Surrendering to Powerlesness: Governing Personal Data Flows in Generative AI

Alejandra Gómez Ortega
Stockholm University
Stockholm, Sweden
alejandra@dsv.su.se

Hosana Morales Ornelas
Delft University of Technology
Delft, Netherlands
h.c.moralesornelas@tudelft.nl

Uğur Genç
Delft University of Technology
Delft, Netherlands
u.genc@tudelft.nl

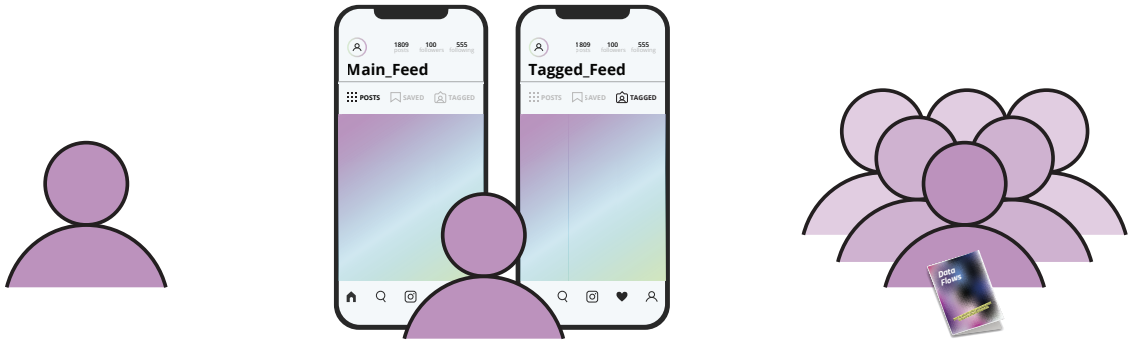

Figure 1: Participants and their involvement in research activities.

Abstract

Personal data flows across digital technologies integrated into people's lives and relationships. Increasingly, these technologies include Generative AI. (How) should personal data flow into and out of GenAI models? We investigate how people experience personal data collection in GenAI ecosystems and unpack the enablers and barriers to governing their data. We focus on personal data collection by Meta, specifically Instagram, in line with their recent policy update on processing user data to train GenAI models. We conducted semi-structured interviews with 20 Latin American Instagram users, based in Europe and Latin America. We discussed the acceptability of their data flowing in and out of GenAI models through different scenarios. Our results interrogate power dynamics in data collection, the (inter)personal nature of data, and the multiple unknowns concerning data and their algorithmic derivatives. We pose provocations around feelings of powerlessness, reframing (inter)personal data, and encountering unknown data and algorithms through design.

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; • **Security and privacy** → *Social aspects of security and privacy*; • **Social and professional topics** → *Government technology policy*.

This work is licensed under a Creative Commons Attribution 4.0 International License.
CHI '25, Yokohama, Japan
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3713504

Keywords

Personal Data; Sensitive Data; Privacy; Data Governance; Generative AI; Social Media;

ACM Reference Format:
Alejandra Gómez Ortega, Hosana Morales Ornelas, and Uğur Genç. 2025. Surrendering to Powerlesness: Governing Personal Data Flows in Generative AI. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3706598.3713504

1 Introduction

"Learn more, create more, do more," reads a header on Meta's website [50] with information about their AI intelligent assistant. This push towards learning/creating/doing more is accompanied by promises of productivity and convenience associated with Generative AI (GenAI). GenAI models are entangled with millions of (personal) data [14]. So much so that among the vast amount of personal information in the training data sets of popular GenAI models, researchers found nude photographs and banking details [33]. Increasingly, GenAI tools are deployed in digital products routinely used by millions, such as Adobe Photoshop, Grammarly, and Instagram. It has opened the possibility of using even more personal data to continue training and developing these models. Does it align with what people (already) expect from personal data collection and use? Data are undeniably central to GenAI, yet these data are often produced through people's actions and interactions. How are their perspectives and wishes accounted for?

We grapple with these questions in the context of Meta's products. These are pervasive in people's everyday lives. In some regions, such as Latin America, mobile data costs for Meta's products are

waived [29], making them the preferred form of communication and access to information. In June 2024, Meta announced an update to their Privacy Policy related to processing user data for GenAI models (Section 2). The update states how Meta will use people's posts, photos, captions, and the messages they send to an AI to train GenAI models. Further, it describes how Meta might process data from users and non-users of Meta's products whose information might be shared by users of Meta on their platforms. Although Meta already collects a plethora of personal data from its users, (potentially) using these data to train GenAI models introduces a new paradigm. Thus, it is a unique opportunity to learn about and foreground people's experiences as the privacy update unfolds.

Re-framing the questions above, does Meta's privacy update align with what people (already) expect from personal data collection and use? How does Meta account for people's perspectives and wishes? Let's say an Instagram user posted a bikini selfie[1] in 2018. Would it be reasonable for her to expect her selfie to be used to train GenAI models? Even more so, when creating her Instagram account, could she have expected that Instagram – now Meta – would use her posts, photos, and captions outside the parameters they were originally shared? What if she posted pictures of her friends and family and vice versa? Does her data spread into other people's data? Moreover, assuming she becomes aware of the privacy update via Meta's notification, she would know *something* about her data being used to train GenAI. Would she know what it *actually* means for her data to be used to train GenAI? What can she do about it?

