Disclosure and Evaluation as Fairness Interventions for General-Purpose AI

Vyoma Raman, Judy Hanwen Shen, Andy K. Zhang, Lindsey Gailmard, Rishi Bommasani, Daniel E. Ho, Angelina Wang

Motivation

- Fairness is ***contextual*** and occurs across levels of analysis
- General-purpose AI (GPAI) ***lacks context*** at deployment time
- Fairness interventions must ***mitigate potential for harm*** and ***support accountability*** in context
- Information-gathering through ***disclosure*** and ***evaluation*** can be effective fairness interventions themselves

Levels of Fairness

Model Level: disparities in outputs, predictions, and representations

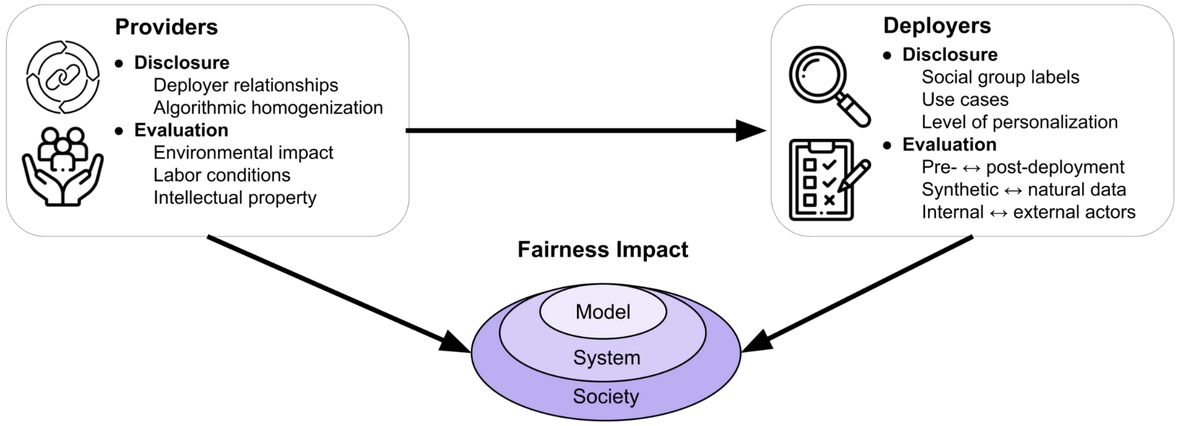
System Level: disparities in decision-making due to interaction of outputs and institutional processes

Society Level: cumulative disparities in social, economic, and health conditions across populations

Actors

System Providers: Make GPAI models accessible to others

System Deployers: Integrate GPAI model into applications



System Providers

Countering Common Intuitions

- Fairness is ***not*** an intrinsic property of a model
- "Eliminate data and model bias": unclear implication without context
- Dual use: utility for valuable tasks
- Limit direct requirements for improving model development practices

Disclosure: Supply Chain Relationships

- ***Which*** deployers are using a model, and ***how***?
- Responsibility allocation based on cause of fairness harms
- Human rights and environmental impacts that are exacerbated for marginalized groups
- Proactive identification of risky use cases
- Identify algorithmic homogenization risks

Evaluation: Development Decisions

- Currently unclear how ***upstream*** development decisions cause ***downstream*** risk across levels
- High-resource developers: invest in AI research establishing these risks
- How biases manifest and cascade across AI lifecycle levels of fairness
- Report correlations between outputs and social attributes

System Deployers

Disclosure: User Interaction

- Unique proximity to usage contexts
- **Social group labels:** make it possible to identify disparities and "hard" data subsets, within reason
- **Personalization:** how individual interaction histories are recorded and used
- **Use cases:** tasks that GPAI systems are used to complete in practice; assists system-level analysis

Evaluation: Diverse Techniques for Rich Analysis

- Identify harm at different levels of fairness

Evaluation	Stage	Data	Actor
Benchmarks	Pre-deploy	Synthetic	Internal
Bug bounty	Pre-deploy	Synthetic	External
Historical data	Pre-deploy	Natural	Internal
Compliance audit	Pre-deploy	Natural	External
Simulations	Post-deploy	Synthetic	Internal
Public red teaming	Post-deploy	Synthetic	External
A/B testing	Post-deploy	Natural	Internal
Incident reporting	Post-deploy	Natural	External

- Deployment stage
  - Before: metric disparities
  - During: unconstrained usage
- Data type
  - Synthetic: controlled perturbation
  - Natural: ecological validity
- Actor
  - Internal: greater access to system
  - External: different assumptions/values

Gathering Information

- Can be an ***active*** intervention, not just a passive one
- Force actors to confront fairness harms: biases visibilized
- (In some cases) corporations have acted to correct their tools
- **Potential drawbacks:** reveal sensitive information; undue burden
- Requires careful scoping

Regulation Scope

Risks of Poor Scoping

- Potential to be ***burdensome*** on small-scale actors, which can stifle competition
- Criteria inadequate to address impacts of real models

Thresholds

- Compute insufficient: ***lacks context*** necessary to estimate fairness harms
- Context requires rich information: ***multiple dimensions*** must be considered together

Fairness Level of Scoping

- System-level lens centers on ***interactions*** between models, data pipelines, institutional processes, and human oversight
- Dimensions: severity, voluntariness, scale, and distribution of harm
- Oversight should be ***proportionate*** to compute resources, labor capacity, and system reach