USENIX

THE ADVANCED COMPUTING
SYSTEMS ASSOCIATION

# "We are not Future-ready": Understanding AI Privacy Risks and Existing Mitigation Strategies from the Perspective of AI Developers in Europe

Alexandra Klymenko and Stephen Meisenbacher, *Technical University of Munich;*
Patrick Gage Kelley, Sai Teja Peddinti, and Kurt Thomas, *Google;*
Florian Matthes, *Technical University of Munich*

https://www.usenix.org/conference/soups2025/presentation/klymenko

This paper is included in the Proceedings of the
Twenty-First Symposium on Usable Privacy and Security.

August 11–12, 2025 • Seattle, WA, USA

ISBN 978-1-939133-51-9

Open access to the Proceedings of the
Twenty-First Symposium on Usable Privacy and Security
is sponsored by USENIX.

# "We are not Future-ready": Understanding AI Privacy Risks and Existing Mitigation Strategies from the Perspective of AI Developers in Europe

Alexandra Klymenko*
*Technical University of Munich*

Stephen Meisenbacher*
*Technical University of Munich*

Patrick Gage Kelley
*Google*

Sai Teja Peddinti
*Google*

Kurt Thomas
*Google*

Florian Matthes
*Technical University of Munich*

## Abstract

The proliferation of AI has sparked privacy concerns related to training data, model interfaces, downstream applications, and more. We interviewed 25 AI developers based in Europe to understand which privacy threats they believe pose the greatest risk to users, developers, and businesses and what protective strategies, if any, would help to mitigate them. We find that there is little consensus among AI developers on the relative ranking of privacy risks. These differences stem from salient reasoning patterns that often relate to human rather than purely technical factors. Furthermore, while AI developers are aware of proposed mitigation strategies for addressing these risks, they reported minimal real-world adoption. Our findings highlight both gaps and opportunities for empowering AI developers to better address privacy risks in AI.

## 1 Introduction

The recent emergence of widely accessible, general-purpose AI systems, including ChatGPT, Gemini, Claude, and Stable Diffusion, has sparked a renewed concern about the privacy risks posed by AI. These risks include models requiring immense amounts of training documents or images [13, 14], models memorizing and reproducing sensitive data [4, 5], model operators logging interactions with potentially sensitive data [8, 17], and potentially harmful applications of models [6, 11]. Along with studying known privacy risks, researchers have also explored mitigations via privacy-enhancing technologies (PETs), yet these investigations often reveal the limitations of such techniques, including balancing the privacy-utility trade-off [11] and justifying computationally expensive methods [37].

In response to the growing list of privacy risks posed by AI, regulators and technologists have worked to systematize known risks and gain perspectives on user concerns about AI. Within the European Union, regulators have codified the risks of AI systems in the upcoming AI Act[1]. Researchers have also come up with their own systematizations of AI risks [8, 11, 33], which can be viewed as a subset of all known AI risks. In addition to enumerating AI privacy risks, these works also recognize the increasing importance of privacy legislation [33] and effective risk management strategies [11]. Finally, studies have also explored user sentiment towards AI and privacy, where users' fears of a decrease in privacy ranks amongst the top two concerns with AI [16]. Missing from these perspectives, however, are the views of developers.

In this work, we explore the current privacy risks of AI, their relative importance, and the efficacy of existing mitigation strategies from the perspective of 25 AI developers in Europe. These participants span the AI development stack, including research, data collection, model training and fine-tuning, and deployment or integration into enterprise systems (termed AI developers, for ease of reference). Through semi-structured interviews, we address three research questions:

**RQ1:** What privacy risks do AI developers believe pose the largest risk to the users, developers, and businesses?
**RQ2:** What mitigation strategies for AI privacy risks have AI developers considered or deployed?
**RQ3:** What factors do AI developers think will influence the future of privacy in AI?

Our findings show that there is little consensus among AI developers on which known AI privacy risks are the most urgent to address. However, developers focused on three high-level categories more than any other: *data management* (e.g., storage of model training data), *data memorization and leakage* (e.g., large language models unintentionally leaking train-

*These authors contributed equally to this work.

[1] https://eur-lex.europa.eu/eli/reg/2024/1689/oj

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.
*USENIX Symposium on Usable Privacy and Security (SOUPS)* 2025.
August 10–12, 2025, Seattle, WA, United States.

ing text), and *misuse via harmful applications* (e.g., deepfakes). Our discussion with developers reveals a variety of factors behind this divergence in risk ranking, including the recency of materialized privacy harms, the user or enterprise context, and ethical factors, among others. In addition, we observe that although general awareness of currently available mitigation strategies among developers is high, widespread adoption of mitigations is minimal beyond select methods, such as anonymization. Overall, developers expressed that they lack sufficient readiness to mitigate future privacy risks, citing reasons such as unpredictability in user engagement with models and lack of clarity coming from regulations. We synthesize these concerns, along with opportunities shared by developers, to understand how we as a community can work towards more privacy-conscious AI development.

We conclude that researchers should strive for greater clarity on how AI privacy risks are understood in practice, and advocate for a greater focus on building awareness of such risks, particularly to reach non-technical audiences such as end users or regulators. We also see that more effort is needed from the research community to make privacy risk mitigation strategies accessible to AI developers, so that advancements in PETs can be translated into practice. To summarize, we uncover the blueprints for an ecosystem in which all parties – AI developers, end users, researchers, and regulators – must work in parallel towards a privacy-friendly future of AI.

## 2 Related Work

**Enumerating AI privacy risks and mitigations.** In recent years, several studies have worked towards systematizing privacy risks in AI [6, 7, 8, 11, 22, 25, 29, 33, 38]. In particular, these works tackle privacy risks from a largely technical perspective, identifying where privacy vulnerabilities may occur in AI systems. As each work presents its own collection of risks, the overall field is disjoint, with no unifying taxonomy. In addition, these works all represent an academic perspective, with no external validation from AI developers (e.g., readiness to mitigate such risks), adding to the lack of practicality.

Recent works have also surveyed existing techniques to mitigate privacy risks in AI systems [7, 8, 33]. Mitigations presented in these works are also often comprehensive yet disjoint, and to the best of the authors' knowledge, none of these works seek to gain practical perspectives on the efficacy of such mitigations. We use these frameworks of risks and mitigations as the foundation for our interview guide, which features a ranking exercise of five high-level privacy risk categories and a graphic of all identified mitigation strategies.

**User and developer perspectives on AI privacy.** On the user side, studies of concerns regarding the transformative nature of AI have been recently conducted with a particular focus on privacy [16], finding that privacy ranks as the second highest concern following job loss. In this study, Kelley et al. [16] observe that nearly half of the roughly ten thousand survey respondents expect privacy to decrease over the next 10 years due to AI, citing reasons of model vulnerabilities, surveillance, unchecked data collection, and others. Such findings highlight the privacy implications of a rapidly developing landscape where AI has become increasingly ubiquitous.

On the practitioner side, previous works have focused on software developers and their privacy perceptions, for example, from the perspective of requirements engineering [27] or privacy information in coding tasks [23]. Hadar et al. [12] conduct interviews with 27 developers and highlight the complex web of cognitive, organizational, and behavioral factors that play a role in privacy decision-making. Others explore the intersection of users and developers, finding that developers often may not align on what user privacy entails [32].

Prybylo et al. [28] conclude that many software teams, regardless of region, are not entirely familiar with privacy concepts such as anonymization, and therefore they must rely on self-teaching and forums for privacy-related questions for their work. Interestingly, though, more than half of the 198 respondents in the study reported using some kind of PETs in their work, pointing to the increasing need for practical privacy protection in software systems. Boenisch et al. [2] reach similar conclusions from a survey with ML practitioners, finding that awareness of security and privacy risks is low, and that privacy protection methods are less well-known than traditional security protections. These studies, however, do not focus specifically on AI developers, and the survey format precludes in-depth discussions on *why* developers may not feel equipped to handle privacy risks. To fill this gap, we focus on developer *awareness* and *perception* of privacy risks in AI.

In studying development teams building AI products, Lee et al. [21] echo many of the same findings regarding unawareness of privacy threats in AI, leading to observations that many developers feel ill-equipped to perform privacy work. Lee et al. conclude by emphasizing the need to improve practitioner awareness of AI-specific privacy risks, and, correspondingly, to bolster their ability to address privacy harms. In our work, we maintain a focus on practitioner perspectives yet shift the target audience from various practitioner roles (designers, product managers, engineers) to exclusively technical practitioners developing, deploying, or integrating AI systems, thereby removing the bias of one business angle. While Lee et al. concentrate on processes, methods, motivators, and inhibitors in privacy work, we focus on awareness and perception of *specific* privacy risks in AI and their mitigations. We build on the focus of Lee et al. on North American companies by exploring the European perspective, where data protection regulations play a significant role alongside impending AI regulations spearheaded by the European Union.

## 3 Methodology

We conducted semi-structured interviews with 25 AI developers in Europe to investigate their perception of AI systems'

privacy risks and potential mitigations. An AI system can be any of the General-Purpose AI Systems (GPAIS), as defined in an early draft of the AI Act[2]:

> *General Purpose AI System (GPAIS): an AI system that is "able to perform generally applicable functions such as image/speech recognition, audio/video generation, pattern detection, question answering, translation, etc." [35], and is able to have multiple intended and unintended purposes. Within the scope of GPAIS, we investigate Large Language Models, Diffusion Models, and Multimodal Models.*

In the remainder of this work, we use the terms *AI* and *AI System* for simplification.

## 3.1 Identifying Privacy Threats & Mitigations

In order to guide the discussion on AI risks and mitigations, we first conducted a literature review to build an overview of existing surveys and taxonomies. The main goal of this review was to collect a set of well-known privacy risks and mitigations to be presented during the interviews, and to systematize dozens of risks under various terminologies in a manner that is practical and comprehensible to participants within the scope of a single interview. The aim of this review was not to create a fully comprehensive privacy risk and mitigation taxonomy, but rather to establish an understandable and accessible basis for investigation in our interview study.