In this paper, we unpack how people experience personal data collection and the enablers and barriers to exercising their rights and governing their data in GenAI ecosystems. We investigate:

> **(RQ1)** How do people experience personal data collection for training GenAI models?
> **(RQ2)** How do people articulate privacy norms around personal data collection for training GenAI models?
> **(RQ3)** Why do people decide (not) to opt out of personal data collection to train GenAI models?

We draw from the Contextual Integrity (CI) scenario-based inquiry [58] and Meta's Privacy Update to develop and discuss up to ten scenarios with varying information flows with 20 Latin American Instagram users, 10 based in the European Union (EU), and 10 based in Latin America (LATAM). We ground these scenarios on participants' own data (i.e., Instagram posts, photos, and their captions). As part of our study design, we supported participants in understanding Meta's privacy update, gaining awareness of their data rights, and exercising their right to opt out of Meta's data processing to train GenAI models if they so wished. Additionally, we created a Zine summarising important concepts around personal data and data governance in GenAI that we distributed to our study participants and their network to invite them to reflect further and discuss.

In sum, we contribute with: (1) a qualitative application of CI and analysis of how people experience the processing of personal data to train GenAI models, including the enablers and barriers to governing their data, (2) the (un)acceptability spectrum, describing the acceptability of personal information flowing into and out of GenAI according to four dimensions: identifiability, privacy, specificity, and labor, (3) three generative provocations that highlight the shortcomings of processing personal data to train GenAI models and open further discussion, and (4) a Zine aimed at the general public summarizing important concepts and discussing our findings in English and Spanish.

[1]We use this and the other examples in this Section as they were discussed with participants during the interviews.

2 Context: Meta's Policy Update

Meta announced an update to their privacy policy [47, Accessed in June 2024], going into effect on June 26, 2024[2]. Users of Meta's products in the EU received a notification informing them of this update (Fig. 2). The main change introduced concerned using the information individuals share on Meta's products to develop and improve AI at Meta. Together with the updated policy, Meta released a document titled: *"How Meta uses information for Generative AI models and features"* [48].

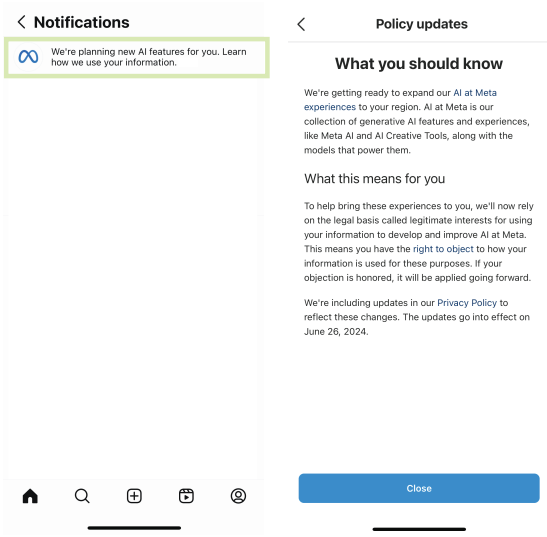

Figure 2: Notification received by Instagram users in EU and the information displayed when clicking on the notification.

According to Meta [47, 48], they use a combination of data to train their AI models. It includes (1) information that is publicly available online, (2) licensed data from other providers, and (3) information shared on Meta's products and services – which includes posts, photos, captions, and messages sent to an AI. Meta states that information publicly available online and licensed data from other providers may include personal information. They clarify that this might impact both Meta users and non-users.

> "Even if you don't use or Products and Services or have an account, we may still process information about you to develop and improve AI at Meta. For example, this could happen if you appear anywhere

[2]At the time of writing, Meta's Privacy Policy is being revised due to concerns expressed by European regulators, as it is not compliant with the GDPR. Meta's privacy policy states: "Based on feedback from regulators, we're delaying our changes to the use of your information to develop and improve AI at Meta. We've reflected this change in our Privacy Policy Update on June 26, 2024" [48, Accessed in June 2024]

> in an image shared on our Products or Services by someone who does use them or if someone mentions information about you in posts or captions that they share on our Products and Services." From *"Where does Meta get training information?"* [48, Accessed in June, 2024]

Meta's legal basis for processing personal information to train GenAI models is legitimate interest [47, Accessed in June 2024]. It applies when personal data processing is necessary to preserve the interests of data controllers (e.g., Meta) and outweighs any risk(s) to individuals. It requires individuals to be informed – although they do not need to consent, and reasonably expect at the time and in the context of collecting personal data that processing for that purpose may occur.

Opting-Out Process

Users in the EU may opt out of data collection by using an online form[3]. It can be accessed through Instagram's notification (Fig. 2). When conducting this study, the form required Meta's users to input their country of residence and email address and fill in a required text field that cites: *"Please tell us how this processing impacts you."*[4] When submitting their request, users read:

> "We'll review objection requests in accordance with relevant data protection laws. If your request is honored, it will be applied going forward. We may still process information about you to develop and improve AI at Meta, even if you object or don't use our Products and Services. For example, this could happen if you or your information: (1) appear anywhere in an image shared on our Products or services by someone who uses them, (2) are mentioned in posts or captions that someone else shares on our Products and Services." From *"Object to Your Information Being Used for AI at Meta"* [51, Accessed in June, 2024]

Users of Meta have perceived the opting-out process as ambiguous and arbitrary. Several artists attempted to follow the process and received a response stating that Meta is *"unable to process the request"* until they submit evidence that their personal information appears in responses from Meta's GenAI [36]. In places where the form is not available, users have resorted to creative ways to attempt to opt-out. Users of Meta in LATAM posted captioned images on Facebook and Instagram stating: *"I do not consent to Meta using the images, videos, audios, and texts uploaded by me to any of their platforms for anything related to AI"* [72].

