Ethical Algo Design in the Generative Era

Generative AI

- What is it?
- E.g. LLMs
- Next-word prediction
- Training & annotation
- History / precursors
- Embeddings
- Autoencoders/decoders

Fairness in GAI

- Compare e.g. group fairness in consumer lending
- Fairness in prompts?
- In tone?
- Biases vs. correlations, use cases

Privacy in GAI

- Regurgitation
- Virtual copying (e.g. code)
- Stylistic mimicry (more later)
- Could do DP model training, but impractical

Toxicity

- Subjectivity & context
- Quotations & censorship
- Offensive opinions
- Nuance & indirection

Hallucinations

- Plausible but verifiably wrong
- E.g. citations

Copyright & IP Concerns

- Stylistic mimicry
- Threat to creative livelihoods?

Plagiarism & Cheating

- College essays, writing samples, etc.
- Tool vs. cheat?
- Author verification

Disruption to Work

- Passing bar exams, MBA courses, etc.
- IP concerns again

Approaches

Toxicity & Fairness

- Training data curation
- Must automate
  - → human annotation (unpleasant)
- Train "guardrail" models
- Apply to training, prompts & output

Hallucinations

- User education
- Independent verification
- External sources (e.g. citation DBs)
- Disclaimers
- Content attribution

Copyright & IP

- Tech, policy & legal mechanisms
- Content attribution & compensation
- Machine Unlearning & model disgorgement
  - DP
  - Sharding

Author Verification

- Train model to detect human/LLM (arms race)
- Redlist/greenlist

Job disruption/creation

- prompt engineer
- "English is the new programming language"

Strongest defense:

Use case specialization