Truth, Trust, and Technology

Mendel Rosenblum

Truth, Trust, and Technology

Optional readings: Operating Systems: Principles and Practice: None

Trust Refresher

- What is Trust?
  'the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party'

  The Ethics of Advanced AI Assistants https://arxiv.org/pdf/2404.16244

  Helpful because it lets us accomplish more than we could on our own

  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 (misplaced trust)
  - Trustor: **over-trust** - individual extends trust beyond reasonable limits (trusting too much)
  - Trustee: **untrustworthiness** - failure to display integrity, reliability, or care (begin unworthy of 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 Misplaced trust

- (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 untrustworthy
- Why over-trust on such a large scale?
  - Hard to figure out who deserves 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 over-trust

Example #1: Social Media Algorithms

- The Problem:
  - 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
  - The platform profits from your confirmation bias
- The harm:
  - Partisan content skews around elections
  - Courts ruled algorithms causes anxiety, depression, body dysmorphia, and suicidal thoughts

Social Media Algorithms Takeaways

- (likes, interaction, sharing) != truth
- “The algorithm showed me this” is not the same as “an editor chose this”
- Not all attention is the kind you want to optimize for
- Who should be responsible?
  - Platforms
  - Government/legislation
  - Individual users
- Key: one actor won't be able to solve. Need coordination.

Example #2: Generative AI (ChatGPT, Claude, etc)

- Generative AI tools can produce useful and insightful information
- The way AI presents information makes people trust it:
  - Authoritative tone with detailed with explanations (Bansal et al. 2021)
  - Lots of concrete “facts” (Bower et al. 2024)
  - Confidence even when wrong
- The Reality:
  - AI systems that hallucinate, particularly unpredictably and without warning are often untrustworthy
    - Particularly for factual accuracy
    - Fact-checking or verification is left to users
- AI hidden inside apps can make people forget where information came from
  - One hallucination can spread and look like many independent sources

Generative AI Takeaways

- Do not trust generative AI for truth without verification!
- Treat output as hypotheses to consider
- All results must be independently validated (must use substitution)

Example #3: Deepfakes & Synthetic Media

- Historically: hard to fabricate convincing photos, videos, audios
- People inferred trust (for good reason)
- New technology enables compelling fakes
    - Old photos can be turned into faked AI videos
        - A video labelled as AI but believed to be real
        - A likely real audio that was dismissed as a deepfake
    - It isn’t just about fakes being believed — it’s about the entire collapse of epistemic trust
        - People stop believing that any evidence reliably tells them what is true.

Synthetic Media Takeaways

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

  "A people that can no longer believe anything cannot make up its own mind…And with such people you can then do what you please" - Hannah Arendt

- Opportunities for innovation:
  - Better ways to label AI-generated content
  - Invest in detection that works
  - Support journalists and trusted institutions
  - Protect people from unauthorized voice and image cloning

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
- Over-trust and untrustworthiness are both major problems
  - Confirmation bias is extremely hard to avoid
- Technology amplifies both problems:
  - Untrustworthy sources appear more credible
  - We over-trust familiar or convenient sources
- There are better and worse ways to decide whom to trust
  - Best hope: institutions with an established record of trustworthiness
  - But, will people trust them?