ML Fairness Communication Best Practices

Scope.

This document describes best practices for communicating on topics related to the inclusion, fairness, and transparency of machine learning algorithms; guidance for creating, reviewing, and sharing documents related to ML Fairness; and when and how to maintain attorney-client privilege.

Introduction.

When we talk about fairness, we can find ourselves getting into some pretty challenging—and often sensitive—topics. These issues are rooted in our individual and collective histories, our cultures, and our unconscious biases, so by definition there is no 'silver bullet'. But when we work toward fairness, we can slow down and be more mindful of our assumptions, with the goal of making conscientious choices. We'll make mistakes along the way, though, even with the best of intentions. And we'll do our best to learn from them. So when unwanted bias manifests in a model to the detriment of a user's experience, let's agree to be constructive, and trust that our peers will respond from a place of understanding. This makes the environment for addressing fairness in our work a whole lot more comfortable and productive for us all. So before we even begin, let's agree to create a safe space together, and understand that this is hard work for all of us.

We can start with some important points to remember:

1. First, please avoid accusing a product or feature. A person worked on that—they put time and energy into it. So let's focus on the user's experience of the product or feature, with the assumption that the person who made it had the best of intentions. Take time to offer specific, constructive observations on how it could be improved, along with feedback that can help make real, positive change happen.
2. Second, fairness is subjective and uniquely experienced by every individual in each case; there is no "perfect". So, let's use the language of "more"—more inclusive; more fair—to frame our work as enabling us to create products that are more useful and more accessible for more people.
3. Lastly, keep in mind that words like "unfair", "prejudicial", or "discriminatory" can also mean or imply specific concepts under the law. We should avoid using these terms when describing our products and services. When writing or presenting about ML Fairness (e.g., tech talk, EngEDU class, workshop facilitation, blog post, etc.), please do your best to respect the implications of using those terms.

Communicating on ML Fairness Topics

Focus on inclusion while remaining objective and neutral.

- Frame communications and publications in terms of machine learning practices that are designed to encourage and increase inclusion. We are always striving to improve user experiences and make the world's information accessible and useful to *everyone*. So it's best to frame our efforts in the context of inclusiveness. Also, our research publications can set a constructive tone with the academic and research community and guide the external discussion with wording that's focused on processes more than outputs, as well as specific evaluation metrics and measurements rather than general statements of unfairness or bias.

- Avoid making guarantees or promises to our users about how inclusive our products and features can or will be. We should not claim in our external communications to be able to "ensure" or

Privileged and Confidential

"guarantee" full inclusion or complete fairness in our products or services. We can always improve our products and services and the user's experience.

- Where possible, consider using generic subgroups when referring to or dealing with sensitive categories like race or gender, especially if the goal is just to identify differences. That said, in some instances, it may be necessary to refer to specific subgroups, and in those cases, please work with pcounsel and Policy on identifying the appropriate groups and usage.
- Remain factual, neutral, objective, and specific in statements about our products. Again, we should be respectful of the consideration and effort that others have put into their work, so offer specific observations and objective feedback that is actionable. In turn, try to avoid feedback that characterizes our products as unfair, unjust, prejudicial, giving preferential treatment of, or discriminatory as those terms can also have a legal meaning or be taken out of context by external parties (e.g., regulators, press, litigants, etc.) as admissions of fault or wrongdoing.
  Pointers:
  - State facts (with working links to sources) rather than opinions.
  - Use objective measures for comparisons and reference observed product behavior against clear criteria and not subjective statements.
  - Make sure statements are accurate and specific.
  - Don't speculate on harm to users. If you think it's necessary to identify the potential effects on individuals as justification for remediation, please include your pcounsel on the communication following the guidance below (including asking for Legal review).

  Examples:
    - *E.g.*, The following statement "offensive content associated with race A is high" could be more specific, like " in Product L, offensive content associated with group A occurs X times more often compared to group B, due to the presence of G and H on publisher websites." The second statement limits the scope to a particular product (not all Google products), gives scale (X times more), identifies the potential data influencing the algorithm, and tries to use a generic subgroup if possible.
    - *E.g.*, The following statement: "targeting those ads to category A is unfair and promotes wrong social norms hurting those users and lowering their self-esteem" is vague, overly broad, assumes a norm, speculates on harm, and could be more specific like: "when advertisers target abc type of ads in Product L based on category X, we observe that less advantageous ads (defined below) are displayed N times more often than advantageous ads, leading to a less desirable user experience because a less advantageous ad doesn't offer xyz." Notice, rather than speculating on how an observed behavior may harm a user personally, the second statement focuses on how we can offer a better user experience in the context of our products. In general, we should try to focus on acceptable product behavior and a good user experience in the context of our product policies when identifying instances that may need remediation. Some educational materials may want to highlight the potential effect certain observed algorithmic behavior may have on individuals to illustrate why remediation is necessary, but those materials should be reviewed by pcounsel and Policy. The second statement also limits the scope to a particular product and gives scale (N times). It would be better to also frame why the ads were less advantageous in the context of our product policies to be more objective.
    - *E.g.*, The following statement "(Image) search results are unfairly amplifying bias in society and negatively affecting attitudes towards M" is vague, overly broad,