3 Related Work

3.1 Personal Data Flows in Generative AI

Personal data corresponds to any information related to an identified or identifiable person. It has multiple forms and formats [81]; within Meta's ecosystem, these include: (1) account information

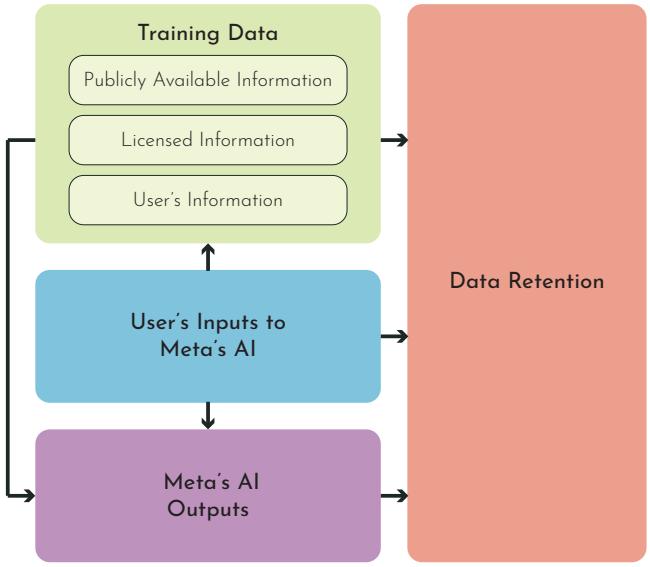

Figure 3: Flow of personal data in Meta's GenAI development process, adapted from Duffourc et al. [25]

(e.g., name, email address), (2) volunteered content (e.g., text, images), (3) digital interactions (e.g., DMs, comments), (4) personal usage (e.g., search queries, time spent) and (5) personal inferences (e.g., inferred interests, inferred ads), among others [32, 81]. Sometimes, personal data is *public*, or publicly available (e.g., email address on a personal blog). This has opened the way for personal data to be (re)used in multiple ways, including scraping it for scientific research and training GenAI models. Whether (re)using publicly available personal data represents a privacy violation is disputable and varies per regulation[5]. On one hand, (re)using personal data could violate privacy expectations. Sharing a selfie on a publicly available blog for a specific purpose (e.g., updating readers) doesn't mean expecting said selfie to be used for other purposes (e.g., training GenAI models). On the other hand, pervasive data collection and consumption have proliferated the view that data shared publicly online is available to anyone [33, 78]. When personal data is not publicly available, it is private personal data (e.g., DMs on Instagram).

GenAI describes AI models capable of generating content such as text, images, sound, and videos. It has become popular and pervasive through models that generate text and images, like ChatGPT and DALL-E. These models are trained with large amounts of data (e.g., text, images) [14] – in most cases data that is publicly available on the internet (e.g., Google's Bard [75], Meta's Llama 2 [49], Stable Diffusion's Midjourney [5]). Some of the data used to train GenAI models is private personal data, posing risks to individuals' privacy and security [28]. Lee et al. [42] proposed a taxonomy of AI privacy risks, including (1) identification, linking data points to an individual's identity; (2) aggregation, combining various pieces of data

[3]Form available at: https://help.instagram.com/contact/233964459562201
[4]At the time of writing, this version of the form has been replaced. Users no longer need to input their country of residence; the text field is no longer required.

[5]In the **EU**, the GDPR considers publicly available data within the scope of personal data, processing such data requires a valid legal basis (e.g., informed consent). Other regulations, such as the California Consumer Privacy Act (CCPA), exclude publicly available information from its scope. Regulations in **LATAM** state that processing publicly available data does not require informed consent.

about a person to make inferences beyond what is captured in the data; (3) distortion, disseminating false or misleading information about people; (4) exposure, revealing sensitive private information, and (5) surveillance, watching, listening or recording an individual's activity; among others.

Duffourc et al. [25] illustrate how personal data are harnessed in various stages of the GenAI development process (Fig. 3), (1) training the model, (2) interacting with the model, (3) generating outputs, and (4) further improving the model. First, models are trained on large datasets primarily constituted by publicly available information, including personal information [43], for instance, "public"[6] social media profiles and personal blogs. More recently, changes in service providers' terms of service have made this usage explicit (Section 2). Second, users of GenAI applications further contribute data in the form of account data (e.g., email address) and their interactions with GenAI models (e.g., prompts and inputs to GenAI). In the case of Meta, it *"could include text, document, images, or recordings"* [48, Accessed in June 2024]. Third, the outputs generated by GenAI models in response to prompts can contain personal data learned through training or provided by individuals (e.g., names and phone numbers) [15, 83]. Golda et al. [28] underline how these outputs, in the form of deepfakes, could harm individuals by using their likeness, voice, or identity without their consent or control. Fourth, personal data collected for developing and improving GenAI models may be retained for further use [25].

Sensitive, private, and public personal data have been found in GenAI training datasets (e.g., [5, 25, 33, 43, 55]) and are likely to continue to be integrated into GenAI due to recent policy updates. Previous research has explored this from the perspectives of privacy and security. We foreground people's perspectives as they navigate this shift in how their data is used.

