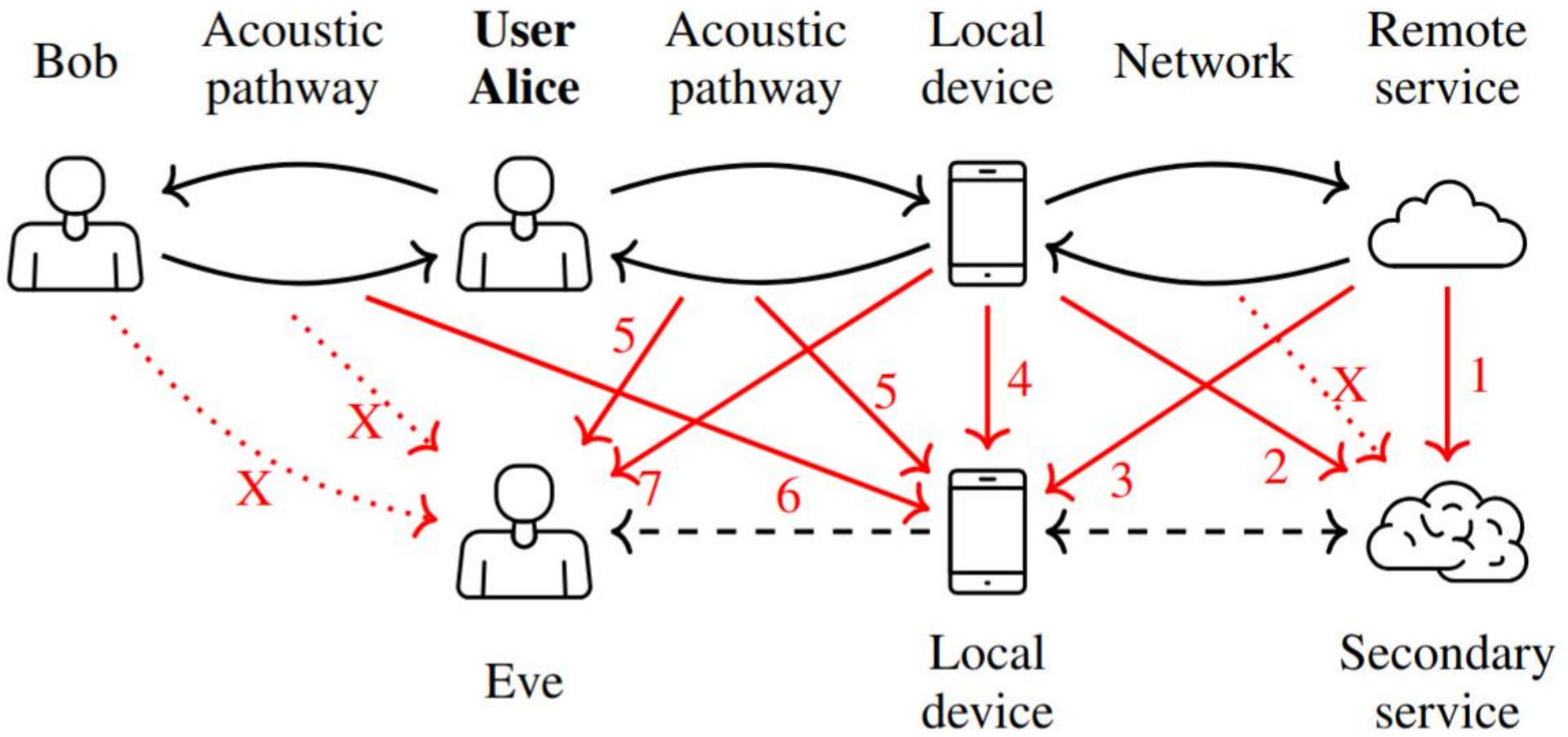


https://presemo.aalto.fi/sigspsc2025

Overview of Privacy in Speech Technology

Tom Bäckström

Aalto University

Background

- Survey talk at Interspeech 2022 extended to article.
- Bäckström, Tom. "Privacy in speech technology." *arXiv preprint arXiv:2305.05227* (2023).
- In review at IEEE for almost 2 years.
  ⇒ Already a few updates.

Contents

- Introduction
  - One more definition of privacy
  - Categories of private information
  - Threats
  - Modelling attacks
- Protections
- Evaluation
- Open questions

Photo by Donald Giannatti on Unsplash

Definitions

(Information) **Security** refers to making sure that **only those with authorization have access = categorical access policy**

Privacy refers to making sure that the level of access to and use **of information is appropriate = contextual access policy**

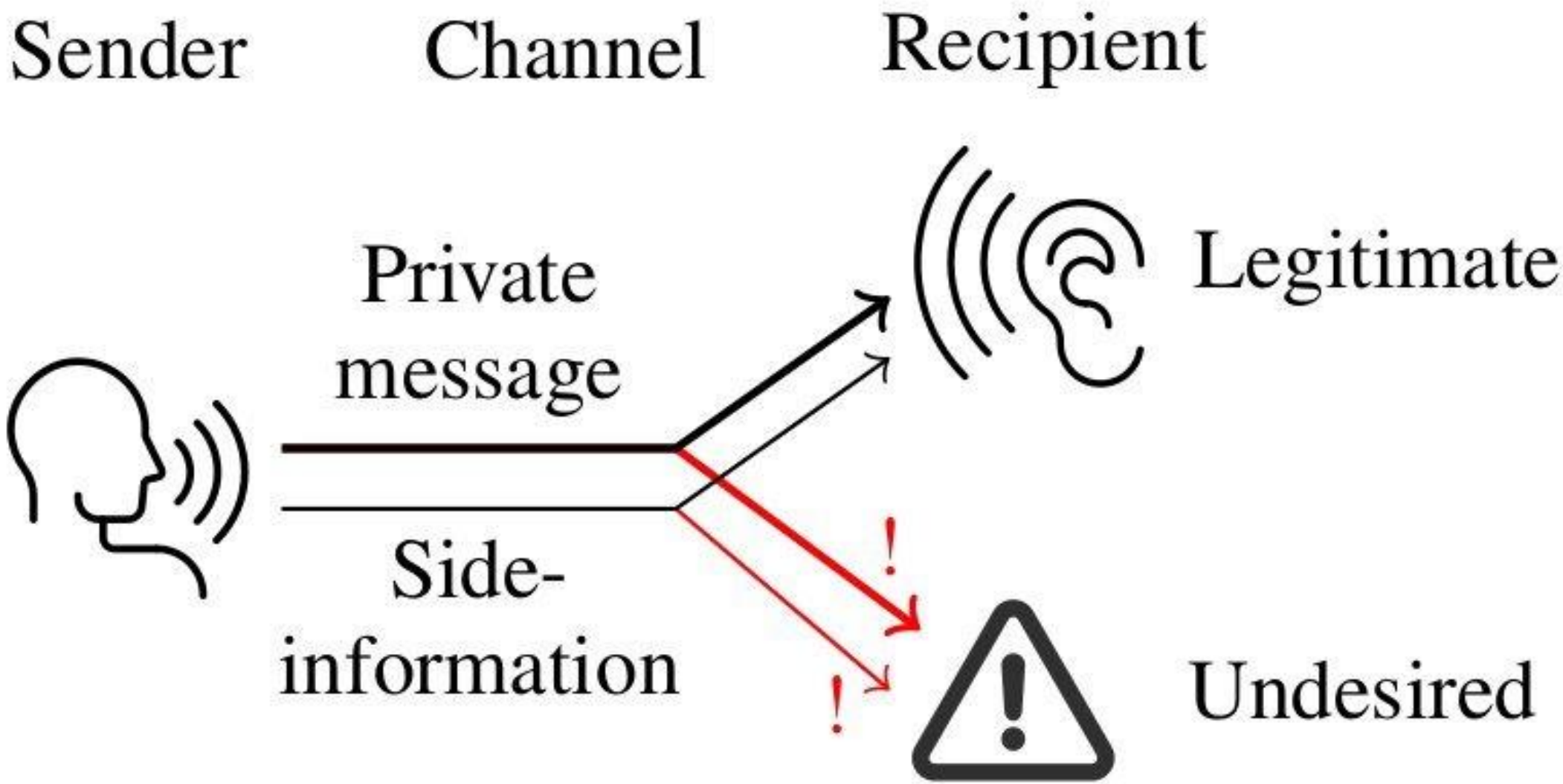
Adapted from Nissenbaum (2004) and discussions with prof Jolly Thomas.

