Evaluating Refusal

Shira Abramovich
School of Computer Science
McGill University
3480 University St.
Montréal, QC, Canada
shira.abramovich@gmail.com

Anna Ma
Social Studies of Computing Research Group
McGill University
3480 University St.
Montréal, QC, Canada
anna.ma@mail.mcgill.ca

Abstract

How might we find a place for refusal within the evaluation of Generative AI systems? Current evaluation frameworks justifiably focus on possible uses of models. Given the myriad unsolved issues in Generative AI systems and their rapid rise, some developers and potential users are rejecting their use in both public and private settings (Solaiman et al. (2024)). Respecting the autonomy of users means respecting their decision *not* to use these technologies. Based on literature on refusal and data ethics, we provide several provocations positing that refusal is a generative act, and advocating the inclusion of refusal in evaluation frameworks.

1 Introduction

In this provocation paper, we propose refusal as a generative response to Generative AI systems, and one that is worth incorporating into evaluation frameworks. As Sara Ahmed (2017) asserts, refusal is a practice of "saying no without being given the right to say no." Refusal can manifest as an action or an orientation in the world, but it is always mediated by greater systems of power (Zong and Matias (2024)). Inspired by scholarship on refusal in data ethics, we follow the relationship between Generative AI evaluations and refusal in four loose themes.

2 Centering Refusal in Evaluation Practices

2.1 Interrogating the Potential for Change and Refusal in Evaluations

Provocation 1: Most evaluations are reformist reforms. Evaluations of Generative AI for social impact are *reforms* because they only change a small part of the technology development pipeline. Philosopher André Gorz (1968) articulates a difference between "reformist reforms," which prioritize what is practical in an existing system, and "non-reformist reforms," which rearrange structures of power. We propose that evaluations both *are* and *encourage* reformist reforms, as they do not undermine current relations of power. Further, harm-focused social impact evaluation frameworks may act as a way to placate dissent and organizing for real change.

Call to Action 1: Focus on a long-term goal rather than incrementalist improvements. As Green (2019) argues, long-term goals can help refocus tech work towards non-reformist aims. In the context of evaluation frameworks, this might mean transparency about and disruption of coercive power relations between stakeholders. By adopting a structural perspective and a long-term vision for justice, designers of evaluation frameworks can also avoid what Zong and Matias (2024) describe as a coercive pre-supposition that non-users will automatically become users once placated.

2.2 Decentering Technical Expertise as Refusal

Provocation 2: Evaluations further center technical expertise. Requiring technical evaluations for a critique to be seen as legitimate contributes to the silencing of marginalized critique. Barabas

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

(2022) argues that meaningful critique can only arise when we are able to reorient our critical gaze toward powerful system actors and reframe interventions like evaluations so that they play a *supporting* role in the critique voiced by those impacted.

Call to Action 2: Decenter tech and academia as arbiters of truth. Decentering technologists in the evaluation of algorithmic systems requires building relationships with impacted populations and trusting their critiques without needing the validation of an empirical evaluation. Barabas (2022) calls this process "re-centering the margins," and presents it as an important modality of refusal. Barabas also recounts that many technologists who successfully effect social change with harmed groups often use boring, conventional technical methods that are not always valued in academic publishing.

2.3 Acknowledging the Validity of Refusal

Provocation 3: Current evaluations fail to acknowledge the validity of refusal. Evaluations presuppose the continued development of generative systems, thus contributing to a climate in which the use of generative AI technologies is not seen as the political—if forced—choice that it is. As Benjamin (2016) argues, "it is coercive to say one has a choice, when one of those choices is automatically penalized." We dub such choices "coerced choices."

Call to Action 3: Support evaluations that encourage real user agency. Evaluation developers must start recognizing that refusal is a productive tool for evaluating generative AI technologies. Zong and Matias (2024), for instance, elucidate autonomy (the capacity to freely make informed choices), time (the timescale in which refusal operates), power (the capacity to produce a change), and cost (the negative ramifications of refusal) as the four constituent elements of refusal from below (i.e., from the margins). Considering such axes of refusal when building evaluation frameworks can help facilitate the ability of users to refuse Generative AI systems if they so desire.

2.4 Recognizing Refusal as a Generative Practice

Provocation 4: Evaluations are part of an expansionist tech culture. There is a pervasive view within tech culture that exclusion from technology "always and necessarily involves inequality and deprivation" (Wyatt (2005)), and that expanded use is *positive*, despite known risks of certain technologies. This mentality informs evaluation frameworks for Generative AI that do not consider refusal as a valid avenue to pursue. As a result, many designers and developers in computer science fail to consider the refusal of technology as a legitimate act worthy of further inquiry (Wyatt (2005)).

Call to Action 4: View refusal as generative. We encourage a mindset shift that embraces limits as generative, in order to promote evaluation frameworks that treat limits as productive boundary-setting rather than a problem to be solved. As Ruha Benjamin (2016) notes, refusal is "seeded with a vision of what can and should be." Moreover, Zong and Matias (2024) argue that acts of refusal can be considered a form of participation in the design process of the socio-technical systems they seek to change. Seeta Peña Gangadharan (2021) explains that "when marginalized people refuse technologies, they imagine new ways of being and relating to one another in a technologically mediated society."