Our search yielded 4 journal papers, 3 conference papers, 4 whitepapers, and 2 preprints. We identified these resources using a methodology proposed by Kitchenham [18], with the following search string (for titles only):

*("privacy" OR "private") AND ("risks" OR "risk" OR "harms" OR "harm" OR "threats" OR "threat" OR "concerns" OR "concern" OR "dangers" OR "danger" OR "protect*" OR "mitigat*") AND ("AI" OR "Artificial Intelligence" OR "GPAIS" OR "General Purpose AI" OR "General Purpose Artificial Intelligence" OR "Language Model" OR "LLMs" OR "Generative AI" OR "Diffusion Model" OR "Multimodal Model")*

We utilized this search string in Google Scholar, which yielded 158 results. We reduced this to 148 documents by restricting our analysis to only papers from 2015 onwards. Next, we screened each document to remove results that: (1) were not accessible via public or institutional login, or (2) were specific case studies, legal opinions, or implementation proposals. We opted for this second criterion in order to emphasize papers of the *survey* nature. Our screening isolated 6 relevant survey papers, from which we performed a forward/backward search and found 7 other relevant works, including gray literature sources [9], leading to the final set of **13** survey papers [1, 6, 7, 8, 11, 24, 29, 30, 33, 34, 37, 38, 39].

From these primary sources, our team worked collaboratively and iteratively to collate individual privacy risks into 5 high-level categories, as shown in Appendix A. Similarly, we reviewed the sources for proposed mitigations, resulting in another 5 high-level categories, as shown in Appendix B.

[2]The final version of the EU AI Act specifies a different, more narrow definition of an AI system.

## 3.2 Interview Design

Below, we introduce the structure of the conducted interviews. The full interview guide can be found in Appendix E.

**Defining AI to participants.** We began our interview with a discussion of the terminology *General-Purpose AI System*, to align all participants on the interview context.

**Background and experience with AI.** We asked with which types of models from our GPAIS definition the participant had the most experience, and inquired about the centrality of AI in their work. This helped to establish the context in which we could analyze all ensuing responses.

**Privacy risk awareness and perception (RQ1).** This section allowed the participants to speak openly about what they believe are privacy risks in AI. We then asked about the role of AI in *creating new risks* or *exacerbating existing ones*. We first let participants reflect on what they believed privacy risks in AI to be, in order to gain an unbiased perspective before presenting our risk categories. We then invited the participant to take part in an interactive *ranking exercise*, where the task was to rank our five risk categories (Table 2 of Appendix A) based on the *urgency to address*, and provide a reasoning for the chosen ranking. Each category was accompanied by a brief description. The order as given in Table 2 mirrors the ordering as presented in the ranking exercise; the order was determined by the AI lifecycle stage. To allow for free and open discussion, we did not restrict the risk context, for example, from the user, developer, or business perspective.

**Mitigation strategy awareness, perception, and readiness (RQ2).** We then switched to mitigations, once again openly asking for any mitigations with which the participant was familiar, in order to understand their general awareness of mitigation strategies. We then presented a structured list of mitigations identified in the literature (see Appendix, Table 3), inviting comments and discussion on the presented strategies. Finally, we asked whether any of the presented mitigations the participant believed to be particularly effective (or not), thus not only assessing familiarity but also perception. As with the presented risk categories, each mitigation strategy was presented alongside a brief definition, in order to facilitate a general understanding for the interview participant.

We also gauged the participant's perception of current readiness to mitigate privacy risks, both personally and generally speaking. We first asked, given current mitigation strategies, which privacy risks the participant believes to be most addressable, and which are currently still challenging or difficult to mitigate. This question was then repeated, but now asking the participant to assess their personal readiness. As a last question, we asked the participant to express what would help to increase their readiness to mitigate privacy risks.

As an additional exercise, we asked each of our participants to partake in an offline survey following the interview. In this survey, we asked the participants to reflect on their familiarity and experiences with each of the 18 presented mitigation

strategies. Specifically, we asked for a self-assessed selection of one of the following options: *never heard of it*, *heard of it*, *considered using it*, *used/using it*.

**Factors influencing the future of privacy in AI (RQ3).** Following both of the above exercises, we asked open-ended questions for the participants to reflect on the factors that might influence the future of privacy in AI.

**Pilot interviews.** After creating the final draft of the interview guide, we conducted two pilot interviews, focusing on identifying any ambiguities in the questions. As a result, no major changes were made to the guide, only minor ordering and stylistic edits. The results of the two pilot interviews (I1, I2) were included in the final analysis and results.

## 3.3 Study Protocol

**Recruitment.** We sought to garner a wide European geographic diversity of participants while maintaining our target audience of technical AI developers. We recruited via snowball sampling through two sources: personal contacts and AI developers via LinkedIn. For the latter, we searched for relevant keywords such as "AI Engineer", "ML Engineer", and "AI Consultant" and then screened promising profiles, particularly to confirm that the candidate fits our profile of a technical stakeholder working in AI (research, development, deployment, or integration). While we reached out to over 150 fitting LinkedIn profiles, we were able to establish formal email contact with 45. From these, we successfully conducted 17 interviews, in addition to 8 personal contacts or referrals.
**Interviews.** We conducted a total of 25 interviews. All interviews were held in English. Two researchers were always present in the interviews. While one researcher was asking the questions, the other would be taking notes and generating potential follow-up questions. All interviews were planned for 60 minutes, and the average length of audio recording was 53 minutes (not including introduction and post-talks). Participants were not compensated for taking part in the study. The interviews were conducted from June to August 2024.
**Demographics.** Participants in our study spanned 12 European countries[3]. Most worked directly on AI systems (14), while the remainder worked on integrating AI features (11). Years of experience working with AI ranged from 1-3 (1), 3-5 (5), 5-10 (5), 10-20 (9), and 20+ (5). Education attainment was generally high: all participants had either a bachelor's degree (1), master's degree (11), or doctorate degree (13). Regarding gender, 21 participants identified as male, 3 as female, and 1 preferred not to say (full details in Appendix Table 4).

## 3.4 Data Processing and Analysis

After transcribing each interview using Otter.ai[4], we conducted a thematic content analysis [3] on the interview data, working collaboratively with our research team. The two primary researchers began coding all transcripts by initially highlighting excerpts of interest. Then, in an iterative process, we grouped selected excerpts together into *themes*, where we unified similar sentiments under a common message. As this was done collaboratively, some themes eventually were merged or deleted. We held regular meetings to synchronize the coding process, and to discuss any possible conflicts or suggestions for edits. In the end, we agreed upon a final set of themes, supported by feedback from our wider research group. Once no new themes were discovered during analysis, the interview study was concluded [31, 36].

With the set of codes, we conducted a *synthesis* phase in which the goal was to construct a coherent narrative from the collection of codes. This was done in two steps, where the two main researchers jointly performed an initial synthesis, followed by a feedback round with all researchers, and then the implementation of feedback to form the final synthesis. This synthesized report forms the basis of our results.

## 3.5 Ethics and Limitations

Before we conducted interviews, we sent an email to potential candidates with the interview consent form, an intake survey with background information, and the full interview guide. We shared the interview guide so that the participants felt prepared to answer our questions. The consent form clearly stated the terms of the interview, most importantly that the participation is voluntary, the interviews would remain confidential, with all PII pseudonymized, and that the participant consented to the recording, transcription, and publication of the results in a pseudonymized form. The participants were given the option to refrain from answering specific questions, as well as from sharing any company details or work practices. This study protocol was approved by the IRB of the Technical University of Munich, with approval #2024-35-NM-BA.

We caution that our participants may not be representative of all European AI developers. Although we defined a clear candidate persona and performed profile screening, the selection process nevertheless followed a somewhat opportunistic approach, where a suitable candidate would immediately be invited once identified. This limitation is partially mitigated by the observed diversity of participants, from the perspective of educational background, industry domain, organization size, experience, and geographic location. We note, however, the relatively high education level of our participants, which could lead to a skewed perspective, as well as the clear geographical focus on Europe. Another limitation is the clear bias in gender distribution, with only 12% women in participation, which lies beneath the current estimate of women in AI (29%[5]). A final limitation is inherent to our chosen method of semi-structured interviews, where willingness to

[3] One participant works for a European company, but is located in India.
[4] The quality of transcriptions was manually verified after each interview.
[5] https://www.randstad.com/randstad-ai-equity/

participate and the resulting interview data are affected by social desirability and self-reporting bias. This bias could have potentially been amplified due to our decision to distribute the interview guide to each participant before the interview.

# 4 Results

We introduce our findings, supported by relevant interview quotes, indicated by the participant ID displayed in Table 4.

## 4.1 Privacy Risks

### 4.1.1 Forces behind privacy tensions with AI

To begin the discussions with our participants, we invited them to share their thoughts on whether and how privacy risks in modern AI systems are different from "general" privacy risks that can arise in any software system.

Participants acknowledged that many of the AI privacy risks in our risk categories are, in fact, not new when compared to privacy risks in other technical systems. Above all, it was seen that where there is data, there is risk, especially when there is some kind of query access to the data, e.g, model prompts (I2), as well as risks inherently stemming from black-box AI models. Especially in the European context, privacy has been a major topic since the introduction of GDPR.

However, there are a wide variety of reasons why developers perceive privacy risks to have been *amplified* by recent advances in AI. While general privacy risks may be well known in popular media and among users, we take a more nuanced view with developers, focusing on specific technical privacy risks. In addition to the general increase in awareness on the topic, we learned of several reasons perceived to be important in the rise of privacy concerns with AI.

**Increased Accessibility.** AI developers perceived that although the risks of AI have existed for some time, these risks have been amplified by an increased accessibility to AI systems, including for non-technical users. Many participants (10/25) pointed to the trend that AI has been made much more accessible, giving examples of the ease of using pay-per-use, user-friendly interfaces or APIs. Others also talked about the "human nature" of these interfaces, thus enabling users potentially to place false trust in AI systems. This was seen as potentially leading to a *"lack of basic understanding"* (I22), contrasted to previously when *"you needed a full team"* (I5) to develop, deploy, and even use AI models. Because of this shift in the way in which AI is provisioned, access to AI systems might be available to those unaware of their privacy implications, leading to uninformed usage and "reckless" behavior, something with which several participants were concerned (6/25). As one participant shared:

> *"The biggest risk is that employees and also end users become more and more uncritical [about] what happens with their own data inside these large language models (...) because they want to have this new functionality."* (I12)

Thus, increased accessibility to AI has exacerbated privacy issues, particularly due to a shift in the user base from research environments to the larger, generally non-technical public.