Categories of private information in speech

- Biological
  - Body characteristics, health, intoxication
- Psychological
  - Emotions, intelligence, education, gender identity
- Message
  - Text, emphasis, style, expression, mannerism
  - Language, accent, skill
- Affiliation
  - Ethnic, national, cultural, religious, political

- Relationship character
  - Hierarchy, familiarity, attraction, intimacy
- Physical environment
  - Background, distance to sensor/in transmission, reverberation
- Hardware & Software
  - Sensor type
  - Codec, enhancement algorithm

- See e.g. Rita Singh, "Profiling Humans from their Voice", Springer, 2019.


Sender
Channel
Recipient
Private message
Side-information
Legitimate
Undesired

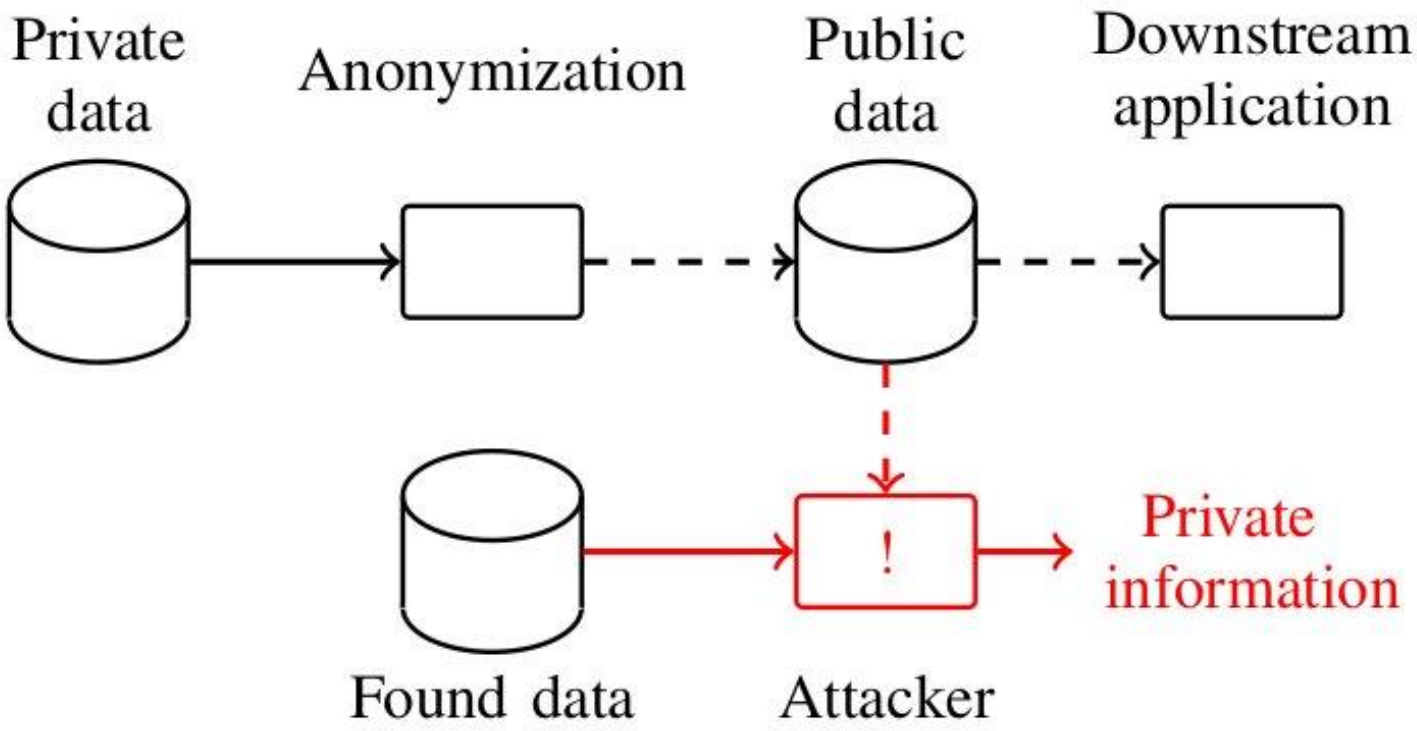
Threats 1/2

- **Price gouging** - Signs of depression or other health problems in users' voices could be misused to trigger an increase in their insurance premiums. Signs of users' emotions could be exploited to offer them products at higher prices.
- **Tracking/stalking** - Voice re-identification could link users across platforms, i.e., from work-related social media to online support groups and dating apps, making it possible to follow them anywhere.
- **Extortion, public humiliation** - Private health problems, and romantic affairs could be detected in the voice and used for blackmail or made public against a user's wishes.

Threats 2/2

- **Algorithmic stereotyping** - Recommender systems based on voice can become biased with respect to age, identity, religion, or ethnicity, in ways that are nearly impossible to monitor.
- **Harassment, inappropriate advances** - Users in chat rooms or virtual reality could be automatically singled out by gender or opinions, making them a target for unwanted attention and harassment.
- **Fear of monitoring** - The subjective feeling of being continuously monitored may cause psychological damage. It may also stifle political expression damaging democratic societies.