3.2 Data Protection, Governance, and Activism in the EU and LATAM

Personal data collection and use are protected across legislations and, in most cases, require explicit informed consent. In the **EU**, the General Data Protection Regulation (GDPR) establishes the legal basis for processing personal data[7], defines a special category of sensitive personal data[8], and provides rights to individuals over their data. According to the GDPR, data subjects (i.e., individuals whose data is processed) have the right to **be informed** on how their data will be used, **access** their personal data, **rectify** inaccurate or incomplete data, **erase** their personal data (i.e., right to be forgotten), **restrict processing** in specific circumstances, **object** the processing of personal data, and **data portability**, obtain a copy of their personal data. Similarly, in **LATAM**, the *Habeas Data* (Latin for *"may you have data")* establishes a protected category of sensitive personal data[9], and provides rights to individuals over their data – stored in databases [61]. *Habeas Data* includes the right of individuals to **be informed** about how their data will be used, **access** their personal data, **update** their data when necessary, **rectify** inaccurate or incomplete data, and **erase** their personal data. These rights, however, are difficult to enforce in practice [61]. Local legislation further protects individuals from how their data are (mis)used (e.g., Ley 25.326 in Argentina [20], Lei 13.709 in Brazil [65], Ley 1581 de 2012 in Colombia [18], Ley Federal de Protección de Datos Personales in Mexico [19], and Ley 29733 in Perú [21]). Most of these enable the processing of personal sensitive data to safeguard the data controller's legitimate interest[10] [72].

Individual rights can be seen as opportunities for empowerment. Vincent et al. [78] propose practices that allow people to influence the performance of data-dependent technologies supported by these rights. These include (1) *data strikes:* withholding or deleting data, enabled by rights to delete; (2) *data poisoning:* contributing harmful data, enabled by rights to rectify and update; and (3) *conscious data contributions:* contributing data to a competitor – enabled by rights to access and data portability. Proactive and reactive practices where individuals can leverage their data (rights) to affect technology design, development, and deployment fall under the umbrella of **data activism** [4, 53]. These practices include a broad range of activities at the individual and collective levels: from individuals using private browsing windows when shopping [77] or going to court to exercise their rights to delete data [44]; to a community of Reddit users changing their subreddits configuration to protest a policy update [54] or a group of people gaining agency over their data and deciding who to share it with through the Solid project [71].

Personal data usage in the GenAI lifecycle is governed by data protection laws such as the GDPR[11] and the Habeas Data. Product-service providers like Meta allow users to delete or restrict their use of personal information to train AI models and exercise their rights under relevant data protection laws. We use Meta as a case study to understand how people experience personal data processing to train GenAI models. We deliberately incorporate data activism into our study design by supporting participants in knowing and exercising their rights if they so wish.

3.3 Data Privacy and Contextual Integrity

Privacy is understood as individuals (re)defining and managing boundaries around their information across multiple contexts (e.g., who accesses content from a private Instagram account). Pervasive personal data collection expands the scope of these boundaries to the realm of data – where individuals, but also product-service providers manage these boundaries. HCI researchers have interrogated people's experiences and concerns around data privacy. Focusing on mobile apps, Shklovski et al. [67] found people to be

[6]The term is in quotation marks as the public nature of this type of content is contested and varies per regulation, as described in Section 3.2.

[7]In the GDPR, the legal basis for processing personal data are contract, legal obligations, vital interests of the data subject, public interest and legitimate interest [27, Art.6])

[8]In the **EU**, GDPR defines sensitive data as a special category of personal data that includes racial or ethnic origin, political opinions, religious or philosophical beliefs; health-related data; and data concerning a person's sex life or sexual orientation, among others [27, Art. 9].

[9]In **Latin America**, sensitive data includes racial or ethnic origin, political opinions, religious or philosophical beliefs, health-related data, data concerning a person's sex life or sexual orientation, and biometric data.

[10]Legitimate interest is the legal basis from which Meta aims to collect personal data from users to train their GenAI models (Section 2).

[11]The Artificial Intelligence Act (AI Act) further regulates AI in the EU by classifying AI applications by their risk.

"creeped out" by the shared data being more than they seem. Similarly, Gómez Ortega et al. [30] and Kurze et al. [39] found people experiencing creepiness and discomfort from the sensitive and intimate information revealed through their data. Creepiness has been accentuated through interactive design projects that invite people to engage differently with their data and critically discuss data privacy [23, 24, 32, 66].

Crabtree and colleagues [17] describe the underlying theories from which privacy is understood; privacy as control, boundary management, and contextual integrity. Privacy as control relates to managing the flow of personal information through activities such as limiting disclosure [80]. Privacy as boundary management [1] relates to selectively disclosing personal information as people move between privacy and publicity according to context and intention [60]. Privacy as Contextual Integrity (CI) [57] frames privacy regarding the appropriateness of information flows according to social or cultural norms and grounded in specific contexts. CI is considered an appropriate framework for understanding privacy norms, especially in HCI research, as it was developed *"in an attempt to understand what people saw threatened by novel sociotechnical practices wrought by a family of technologies, including computers, digital networks, information systems, databases, communications media, electronic hardware, and software"* [58].

