Polarization Traps: A Formal Model of the Epistemic Structure of Moral Polarization

Anonymous Author(s)
Affiliation
Address
email

Abstract

As individuals, we form our beliefs and opinions through exchange with other individuals and through outside influences. This often leads to groups of individuals holding relatively homogeneous positions that are diametrically opposed to other groups' opinions, ending in polarized states. Internet platforms add an additional dimension to this question, as the ways in which the platform curates and disperses information can intensify or mitigate polarization. In this work, we flesh out a proposal to develop theoretical models and algorithmic solutions for this phenomenon by linking philosophical and political literature to literature in statistics, opinion dynamics, and social learning. In this report, we narrow the scope of the problem we wish to study and define our goals for future work.

1 Introduction

A fundamental question in social epistemology is how individuals' opinions are influenced by others with whom they interact and how they contend with information that differs from, and even contradicts, their existing world view. In practice, individuals often form into groups based on shared views (homophily) and interact mostly with people within their group. As a result, information flow is limited to people who are already filtering it through similar lenses, and it is much more difficult for people to drastically change their world view, even in the face of compelling evidence.

We can see this process occur in many settings, ranging from harmless to capable of destabilizing fundamental institutions. For instance:

1. We grew up avoiding eating a certain vegetable. No one in our extended family eats it, and we have never gotten a clear answer about why. Whenever we go to a restaurant and see it on the menu, we do not order it. One day in adulthood, we are eating dinner at a friend's house, and we express how much we like this particular dish. The friend lists the ingredients and the particular vegetable happens to be in it. We realize that we quite like that vegetable and are now open to trying foods that happen to have it, even if we are not yet ready to purchase and prepare it ourselves.
2. We are watching a TV show and find ourselves loyal to a particular character. We join an online fanpage for that character. We engage in discussions about why that character is the best one in the show. Then, our friend watches the show and decides they dislike the character. We share with them all of our rationale for liking the character, and we feel particularly validated in liking that character because there is a huge community of people online who like that character. However, our friend fails to be convinced. We then feel we cannot talk about the show with that friend. As a result of disengaging with people who

Submitted to the NeurIPS 2025 Workshop on Algorithmic Collective Action. Do not distribute.

have different opinions than us about the character, we will likely never change our opinion about the character.

3. We borrow this final example from [9]. "A toy example: suppose I am a cult leader, and I have taught my followers to believe that every human except the members of our group has been infested and mind-controlled by alien ghosts from Mars. I also teach my followers that these alien ghosts from Mars hate our group for knowing the truth, and so will constantly seek to undermine our knowledge of their existence through mechanisms like calling us a 'cult' and calling us lunatics. ... Suppose that I tell my followers to expect outsiders to falsely claim that there are no ghosts from Mars. When my followers do confront such contrary claims from outsiders, those contrary claims are exactly what they expected to hear. Thus, new contrary testimony is neutralized, because it was predicted by past beliefs. ... There is also a secondary effect. When my followers hear exactly what I predicted, then my claims have been verified, and so my followers will have some reason to increase their trust in me."

Humans naturally, sometimes even accidentally, filter opposing views from their epistemic environment, and we can trace many potential causes. Part of the reason is because there are rewards for conformity. Another reason is that we evaluate opinions as valid when more people hold them, without accounting for the fact that the source of the idea might be the same for all of them. In the examples above, individuals have different capacities for changing their minds, and the community with which they share their beliefs has different impacts on how they react to information that goes against their belief.

More recently, we have begun to witness polarization occur in communities on the internet, where people adopt separate and extreme beliefs. This is often a function of the epistemic environment generated by algorithmic content recommendation, i.e., optimizing for clicks necessarily shows people more of the content they are already clicking on. This leads to the development of filter bubbles. Another driving phenomenon is the "outrage machine," dubbed as such by author Tobias Rose-Stockwell, who argues that it is expeditious for content creators to appeal to our angry and tribalistic instincts [10]. The specific platform can influence the degree to which polarization occurs, both through the kinds of people using the platform and through systemic design choices that influence the way users engage with content.