assumes an abstract harm, and could be more specific like "(Image) search results for [text] showed race A in X cases, while the results for T showed race B in Y cases, leading to a less favorable user experience for race B users because of the greater association with [text]. This may be caused by xyz content appearing on abc sites." The second statement makes the observation in a factual statement in the context of a good/bad user experience and identifying the potential external cause.

- Be respectful and thoughtful in your choice of words when describing data or observed product behavior, especially when it involves sensitive categories of data like race, gender, religion, and ethnicity.

Creating, Reviewing, and Sharing Documents Related to ML Fairness

Please ask pcounsel for legal review of your documents when appropriate.

- Try to give Legal reasonable time to review your document. Examples of documents that should be reviewed by pcounsel include drafts of: definitions, policy, terminology, case studies, examples of incidents, decision steps for remediation, evaluation checklists, incident reporting tools, and comms.
- Mark drafts as "Privileged and Confidential," both in the title and header of the doc, which can help our litigation team identify and filter the document so we don't accidentally turn it over to the other side in a litigation matter or regulatory investigation.
- Grant pcounsel Editing or Suggestion rights to the doc and share the doc directly with each individual (not with aliases or a link accessible to all of Google).
- Share the doc with pcounsel with a statement that you are seeking Legal review of the doc and follow the tips below for maintaining privilege on the doc. The doc won't be privileged until you seek advice from Legal, so considering sharing sensitive drafts early.
- If you think a draft is final, please confirm the draft with pcounsel and PR and discuss your intended audience and channel(s) for wider distribution. Then share a copy of the doc that doesn't include the legal comments or advice.

Maintaining Attorney-Client Privilege

The privilege only protects communications to and from lawyers for the purposes of seeking legal advice.

- Legal privilege keeps communications between an attorney and client secret, and when we are confronted with a legal demand for such communications (e.g., when we receive a litigation discovery request or a demand that the lawyer testify under oath) we can assert the privilege as a defense to providing that communication.
- Always include pcounsel on the to: line when seeking legal advice or when passing along legal advice you received. Just adding pcounsel to an email or document doesn't guarantee that it will be protected by the privilege, and it doesn't make the privilege apply retroactively to earlier emails or drafts. Also, be careful of threads that start diverging from the original legal question. If you start adding new people to the thread (e.g., to address a separate topic) or forward it to others, it could break privilege.
- Clearly indicate you are seeking legal advice in your communications, both by marking the email as "Privileged and Confidential", "Privileged", or "Attorney-Client Privilege." and by indicating in the email "Adding Legal for their advice", "Seeking your legal review" or something similar when you include pcounsel. There are no magic words that will always make a document privileged, but it's helpful to identify documents that might contain privileged material.

ML Fairness Communication Best Practices

- Limit the distribution of privileged emails/documents to a need-to-know audience internally. Do not copy aliases unless absolutely necessary to convey the information. Never send privileged emails outside of Google (except to Google's outside lawyers), which will break the privilege.
- Ask pcounsel to share Legal advice, and don't summarize or comment on the legal advice you receive with others.
- Always communicate with care. The privilege is not absolute, isn't recognized in all jurisdictions, and only applies to legal advice (not business advice). So when communicating about particularly sensitive topics like these, imagine someone outside of Google eventually reading your email, presentation, or other communication (e.g., press, lawyers, etc.) and how your language might come across to them.
    - Not all countries recognize attorney-client privilege with respect to in-house attorneys. For example, the EU generally does not recognize attorney-client privilege to include in-house counsel and China and Japan generally don't recognize the privilege at all. This means that documents that are privileged in the U.S. may have to be produced to foreign government agencies in their investigations. (Note, communications with outside counsel in the EU giving or getting legal advice are privileged in the EU, and should be clearly marked as "Outside Counsel Advice.")
- Don't offer legal opinions in communications or speculate on the outcome of ongoing investigations or litigations, whether they involve Google or other companies. If you have questions about legal matters, please ask your pcounsel.
- When in doubt, consider an in-person/VC meeting with pcounsel rather than email, especially to discuss sensitive topics like:
    - instances of potential harm to our users or discrimination in our products; and
    - how to choose labels for sensitive categories like race, gender, ethnicity, and religion.

  Also, keeping your pcounsel up to date on activities, like sprints, case studies, etc., can help ensure that an attorney is present to handle legal questions or maintain privilege on the discussions.