As proposed by Nissenbaum [57], information flows are described according to five parameters: (1) **subject** of the information, (2) **sender** of the information, (3) **attribute**, describing the type of information, (4) **recipient** of the information, and (5) **transmission principle**, stating the condition under which the information flow is permitted. For example, an Instagram user (subject) might be comfortable with Meta (sender) sharing her Instagram data (attribute) with a third-party service (recipient) if she has authorized it (transmission principle); but not for Meta using them for training GenAI models (a different transmission principle and privacy violation). Applications of CI in HCI primarily employ quantitative methods to identify the appropriateness of information flows [38], operationalized in large-scale surveys (e.g., [3, 69]) where information flows are illustrated through scenario-based vignettes with varying parameters (e.g., a different recipient or transmission principle). There are a few applications of CI in qualitative research. Gómez Ortega et al. [30] used data representations to introduce and discuss concrete CI scenarios during interviews. While Bowser et al. [8] and Kumar et al. [37] conducted interviews discussing privacy with citizen scientists and families, respectively, and analyzed the data considering the CI parameters. Kumar et al. [38] advocate for engaging with CI in qualitative research to move beyond identifying privacy concerns towards determining how to respond to those concerns.

We expand on the HCI research that explores and re-configures people's relationship with their data, by focusing on how people experience their data flowing into and out of GenAI models at Meta. We continue to build on and illustrate the qualitative applications of CI to understand the appropriateness of information flows and envision ways to respond to them.

4 Methodology

We unpack how people experience personal data collection and the enablers and barriers to exercising their rights and governing their data in GenAI ecosystems. We scoped our research questions in Meta's product portfolio as a response to Meta's policy update regarding the collection and use of user data to train GenAI models. We further scoped our research by focusing on adult (18+) Latin American Instagram users (Section 4.3). We focused on Instagram as it is one of the most used online platforms by (young) adults in Meta's product portfolio. We initially hoped to explore Instagram, Facebook, and WhatsApp, but we determined that data collected through WhatsApp was too sensitive to disclose to researchers and found that our participants were (mostly) inactive on Facebook.

We strove to leverage our research to support individuals engaging in data activism [4, 53, 78]. To do so, we created spaces during the interview to introduce and discuss Meta's privacy policy and increase participants' awareness of the notion of personal data and how Meta collects and uses their data. We sensitized participants to their rights under local regulations and supported them in opting out of Meta's training GenAI models with their data if they wished. Moreover, we translated our research and findings into an informative Zine. We shared it with participants, who distributed it broadly and used it to continue the discussion.

4.1 Procedure: Contextual Integrity Scenarios

Inspired by the Contextual Integrity (CI) scenario-based inquiry (Section 3.3), we conducted semi-structured interviews with 20 Latin American Instagram users, 10 residing in the EU and 10 in LATAM. During the interviews, we interrogate different scenarios with varying information flows in the context of GenAI at Meta.

Stage 1: Defining Information Flows from Policy Analysis. We defined the information flows through an analysis of Meta's policy update going into effect on June 26, 2024 [47, Accessed on June 2024] and Meta's *"How Meta uses information for generative AI models and features"* page [48, Accessed on June 2024]. We analyzed these documents by annotating the pertinent CI parameters (i.e., *subject*, *sender*, *attribute*, *recipient*, and *transmission principle*) in each section and identified the possible variations in each parameter. For instance, Meta states that *"Even if you don't use our products and services or have an account, we may still process information about you to develop and improve AI at Meta."* meaning the *subject* of the information can be the (main) user, other users, and other people (i.e., not users of Meta's products). Table 1 presents an overview of our analysis with salient excerpts from Meta's documents.

Due to our focus on Instagram and intending to discuss a variety of posts, we expanded on the *attributes* (i.e., photos and captions) through a categorization proposed by Hu et al. [34]. They identified eight different types of photos on Instagram, including (1) selfies (i.e., photos of one person); (2) friends (i.e., photos with at least two people); (3) activities (i.e., photos of activities or places where activities happen); (4) food, (5) gadgets, (6) captioned photos, (7) pets, and (8) fashion. Our resulting *attributes* comprise (1) selfies, (2) friends, (3) activities, (4) other (i.e., photos belonging to any of the other categories), and (5) information that is publicly available online. The resulting *senders* correspond to the participants themselves, as main users of the account, and the users who tagged them on their posts. Figure 4 illustrates the resulting ten information flows. For each one, we invited participants to discuss the following question:

Table 1: CI parameters mapped along Meta's Privacy Policy Update on June 26, 2024

CI Parameters	*Excerpt from Meta's Documents*	*Variations*
(1) *Subject* **(2)** *Sender*	"Even if you don't use our products and services or have an account, we may still process information about you to develop and improve AI at Meta. For example, if you appear in an image shared in our products and services by someone who does use them or if someone mentions information about you in posts or captions."	User, other users, other people.
(3) *Attribute*	"Since it takes such a large amount of data to teach effective models, a combination of sources are used for training. We use information that is publicly available online and licensed information. We also use information shared on Meta's Products and services. This information could be things like posts or photos and their captions."	Posts, photos, and captions on Meta's products, publicly available information online.
(4) *Recipient*	"We share certain information with: (1) advertisers who show ads on our products, (2) businesses we hire to market our products for us, (3) businesses we hire to do things like offer customer service or conduct surveys, (4) researchers who use it to do things like innovate, advance technology, or improve people's safety. We don't sell your information, and we never will."	Meta, third-parties.
(5) *Transmission Principle*	"We believe the use of this information is in the legitimate interests of Meta, our users, and other people. You have rights related to how your information is used for AI at Meta. This includes the right to object."	Legitimate interest (i.e., no explicit consent), opt-out.

