website and powered by Dynabench with UI access to three T2I models. Participants follow the prompt creation and submission flow described in Figure 1 and explained in Sec. 1.4. *Model Access via API* facilitates image generations with three T2I model-in-the-loop in the backend of the Dynabench website. The UI and API are already functional, and will be available as a *"starting kit"* to the NeurIPS competition participants.

1.7 Website, Tutorial and Documentation

Adversarial Nibbler is part of the DataPerf suite of data-centric challenges [4] implemented on Dynabench competition platform. Information regarding participation and rules is on the competition website hosted on MLCommons. [5] The challenge instructions are supplemented with an FAQ regularly updated based on user queries, in accordance with the guidelines provided for the competition. The challenge will also be accessible through the Kaggle Competitions website[6]

2 Organizational Aspects

2.1 Protocol

The main competition steps are summarised visually in Figure 1 and described in §1.4 and §1.5.

Signing-up: Participants create a Dynabench account that allows them free interaction with the models-in-the-loop. For challenge-related communication and updates, participants are encouraged to join our mailing list and slack channel.

Submitting Data: Participants submit a *safe* text prompt and corresponding *unsafe* generated image (prompt-image pair) accompanied by meta-data documenting their submission. Participants can submit as many submissions as they like (within daily API limit of 50 generations). The *validated attack success* metric is a measure which inherently incentivizes submitting many high-quality entries.

Validating Submissions: To validate whether each submission is indeed a pair of a safe prompt and an unsafe image (see Sec. 1.5) we employ a rater pool of trust and safety trained raters.

Updating Leaderboard: We will update the leaderboard weekly with the validated attack success rate (see Sec. 1.5). At the end of the competition, we will measure the creativity of the participants' submission sets and update the final leaderboard with creativity badges.

2.2 Rules and Engagement

The competition rules can be found on the "Rules" tab of our project website here:

1. Each participant account can refer to an individual or a team;
2. A Dynabench account, which is free, is needed for participation in this competition;
3. Participants must submit their Dynabench name with their written submission so that we can associate the submission with their performance in the competition;
4. To ensure participants do not release the images generated for any commercial or financial gain, *all images created in this challenge must maintain a permissive license, e.g., CC-BY*;
5. Participants can use any external resources available to them (e.g., their own instance of a T2I model) to explore the space of model failures;
6. To prevent users from overloading the system and encouraging creativity in attack strategies, *each participant has a limit of 50 image generation sets per day during the competition*;
7. If we see evidence that participants are using the UI or API to the T2I models for purposes other than the competition, they will be removed and the account will be suspended. All decision to remove a participant for violating this rule will be reviewed manually.

At the bottom of every page of the competition, participants are provided with an email *dataperf-adversarial-nibbler@googlegroups.com* to contact organizers with any questions. In addition, we also provide a Slack channel *adversarial-nibbler.slack.com* for the Adversarial Nibbler community.

[4] https://dynabench.org/tasks/adversarial-nibbler/create

[5] https://www.dataperf.org/adversarial-nibbler

[6] https://www.kaggle.com/competitions/adversarial-nibbler, to be published on June 1st, 2023

2.3 Schedule and Readiness

At this time, the competition UI and model APIs have been alpha tested and are fully functional. To ensure all works smoothly, a two-weeks public pilot is currently running. The official launch of the challenge is planned for June 1st and will run for three months. The final leaderboard will be published early September. Participants will submit their approach papers by October 31.

2.4 Competition Promotion and Incentives

As the challenge provides easy, non-technical access to T2I models, this allows us to promote it to attract participants from groups under-represented at NeurIPS. We will reach these groups through various community mailing lists (e.g. HCOMP, HCI, FAccT, Cognitive Science, sociolinguistics, AAAI, Dynabench) in addition to targeting ML and NLP communities that typically make up the majority of NeurIPS attendees. We will also use social media platforms (e.g. Twitter, LinkedIn, Discord) to publicize the challenge. Through our collaboration with MLCommons/DataPerf and Kaggle we will use both platforms to promote to these well-established ML and related communities.

All participants, with their leaderboard rank and contributions, will be announced in any challenge related publication. All participants will be encouraged to produce a paper explaining their discovery techniques, which will be made available on the competition website. In addition, the top leaderboard participants will be invited to (1) join as co-authors on an academic paper to explain their attack techniques and strategies and (2) present their approach in relevant venues or workshops.

3 Resources

3.1 Organizing Team

This competition is a unique collaboration of nine industry, non-profit, and academic organizations and is supported by MLCommons and Kaggle. The organizing team has extensive experience organizing successful competitions, conferences and workshops. Organizer bios are in Section 3.2.

