PRIVILEGED AND CONFIDENTIAL

COMMS DOC: ML FAIRNESS

POC: [redacted] and [redacted]

GOALS

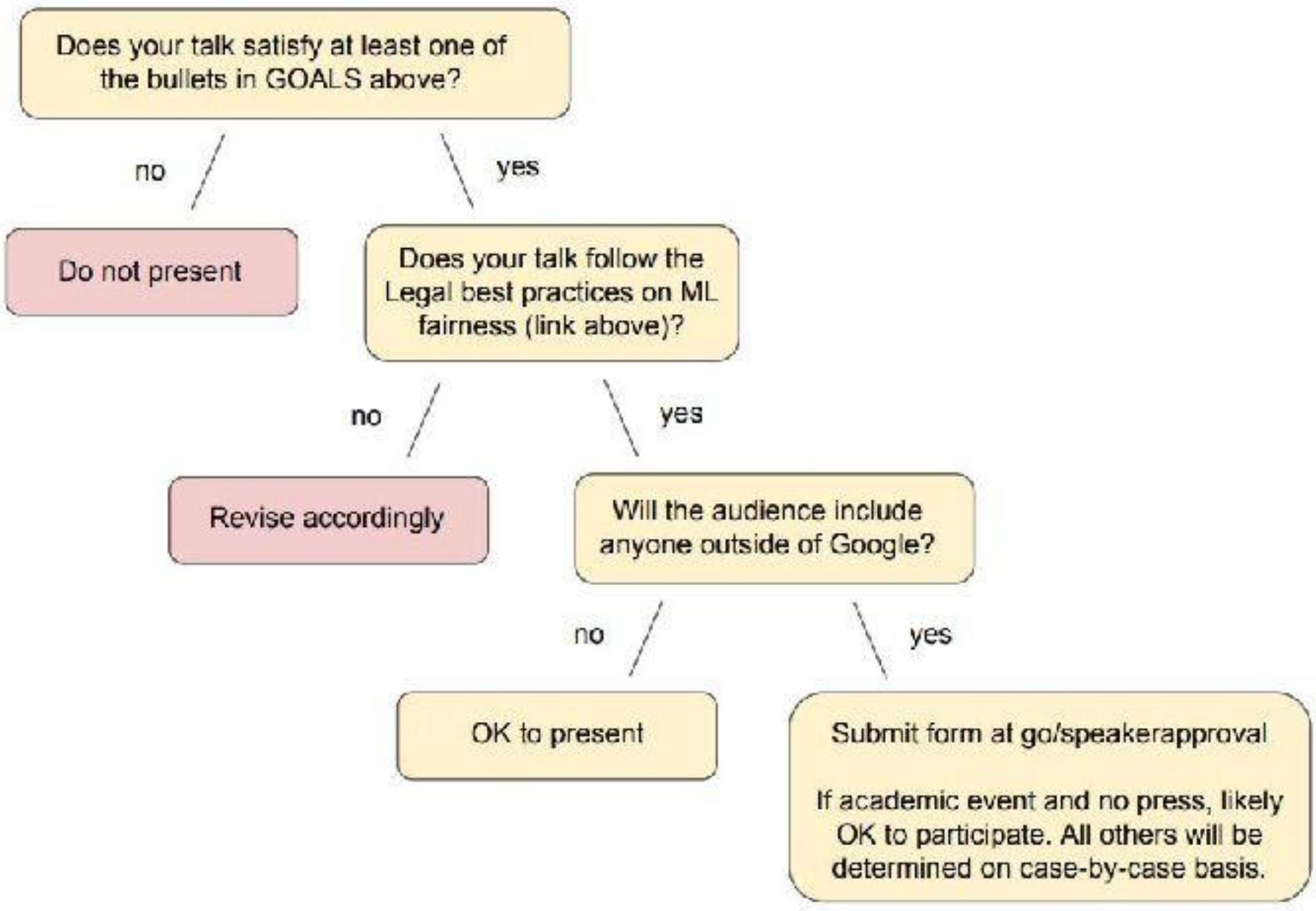
As an AI-first company, Google aims to develop the benefits of machine learning for everyone. Building inclusive algorithms, datasets, and products is crucial to this mission.

Communications on the topic should do one or more of the following:

- Show that Google has diverse teams working on machine learning
- Share models, tools, datasets, and research that other organizations and individuals can use to inform their own efforts
- *[INTERNAL ONLY]* Outline product challenges and internal processes to ensure Google products are inclusive

All communications should follow Legal best practices on ML fairness.

CAN I GIVE A TALK ON THE TOPIC?