Attack / Risk model



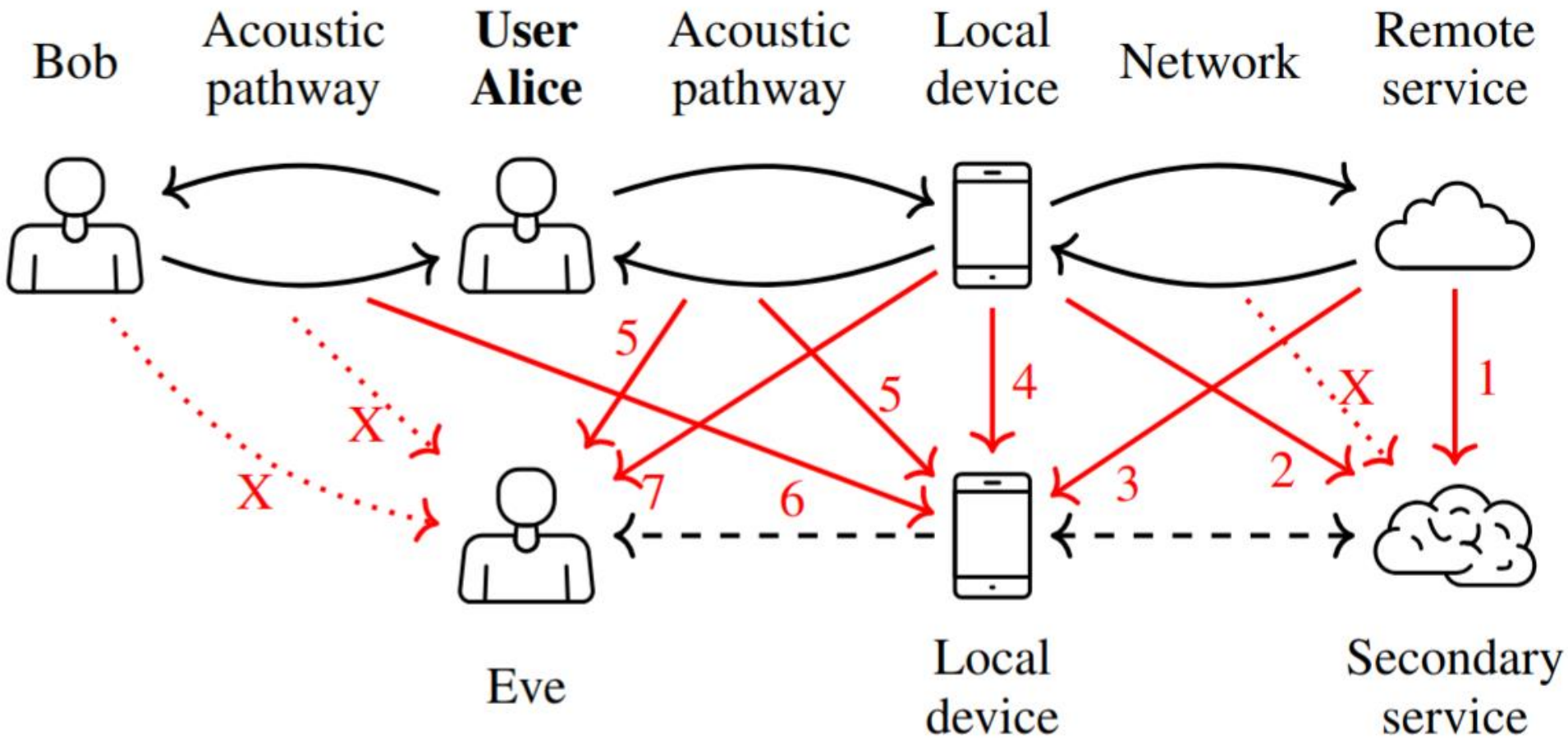
Based on Maouche, M., Srivastava, B.M.L., Vauquier, N., Bellet, A., Tommasi, M., Vincent, E. (2020) A Comparative Study of Speech Anonymization Metrics. Proc. Interspeech 2020, 1708-1712, doi: 10.21437/Interspeech.2020-2248

Attack model / Scenario of Use

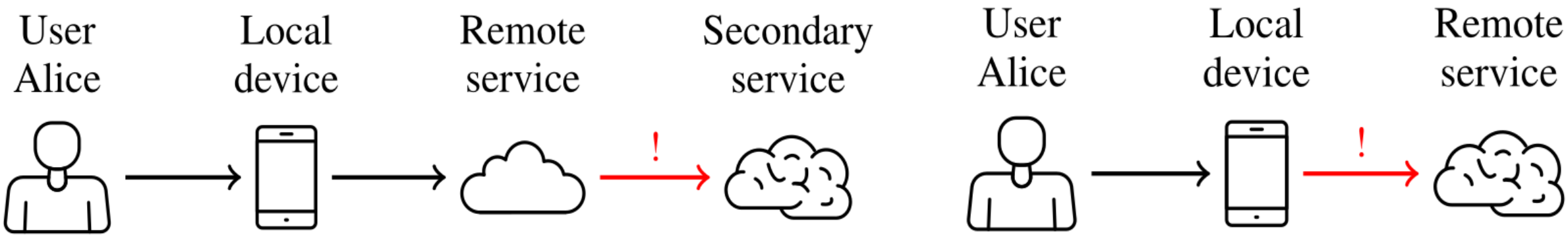
Attacker Model (Attack on Privacy)		Protector Model (Protection of Privacy)	
Objective	The information about the target speakers the attacker seeks to obtain.	Objective	**Defense objective**: The information in the spoken audio that must be protected. **Utility objective**: The information in the spoken audio that must be retained.
Opportunity	The access of the attacker to the target, including availability of target speakers' audio. Also, access to (and any knowledge about) the protection.	Opportunity	The possibilities available for protection (i.e., countermeasures). Also, access to (and any knowledge about) the attack.
Additional Resources	The knowledge, data, and compute available to the attacker to carry out the attack (beyond the access to the target specified under Opportunity).	Additional Resources	The knowledge, data, and compute available to the protector to defend against attack (beyond the access to the attack specified under Opportunity).