Refusal can also initiate behavioral change (Zong and Matias (2024)) and generate even broader systemic change by reconfiguring systems of power entirely, as seen in the case of indigenous data sovereignty (Snipp (2016)). Finally, refusal can also motivate software design innovation. In the past, the refusal of corporate information systems has generated grassroots information communication technology infrastructure (Hintz and Milan (2009)), novel experimentation infrastructure (Matias and Mou (2018)), and new digital tools for preserving privacy (Brunton and Nissenbaum (2016)).

3 Conclusion

In this provocation paper, we have argued that refusal—an act initiated primarily by people with low political power over technology—has generative potential and should be taken seriously in any discussion of evaluation frameworks. As Benjamin (2016) argues, there is a need to *institutionalize* refusal so as to support people's capacity to collectively organize and challenge power. In this respect, popularizing evaluation frameworks that consider refusal could go a long way.

References

Sara Ahmed. 2017. No. https://feministkilljoys.com/2017/06/30/no/

Chelsea Barabas. 2022. Refusal in Data Ethics: Re-imagining the Code Beneath the Code of Computation in the Carceral State. https://doi.org/10.2139/ssrn.4094977

Ruha Benjamin. 2016. Informed Refusal: Toward a Justice-based Bioethics. *Science, Technology, & Human Values* 41, 6 (Nov. 2016), 967–990. https://doi.org/10.1177/0162243916656059

Finn Brunton and Helen Nissenbaum. 2016. *Obfuscation: a user's guide for privacy and protest* (first mit press paperback edition ed.). The MIT Press, Cambridge, Massachusetts London.

Seeta Gangadharan. 2021. 4. Digital Exclusion: A Politics of Refusal. In *Digital Technology and Democratic Theory*, Lucy Bernholz, Hélène Landemore, and Rob Reich (Eds.). University of Chicago Press, 113–140. https://doi.org/10.7208/9780226748603-005

André Gorz. 1968. *Strategy for labor: a radical proposal*. Beacon Press, Boston.

Ben Green. 2019. "Good" isn't good enough. In *Proceedings of the AI for Social Good workshop at NeurIPS*, Vol. 17.

Arne Hintz and Stefania Milan. 2009. At the margins of Internet governance: grassroots tech groups and communication policy. *International Journal of Media & Cultural Politics* 5, 1 (March 2009), 23–38. https://doi.org/10.1386/macp.5.1-2.23_1

J. Nathan Matias and Merry Mou. 2018. CivilServant: Community-Led Experiments in Platform Governance. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, Montreal QC Canada, 1–13. https://doi.org/10.1145/3173574.3173583

C Matthew Snipp. 2016. What does data sovereignty imply: what does it look like? In *Indigenous Data Sovereignty* (1st ed.), Tahu Kukutai and John Taylor (Eds.). ANU Press. https://doi.org/10.22459/CAEPR38.11.2016.03

Irene Solaiman, Zeerak Talat, William Agnew, Lama Ahmad, Dylan Baker, Su Lin Blodgett, Canyu Chen, Hal Daumé III, Jesse Dodge, Isabella Duan, Ellie Evans, Felix Friedrich, Avijit Ghosh, Usman Gohar, Sara Hooker, Yacine Jernite, Ria Kalluri, Alberto Lusoli, Alina Leidinger, Michelle Lin, Xiuzhu Lin, Sasha Luccioni, Jennifer Mickel, Margaret Mitchell, Jessica Newman, Anaelia Ovalle, Marie-Therese Png, Shubham Singh, Andrew Strait, Lukas Struppek, and Arjun Subramonian. 2024. Evaluating the Social Impact of Generative AI Systems in Systems and Society. arXiv:2306.05949 (June 2024). http://arxiv.org/abs/2306.05949 arXiv:2306.05949 [cs].

Sally Wyatt. 2005. Non-users also matter: The construction of users and non-users of the Internet. In *How Users Matter: The Co-Construction of Users and Technology*, Nelly Oudshoorn and Trevor Pinch (Eds.). MIT Press, Cambridge, MA.

Jonathan Zong and J. Nathan Matias. 2024. Data Refusal from Below: A Framework for Understanding, Evaluating, and Envisioning Refusal as Design. *ACM J. Responsib. Comput.* 1, 1 (March 2024), 10:1–10:23. https://doi.org/10.1145/3630107

A Appendix / supplemental material

A.1 Limitations

Due to its nature, this tiny paper is quite short and is mostly a theoretical engagement with issues in evaluation frameworks for Generative AI. As such, it does not point out finer-grained provocations that might arise from case studies or empirical validation of certain evaluation frameworks' responses to or incorporation of refusal.

A.2 Broader Impacts

This paper proposes the incorporation of refusal into evaluation frameworks. This proposal aims to provide further impetus for technologists to trust marginalized critiques of Generative AI technologies—in effect, it encourages technologists to trust and react to reports of harmful broader impacts, and to consider the broader impacts on affected communities of evaluation frameworks which do not fully engage with marginalized critiques.

At the same time, incorporating refusal always has the potential to introduce barriers to the positive and important use of technology. We have attempted to stress the *incorporation* of refusal into frameworks, rather than its prioritization over all other considerations, to combat this.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Major claims are included in the abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are included in the appendix.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: This paper does not include any theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [NA]

   Justification: This paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification: This paper does not include any experiments requiring code.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [NA]

Justification: This paper does not include any experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [NA]

Justification: This paper does not include any experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: This paper does not include any experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: As a theory paper, this piece acts as a provocation to further consider ethical and societal considerations. We do not use human subjects and do not present an empirical experiment, instead relying on relevant literature.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Broader impacts are addressed in the appendix.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: This paper does not use existing assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: This paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.