KEY TALKING POINTS

As an AI-first company, Google aims to develop the benefits of machine learning for everyone. Building inclusive machine learning algorithms is crucial to this mission. We've been doing a lot of work in this area, including:

Fairness:

- Equality of opportunity (Hardt et al. NIPS 2016)
- Adding constraints into training (Cotter et al. NIPS 2016)
- Designing fair auctions (Bateni et al. EC 2016, Goel et al. LIPics 2016)
- Using machine learning to help bust gender bias in media (Geena Davis Inclusion Quotient)

Explainability / interpretability:

- Designing transparent machine learning (Gupta et al. JMLR 2016, Gupta et al. NIPS 2016)
- Visualizing what an ML system is learning ("interlingua" in multi-lingual neural translation , Smilkov et al., 2016 and open-sourced tool)
- The ability to debug a deep learning system (Sundararajan et al., ICML 2017)

Openness:

- Democratizing use of machine learning through education tools like TensorFlow playground and AIY projects
- Hosting a wide variety of machine learning interns, residents, and professors each year with great outside perspectives and opportunities to exchange ideas
- Open-sourcing our ML library TensorFlow so it's easier for everyone to participate in this process, including dozens of open-source add-ons for particular tasks
- Sharing dozens of datasets across a range of domains, from images to videos to text to speech

COMMUNICATIONS CALENDAR

Date	Channel and message(s)
Jun 10, 2017	ML Fairness at Legal Hot Topics [*INTERNAL ONLY*]; talking points, deck
Jul 2017	Science magazine on ML interpretability; briefing doc
July 10, 2017	PAIR launch; comms doc
Aug 17, 2017	TGIF on fairness in ML systems
Aug 2017	NYT Magazine on ML interpretability
Aug 2017 (mid-month)	ML education externalization efforts; go/mle-status
Aug 29, 2017	R/MI Inclusion Summit
Aug - Dec 2017	Various research papers posted
Sep 2017 TBD	ML Fairness at TGIF [*INTERNAL ONLY*]
Sep 2017	ProFair policies and process in TL;DR [*INTERNAL ONLY*]
Nov 2017	Research at Google Conference [*INTERNAL ONLY*]

	Template deck will include a slide to outline key actions in that project towards inclusion

FAQS

What is Google doing to make sure the machine learning tools you develop are inclusive?
As an AI-first company, Google aims to develop the benefits of machine learning for everyone. Building inclusive machine learning algorithms is crucial to this mission. We've been doing a lot of work in this area, including:

- **Fairness:** equality of opportunity (Hardt et al. NIPS 2016), adding constraints into training (Cotter et al. NIPS 2016), designing fair auctions (Bateni et al. EC 2016, Goel et al. LIPics 2016), using machine learning to help bust gender bias in media (Geena Davis Inclusion Quotient)
- **Explainability / interpretability:** designing transparent machine learning (Gupta et al. JMLR 2016, Gupta et al. NIPS 2016), visualizing what an ML system is learning ("interlingua" in multi-lingual neural translation, Smilkov et al., 2016 and open-sourced tool), and the ability to debug a deep learning system (Sundararajan et al., ICML 2017)
- **Openness:** democratizing use of machine learning through education tools like TensorFlow playground, hosting a wide variety of machine learning interns, residents, and professors each year with great outside perspectives and opportunities to exchange ideas, open-sourcing our ML library TensorFlow so it's easier for everyone to participate in this process.

[INTERNAL ONLY] A number of people from R/MI, Privacy, Policy, Legal, Communications, People Ops, and various product areas are working together on many of these issues in the ML Fairness project. Check out go/ml-fairness to learn more, and to get involved.

Neural net models are a black box, right? And there's a tradeoff between interpretability and accuracy?
First of all, humans aren't very good at explaining their decision-making either!

Second, in machine models, complexity and interpretability are not necessarily opposed. It's true in general that complex neural net models can achieve more accurate results, and can be harder to interpret, than simpler models. However, seemingly simple models like regression, especially when used in systems at scale, can also be difficult to interpret. For example, correlations between variables, variables of different units and magnitudes, and larger systems made of many models chained together can all make it challenging or even infeasible to untangle the effects of individual variables in "simple" models.

Second, neural net models are not inherently uninterpretable—we are just still developing the tools to probe and understand them. We are making progress, for example designing transparent machine learning (Gupta et al. JMLR 2016, Gupta et al. NIPS 2016), visualizing what an ML system is learning ("interlingua" in multi-lingual neural translation, Smilkov et al., 2016 and open-sourced tool), and the ability to debug a deep learning system (Sundararajan et al., ICML 2017).

What Google programs exist to increase my awareness and sensitivity to diversity and inclusion issues?
[INTERNAL ONLY]

- Unconscious Bias and Bias Busting training (go/unbiasing)
- Sojourn—a new course offering that aims to improve the racial and gender climate at Google (go/sojourn; in 2017 only available in the Bay Area; in 2018 plans to scale to 5-7 other US offices)
- Integrating Inclusion (I^2) - go/intergratinginclusion