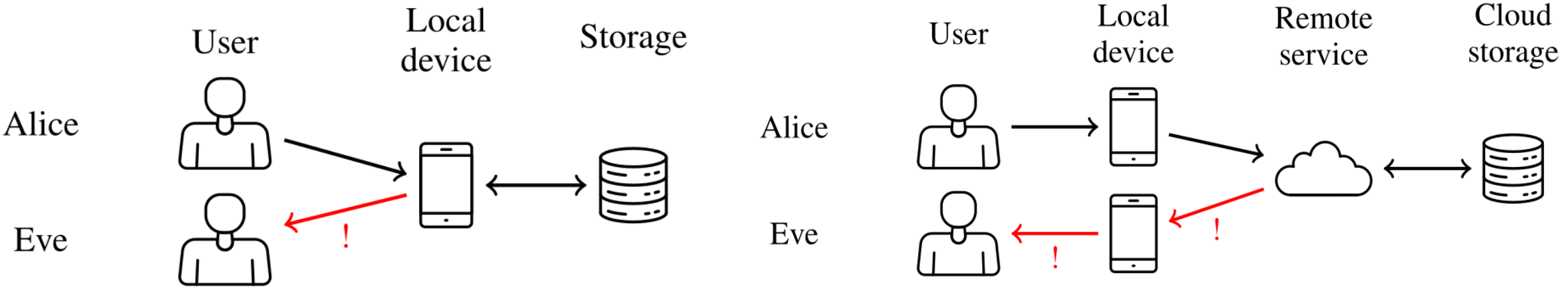
From Rahman, M.U., Larson, M., Bosch, L.t., Tejedor-García, C. (2024) Scenario of Use Scheme: Threat Modelling for Speaker Privacy Protection in the Medical Domain. Proc. 4th Symposium on Security and Privacy in Speech Communication, 21-25, doi: 10.21437/SPSC.2024-4

Attack surfaces



Examples of specific scenarios





Protections

Photo by Jamie Taylor on Unsplash

Protections

1. Information isolation (or data minimization[1])
2. Secure processing
3. Privacy-preserving architectures
4. Acoustic interventions[2]
5. Improving performance

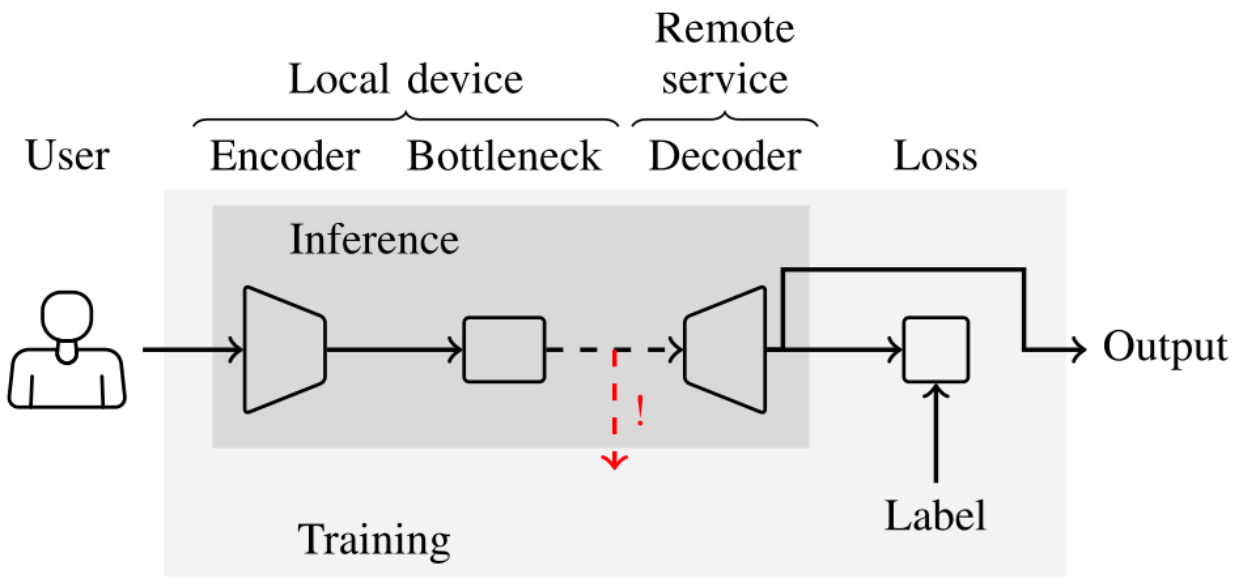
[1] As per a suggestion by Nicholas Evans

[2] To align with vocabulary used in video processing; see Bäckström, Tom, Siddharth Ravi, and Francisco Florez-Revuelta. "Privacy preservation in audio and video." Privacy-Aware Monitoring for Assisted Living. Springer, 2024.

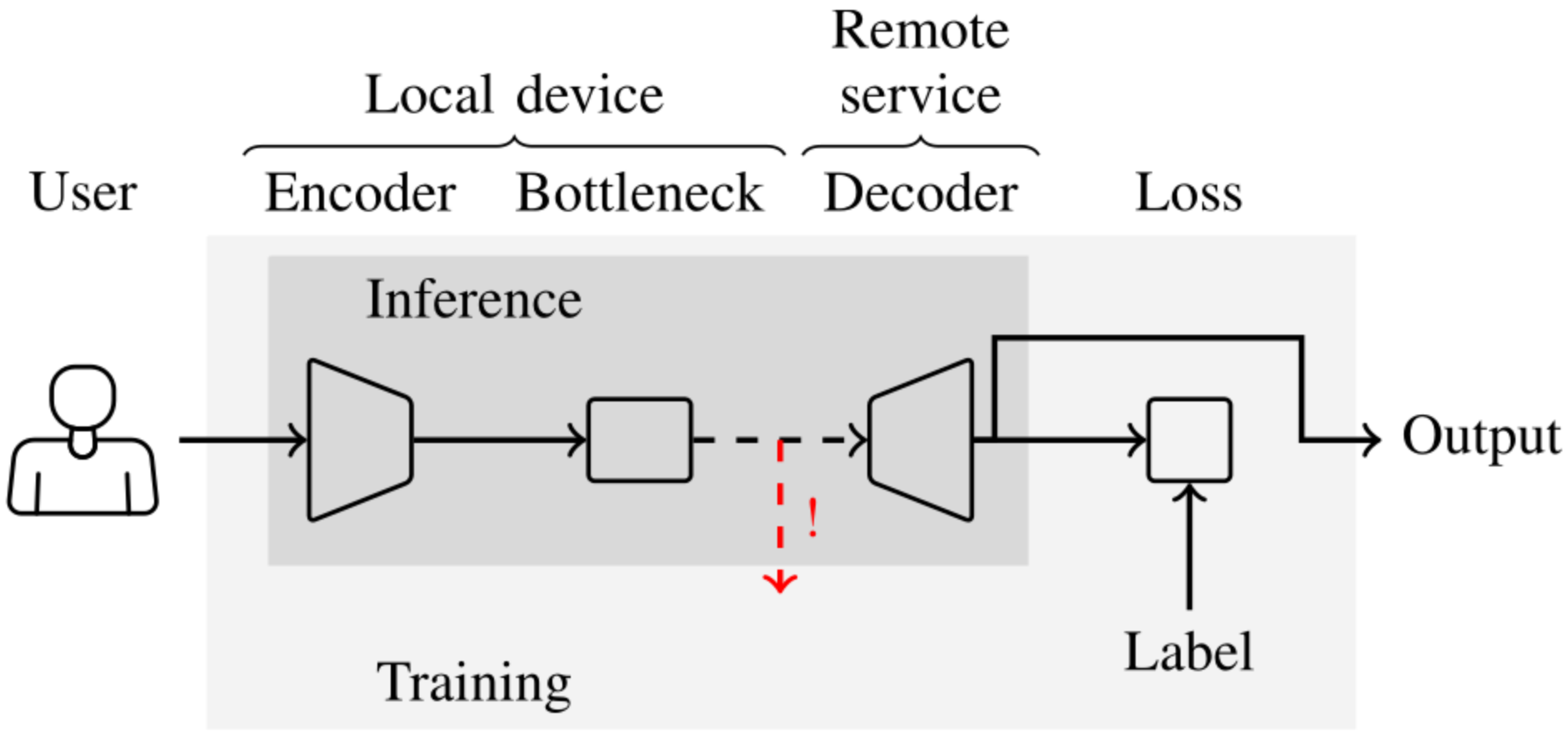
Photo by Jamie Taylor on Unsplash

Information isolation

- Removing or replacing private, while preserving desired information, using:
  1. Information bottleneck / funnel
  2. Adversarial training
- Both approaches can be parallelized for *disentanglement*.