> How acceptable is it for *[sender]* to allow *[attribute]* to be used by Meta to train their GenAI models?

Generally, CI scenario-based inquiries explore and vary five parameters: subject, sender, attribute, recipient, and transmission principle. We varied only the sender and attribute since the subject (i.e., participant and Instagram user), recipient (i.e., Meta), and transmission principle (i.e., Meta's policy update) remain the same. We removed the subject and transmission principle from the wording to limit complexity and length – they are introduced and explained before discussing the scenarios.

Stage 2: Selecting Information Flows from Participant's Data. In preparation for each interview, we asked participants for their Instagram usernames and relevant links with information about them that were publicly available online (e.g., personal page on employer's website, personal blog). We followed them from an empty Instagram account created only for this research to access their posts. At this stage, participants consented to the first author selecting posts from their main Instagram feed and tagged posts feed to discuss during the interview. The first author selected examples from the participant's main feed and tagged posts feed based on the attributes identified during the policy analysis. We did not store any posts and unfollowed participants after each interview. Each participant was only invited to see and discuss their data.

Stage 3: Discussing Information Flows with Participants. We conducted semi-structured interviews[12] with participants where they reflected on the acceptability of the various information flows – introduced with concrete examples from their Instagram data (Fig. 5). The interviews consisted of five parts:

(1) **General Knowledge of GenAI:** Participants discussed their familiarity and experiences with GenAI tools (e.g., Chat-GPT, DALL-E) to gain insights into their understanding of these.
(2) **General Knowledge of Data (Governance) in Meta:** Participants discussed their understanding of Meta's data practices. We asked them how familiar they were with Meta's privacy update and sensitized them to Meta's data practices and policy update by showing them key sections of their Privacy Policy.
(3) **Information Flows:** We introduced the CI structure (Section 4.1). Participants discussed the acceptability of up to ten information flows – depending on the availability of the different attributes in their data. We grounded Meta's policy update in concrete examples from participants' data by introducing each *attribute* through their main and tagged posts (Fig. 5). They could see the photos, captions, and comments for each post. For each *attribute*, we asked them to discuss the acceptability of it being used by Meta to train GenAI models. After discussing all attributes, we invited participants to summarize their privacy norms.

[12]The complete interview protocol in English and Spanish is available in the supplementary material.

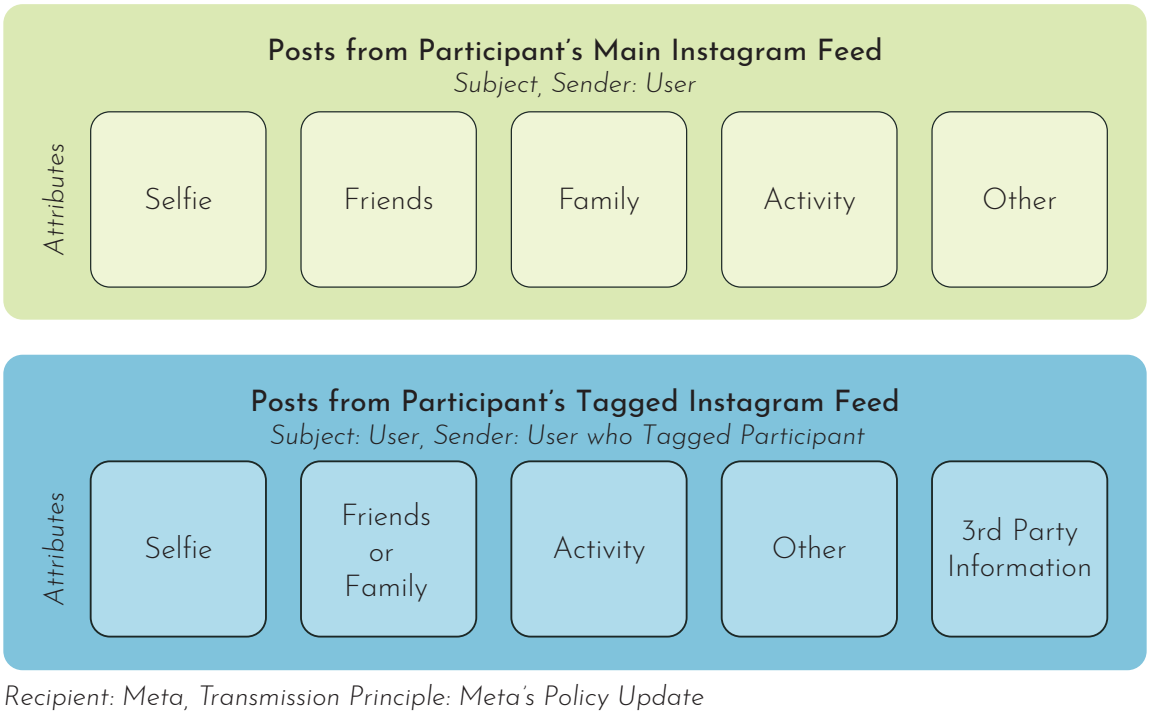

Figure 4: Ten information flows discussed with participants during the interviews.

