Truth, Trust, and Technology

CS 111 Ethics

Trust Refresher

- **What is trust?**
  - An unquestioning attitude
  - Beneficial because it extends agency
- **Ways to establish trust**
  - Assumption (weak, risky)
  - Inference (most powerful)
  - Substitution (build on something else you trust)
- **Trust is essential but risky**
  - Agential gullibility: misplaced trust

Societal Conflicts

- **Trust plays a key role in our country's divisions**
- **Different groups have conflicting beliefs about basic facts:**
  - Who won the election?
  - Is the economy getting better or worse?
  - Is crime rising or falling?
  - Is climate change happening? If so, are humans responsible?

There is only one truth: tens of millions of people are wrong!!

Large-Scale Agential Gullibility

- **(What I Believe) >>> (What I Perceive)**
  - Individuals don't have resources to answer questions ourselves
  - We must choose to trust information/conclusions from others
- **Different groups trust different sources on key issues of fact**
- **Some of these sources must be wrong: agential gullibility**
- **Why agential gullibility on such a large scale?**
  - Hard to reliably infer trust
  - Error-prone inference techniques:
    - Confirmation bias: "I trust this source because it validates my beliefs"
    - False trust in numbers: "Lots of people are saying this, so it must be true"

Technology is exacerbating agential gullibility

Example #1: Facebook

- **Attention => $$**
- **Reinforcing biases and fears increases attention (users aren't interested in conflicting views/data)**
- **Result: users see *lots* of material confirming their beliefs**
- **Different users see different material**
- **Facebook profits from your confirmation bias**

Takeaways:

-  **!= truth**
- **Optimizing for attention leads to bad places**

Example #2: ChatGPT

- **Generative AI tools can produce useful and insightful information**
- **ChatGPT presentation causes people to infer trust:**
  - Authoritative, with explanations (Bansal et al. 2021)
  - Lots of concrete “facts” (Bower et al. 2024)
- **But, ChatGPT hallucinates; no reason to trust!**
- **Embedding ChatGPT in other apps obscures origin of information**

Takeaways:

- **Do not trust ChatGPT for truth!**
- **Treat output as hypotheses to consider**
- **All results must be independently validated (must use substitution)**

Example #3: Deepfakes

- **Historically: hard to fabricate convincing photos, videos, audios**
- **People inferred trust (for good reason)**
- **New technology enables compelling fakes**

Takeaways:

- **Must unlearn trust in photos, videos, and audios**
- **Do not trust without additional validation**

Small-Group Discussions

- **What observables can be used to separate trustworthy information sources from untrustworthy ones?**
  - Indicators suggesting trustworthiness
  - Indicators suggesting untrustworthiness
  - How to prevent confirmation bias?
- **Discuss in groups of 2-3**

Conclusions

- **Trust is at the heart of our societal divisions**
- **Deciding whom to trust is becoming more difficult**
  - Confirmation bias is extremely hard to avoid
- **Technology makes the problems worse**
- **There are objective factors you can use to infer trust (but it's not easy!!)**

- **Best hope: institutions with an established record of trustworthiness**
- **But, will people trust them?**