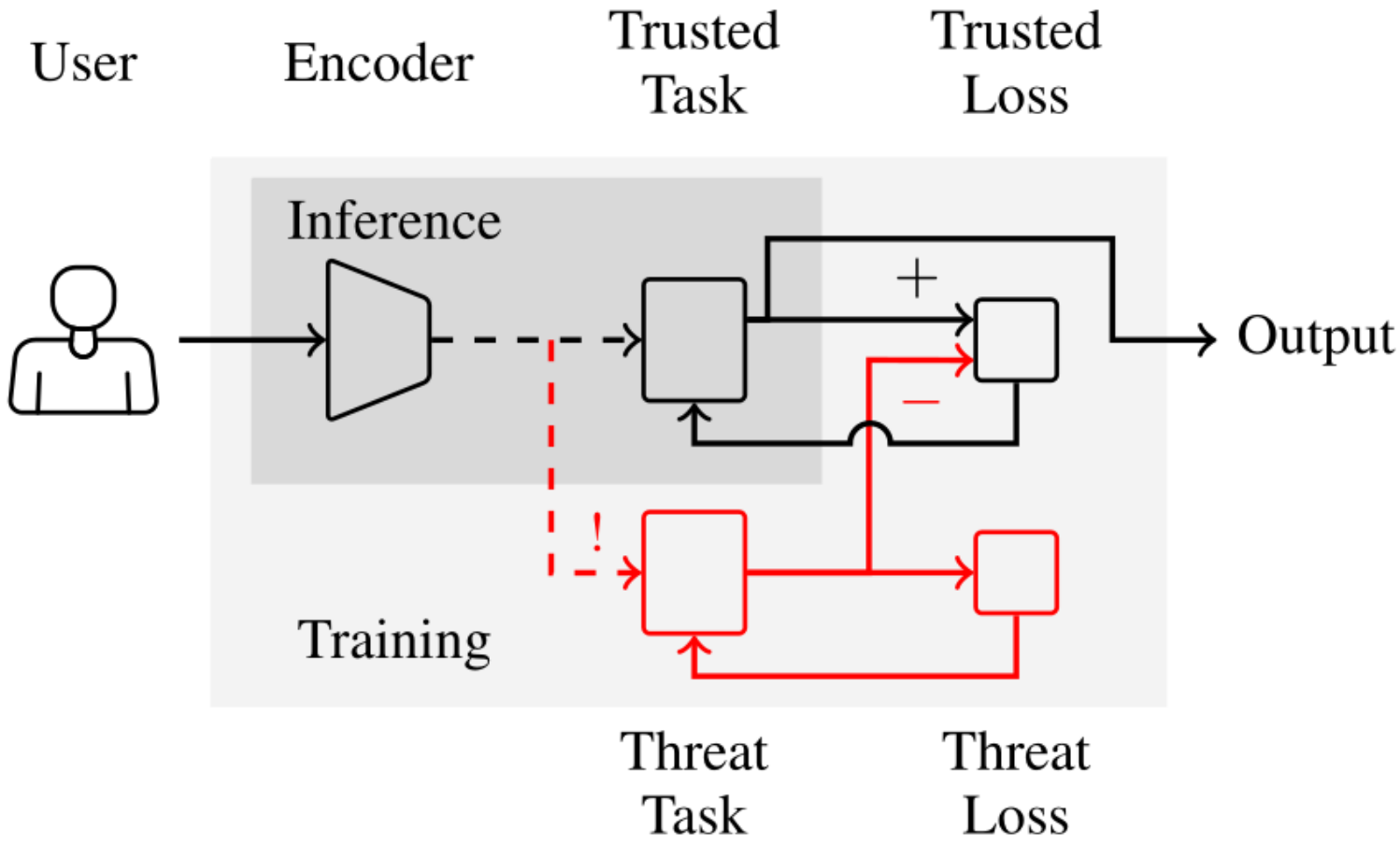


Information bottleneck / funnel



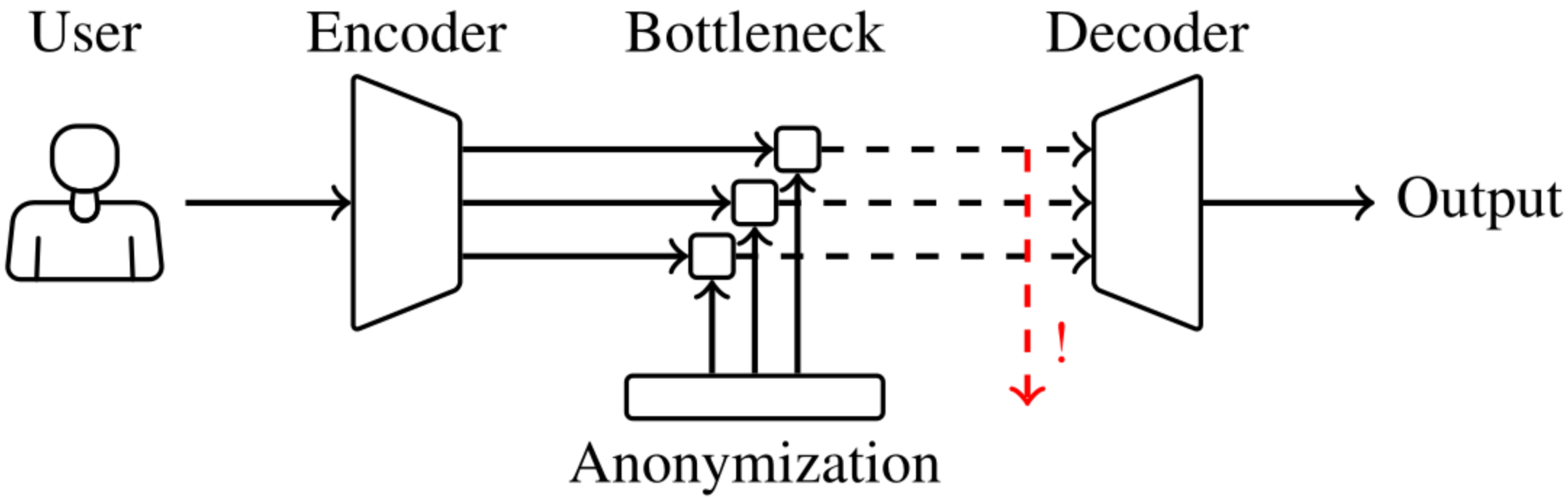
Potential theoretical guarantees on effectiveness of protection.