Note that there are at least three different settings in which we might be interested in studying the kind of polarization we have described: (1) people naturally form into groups based on their interactions and then do not trust those outside; (2) collaborative filtering / recommender systems cause people's information to be limited; (3) online platforms with limited algorithmic filtering (e.g., on a specific reddit thread) still witness individuals adopting extreme opinions.

Our main goal is to theoretically model the development of these polarized information bubbles and design interventions that bring us out of these fractured information environments. Importantly, we are not just interested in exposing people to new information (which can easily be achieved through randomly incorporating different views); we are also interested in encouraging users to critically engage with new perspectives. Our work will focus on the third setting described above, as it has similar dynamics to the first with interventional strategies that can be designed specific to the internet setting. Further, we can consider mitigation strategies stemming either from the platform side or the user side, including potential collective action strategies.

In the rest of this report, we first summarize a paper with clarifying qualitative development. Next, we use this qualitative description to specify the setting we wish to study in subsequent work. Then, we survey related technical results and discuss the nature of and goals for interventions. This report summarizes our literature review and qualitative analysis, culminating in a concrete project proposal. Thus, we conclude with a detailed discussion of next steps.

1.1 Qualitative Scaffolding

In his paper "Echo Chambers and Epistemic Bubbles", [9] draws a distinction between "epistemic bubbles" and "echo chambers." He points out that in an epistemic bubble, certain perspectives are omitted due to self-selection or even by accident. In these cases, we can introduce new perspectives by just exposing individuals in the bubbles to new ideas. On the other hand, echo chambers occur when trust of people outside the chamber is undermined. He writes:

> " An epistemic bubble is an epistemic structure emerging from the informational architecture of communities, social networks, media, and other sources of information and argument. It is an impaired informational topology — a structure with poor connectivity. An echo chamber, on the other hand, is an epistemic structure created through the manipulation of trust; it can exist within a healthy informational topology by adding a superstructure of discredit and authority. I hope to show, contra the recent focus on epistemic bubbles, that echo chambers pose a significant and distinctive threat – perhaps even a more dangerous one – that requires a very different mode of repair." ([9])

As we will see later, plenty of work in theoretical computer science and statistics has focused on modeling the polarization phenomenon. The fraction that considers interventions has focused on epistemic bubbles. Our goal is to study echo chambers. In more detail, Nguyen writes:

> "I use "echo chamber" to mean *an epistemic community which creates a significant disparity in trust between members and non-members.* This disparity is created by *excluding non-members through epistemic discrediting*, while simultaneously *amplifying insider members' epistemic credential*. Finally, echo chambers are such that *in which general agreement with some core set of beliefs is a pre-requisite for membership*, where those core beliefs *include beliefs that support that disparity in trust*." ([9], emphasis Nguyen's.)

He also goes on to note that, "once the discrediting beliefs are in place, the ensuing beliefs and action of the echo chambers' members are surprisingly close to rational" [9]. This is reminiscent of the epistemic structure of pessimism traps [7, 2].

2 Setting

Through analysis of our motivating examples and several works of qualitative scholarship, we have identified several desiderata for our theoretical model. In our model, individuals will interact with other individuals in rounds on some platform. The platform might affect the dissemination of their expression: it could either be a neutral platform that minimally intervenes just facilitates conversation (e.g., in-person), or it could be an adversarial (e.g., engagement-farming, outrage-machining) platform. Thus, the platform could potentially manipulate the likelihood of transmission of certain messages. Individuals update their beliefs through interactions and through outside influences (both have been studied in polarization literature, e.g. [3, 6]). Now, we describe the phenomena our model should be able to replicate in order to be a faithful one in which to study and implement interventions:

1. Once polarization has occurred, in the absence of external interventions (i.e., with just individuals interacting with each other, even off platform), there should be no evidence that allows for updating one's beliefs away from the polarized state. In the three examples above, we see the following lack of belief updates: (1) an individual is unlikely to experience an interaction with the item (the vegetable) that could change their mind, (2) an individual lacks interactions with people with different beliefs (about the favorite character) due to avoiding the topic, (3) an individual has actively cultivated a distrust of people with different beliefs (about alien ghosts from Mars).
2. If agents have neutral trust between them, then adding a connection (graph edge) between one agent and an agent with different belief should help shift opinions.
3. If active distrust has been cultivated, then adding an edge between one agent and an agent with different belief should either have no effect, cause breaking of the connection (graph edge), or even intensify opinion.