**Increased Capabilities.** Many participants cite the considerably more advanced technical capabilities of modern AI systems as an amplifier of privacy risks (10/25). Older AI models also posed privacy risks, *"but with the bigger models, you have more weights, more parameters, and I guess this increases the ability to memorize training data"* (I10). As a final piece to the puzzle, I23 points out that with such improvements, we can now leverage more data than before:

> *"For instance, if I went back 10 years ago, I could have easily had a crawler and still do the same thing. But what would I do with that data, there was no algorithm or there was no model. So there was no incentive to do that."* (I23)

In this, we see a progression towards increased surface for privacy risks to transpire, rooted in the capabilities enabled by better and larger AI systems trained on increasingly larger data. Better AI systems also open the door to more capable malicious users, a point emphasized by a number of participants (4/25), as well as less concern about privacy in favor of functionality: *"The benefits are so high in certain use cases that people are using it with less sensitivity"* (I7).

**Increased Use of Data.** Advancements in AI training and deployment have allowed for the usage of larger amounts of data to train better models. The risks of the sheer size of such data in comparison to classic methods are made clear:

> *"When we talk about classical ML training, you might be talking about a few thousand, tens of thousands of records. And here you're talking about so many terabytes of data. So you don't know what is actually going into there."* (I11)

An important extension emphasized by multiple AI developers (7/25) pertains not only to the size, but also to the *value* of data. Specifically, data being sent by users to AI systems is both more sensitive and more valuable *"because the data that you send to an LLM is, by definition, good training data"* (I7). Pointing to this fact, participants listed several privacy-related consequences, such as data leakage and profiling.

**Key Finding:** While many of the interviewed developers saw that some privacy risks are not new with AI, they pointed to the key *amplification* factors of increased *accessibility*, *capabilities*, and *use of data* that have been enabled by recent AI advances.

### 4.1.2 Perceptions of AI privacy risks

**Misuse via Harmful Applications.** Many participants acknowledged that misuse is dangerous to privacy, often citing

| | No. of times ranked as: | | | | | Median | Average | Standard |
|---|---|---|---|---|---|---|---|---|
| | #1 | #2 | #3 | #4 | #5 | rank | rank | deviation |
| **Misuse via Harmful Applications** | 7 | 6 | 4 | 3 | 5 | 2 | 2.72 | 1.51 |
| **Data Management** | 7 | 5 | 4 | 5 | 4 | 3 | 2.76 | 1.48 |
| **Data Memorization and Leakage** | 7 | 4 | 2 | 6 | 6 | 3 | 3.00 | 1.61 |
| **Unintended Downstream Usage** | 3 | 5 | 6 | 5 | 6 | 3 | 3.24 | 1.36 |
| **Adversarial Inference and Inversion** | 1 | 5 | 9 | 6 | 4 | 3 | 3.28 | 1.10 |

Table 1: AI developers' ranking of AI risks based on their potential harm and urgency to address (highest first).

deepfakes, fake news, misinformation, propaganda, impersonation, and others. These types of threats were seen to be of great societal importance, especially in the modern day. In addition, such risks were tied to the idea of mental health, which can be perceived to be "privacy-adjacent". In this light, the risk of misuse was seemingly the most "tangible" to many of the participants. They were able to give concrete examples of how this might affect people directly, such as through political deepfakes or family relative impersonation.

> *"I find this whole news and misinformation and fake news, deep fakes, very problematic. And it's not just us technical people that are affected by this."* (I10)

While the discussions led to many insights regarding the ethical use of AI, these were considered outside the scope of this work focused on privacy risks. In essence, most of the participants pointed to the harms potentially resulting from the misuse, rather than the risk of misuse happening itself. In contrast to other discussed risks, where the harms described by the participants were largely confined to privacy harms, the concerns expressed under this risk category often extended beyond privacy to more general societal harms (e.g., the threat of misinformation and propaganda). This observation implies that the relatively high ranking of the risk of misuse may be entangled with broader concerns with AI, including privacy.

**Data Management.** Many participants saw proper data management as a crucial first step in AI pipelines, stating that mitigation of risks here is the best mitigation for other risks such as data leakage and adversarial attacks. It was clear to many participants that the scale at which data is being collected necessitates strong data management practices. Thus, the risk lies in the fact that *"if the baseline is problematic, then you're just making a big problem for the system itself"* (I6). Another participant highlights the danger of large-scale data collection without proper processes: *"Companies nowadays, they just want to collect more and more data, just following the high-level understanding that more data is better."* (I25)

Nevertheless, others expressed that data management might not be as important as other AI privacy risks because it is not specific to AI, cannot prevent errors in training procedures, is not relevant if models are trained on short-term stored (or streamed) data that is deleted post-training, and is a non-issue if the data itself was collected improperly. It was particularly stressed by I24 that *"some of the what's captured in the model ends up being more important [than data management] because the model lives longer"* (I24), presenting an opposite perspective to data management being *"at the core"* (I13).

**Data Memorization and Leakage.** The concepts of data memorization and leakage were familiar to many participants, demonstrated by the repetition of popular examples from recent related papers. Acknowledging that memorization is an issue, participants hinted at the "garbage in, garbage out" principle, saying *"don't put PII if you don't want them outputted. Simple as that."* (I22). However, others saw the issue as more complex, expressing that it may be very difficult to determine if privacy has been compromised if we are not even sure what has been encoded in a model. Furthermore, controlling data leakage is *"really very hard to control for developers"* (I17). Still, data leakage was not seen across the board as dangerous, as it might not happen often and *"the difference between data leakage and hallucination... most of the times you can't really see that as an end user"* (I18).

**Unintended Downstream Usage (of Sensitive Prompts and User Metadata).** The notion of unintended downstream usage attracted the least comments, yet it was seen as *"specific to AI systems"* (I2). Placing sensitive information into prompts could pose privacy risks, and furthermore, carelessness in the processing of prompting can lead to important (company) information being leaked. Interestingly, also considered under this category was profiling via user metadata, as well as the uncertainty in many AI systems as to how the system responds to prompts, e.g., what the system prompt is.

**Adversarial Inference and Inversion.** Adversarial inference and inversion attacks were at times tied to data memorization and leakage risks (4/25), pointing to their intertwined nature. In some cases (5/25), participants commented on the fact that such attacks are quite hard to execute and are not that common, and therefore are not as urgent, *"not because it's less risky, but because at least nowadays, you need some skills to do that"* (I13). The same participant ranks this risk as the least urgent, saying that since most people have good intentions, we should focus on the risks that affect this majority, i.e., unintentional privacy risks. From another perspective, context is very important for adversarial privacy risks, and such risks might not be applicable in some settings:

> *"If it is a public system, probably there will be a lot of adversarial attacks (...) But at least in my experience, for more controlled environments, it's not a big issue."* (I16)

Beyond context, there was potentially less urgency placed in adversarial risks due to a lack of clarity on what the risk entails, suggesting the importance of the *tangibility* of privacy risks, which may impact their perceived urgency.

#### 4.1.3 Ranking criteria for privacy risk importance

In the ranking exercise of each interview, we asked the participant to rank the five privacy risk categories of Table 2 in

terms of importance, or rather, *urgency to address*. We observed little unified consensus on risk importance (see Table 1), albeit with a clear top-3. We also asked the participants for their reasoning behind their ranking, which led to some salient *reasoning patterns* behind perceived risk importance.

- **Cause and Effect**: If a risk is the impetus for other risks, then it is perceived to be more important. Most notably, five participants believed that improper data management may realize other privacy risks.
- **Intent**: If a risk transpires *unintentionally* (accidentally, without malicious intent), this was perceived by some to be worse than harms caused intentionally.
- **Difficulty**: Some participants stated that if a risk is technically more advanced, requiring a truly capable adversary, then this is perceived to be *less* urgent. This is tied to the fact that such attacks may be less *likely*.
- **Context is Key**: An important distinction is the enterprise (company) vs. individual (role) context, e.g., risks related to data management might be more relevant in the enterprise context. I25 asserts that *role* is important: different levels of hierarchy have different priorities.
- **Existence of Solutions**: Sometimes, participants ranked privacy risks lower when they felt that solutions already existed to help mitigate them. Most notably, the risks of data management and data memorization and leakage were sometimes cited as less urgent for this reason.
- **Recent Privacy Incidents**: In select cases, the higher ranking of a risk was accompanied by the recounting of a recent incident, showing the linkage between risks and incidents in the minds of developers.

**Key Finding:** Beyond individual insights on AI privacy risks, we learn of a number of *reasoning patterns* behind perceived risk prevalence, ranging from *technical difficulty* to *risk context*.

#### 4.1.4 How addressable are these risks?

We asked *how addressable* the presented privacy risks were. In response, some participants were confident that certain privacy risks can be addressed given current mitigation strategies, saying that it is a matter of effort. Other participants, however, were less optimistic, pointing out that certain risks are quite unpredictable and, therefore, difficult to mitigate. For example, data management was often perceived positively:

> *"Data management is the one we're most able to address, and that was also a factor in me putting it as the least urgent (...) because we do know how to do data governance."* (I24)

I25 echoed this sentiment, saying that with *"clear policies, clear guidelines"* on both the technical and management level, mitigation of risks is manageable. Similarly, it was perceived that data memorization and leakage can be well addressed if companies make a choice to implement techniques such as Differential Privacy, and *"are not in a hurry to push out models"* (I23). One opinion states that many risks are mitigatable:

> *"I think all the technological ones with a bit of effort, we are able to mitigate. Incorporating these technologies is just a matter of effort."* (I13)

Many of the risks believed to be *less* mitigatable revolve around a common theme: *that which cannot be easily predicted or controlled*, or rather, *"any risk which is fuzzy"* (I5). Explicit examples of such risks, as given by the participants, include unpredictable prompts (I2, I10) and unpredictable users (I18, I24), lack of explainability (I7, I15, I17), and difficulty in detecting what PII is (I16). A particularly interesting aspect is the lack of clarity surrounding the legal requirements with respect to AI systems (I10, I13), which will be discussed in Section 4.1.5. Also of note is the fact that data management (I7, I15) and data memorization (I24) were also listed as least addressable, further highlighting the lack of consensus.

#### 4.1.5 Technical, legal, and ethical risks

When asking generally about the relationship between the three aspects of privacy risks (technical, legal, and ethical), we received "orderings" of their relative importance.

Some participants (3/25) see the privacy problem foremost as a technical problem requiring technical solutions, as *"first comes the technologies, the trends, the potential, and then the rights"* (I14). Other participants expressed that regulations lag behind technical advancements, and therefore, the technical nature is more important. As we only interviewed technical roles, these opinions can naturally be biased.