(4) **Whether to Opt-Out:** Participants reflected on whether they wanted to opt-out from Meta's GenAI data processing or continue opting-in. We emphasized that participants were free to choose and provided detailed information about both options. For those who chose to opt-out, we explained the process, followed it along with them, and invited them to reflect on it.
(5) **Reflection:** Participants reflected on the interview by identifying general principles or rules they use to determine what's acceptable regarding data collection and use, and more broadly, discussing the potential benefits and harms from allowing product and service providers to use their data to train GenAI, and envisioning what the 'ideal' system would be.

Stage 4: Broadening the Discussion through a Zine. To continue the conversation with participants beyond the bounds of our study, we created a Zine[13] summarising important concepts around personal data and data governance in GenAI systems and inviting individuals to reflect on these through our findings. We distributed the Zine to participants, who used it to discuss with friends and families. They distributed it further within their communities.

4.2 Ethical Considerations

The research activities described above were reviewed and approved by our institution's Human Research Ethics Committee (HREC). The principle guiding our research activities was informed consent. We invited participants to actively and explicitly consent to the different research stages and activities (i.e., sharing their Instagram account handles and participating in the interviews) and (re)evaluate their participation if necessary. We cared for our participants' data by not storing any of their posts, accessing them only through private browser windows, and deleting our search history after each interview. Upon discussion with the HREC, we decided not to compensate our participants due to the ethical precedent of compensation limiting their ability to voluntarily consent [59, 62] – especially as we were accessing their (sensitive) data [31]. Following feminist research practices [31], we strove to find ways – other than financial incentives – for participants to gain value from participating in our research, such as the Zine.

4.3 Participants

We interviewed 20 adult active Latin American Instagram users, half of whom (10) resided in the EU and the other half (10) in LATAM. We focused on Instagram as it is one of the most used online platforms by (young) adults across Meta's products. We included Latin American migrants residing in the EU as they are protected by the GDPR and inhabit a digital space strongly shaped by Meta to stay close to their friends and families elsewhere. Further, we included Latin Americans residing in LATAM[14] as a direct response to Meta's Privacy Policy underlining the relational nature of data – where people's social relationships can also expose them to their data being used to train GenAI outside regulatory borders. Here,

[13] Zine in English and Spanish are available in the supplementary material.

[14] When designing our study, it was unclear whether opting out was possible for participants in LATAM. Previous research has shown how some of the rights established in the GDPR, such as the rights of access and data portability, are available worldwide since international companies rarely limit them by geography (e.g., [9, 30]).

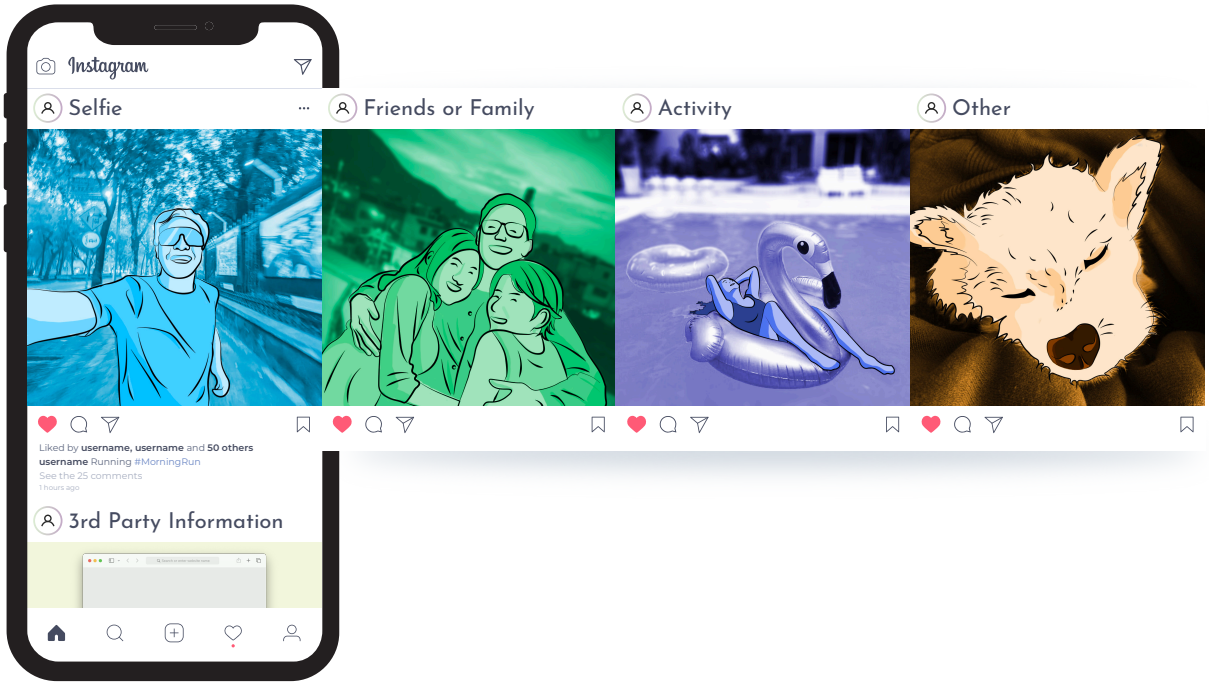

Figure 5: Examples of information flows from participants' data. Amalgamated from several participants to preserve anonymity. Shared with permission from participants.

it is important to note that Meta plays an important role in the mobile data space in some Latin American countries, including Brazil, Colombia, and Mexico, where Meta's products are often *free* to use – i.e., mobile data consumed while using Meta products is not charged for. We searched for participants by advertising our call for participation through Meta's products, including Facebook groups, Instagram, and WhatsApp.

