 - (b) Would you feel comfortable including this limitation in your paper? Why or why not?
 - (c) Would this limitation be useful to state in your paper? Why or why not?
 - (d) Were there other reasons you chose to not include this in the original paper (e.g., practical reasons like page limits)?

Section #3

Goal: Improve the tool and make it more useful for ML researchers

- (1) Improving the taxonomy:
 - (a) Which limitation types are missing / need improvement from this tool?
 - (b) Which decision-making types are missing / need improvement from this tool?
 - (c) How can the “steps to writing a limitation” be improved? What worked or didn’t work for you? What was most/least helpful?
- (2) When would it be helpful for you to use a tool like this (a more finalized and robust version) for your ML research (e.g., before research, during research, during writing, before submission)?
- (3) In general, when do you think limitations should be stated in an ML research paper (e.g., beginning, middle, end)?
- (4) Do you think that limitations are better expressed in one section? Or better disseminated throughout the entirety of a paper (e.g., when the limitation arose as a new idea is introduced in the paper)?
- (5) In what ways did this tool confuse you or seem counterintuitive?
- (6) In what ways could this tool be modified to be more useful to you as you conduct your own ML research?

D STAGE 2 INTERVIEW PROTOCOL

As described in Section 3, stage 2 of our study involved conducting interviews with six researchers who were knowledgeable about ML and also experts in more sociotechnical fields. Our goal in including these participants was to surface any community norms or assumptions that might have been taken for granted by ML researchers, but would stand out to ML-adjacent researchers. These participants were asked to provide feedback on V2 or V3 of REAL ML.

Section #1

Goal: Get background information about participant and their views on limitations

- (1) What areas of research do you typically focus on? What is your relation to the field of ML? What types of conferences do you publish in?
- (2) What has your experience been towards limitations in your discipline(s)?
- (3) What is your interpretation of “limitations” as they apply to ML research?
 - (a) What do you think are some of the impact(s) of not reporting limitations in ML papers?

- (4) How do you define the word limitation?

Section #2

Goal: Run through the tool and get feedback from participant

- (1) Feedback on limitations table
- (2) Feedback on sources
- (3) Feedback on tips/refining:
 - (a) In the event of page limitations: Which limitations should be the focus? Which aspects of limitations should be the focus? (Feedback on the 4 qualities)
 - (b) When should limitations be in a paper?
 - (c) How should limitations be in a paper?
- (4) Feedback on the table deliverable (other deliverables / output that would be better?)
- (5) General thoughts / feedback on this tool?

E STAGE 3 SURVEY PROTOCOL

As described in Section 3, stage 3 of our study involved conducting an online survey with four ML researchers. We sent a copy of REAL ML to each participant via email and asked them to use it on their own, without any additional guidance. After using REAL ML, each participant provided feedback via an online survey, included below.

- (1) Please paste your final limitations section that you wrote at the end of the tool for your current ML research paper in the space below.
- (2) Please rate your level of agreement or disagreement with the following statements about the tool (Strongly disagree, Disagree, Neither agree nor disagree, Agree, Strongly agree):
 - The tool fostered a greater appreciation of the value of identifying, disclosing, and discussing the limitations of my research.
 - The tool helped me identify limitations that I would not have identified otherwise.
 - I felt better prepared to write about these limitations than I would have been without the tool.
 - I was more willing to disclose and discuss the limitations in my paper than I would have been without the tool.
 - I would use the tool again when writing research limitations of my work.
- (3) Did you choose not to disclose any identified limitations in your writeup? If so, can you tell us about them and explain why you chose not to disclose them?
- (4) Which audience(s) did you have in mind when you prepared your limitations section with this tool?
- (5) Did the tool help you uncover any limitations you weren't previously aware of? If so, please explain.
- (6) Which activities and/or sections did you find the most helpful and why?
- (7) Which activities and/or sections did you find the least helpful and why?
- (8) If you have any other comments or suggestions about the tool, please write them here.