Focusing on ethical risks of AI was perceived by some participants to be of higher importance than legal ones (4/25), as they capture *"concerns about people losing their rights more than the legal"* (I24). Nevertheless, they are all interlinked:

> *"[T]he societal impact of it is, in the end, what matters. Technology I would see more as a means to get there. And the same thing for regulation, they should all make sure it has a positive effect on society."* (I18)

Some participants agreed that responsible AI should be reinforced in developers, such that they are *"responsible for what they are doing (...) [since] it is for the common good"* (I4).

Finally, a number of participants (7/25) emphasized that ethical thinking and decision-making are more important than technical risks, because they guide the way in which technology is designed: *"It needs to start with making a good ethical decision in the beginning."* (I3). Moreover, in comparing technical risks versus societal risks, I15 expresses that it is hard to solve privacy risks, which are inherently human-oriented, with technical means, and I14 echoes this sentiment:

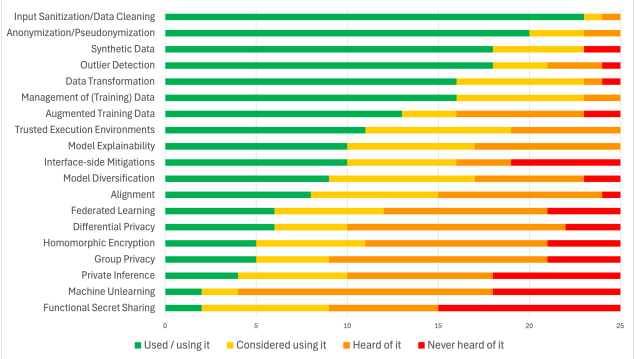


Figure 1: Participant self-ranking of their mitigation familiarity and experience in adopting these techniques.

> *"There are many factors that must be counted when solving technical problems, but they can be solved by reasoning, but (...) people are not problems you can solve."* (I14)

Regarding the potential harms of privacy risks, I23 says that technical risks may result in financial loss, *"but if you start molding a society in a certain way, there is no coming back"*. This shows that some developers have extrapolated the implications of privacy risks well beyond technical vulnerabilities.

**Key Finding:** Developers perceive the technical, legal, and societal aspects of AI privacy risks to be intertwined, with particular concerns linking privacy risks to broader ethical considerations.

#### 4.1.6 Other risks

Some participants (7/25) also suggested risks that were not directly captured under our risk categorization. The reliance of some AI providers on *third-party services*, including open-source libraries, was seen as potentially dangerous. Other participants brought up the risk of *bias* as adjacent to privacy. Finally, the risk of *not having established processes or standards* for developing or deploying AI systems was seen as a danger by some developers, as opposed to fields with standardized roles such as the medical or legal domains.

### 4.2 Privacy Risk Mitigation Strategies

We gained insights into the perceived effectiveness of mitigations as a whole, as well as individual perspectives on the strategies. To guide the discussion, we include in Figure 1 the results of our supplemental survey asking participants about their level of familiarity with each of the 19 mitigations.

#### 4.2.1 Which mitigations are most (or least) effective?

We wanted to know what AI developers thought were the most effective current mitigation strategies, whether it be the ones presented directly to them (Table 3) or otherwise. Instead of asking for comments on each mitigation strategy, we left the discussion open to the participant, and we asked instead for comments on the most familiar techniques, as well as those believed by the participant to be most or least effective. Generally speaking, there was no major consensus on which mitigation strategies are the "most effective" in mitigating privacy risks. When posed with this question, developers had varying responses, as well as "it depends" types of answers.

Many participants (7/25) stated explicitly that data sanitization techniques are the most effective in mitigating privacy risks in AI systems, aligning with the high amount of reported usage in Figure 1. These developers saw it as straightforward techniques to remove PII, while others (3/25) saw general organization measures, including proper decision-making, as the best techniques for privacy risk mitigation: *"the biggest impact you can make is reduce the number of bad decisions that someone in an organization makes"* (I3).

From a technical standpoint, smaller subsets of participants each supported different mitigation strategies as being very effective, from Differential Privacy (I1, I2), to Synthetic Data (I1, I5, I19, I20), to Federated Learning (I21). Additionally, some named techniques that we did not present, such as strong data policies (I16) or active monitoring systems (I18). Nevertheless, I13 points out that no single mitigation strategy will be a silver bullet: *"in general, I think you need to use a combination of those. And it depends on the use case."*

Notably, developers had fewer opinions on what the "least effective" mitigations might be, often not directly answering the question. Some isolated opinions included explainability (contradicting research, I21), organizational (hard to implement, I21), and Trusted Execution Environments (not practical at scale, I24). This could show that while developers may be generally familiar with mitigations, they are not able to judge their efficacy easily, or rather, to explain the downsides.

#### 4.2.2 Perceptions of individual mitigation strategies

We present insights gained from participants on each of the 19 mitigation strategies, delineated by the five mitigation *categories* (Table 3). For each mitigation, we list how many participants provided substantial comments on the technique.

**Organizational.** Management of Training Data (4/25) was stressed as an important organizational process, especially with the GDPR. However, it was noted that there exists a gap between research and practice, with a lack of good data management practices from research.

While Alignment (7/25) was seen by one to be quite important, most of the feedback focused on limitations, as well as it often not being a priority for organizations. The challenges of over-alignment and proper evaluation were also emphasized.

In a similar way, the discussion on Trusted Execution Environments (3/25) focused on whether they can be implemented in a cost-effective way, and the dependence on organizational infrastructure to determine when hardware might be secure.

**Data Preprocessing.** While the strategy of Input Sanitization / Data Cleaning (11/25) attracted a great deal of attention, there was a split opinion on its effectiveness. While some championed it as "essential" and "key" (7/25), others emphasized that truly sanitizing data is more difficult than it seems, especially when considering the involvement of third parties or when dealing with highly unstructured data formats.

A similar sentiment was observed in regard to Anonymization / Pseudonymization (6/25). Some participants used the language that such techniques are "common" and should be used "by default". A number of participants (4/25), though, emphasized limiting factors in true anonymization, such as choosing the right parameters, high cost, and incorrect implementations leading to faulty anonymization.

Outlier Detection (5/25) was perceived as directly related to privacy, and several participants regarded it as a support tool for other mitigations such as Differential Privacy.

**Data Augmentation.** The number of supporters for Differential Privacy (9/25) was significant, with several alluding to it being the *"gold standard"* (I10). These participants pointed to experiences in implementing it for AI applications, and I2 credits the maturity and availability of open-source resources, including a wide body of academic literature, as being helpful in facilitating adoption. At the same time, some developers mentioned challenges, such as in measuring privacy (I13), finding an acceptable privacy-utility trade-off (I10), and implementing it in already noisy domains (I7).

The notion of Group Privacy (3/25) attracted little attention, with the majority of the participants not being familiar with concepts such as k-anonymity. Interestingly, it was noted that such techniques are rarely used in production (I1), and two candidates (I10, I13) made a direct comparison to Differential Privacy, explaining its merits compared to Group Privacy.

Data Transformation methods (3/25) were met with skepticism, with participants pointing out that *"you can partly obfuscate the data, but not fully"* (I22), and it is unclear whether these actually fulfill the purpose of privacy protection (I7).

Although multiple participants commented on Homomorphic Encryption (7/25), the perceptions of this technology were overwhelmingly negative, with several (4/25) concerns about its unclear applicability to AI models. In addition, the high computational cost and slowness of Homomorphic Encryption were emphasized by I1, and I22 indicated a lack of standardization, making practical applicability difficult.

Functional Secret Sharing (1/25) received the least feedback out of *any* mitigation strategy, and we noticed that the terminology itself was quite unclear to many participants. Only one participant (I10) was able to make a connection to a particular technique, i.e., Secure Multi-Party Computation.

In contrast, Synthetic Data (15/25) was touted for being "commonly" and "widely" used (I8, I17, I21) for training purposes, as well as helpful in privacy-critical scenarios, such as with medical data (I20, I22), or particularly in combination with Differential Privacy (I10, I13, I25). Some, however, were not so convinced by the privacy-preserving capabilities of synthetic data since it has no *"provable guarantees"* (I10) and is generated *"from the original data, which doesn't really solve the privacy problem"* (I7). Furthermore, I20 and I21 note that synthetic data may still leak sensitive characteristics of the original data. Common challenges were also cited, such as lack of variability, difficulty in controlling synthetic data generation, and quality of models trained on synthetic data.

**Private Training.** Although several participants commented on Augmented Training Data (4/25), this was often in the context of model training, and no connections were made to privacy specifically, such as with adversarial training.

Similarly, Model Diversification (5/25) was familiar to some for model robustness and improving performance, but not for privacy: *"I've never thought about it in terms of privacy improving."* (I5). Only I24 referenced this technique as a privacy risk mitigation, although with skepticism.

Federated Learning (9/25) was a familiar concept, but few had practical experiences. Multiple participants (5/25) pointed out challenges in practical applicability, particularly with defining the threat model. I7 advocated for Federated Learning as an efficiency improvement rather than a privacy risk mitigation strategy, while I21 supported its use to comply with European regulations, especially with medical data.

**Post Hoc Privatization.** Model Explainability (7/25) was perceived as promising, yet still remaining very much in research. Explainable models were tied to privacy in the way that they increase transparency on *"what type of information is actually stored"* (I23). Nevertheless, I7 and I16 doubted the feasibility of explainability given current techniques.

Interface-side Mitigations (7/25) were largely interpreted to be guardrails implemented in AI systems, seen to be very relevant by the participants familiar with them. The main challenge, though, becomes how to develop guardrails in unison with rapidly evolving technologies: *"This technology is growing so fast that guardrails are not catching up yet."* (I8). Deploying guardrails at scale was named as a challenge (I22), and the main limitation remains that *"no matter how many guardrails you make, there will always be cases where you will have issues, where you won't be able to guard well"* (I23).

While Machine Unlearning (7/25) was agreed upon to be interesting and potentially useful, several participants (3/25) noted its complexity to implement and its lack of reliability in its current form. Furthermore, I1 and I7 both provide the insight that if the need for unlearning arises, *"the damage is already done"* (I1), making it a *"last resort"* option (I7).