Adversarial approach

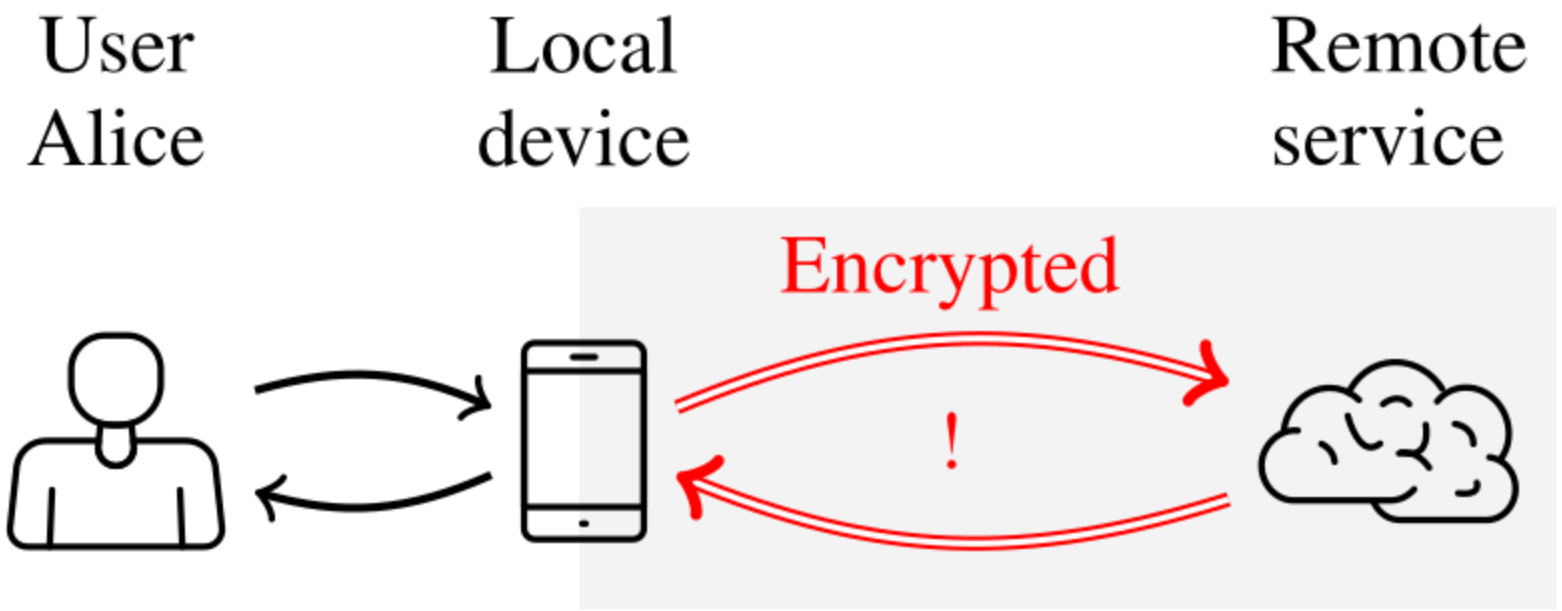


Depends on effectiveness of the attacker!

