

Privacy in training and using AI

kyungsinpark@korea.ac.kr

KS Park

Open Net/ Korea University

What is AI?



- Currently, a stochastic machine that analyzes across vast sets of human-behavioral data and regurgitates the most statistically probable response to a genuine human prompt.
- Learns "things" like a child learns: lots of cat photos v. non-cat photos but without ever defining felinity
- Does not really "learn things" as in "understanding" but only **mimics human cognition**
- Alpha Fold: trains on known (to humans) folded structure from known amino acid sequences and mimics the relationship in deducing unknown folded structure from known amino acids. (knowing structure → knowing function)
- Zero creativity! Just statistics. But hugely and hugely helpful.
- Ben Afflek: "AI results converge on mediocrity". Obvious cz it **averages** over vast sets of human behavior.

Training – Inference - Feedback

- Training
- Inference
- Feedback

Training and privacy

- Theoretically not privacy-infringing b/c tokenization
- “Cats hate dogs” → “Cats”, “Hate”, “Dogs”, relational data among the three
- Only relationship between sub-parts of data is remembered, not data itself
- Example: pointy ears * “cat” vs point ears * “dog”
- Problem of **overfitting**: Too many photos showing Harry Potter for wizard students. AI now thinks all wizard students are Harry Potter
- Same thing can happen in privacy

Case of Lee Luda

Scatter Lab's **Science of Love**: collects and analyzes private conversations to give relationship advices

Lee Luda: AI trained upon Science of Love data answers in conversational mode

Problem: Real addresses of the people are shown (overfitting)
Personal data? Are phone numbers without any other data?

What semantic value: in addresses exposed in Lee Luda?

Safety measure: Anonymization!

Be careful: Overzealous protection can backfire.

Korea: Complete ban on re-identification may shut down data subjects rights of access, portability, halting processing, etc.



Inference and Privacy

- Input: human prompt and other contextual information made available to AI (e.g., public web)
- Output: stochastic result produced by AI **to the user who gave human prompt**
- Q: how long should AI retain the human prompt (maybe sensitive data)? Is it privacy-infringing if kept too long and used as contexts for later prompts? Do we need new norms?
- Case of **targeted advertising on third party website behavioral data**
- We already decided that **using my personal data on future me can violate my privacy**

Feedback

- One user's feedback used as training data for AI which will do inference for other users
- Example: Rare disease patient's private conversation transcripts used to retrain AI for conversing with other patients
- Privacy Q: Possibility of overfitting, possibility of re-identification
- Hypothetical: Modified Lee Luda

Conclusion

- Do we need AI-specific or agentic-AI-specific privacy law or design?
 - Agentic AI making \$500 purchases that user would not have approved. Is it a privacy problem? → **Singularity control** – IRB (institutional review boards used for bioethical research)
 - Why are we concerned about privacy so much?
- **Data monopoly regulation:**
- X - doubling down on IP
 - O – data portability
 - O – open (government) data
 - O – open source AI