Two participants mentioned Private Inference (2/25) as a *"necessity in all deployed systems"* (I7) and *"ironclad solution by design"* (I22); few insights beyond this were gained.

**Key Finding:** AI privacy risk mitigation strategies received varying degrees of attention in the interviews, with popular choices such as *Differential Privacy* and *Synthetic Data*, alongside lesser-known mitigations such as *Functional Secret Sharing*.

#### 4.2.3 Other mitigation strategies

Our participants brought new mitigation strategies to light:

- *Organizational Measures*: Multiple participants (6/25) mentioned organizational measures as effective, despite serving in a technical role. Examples include trainings, guidelines, and strong data governance.
- *Output Sanitization*: I11 mentions output sanitization, particularly in the form of hallucination control and ensuring that no PII is contained in the outputs of models.
- *Monitoring*: I18 and I19 were proponents of continuous monitoring, to be used in conjunction with guardrails.
- *Model-specific Mitigations*: Two participants mentioned model-specific mitigations, such as privacy-conscious LLMs or smaller, specialized models that reduce the risks brought about by large, data-hungry models.
- *Auditing*: I.e., stricter audits for AI providers: *"We have audits for finance, right? (...) We don't have audits for AI (...) So we need to move towards where we actually audit all these companies for risk and safety."* (I23).

### 4.3 The Future of Privacy in AI

Finally, we explored with participants what stakeholders should be involved in safeguarding AI privacy as well as socio-technical opportunities for improving AI privacy.

#### 4.3.1 Key stakeholders for safeguarding AI privacy

We asked the participants which entity is ultimately responsible for safeguarding privacy in AI.

- **The Government and Regulatory Bodies (9/25)**: these authorities can best protect privacy through laws and regulations. In addition to "restricting", it was suggested that governments should also *"open opportunities"* (I21), such as to learn about the responsible use of AI.
- **AI Providers (8/25)** as the organizations *"holding the key"* (I16) to AI, these entities have a responsibility to increase the privacy, transparency, and trust in their models, as well as to give users better control of their privacy.
- **End Users (4/25)**: some developers suggested that the responsible use of AI is *"ultimately, the responsibility of the user"* (I17), and the pressure to preserve privacy must also come from end users, since *"if the customer is going to pay only for a privacy-friendly product, then there is no other choice for the companies."* (I25).
- **Academia (4/25)**: the role of universities and other educational institutions was seen as important to *"reach the masses (...) to explain things in a balanced and objective way"* (I14), as well as to continue advancing privacy research efforts, with a focus on practical applications.

#### 4.3.2 Opportunities for improving AI privacy

Some participants were not very optimistic about addressing or improving the privacy concerns around AI:

> *"We are not future-ready. I don't know if anyone is future-ready, and or even the short future-ready. We are just trying to save the day."* (I20)

However, many participants shared ideas on what would help reinforce privacy-conscious, responsible AI in the future. They also expressed clear needs that would help them feel better equipped to mitigate AI privacy risks going forward.

**Regulations, Standards, and Guidelines (14/25).** Many participants hoped for a future with higher regulation of training data, more transparency, and increased safety checks when AI models are released. Especially in enterprise contexts, legal compliance was seen as the *"highest concern"* (I11). Furthermore, some participants criticized current regulations for not being well-equipped to handle AI cases, and that privacy rights might not be protected well (I24).

Among the main expectations regarding privacy in AI is the development of clearer, stricter regulations – an "AI equivalent of GDPR" (I5). Similarly, standardization would be an advancement in making it *"an official profession to be able to work with these models"* (I18). It was emphasized, however, that it is important that regulations do *"not hinder innovation itself, but hinder malicious use of it"* (I2), and that *"bad guys do not follow regulations, but they are free to use AI"* (I17), reinforcing the belief that regardless of the regulations, malicious actors will find a way to circumnavigate safeguards:

> *"There will always be bad actors. If you try to regulate them, it won't work. (...) Your best option is to regulate big companies. That's the only thing you can do."* (I23)

In terms of supporting AI developers in mitigating AI privacy risks in their work, a wish for official regulatory guidance was expressed, since as I7 put it, *"what's missing is clear direction and clear advice, what is okay, what is not okay."* (I7)

**Tools and Testing Systems (11/25).** Many participants pointed to the lack of tools available to detect privacy risks, as well as to implement and evaluate privacy mitigations. Privacy threat modeling was explicitly mentioned twice (I13, I24). Notably, it was mentioned that misuse is important to address, but monitoring and detecting it is difficult.

**Education and Awareness (10/25).** The need for better education on privacy in AI was advocated for, particularly among end users, developers, and legal professionals. Many participants supported the idea that once awareness of privacy risks is made clear, then more informed action can be taken.

This was particularly tied to the discussion of ethical aspects of AI privacy risks, which was seen to be exacerbated by the current *"wild west"* (I19) in AI. In this, developers highlighted the point that teaching responsible AI is paramount.

*"People who deal with it every day need to explain it in simple terms, not necessarily technical terms. (...) You have to lower the hurdle."* (I14)

**Transparency (6/25).** Many participants called for higher levels of transparency in the AI field, particularly from the creators of AI systems. This could entail clear documentation on the data processing, steps taken to train models, as well as interface-side privacy disclaimers in simple language.

**Interdisciplinary Collaboration (5/25).** Other participants supported collaboration between practitioners of varying backgrounds, both legal and technical, to help build up competencies and reduce miscommunication. This would help to *"get support from different teams or different other groups which work with specific things in AI such as privacy"* (I6). Another example given was the sharing of experience reports.

*"I think lots of bad decisions that happen along the way are due to some form of miscommunication or things that happen between people with different backgrounds who are maybe not able to fully understand each other."* (I3)

**Key Finding:** The interviewed developers indicated a number of key stakeholders in AI privacy, particularly the government and AI providers. Together, these stakeholders are tasked with providing clearer guidelines and more usable mitigation tools, with the goal of increased awareness, transparency, and collaboration.

## 5 Discussion

We reflect on our findings, re-examining the lessons of our results in the context of our original research questions.

### RQ1: Navigating the Sea of Privacy Risks

Our discussions with 25 AI developers reveal that there is surprisingly little agreement within the technical community over which AI privacy risks are the most pressing. In fact, out of the 25 ranking exercise responses, we observed 23 unique orderings, supporting that there is indeed a clear lack of consensus. Interestingly, the risks of data management, data memorization, and misuse ranked amongst the top 3, each with seven #1 votes, yet they also appeared as last in the ranking a number of times (4, 4, and 5, respectively). This implies that although certain risks are generally perceived as more urgent than others, this is not uniformly the case.

The variety in responses raises potential concerns. The notion that ethical considerations currently trail behind the proliferation of AI was expressed often, adding to a seeming confusion over what risks are most important. Additionally, we observe little cohesion between rankings even when isolating attributes such as educational level, experience, or company size. This points to an industry-wide lack of unification on privacy risk prevalence, displaying that companies may scope privacy risks quite differently or, rather, may not be scoping privacy risks at all. Although this is natural given the variety of business contexts and the relative nascence of AI privacy risk knowledge, these findings call for a more comprehensive understanding among developers and companies, specifically on how to prioritize privacy risk assessments and how one may differentiate between different risk classes.

We learn that assessing privacy risks is currently very personal – reasons behind a specific ranking are often attributed to an internal hierarchy influenced by some self-guided reasoning pattern, which can be especially influenced by a developer's specific work context. This resonates with previous findings on general privacy work among developers [21, 26], and suggests a lack of clear guidelines from companies or authorities, which forces developers to navigate the sea of privacy risks on their own. Moreover, we find that in lieu of clear guidelines, developers often fall back to internal reasoning patterns on how to assess and scope privacy risks. However, such reasoning may become clouded by confounding factors, such as the uncertainty between privacy risks and model hallucinations (I18). While this suggests the need for more guidance from researchers and authorities, it also points to one potential benefit: allowing developers to explore novel mitigation techniques and build up best practices.

Given this current state, regulators looking to the industry for best practices on AI privacy risk mitigation may be left without clear answers. Thus, attempts to regulate AI without validated industry standards and accepted mitigation strategies may lead to premature regulations that force AI developers into non-optimal solutions, thereby creating more friction.

### RQ2: Mitigations Abound, but Adoption is Lagging

In discussing current mitigation strategies for AI privacy risks, we learned that general awareness of the 19 presented techniques is high, and beyond this, developers shared a number of other strategies they believed to be mitigations. Certain mitigation strategies received considerable attention, such as Synthetic Data and Differential Privacy, while others garnered nearly no comments (e.g., Functional Secret Sharing). Participants also mentioned mitigations not popularly covered in the literature, such as active monitoring for quality assurance. These findings show a development in the conclusions drawn by Boenisch et al. [2], suggesting that developer awareness of privacy mitigations has increased in the past few years.

We observed that some strategies received quite a deal of attention, such as Alignment, Synthetic Data, and Differential Privacy. This attention, one can argue, echoes current research and media attraction to such topics, implying that mitigation awareness may coincide closely with prevalence in academic literature or popular media. This points to the reach of academic publications, as well as exhibits the degree to which developers may relate to popular media topics.

Much of the commentary on mitigations also sheds light on the "go-tos" in the industry, such as the blanket terms of

anonymization and data cleaning, which were largely met with approval and reported to be used often. In contrast, many of the "advanced" techniques, including those in Data Augmentation, Private Training, and Post Hoc Privatization, were often perceived to have significant shortcomings, which hinder practical adoption, such as high costs, slow speeds, and/or loss of utility or functionality, a finding that echoes previous work [19, 20]. Thus, a clear divide was formed between the widely accepted, yet less advanced mitigations and those that show high promise but low widespread practicality.

The implications of this divide become significant when addressing the practical adoptability of advanced PETs such as Differential Privacy, where it will become important to distinguish clear practical limitations and resistance to change from more familiar "anonymization techniques" [10]. Similarly, the case of Synthetic Data teaches us that wide application does not equate to sound privacy protection, as the promise of this technique was met with equally as many concerns over its ability to provide actual privacy guarantees. This comes into contrast with other advanced techniques, which achieved little recognition from any of the interviewed developers. Furthermore, the question of familiarity may be influenced by the potential *use cases* of a mitigation strategy, where a lesser-known strategy may be linked to its limited applicability to specific or niche AI use cases, e.g., with machine unlearning.