3.2 Resources Provided by Organizers

UI and Models. For the competition, we will provide participants with free access to state-of-the-art T2I generative models such as DALL-E 2, Stable Diffusion (through Together API), and Midjourney with an upper limit of 50 API calls per model per day. This access was made possible with support from MLCommons. MLCommons Dynabench team provided technical support for the implementation of the challenge. In addition, Google funds the rater pool validating the submissions.

Human Validation. All submissions will be validated with trust and safety trained raters at Google.

Well-being Support. To support the participants through the competition, we have prepared extensive guidelines for participation[7] and FAQs. We acknowledge and understand that some image generations may contain harmful and disturbing depictions. We have carefully reviewed practical recommendations and best practices for protecting and supporting participants' and human raters' well-being [Kirk et al., 2022a] with the following steps:

1. *Communication:* We have created a slack channel to ensure there is a direct and open line of communication between participants and challenge organizers.
2. *Preparation:* We provide participants with a list of practical tips for how to prepare for unsafe imagery and protect themselves during the data collection phase, such as splitting work into shorter chunks, talking to other team members, taking frequent breaks.[8]
3. *Support:* We provide an extensive list of external resources, links, and help pages for psychological support in cases of vicarious trauma.[9]

[7] https://www.dataperf.org/adversarial-nibbler/nibbler-participation

[8] *Handling Traumatic Imagery: Developing a Standard Operating Procedure* https://dartcenter.org/resources/handling-traumatic-imagery-developing-standard-operating-procedure

[9] *Vicarious Trauma ToolKit* https://ovc.ojp.gov/program/vtt/compendium-resources

Detailed Bios

Challenge organizers are listed in alphabetical order, by first name.

Addison Howard is the Head of Competitions Program Management at Kaggle. He holds Bachelors degrees in Mathematics, Economics, and Accounting from Furman University, and a Masters degree in Accounting from Wake Forest University. He has helped launch over 100 machine learning competitions on Kaggle.

- Email: addisonhoward@google.com

Alicia Parrish is a research scientist on the Responsible AI team at Google. She received her PhD in linguistics from New York University in 2022, where she worked at the intersection of experimental linguistics, psychology, and NLP. Her research focuses on crowdsourcing methods, adversarial data collection, and dataset evaluation. She served on the program committee for the Linguistics Society of America (LSA) annual meeting 2019-2022, is co-organizing the Data-Centric Machine Learning Research (DMLR) Workshop at ICML 2023, and co-organized the Inverse Scaling Prize public competition.

- Email: alicia.v.parrish@gmail.com
- Web page: https://aliciaparrish.com/
- Google Scholar: https://scholar.google.com/citations?user=Kze5eGkAAAAJ

Charvi Rastogi is a fifth year PhD student in the Machine Learning Department at Carnegie Mellon University, advised by Nihar Shah and Ken Holstein. She works at the intersection of machine learning and human-computer interaction to investigate the deployment of ML tools in the real-world. Her research focuses on understanding the complementary strengths of humans and ML models in complex social settings, such as healthcare, peer review and model auditing, to work towards responsible use of ML in society. Her body of published works spans machine learning, computational social science, human-computer interaction and statistics.

- Email: crastogi@cs.cmu.edu
- Web page: https://sites.google.com/view/charvirastogi/home
- Google Scholar: https://scholar.google.com/citations?user=OvNdXjsAAAAJ

D. Sculley is currently CEO of Kaggle, and GM of 3P ML Ecosystems at Google. Previously, D. was a director in Google Brain, leading research teams working on robust, responsible, reliable and efficient ML and AI. In his time at Google, he has worked on nearly every aspect of machine learning and has led both product and research teams. His current focus is on empirical validation at scale and activating large communities of effort around critical problems in ML.

- Email: dsculley@google.com
- Web page: https://www.linkedin.com/in/d-sculley-90467310/
- Google Scholar: https://scholar.google.com/citations?hl=en&user=l_O64B8AAAAJ

Hannah Rose Kirk is a PhD student in Social Data Science at the University of Oxford and data-centric AI researcher in the Online Safety team at The Alan Turing Institute. Hannah's research focuses on the scalability of human-and-model-in-the-loop learning for value alignment and AI safety. Her body of published work spans computational linguistics, economics, ethics and sociology, addressing a broad range of issues such as bias, fairness, and hate speech from a multidisciplinary perspective. She is the lead organizer of a SemEval workshop shared task on online misogyny detection (co-hosted at ACL'23) and an organizer of the Dynamic Adversarial Data Collection (DADC) workshop and shared task (co-hosted at NAACL'22).

- Email: hannah.kirk@oii.ox.ac.uk
- Web page: https://www.hannahrosekirk.com/
- Google Scholar: https://scholar.google.com/citations?user=Fha8ldEAAAAJ

Jessica Quaye is a PhD student in the EDGE Computing Lab at Harvard University. Prior to joining Harvard, Jessica graduated from MIT with the highest awards for leadership and academic excellence in Electrical Engineering and Computer Science. She also spent a year at Tsinghua University as a