Disentanglement

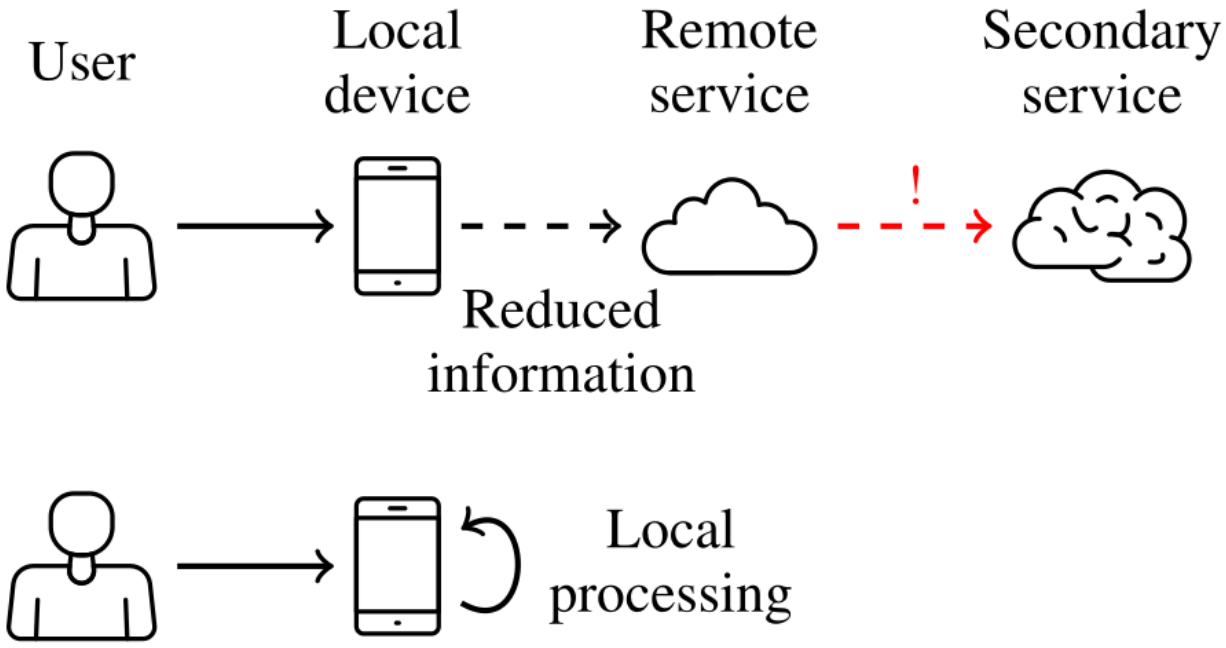


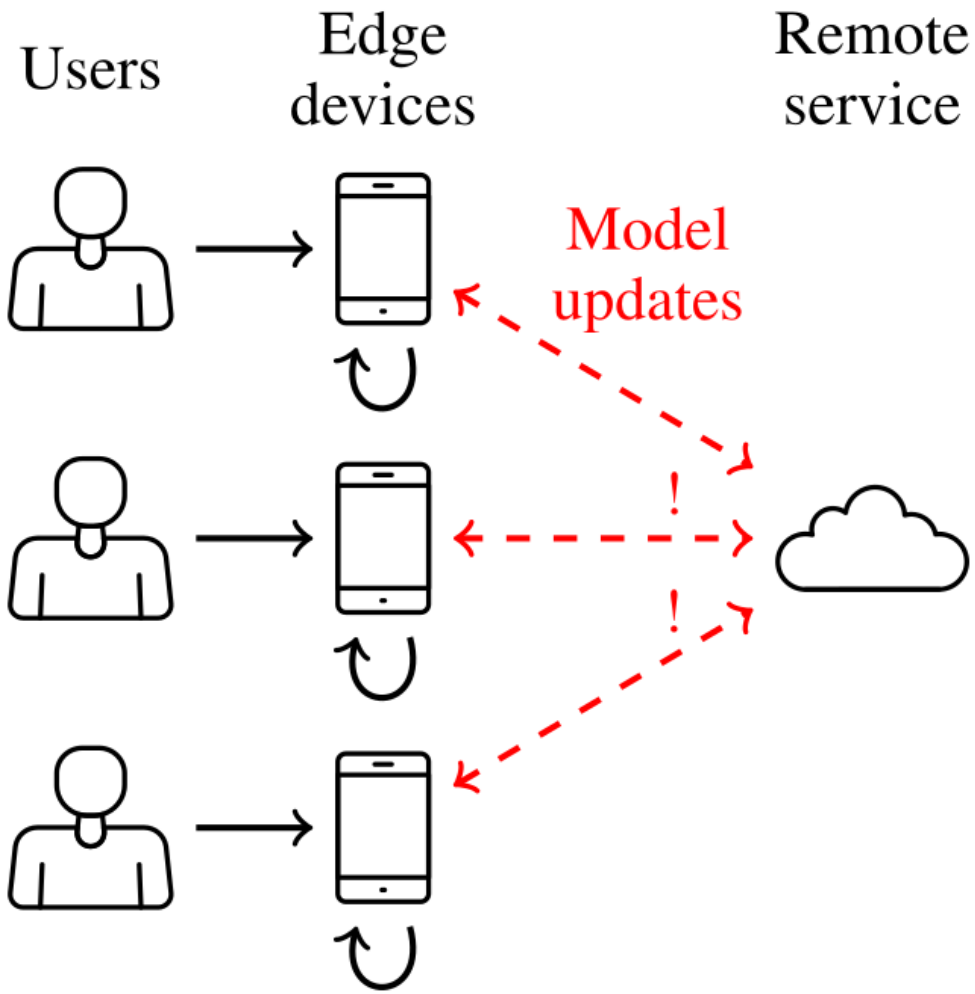
Secure processing



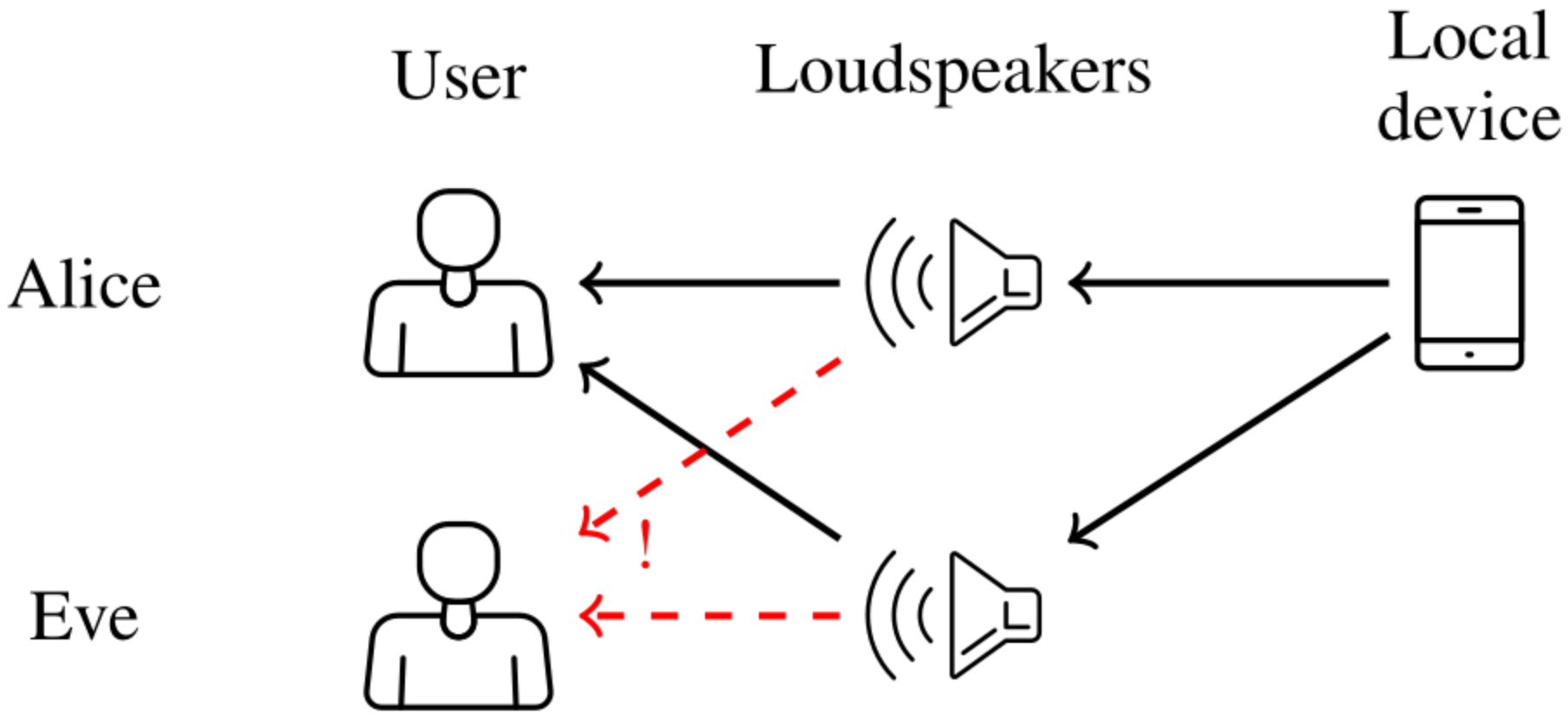
Large overhead in computations and transmission.