As such, we see it as crucial to raise awareness of the wide range of privacy risk mitigation strategies to more developers. Beyond this, it is important that the burden of learning about new mitigations is not placed solely on individual developers, but rather the discussion should involve more people, whether through the establishment of community norms or fostering interdisciplinary collaboration. In addition, regulators and other supervisory authorities, such as standards organizations, could work to provide guidance and recommendations on the use of state-of-the-art mitigations to support privacy-conscious AI development. These steps can help developers reach a better consensus on how to reason about privacy risks in AI.

#### RQ3: Becoming Future-ready

In the interviews, we received various recommendations on how to advance understanding of privacy risks and adoption of mitigation strategies. These include clear calls to action for key stakeholders involved in the AI pipeline, from researchers and developers, to enterprise leadership and end users.

Many of the interviewed developers believed that AI providers, or more generally those with the resources to develop, train, and deploy large-scale AI systems, are the ones that "hold the keys" to developing more privacy-conscious AI. This, however, will not be done without a clear demand from enterprise customers and end users to make privacy a priority, thereby incentivizing privacy to ensure the continued success of an AI product or service; this, however, was seen as the "happy path", or more optimistic view of the future.

As noted by participants, privacy-conscious demand from users may be hindered by their focus on functionality over privacy, and a greater privacy awareness among users will not come without intervention. Similarly, in the enterprise context, a lack of awareness of the privacy risks when entrusting company data in the hands of AI providers may pose a risk to further adoption. Raising awareness thus becomes the task of researchers and educators, where university curricula and other programs should prioritize privacy and other ethical considerations in AI education, similar to calls emphasizing privacy education when teaching software development [28].

The final point that reached consensus is that governments and other supervisory authorities can directly influence ethical AI practices within companies by enforcing strict regulations and providing official guidelines that promote the use of advanced PETs. This influential role is emphasized by previous work [19], requiring collaboration with researchers for guidance on technological solutions for privacy risk mitigation.

In this, we uncover the blueprints for an ecosystem in which all parties – AI developers, end users, researchers, and regulators – must actively participate, where privacy is simultaneously incentivized and enforced. Only with these forces working in parallel will there be an advancement towards a future where AI privacy risks are well-defined, mitigations are not only promising but also offer an attractive privacy-utility trade-off, and developers and users alike feel equipped to contribute to a more privacy-friendly future of AI.

## 6 Conclusion

We interviewed 25 AI developers to understand their perceptions of known privacy risks in AI, as well as their familiarity with current strategies to mitigate these risks. In exploring the perceived importance of these risks, we learned that there exists a multitude of reasoning patterns as to why one risk may be more urgent than another. We also observed that while general familiarity with mitigation strategies is strong, few developers in our study could point to concrete cases in which more advanced privacy techniques (e.g., Differential Privacy) are currently employed. Developers shared that building more privacy-conscious AI systems necessitates the involvement of large AI providers, end users, researchers, educational institutions, and governmental bodies. As a whole, our findings highlight the need for better awareness of privacy risks, improved privacy-enhancing tooling in AI, and greater cooperation in order to improve AI privacy for users and enterprises.

## Acknowledgments

This research is based on work that has been partially funded by Google. Any opinions, findings, conclusions, or recommendations expressed in this work are those of the authors and do not necessarily reflect the views of Google.

## References


[1] Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, CCS '18, page 2154–2156, New York, NY, USA, 2018. Association for Computing Machinery.

[2] Franziska Boenisch, Verena Battis, Nicolas Buchmann, and Maija Poikela. "I never thought about securing my machine learning systems": A study of security and privacy awareness of machine learning practitioners. In *Proceedings of Mensch Und Computer 2021*, MuC '21, page 520–546, New York, NY, USA, 2021. Association for Computing Machinery.

[3] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101, 2006.

[4] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 267–284, Santa Clara, CA, August 2019. USENIX Association.

[5] Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650. USENIX Association, August 2021.

[6] Norwegian Consumer Council. Ghost in the Machine: Addressing the Consumer Harms of Generative AI. *Norwegian Consumer Council, June*, 2023.

[7] James Curzon, Tracy Ann Kosa, Rajen Akalu, and Khalil El-Khatib. Privacy and artificial intelligence. *IEEE Transactions on Artificial Intelligence*, 2(2):96–108, 2021.

[8] Federal Office for Information Security. Generative AI models - opportunities and risks for industry and authorities. 2024.

[9] Vahid Garousi, Michael Felderer, and Mika V Mäntylä. Guidelines for including grey literature and conducting multivocal literature reviews in software engineering. *Information and software technology*, 106:101–121, 2019.

[10] Gonzalo M Garrido, Xiaoyuan Liu, Florian Matthes, and Dawn Song. Lessons learned: Surveying the practicality of differential privacy in the industry. *Proceedings on Privacy Enhancing Technologies*, 2:151–170, 2023.

[11] Abenezer Golda, Kidus Mekonen, Amit Pandey, Anushka Singh, Vikas Hassija, Vinay Chamola, and Biplab Sikdar. Privacy and security concerns in generative AI: A comprehensive survey. *IEEE Access*, 2024.

[12] Irit Hadar, Tomer Hasson, Oshrat Ayalon, Eran Toch, Michael Birnhack, Sofia Sherman, and Arod Balissa. Privacy by designers: software developers' privacy mindset. *Empirical Software Engineering*, 23:259–289, 2018.

[13] Priyank Jain, Manasi Gyanchandani, and Nilay Khare. Big data privacy: a technological perspective and review. *Journal of Big Data*, 3:1–25, 2016.

[14] Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. Knowledge unlearning for mitigating privacy risks in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14389–14408, Toronto, Canada, July 2023. Association for Computational Linguistics.

[15] Bargav Jayaraman and David Evans. Are attribute inference attacks just imputation? In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pages 1569–1582, 2022.

[16] Patrick Gage Kelley, Celestina Cornejo, Lisa Hayes, Ellie Shuo Jin, Aaron Sedley, Kurt Thomas, Yongwei Yang, and Allison Woodruff. "There will be less privacy, of course": How and why people in 10 countries expect AI will affect privacy in the future. In *Nineteenth Symposium on Usable Privacy and Security (SOUPS 2023)*, pages 579–603, Anaheim, CA, August 2023. USENIX Association.

[17] Sunder Ali Khowaja, Parus Khuwaja, Kapal Dev, Weizheng Wang, and Lewis Nkenyereye. ChatGPT needs SPADE (sustainability, privacy, digital divide, and ethics) evaluation: A review. *Cognitive Computation*, pages 1–23, 2024.

[18] Barbara Ann Kitchenham, David Budgen, and Pearl Brereton. *Evidence-Based Software Engineering and Systematic Reviews*, volume 4. CRC press, 2015.

[19] Alexandra Klymenko, Stephen Meisenbacher, Luca Favaro, and Florian Matthes. On the integration of privacy-enhancing technologies in the process of software engineering. In *Proceedings of the 26th International Conference on Enterprise Information Systems - Volume 2: ICEIS*, pages 41–52. INSTICC, SciTePress, 2024.

[20] Patrick Kühtreiber, Viktoriya Pak, and Delphine Reinhardt. "A method like this would be overkill": Devel-

opers' perceived issues with privacy-preserving computation methods. In *2023 IEEE 22nd International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, pages 1041–1048, 2023.

[21] Hao-Ping (Hank) Lee, Lan Gao, Stephanie Yang, Jodi Forlizzi, and Sauvik Das. "I don't know if we're doing good. I don't know if we're doing bad": Investigating how practitioners scope, motivate, and conduct privacy work when developing AI products. In *33rd USENIX Security Symposium (USENIX Security 24)*, Philadelphia, PA, August 2024. USENIX Association.

[22] Hao-Ping (Hank) Lee, Yu-Ju Yang, Thomas Serban Von Davier, Jodi Forlizzi, and Sauvik Das. Deepfakes, phrenology, surveillance, and more! a taxonomy of ai privacy risks. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery.

[23] Xinyao Ma, Ambarish Gurjar, Anesu Chaora, and L Jean Camp. Programmer's perception of sensitive information in code. 2024.

[24] Gary McGraw, Harold Figueroa, Katie McMahon, and Richie Bonett. An architectural risk analysis of large language models: Applied machine learning security. *Berryville Inst. Mach. Learn.(BIML), Berryville, VA, USA, Tech. Rep*, 2024.

[25] Stephen Meisenbacher, Alexandra Klymenko, Patrick Gage Kelley, Sai Teja Peddinti, Kurt Thomas, and Florian Matthes. Privacy risks of general-purpose ai systems: A foundation for investigating practitioner perspectives. *arXiv preprint arXiv:2407.02027*, 2024.

[26] Leysan Nurgalieva, Alisa Frik, and Gavin Doherty. A narrative review of factors affecting the implementation of privacy and security practices in software development. *ACM Comput. Surv.*, 55(14s), July 2023.

[27] Mariana Peixoto, Dayse Ferreira, Mateus Cavalcanti, Carla Silva, Jéssyka Vilela, João Araújo, and Tony Gorschek. On understanding how developers perceive and interpret privacy requirements research preview. In *Requirements Engineering: Foundation for Software Quality*, pages 116–123, Cham, 2020. Springer International Publishing.

[28] Maxwell Prybylo, Sara Haghighi, Sai Teja Peddinti, and Sepideh Ghanavati. Evaluating privacy perceptions, experience, and behavior of software development teams. In *Twentieth Symposium on Usable Privacy and Security (SOUPS 2024)*, pages 101–120, Philadelphia, PA, August 2024. USENIX Association.

[29] Md Mostafizur Rahman, Aiasha Siddika Arshi, Md Mehedi Hasan, Sumayia Farzana Mishu, Hossain Shahriar, and Fan Wu. Security risk and attacks in AI: A survey of security and privacy. In *2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)*, pages 1834–1839. IEEE, 2023.

[30] Maria Rigaki and Sebastian Garcia. A survey of privacy attacks in machine learning. *ACM Computing Surveys*, 56(4):1–34, 2023.

[31] Benjamin Saunders, Julius Sim, Tom Kingstone, Shula Baker, Jackie Waterfield, Bernadette Bartlam, Heather Burroughs, and Clare Jinks. Saturation in qualitative research: exploring its conceptualization and operationalization. *Quality & quantity*, 52:1893–1907, 2018.

[32] Awanthika R. Senarath and Nalin Asanka Gamagedara Arachchilage. Understanding user privacy expectations: A software developer's perspective. *Telematics and Informatics*, 35(7):1845–1862, 2018.