Thus, the first important phenomenon that we should see in a faithful model is the lack of evidence available to change one's beliefs once polarization has occurred. Interestingly, a similar situation surrounding the evidential basis for a decision has been described by [7] in her description of pessimism traps. In particular, she argues that a pessimism trap develops when an individual takes a decision and therefore loses access to the evidence that would have been provided by the opposite action. Recent theoretical work [2] has indeed faithfully modeled pessimism traps using an opinion dynamics model, suggesting this is a valuable tool for this "not-even-bandit" feedback setting.

Finally, recall that Nguyen formalizes the distinction between echo chambers and epistemic bubbles, which is important to model. Thus, stated explicitly, we wish to develop a model where under neutral trust parameters, we are in the epistemic bubble setting, whereas in the presence of distrust of outsiders and increased trust of insiders, we end up in echo chambers.

2.1 Mathematical Tools and Related Theoretical Models

We will extend opinion dynamics / social learning models to study this. The general area of opinion dynamics and social learning provide mathematically-tractable models in which to reason about (1) individuals' connections to each others; (2) update rules for an individual's beliefs as a function of those to whom they are connected and any exogenous influences. In turn, they study questions like (1) convergence: does an agent's belief reach steady-state?; (2) agreement: does an agent's belief match another agent's belief?; (3) learning: does an agent's belief match the ground truth? Some works that survey this are [8, 1]. The results in these works are theoretical, showing statistical proof that if the respective conditions are met, outcomes are guaranteed.

Within this framework, many works have studied the development of polarization. These works generally differ in the update rules: in some, individuals update their beliefs by averaging their neighbors' opinions (DeGroot model); in others, they choose a (uniformly) random neighbor whose opinion to adopt (voter model). [3] have introduced an extension of these models that capture the fact additional behavioral patterns, including stubbornness (never changing opinions) or *biased* adoption. Some works have also considered situations where opinions are only affected by the actions of "influencers" [6, 4].

These models capture interesting things about the potential ways in which agents update their beliefs and the respective works analyze conditions under which consensus and polarization occur. We are interested in at least combining, if not unifying, these models: we wish to study interactions between individuals and also those with extra influence. We also want to vary the nature of the platform.

Finally, some works consider interventions. Mostly, these interventions are in the vein of adding edges to the graph [5, 11]. As discussed above, such an intervention would help break out of *epistemic bubbles* but not *echo chambers*.

2.2 Interventions and Influencing Algorithmic Outcomes

What is an Intervention? In this work, we take an intervention to be anything that can be "implemented" (either by the platform or the agents) that causes disruption of epistemic bubble or echo chamber. On the platform side, the platform may either change the way they present information *without imposing their views* (e.g., via randomness) or adjust the update calculus for an agent (e.g., add noise to reward or discount the past). Interventions on the platform end are ones that prevent negative outcomes that result from a small group of individuals flooding the platform with content that reflects a particular agenda. On the individual side, an intervention involves an agent changing their update structure, either by changing those who affect their opinions or by changing the reward for themselves. If the platform is more adversarial, e.g., the platform farms for engagement or outrage, then the intervention will be something that empowers good-faith agents to engage in discourse with people with varying opinions. Here, the collective action by human agents to have liberal discourse is *aided* by the intervention. Our intended model will allow us to study both kinds of platforms and therefore both kinds of interventions.