Privacy-preserving architectures



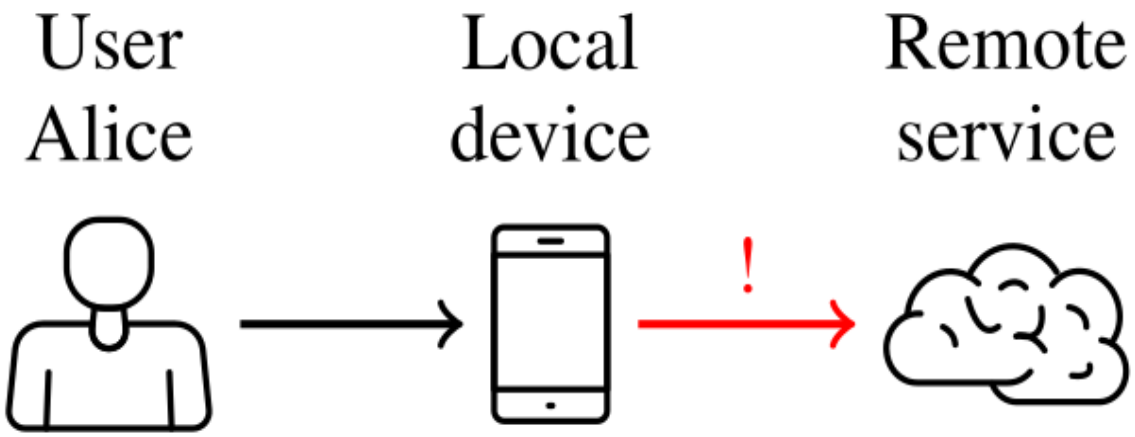


Acoustic interventions



Improving performance

- E.g., **improved wakeword spotting** reduces false activations.



Evaluation

Photo by Siora Photography on Unsplash

Evaluating performance

User-centric performance:

	Performance	
	Objective	**Subjective**
Privacy	The consequences and likelihood of attacks	Users' perception and experience of privacy
Utility	Performance in the downstream, trusted task	User experience (UX) and quality of experience (QoE)

Community and society-level effects currently underappreciated!

Objective measures of privacy

Empirical

- Equal error rate (EER)
- Application-independent log-likelihood-ratio cost function $C_{llr}^{\min}$
- Expected privacy disclosure
- Worst-case privacy disclosure

Theoretical

- Information measured in *bits* $\epsilon = log(M/N)$, where M,N are the population sizes before and after a leak
- Log-likelihood ratio (LLR)
- *k*-anonymity

Theoretical measures are approximated by empirical quantities, making the distinction fuzzy.

Subjective measures of privacy

- Objective privacy is a prerequisite
- No matter what, some people will be paranoid and some oblivious.
- User-interface design can help align perception with reality.
    - Over-promising is a dark design pattern!

⇒ **User evaluation**

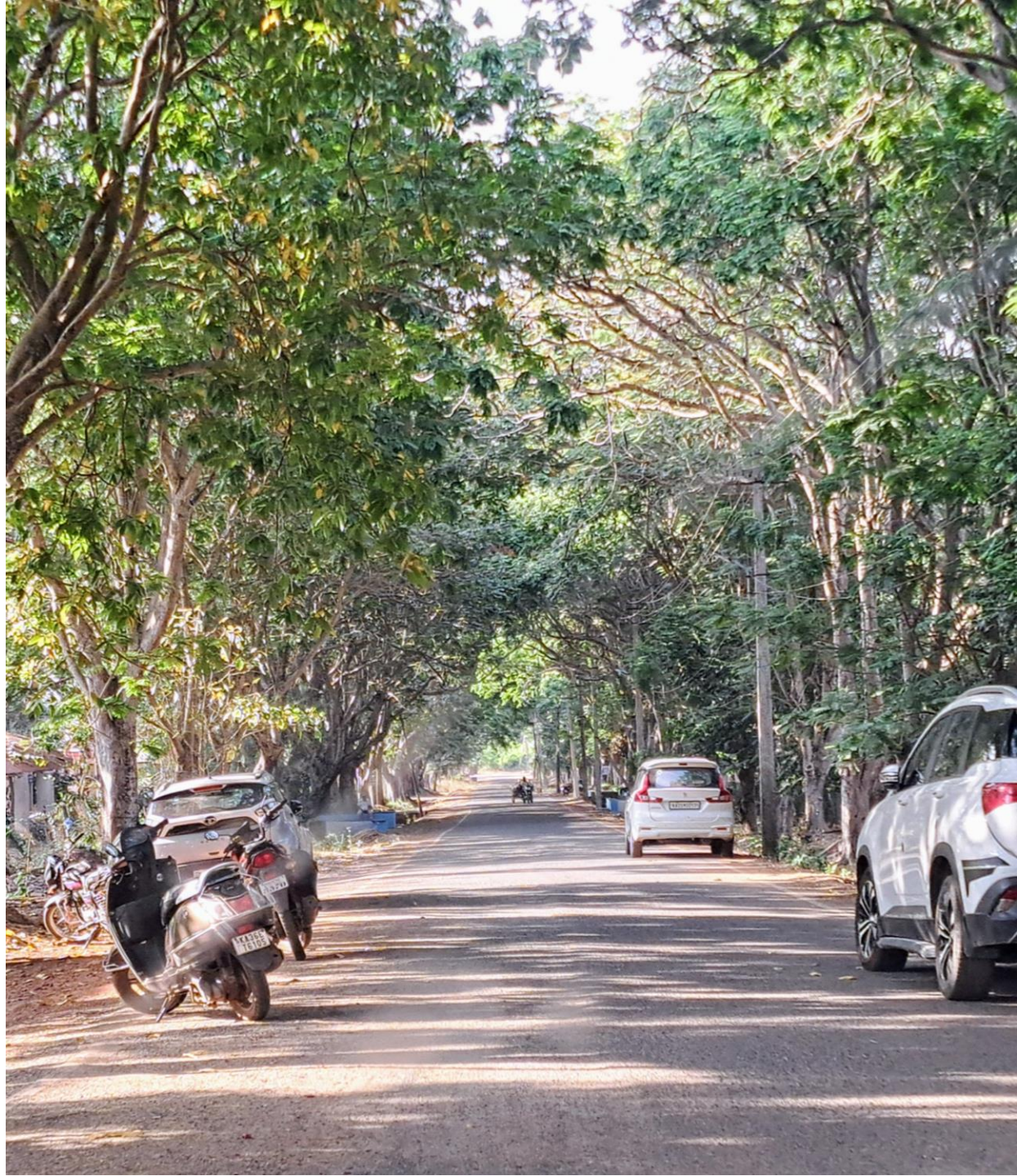
- Questionnaires
- Co-design

Legal landscape

- GDPR in Europe and CCPA in California
  - Good steps forward
  - Practical implementation is still open
- Main weakness
  - Based on categorical concepts (identifiable / not identifiable).
  - Science is based on statistical evidence (probabilities).

Open questions

- Consent, control, and monitoring
- Theoretically valid practical metrics
- Collective privacy[1]
- Disentanglement
- Perception, experience and design of privacy



[1]See e.g. Taylor, Linnet, Luciano Floridi, and Bart Van der Sloot. "Group privacy." New challenges of data technologies. Cham: Springer (2017).

The end – to be continued

Bäckström, Tom. "Privacy in speech technology." arXiv preprint arXiv:2305.05227 (2023).

Aalto University