[33] Sakib Shahriar, Sonal Allana, Mehdi Hazrati Fard, and Rozita Dara. A survey of privacy risks and mitigation strategies in the artificial intelligence life cycle. *IEEE Access*, 2023.

[34] Victoria Smith, Ali Shahin Shamsabadi, Carolyn Ashurst, and Adrian Weller. Identifying and mitigating privacy risks stemming from language models: A survey. *arXiv preprint arXiv:2310.01424*, 2023.

[35] European Union. Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. 2021. https://data.consilium.europa.eu/doc/document/ST-14278-2021-INIT/en/pdf.

[36] Cathy Urquhart. Grounded theory for qualitative research: A practical guide. 2013.

[37] Apostol Vassilev, Alina Oprea, Alie Fordyce, and Hyrum Anderson. Adversarial machine learning: A taxonomy and terminology of attacks and mitigations. Technical report, National Institute of Standards and Technology, 2024. https://csrc.nist.gov/pubs/ai/100/2/e2023/final?ref=studygrc.com.

[38] Biwei Yan, Kun Li, Minghui Xu, Yueyan Dong, Yue Zhang, Zhaochun Ren, and Xiuzheng Cheng. On protecting the data privacy of large language models (LLMs): A survey. *arXiv preprint arXiv:2403.05156*, 2024.

[39] Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, 4(2):100211, 2024.

# A AI Privacy Risks

| |
|---|
| **(R1) Data Management** [33]<br>As the training and fine-tuning of AI requires massive amounts of data to be accurate and robust, large amounts of data must be collected and stored, posing privacy risks to those from whom the data originated. |
| **(R2) Data Memorization and Leakage** [8, 29, 33, 34, 37]<br>Larger models have the tendency to "memorize" some of their training data, particularly the data that is more unique or rare occurring. Data leakage occurs when models unintentionally expose their memorized data, and in the case of prompting, malicious users may be able to reconstruct or extract data with careful prompting. |
| **(R3) Unintended Downstream Usage (of Sensitive Prompts and User Metadata)** [8, 38]<br>User inputs and queries may be used for originally unintended purposes, for example, building up user profiles based upon contextual information, or sensitive company data being captured and used downstream for model fine-tuning. |
| **(R4) Adversarial Inference and Inversion** [8, 15, 29, 30, 33, 34, 37, 38, 39]<br>A class of attacks against AI referred to as inference attacks include an adversary trying to distinguish whether a particular instance (member) was in the training data, attempting to recover attributes of the training data, making broader inferences about the underlying data, or reverse-engineering model-specific parameters. |
| **(R5) Misuse via Harmful Applications** [6, 11, 24, 39]<br>The predictive and generative power of AI opens the door for misuse in downstream applications, where malicious users leverage AI for nefarious purposes (e.g., deepfakes). |

Table 2: AI privacy risks identified through the literature review, as introduced to the participants.

# B Mitigations for AI Privacy Risks

| **Organizational** |
|---|
| **(M1.1) Management of (Training) Data** [8, 11]<br>Mitigating privacy risks begins with the responsible handling of training data, for example in trusted and secured data warehouses, as well as proper data governance structures. |
| **(M1.2) Alignment** [8]<br>An important organizational measure involves the aligning of AI to human values, a step that can help to ensure trustworthiness and reduce the risk of adversaries compromising models. |
| **(M1.3) Trusted Execution Environments** [7, 38]<br>Secured hardware environments where data processed within cannot be read or tampered with by outside parties or code. |
| **Data Preprocessing** |
| **(M2.1) Input Sanitization/Data Cleaning** [7, 8, 29, 34, 38, 39]<br>When preparing data for model training, a wise preprocessing step includes data cleaning, in which explicit sensitive information is removed, such as phone numbers. |
| **(M2.2) Anonymization/Pseudonymization** [8]<br>Similarly, explicit personally identifiable information (PII) can be removed via anonymization or pseudonymization techniques. |
| **(M2.3) Outlier Detection** [24, 33, 37, 39]<br>Outlier detection methods aim to identify and detect outliers in the training data, as these datapoints might be more readily memorized. |
| **Data Augmentation** |
| **(M3.1) Differential Privacy** [7, 8, 11, 34, 37, 38, 39]<br>Differential Privacy is a mathematically grounded notion of privacy that usually involves adding random noise to query outputs to inject plausible deniability into computations on potentially private data (attributes). |
| **(M3.2) Group Privacy** [6, 7, 37]<br>The notions of k-anonymity, l-diversity, and t-closeness concern themselves with group privacy, where the goal is to provide a certain level of indistinguishability between members of a dataset. |
| **(M3.3) Data Transformation** [7, 33]<br>Storing data in "transformed" forms, such as in embedding format or in reduced dimensions, may help to obfuscate raw sensitive data. |
| **(M3.4) Homomorphic Encryption** [11, 38]<br>Advanced privacy-preserving techniques based on encryption can be leveraged such that computation on raw user data is not necessary. |
| **(M3.5) Functional Secret Sharing** [38]<br>In some privacy-preserving mechanisms, user data is "split" among a pool of users, such that no single data point must be shared in full. |
| **(M3.6) Synthetic Data** [7]<br>Rather than use the original training data, some training strategies may opt to use synthetically generated data, which ideally shares a similar distribution of the original data. |
| **Private Training** |
| **(M4.1) Augmented Training Data** [7, 8, 29]<br>Adding augmented samples to training data can help to improve model robustness, thereby reducing the risk of privacy breaches. |
| **(M4.2) Model Diversification** [29]<br>Some training procedures may opt to train a variety of models (or parameters) to improve robustness in decision-making and to mitigate any vulnerabilities of one single model. |
| **(M4.3) Federated Learning** [11, 29, 38]<br>The training of models is performed locally in a distributed fashion, where only model updates are shared with a central aggregator. |
| **Post Hoc Privatization** |
| **(M5.1) Model Explainability** [8, 11, 29, 33]<br>Focusing on the development of explainable models can detect model vulnerabilities, as well as verify the integrity of a model. |
| **(M5.2) Interface-side Mitigations** [8, 34, 37]<br>Safeguarding the interface between users and models includes detecting suspicious queries, limiting the number of queries, and building guardrails. |
| **(M5.3) Machine Unlearning** [8, 34, 38]<br>The ability to remove a single user's contributions to a training data set upon request, without the need for complete retraining. |
| **(M5.4) Private Inference** [8]<br>Augmenting model inference (computation on unseen data) with privacy-preserving solutions such as Differential Privacy or Homomorphic Encryption adds an extra layer of privacy protection for users inputting potentially sensitive data. |

Table 3: Mitigations for AI privacy risks identified through the literature review, as introduced to the participants.

## C Recruitment Message

The following message was used to contact and recruit the interview participants described in this work. Where applicable, we personalized this message to tailor to a specific interview candidate.

---

Subject: **Privacy Risks in General-Purpose AI**

Dear [TITLE][NAME],

I hope this message finds you well.

[One-line introduction of the message author.]

We are currently conducting a study investigating the practical perspectives on Privacy Risks in General-Purpose AI Systems. For this, we are searching for technical experts in the field of AI to share their opinions on the topic.

Based on your impressive professional background and your current role as [ROLE], I believe you could bring very valuable insights to our study and would like to invite you to participate in a 1-hour Zoom interview at a time that is convenient for you.

I would really appreciate it if you could participate, and I look forward to hearing from you!

Best regards,

[Email-style signature.]

---

## D Interview Participant Demographics

| ID | Position | Education | Industry Domain | Org. size | Country | Exp. (years) | AI-centered |
|---|---|---|---|---|---|---|---|
| I1 | Senior Researcher | Doctorate | Information Technology | Large | Germany | 3 - 5 | No |
| I2 | Senior Researcher | Doctorate | Information Technology | Large | Germany | 5 - 10 | No |
| I3 | Generative AI Engineer | Doctorate | Information Technology | Small | Sweden | 3 - 5 | Yes |
| I4 | Senior AI Developer | Doctorate | Information Technology | Small | Spain | 5 - 10 | Yes |
| I5 | Software Architect | Doctorate | Information Technology | Large | Germany | 10 - 20 | No |
| I6 | AI Engineer | Master's | Information Technology | Medium | Germany | 1 - 3 | Yes |
| I7 | CEO | Doctorate | Information Technology | Small | Germany | 10 - 20 | Yes |
| I8 | MLOps Engineer | Master's | Information Technology | Medium | Ukraine | 5 - 10 | No |
| I9 | Managing Director | Master's | Finance | Large | Germany | 20+ | No |
| I10 | Senior Researcher | Doctorate | Information Technology | Large | Germany | 10 - 20 | No |
| I11 | Head of AI & Advanced Analytics | Master's | High tech, Telecom | Large | India | 10 - 20 | Yes |
| I12 | Head of the Innovation Department | Bachelor's | Finance | Large | Germany | 20+ | No |
| I13 | Senior Privacy Engineer | Master's | Advertising | Small | Italy | 3 - 5 | Yes |
| I14 | Deputy Head of Department | Master's | Information Technology | Large | Germany | 10 - 20 | No |
| I15 | Principal AI Consultant | Doctorate | Information Technology | Medium | Croatia | 10 - 20 | No |
| I16 | Senior AI Scientist | Doctorate | Information Technology | Large | Sweden | 10 - 20 | Yes |
| I17 | AI Engineer and Researcher | Master's | Electronics | Medium | Ukraine | 10 - 20 | No |
| I18 | Founder / ML engineer | Doctorate | Information Technology | Micro | Netherlands | 10 - 20 | Yes |
| I19 | Lead Data & AI Consultant | Doctorate | AI Consulting | Small | Switzerland | 20+ | Yes |
| I20 | Head of AI Division | Doctorate | Information Technology | Small | Germany | 20+ | Yes |
| I21 | AI Engineer | Master's | Information Technology | Medium | United Kingdom | 3 - 5 | Yes |
| I22 | AI Engineer | Master's | Information Technology | Small | France | 20+ | Yes |
| I23 | Senior AI Engineer | Master's | Information Technology | Micro | Belgium | 3 - 5 | Yes |
| I24 | Senior Privacy Engineer | Master's | Information Technology | Large | Austria | 5 - 10 | Yes |
| I25 | Senior Privacy Engineer | Doctorate | Automotive | Large | Germany | 5 d- 10 | No |

Table 4: Interview participant demographics. *AI-centered* denotes whether AI is integral to the main product or service of the participant's company.