Intervention Desiderata We generally consider strategies that encourage people to share information / opinions without as much influence from ideas that have already been "normalized." We wish to avoid epistemic platforming of ideas that should not ever be normalized. We are interested in endogenous interventions which do not require buy in from any powerful agency whose interests might be at odds with liberal discourse. One particularly interesting question is to what degree we can *decentralize* these interventions, so that the platform where the intervention is implemented does not imposing a moral framework and instead allows for engagement in liberal discourse.

A Note on Societal Impact Our problem is extremely societally-relevant, and we acknowledge that any algorithmic interventions we propose in future work will come with limitations with which we will engage. Our goals of not platforming things that should never be platformed while maintaining

liberal discourse are not easy goals to achieve, but they will guide our work at every step. As we develop a formal model, it will be important to keep in mind that our results and guarantees hold *in that model*, when the assumptions of that model are met. Thus, they might not immediately extend to particular real-world situations of interest.

3 Discussion and Next Steps

Our next steps are to devise a tractable theoretical model based on the theoretical models discussed above. This will involve extending existing models so they can be "stitched together" in the interpolation regimes. From there, we can think about interventions in various settings, including on- and off-platform and human- and algorithm-initiated.

References

[1] Sushil Bikhchandani, David Hirshleifer, Omer Tamuz, and Ivo Welch. Information cascades and social learning. *Journal of Economic Literature*, 62(3):1040–1093, 2024.

[2] Avrim Blum, Emily Diana, Kavya Ravichandran, and Alexander Tolbert. Pessimism Traps and Algorithmic Interventions. In Mark Bun, editor, *6th Symposium on Foundations of Responsible Computing (FORC 2025)*, volume 329 of *Leibniz International Proceedings in Informatics (LIPIcs)*, Dagstuhl, Germany, 2025. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

[3] Abhimanyu Das, Sreenivas Gollapudi, and Kamesh Munagala. Modeling opinion dynamics in social networks. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 403–412. ACM, 2014.

[4] Jason Gaitonde, Jon Kleinberg, and Éva Tardos. Polarization in geometric opinion dynamics. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, EC '21, pages 499–519. Association for Computing Machinery, 2021.

[5] Kiran Garimella. *Polarization on Social Media*. Aalto University, 2018. ISSN: 1799-4942 (electronic).

[6] Jan Hązła, Yan Jin, Elchanan Mossel, and Govind Ramnarayan. A geometric model of opinion polarization, 2021.

[7] Jennifer Morton. Resisting pessimism traps: The limits of believing in oneself. *Philosophy and Phenomenological Research*, 104(3):728–746, 2022.

[8] Elchanan Mossel and Omer Tamuz. Opinion exchange dynamics. *Probability Surveys*, 14, 2017.

[9] C. Thi Nguyen. ECHO CHAMBERS AND EPISTEMIC BUBBLES. *Episteme*, 17(2):141–161, 2020.

[10] Tobias Rose-Stockwell. *Outrage Machine: How Tech Amplifies Discontent, Disrupts Democracy-And What We Can Do About It|Hardcover*. Grand Central Publishing, 2023.

[11] Miklos Z. Rácz and Daniel E. Rigobon. Towards consensus: Reducing polarization by perturbing social networks. *IEEE Transactions on Network Science and Engineering*, 10(6):3450–3464, 2023.

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: This report should be read as a project proposal. We have surveyed a broad range of works in philosophy, political theory, statistics, and computer science to narrow the scope of the problem we wish to address and tools with which to do it.

   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: [NA]

   Justification: As this report concretely identifies the problem we wish to study among a broad range of possible problems, we have not yet finalized our formal approach. Thus, it is not yet relevant to discuss limitations of the approach.

   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 report does not yet contain 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 report does not yet contain experimental results.

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: No experiments requiring code / data.

   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, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: No 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: No 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: No 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: -

   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: Any algorithms developed toward the problem we identify in this work have clear potential for positive impact; the details of potential negative impacts would depend on the specific future directions taken by the project.

    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: Not relevant.

    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: No such 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: No such 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: No such experiments.

    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: No such experiments.

    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.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: No such use.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.