## E Interview Guide

**Disclaimer**
By taking part in this interview, you agree for the audio of the interview to be recorded and transcribed for further analysis. The interview transcripts will not be shared with anyone outside of our immediate research team. You also agree for direct quotes from the interview to be used for publication, and that such quotes will be attributed in a pseudonymized form. No PII or any other personal data will be shared or attributed. Please confirm your consent to these terms.

**Definitions**
General Purpose AI System (GPAIS): an AI system that is able to perform generally applicable functions such as image/speech recognition, audio/video generation, pattern detection, question answering, translation, etc., and is able to have multiple intended and unintended purposes. Within the scope of GPAIS, we investigate Large Language Models, Diffusion Models, and Multimodal Models.

**AI Background**

1. *With which of the abovementioned model types do you have (the most) experience?
2. Do you use AI systems in your role at work?
   - If yes, can you describe in what capacity?
   - Have you trained or fine-tuned AI models?
   - Which models have you used?
   - What applications have you integrated AI into?
3. How central is AI to the product you work on?

**Risk Awareness + Mitigation Strategies**

4. *What do you think are some of the privacy risks of General-Purpose AI?
5. Are these risks general privacy risks, or specific to GPAIS? Are any risks amplified by GPAIS?
6. [Ordering Exercise]
   *In your opinion, which of these poses the greatest risk?*
7. How would you rate the importance of other non-technical privacy risks, in particular legal/regulatory and ethical/societal risks?
8. Can you think of any other privacy risks in GPAIS that we have not covered in the previous question?
9. *Are you familiar with any mitigation strategies for the risks we have just covered?
   - Have you implemented any of them?
     - If yes, can you describe your experience with this? Were there any challenges?
     - If not, what were the blockers?
   - Who normally implements these measures? At which stage? [Development, Deployment, Application]
10. Are you aware of any other mitigation strategies that could be employed?
11. Out of the mitigation strategies mentioned, which do you believe to be most effective?

**Privacy Risk Mitigation Readiness**

1. *In your opinion, which privacy risks do you feel can currently be best mitigated given current mitigation strategies?
   (a) Which risks are more difficult or not currently possible to mitigate?
2. *Which of these risks do you personally feel best equipped to address? Why?
3. What would help you feel more equipped?
4. How do you learn about new technologies for privacy risk mitigation?

**Looking Forward**

1. *What do you envision the next few years will look like with regard to privacy and General-Purpose AI?
   (a) Do you think privacy will become a bigger/smaller issue in the near future with respect to the continued development of General-Purpose AI?
   (b) How do you think these issues will be handled? (reduced adoption, increased mitigation, etc.)
   (c) What would make addressing these issues easier/harder?
   (d) What would make people more confident in AI systems with respect to privacy? Who should be responsible for this?

**Other**

1. Is there any other aspect of this topic we may have missed but you feel is important to discuss?
2. Can you refer anyone who would also be able to contribute to this discussion?

## F Risk Ranking Survey

**Privacy Risks of General-Purpose AI Systems**

1. **Data Management**
   As the training and fine-tuning of GPAIS requires massive amounts of data to be accurate and robust, large amounts of data must be collected and stored, posing privacy risks to those from whom the data originated.
2. **Data Memorization and Leakage**
   Larger models have the tendency to "memorize" some of their training data, particularly the data that is more unique or rare occurring. Data leakage occurs when models unintentionally expose their memorized data, and in the case of prompting, malicious users may be able to reconstruct or extract data with careful prompting.
3. **Unintended Downstream Usage of Sensitive Prompts and User Metadata**
   User inputs and queries may be used for originally unintended purposes, for example building up user profiles based upon contextual information, or sensitive company data being captured and used downstream for model fine-tuning.
4. **Adversarial Inference and Inversion**
   A class of attacks against GPAIS referred to as inference attacks include an adversary trying to distinguish whether a particular instance (member) was in the training data, attempting to recover attributes of the training data, making broader inferences about the underlying data, or reverse-engineering model-specific parameters.
5. **Misuse via Harmful Applications**
   The predictive and generative power of GPAIS opens the door for misuse in downstream applications, where malicious users leverage AI for nefarious purposes (e.g., deepfakes).

*[Sorting Exercise]*
In your opinion, which of these risks is most urgent to address / most harmful?

- Data Management
- Data Memorization and Leakage
- Unintended Downstream Usage of Sensitive Prompts and User Metadata
- Adversarial Inference and Inversion
- Misuse via Harmful Applications

## G Codebooks

We make our codebooks public, i.e., resulting from the conducted thematic content analysis on the interview data. We split our analysis results into three codebooks: *Risks*, *Mitigations*, and *Future*, corresponding to findings related to each of our three research questions. The codebooks can be found in PDF form at: https://drive.google.com/drive/folders/1BPTOF_JOIsSPqK4WJ-uVaPXdoiIIcVxk?usp=sharing

## H Mitigations Survey

**Mitigations for Privacy Risks in General-Purpose AI Systems**

Thank you once again for taking part in our interview study! We kindly ask you to help us structure some of the feedback we received during the interviews regarding the possible mitigations for the privacy risks of General-Purpose AI systems.

In particular, we will ask you to assess your familiarity and experience with the mitigations that we presented during the interview, scored on a 4-point scale: *never heard of it*, *heard of it*, *considered using it*, *Used/using it*.
NOTE: Please respond in the way applicable at the time of the interview, i.e., if you never heard of a mitigation before our interview, please select *never heard of it*. In cases of uncertainty, please just pick the most fitting response!

The remainder of this form consists of five sections that list the different mitigations we discussed, grouped into five categories: *Organizational*, *Data Preprocessing*, *Data Augmentation*, *Private Training,* and *Post Hoc Privatization*.

The survey should take no longer than 5 minutes to complete.

Before proceeding, please indicate your name - this is just for us to keep track of who submitted the form, of course, these results will also be pseudonymized!

Thank you very much!

1. *Your name

**Organizational**

- **Management of (Training) Data**: Mitigating privacy risks begins with the responsible handling of training data, for example in trusted and secured data warehouses, as well as proper data governance structures.
- **Alignment**: An important organizational measure involves the *aligning* of GPAIS to human values, a step that can help to ensure trustworthiness and reduce the risk of adversaries compromising models.
- **Trusted Execution Environments**: Secured hardware environments where data processed within cannot be read or tampered with by outside parties or code.

2. Please assess your familiarity with the above privacy risk mitigation approaches:
   *[Response options: never heard of it, heard of it, considered using it, Used/using it.]*
   (a) *Management of (Training) Data
   (b) *Alignment
   (c) *Trusted Execution Environments

**Data Preprocessing**

- **Input Sanitization/Data Cleaning**: When preparing data for model training, a wise preprocessing step includes data cleaning, in which explicit sensitive information is removed, such as phone numbers.
- **Anonymization/Pseudonymization**: Similarly, explicit personally identifiable information (PII) can be removed via anonymization or pseudonymization techniques.
- **Outlier Detection**: Outlier detection methods aim to identify and detect outliers in the training data, as these datapoints might be more readily memorized.

3. Please assess your familiarity with the above privacy risk mitigation approaches:
   *[Response options: never heard of it, heard of it, considered using it, Used/using it.]*
   (a) *Input Sanitization/Data Cleaning
   (b) *Anonymization/Pseudonymization
   (c) *Outlier Detection

**Data Augmentation**

- **Differential Privacy (DP) (Randomized Response)**: DP is a mathematically grounded notion of privacy which usually involves adding random noise to query outputs to inject plausible deniability into computations on potentially private data (attributes).
- **Group Privacy**: The notions of k-anonymity, l-diversity, and t-closeness concern themselves with group privacy, where the goal is to provide a certain level of indistinguishability between members of a dataset.

- **Data Transformation**: Storing data in "transformed" forms, such as in embedding format or in reduced dimensions, may help to obfuscate raw sensitive data.
- **Privacy-Preserving Data Aggregation / Homomorphic Encryption**: Advanced privacy-preserving techniques based on encryption can be leveraged such that computation on raw user data is not necessary.
- **Functional Secret Sharing**: In some privacy-preserving mechanisms, user data is "split" among a pool of users, such that no single data point must be shared in full.
- **Synthetic Data**: Rather than use the original training data, some training strategies may opt to use *synthetically generated* data, which ideally shares a similar distribution of the original data.

4. Please assess your familiarity with the above privacy risk mitigation approaches:
   *[Response options: never heard of it, heard of it, considered using it, Used/using it.]*
   (a) *Differential Privacy
   (b) *Group Privacy
   (c) *Data Transformation
   (d) *Privacy-Preserving Data Aggregation / Homomorphic Encryption
   (e) *Functional Secret Sharing
   (f) *Synthetic Data

**Private Training**

- **Augmented Training Data**: Adding augmented samples to training data can help to improve model robustness, thereby reducing the risk of privacy breaches.
- **Model Diversification**: Some training procedures may opt to train a variety of models (or parameters) to improve robustness in decision-making and to mitigate any vulnerabilities of one single model.
- **Federated Learning**: The training of models is performed locally in a *distributed* fashion, where only model updates are shared with a central aggregator.

5. Please assess your familiarity with the above privacy risk mitigation approaches:
   *[Response options: never heard of it, heard of it, considered using it, Used/using it.]*
   (a) *Augmented Training Data
   (b) *Model Diversification
   (c) *Federated Learning

**Post Hoc Privatization**

- **Model Explainability**: Focusing on the development of *explainable* models can detect model vulnerabilities, as well as verify the integrity of a model.
- **Interface-side Mitigations**: Safeguarding the interface between users and models includes detecting suspicious queries, limiting the number of queries, and building guardrails.
- **Machine Unlearning**: The ability to remove a single user's contributions to a training data set upon request, without the need for complete retraining.
- **Private Inference**: Augmenting model inference (computation on unseen data) with privacy-preserving solutions such as DP or HE adds an extra layer of privacy protection for users inputting potentially sensitive data.

6. Please assess your familiarity with the above privacy risk mitigation approaches:
   *[Response options: never heard of it, heard of it, considered using it, used/using it.]*
   (a) *Model Explainability
   (b) *Interface-side Mitigations
   (c) *Machine Unlearning
   (d) *Private Inference