Participants' ages ranged from 21 to 61 (median = 30). Twelve participants identified as women and eight as men. They were from Brazil, Colombia, Mexico, Peru, and Puerto Rico. Most participants reported using Meta products for a long time (i.e., more than 10 years). All participants reported frequently using Instagram, opening the app at least once a day, and posting stories and posts on their feeds at least once a month. We conducted interviews in person and online, in English and Spanish. The interviews lasted between 35 and 75 minutes and were audio recorded.

4.4 Reflexive Thematic Analysis

The first author made a transcript of the 20 interviews using MS Office 365, then manually reviewed and edited them. We analyzed the transcripts through reflexive thematic analysis [11, 12]. We performed several iterations of coding, interpreting, and discussing[15]. We leveraged Atlas.TI and physical materials to support the process. The first two authors read the transcripts to familiarize with the data. We coded two interviews individually and then discussed and interpreted the codes together. With this activity, we aimed to understand our different perspectives on the data and start to develop a shared understanding. We continued iteratively coding, interpreting, and discussing the codes until we concluded the coding process. We created 363 unique codes that we iteratively refined. The three authors then discussed and generated the tentative themes. We generated three themes: **Surrendering to Powerlessness**, **Unknowns in Gen AI**, with the sub-themes: *Expectations of Governance*, *Unknown(s) in GenAI*, and *Acceptable, Ambiguous, and Unacceptable Information Flows*, and **Uncertainties about Data Governance**. We further defined each theme through the writing process. Additionally, we report on participants' perceptions in the EU and LATAM.

Positionality. We are migrants (from Colombia, Mexico, and Turkey) residing in the EU and LATAM. We are active users of Meta's products. Two of us have opted out of Meta's processing of our data to train GenAI models; one is willing to opt out but doesn't consider it a pressing issue. Our research interests and personal experiences influence our focus on personal data governance.

5 Findings

5.1 Surrendering to Powerlessness: Individual Experiences of Personal Data Collection for GenAI

Most participants (16/20) had no prior knowledge of personal data processing in Meta and were unaware of Meta's policy update. Among those participants in the EU – who should have received a notification – only three were somewhat aware of Meta's privacy

[15]We are proficient in English and Spanish, so we did not translate the interviews for analysis, we constructed all the codes and themes in English – we only translated the representative quotes reported in this paper.

update. They became aware through the news: *"I heard something about it, but very briefly, in the news. It was not very clear to me, but I heard something about it"* (P8). For most participants, the interview involved gaining new knowledge and awareness about Meta's practices and being confronted with the (uncomfortable) feelings derived from these. We synthesized the nuances of going through this process into a theme: **surrendering to powerlessness**.

Powerlessness first manifests during the interview as participants recognize their little understanding of Meta's data processing and the impotence and discomfort that might result from an increased understanding; *"I have no idea, and I think I would panic if I find out"* (P4). Participants feel there are no (adequate) resources for them to expand their knowledge. They expect and have normalized the inaccessibility of the information on privacy policies, notifications, and other (un)informative information sources. Similarly, they expect and have normalized that service providers benefit – and profit – from processing their data. Moreover, they expect and have normalized not having agency in how their data is processed.

> "To me, the idea of telling Meta «I don't want my private information [used] to train AI» is basically pointless. I feel they don't care. It is possible to do it, but they don't follow it." (P3)

Powerlessness continues to grow during the interview as participants are sensitized to Meta's Privacy Update. Most participants perceive it as overwhelming, scary, and abusive: *"I feel violated, I feel abused, although my profile is public, so it is a contradiction"* (P4). Additionally, they are uncertain about what it actually means. Here, powerlessness manifests from uncertainty, as participants express uncertainty and doubt about Meta using their posts, photos, and captions to train Gen AI models (Section 5.2.2). Uncertainty also stems from potential vulnerabilities that participants – and society in general – might be exposed to from GenAI outputs involving their personal information, such as hallucinations, deepfakes, and fake news.

> "Having my swimsuit photos used to generate a nude [picture], for example, a nude of me, and having it spread everywhere, like all that online stalking, that would be something that would freak me out, you know." (P10)

The different layers of uncertainty create fear, discomfort, and creepiness. These feelings briefly prompted participants to delete some of their posts, and even reconsider using Meta's products: *"Right now, I feel I should close my Instagram; this does not feel good."* (P1). Nonetheless, participants acknowledge how difficult – impossible – it would be to stop using Meta's products and services, even if doing so can feel uncomfortable and abusive. Thus, powerlessness comes across again as the inability to act or change.

> "Anyway, I'm not going to read that [Meta's Privacy Update] and feel uncomfortable knowing that I will not stop using Instagram because they use my data to train their models." (P4)

Participants feel powerless, and then the surrendering comes. Participants accept and acknowledge the status quo and feel they have no option but to give their data away for Meta's GenAI models and whatever other purposes, *"I accepted it. I know the rules and that it could happen. So if it happens, then, well, yeah, I don't like it. But I have to accept it"* (P3). Here, it is paramount to underline the importance of Meta's products in people's personal lives and livelihoods. Not surrendering personal data to Meta, by not using Meta's products, could hamper their lives and livelihoods. Moreover, it is ineffective at the individual level – everyone else uses Meta's products. Surrendering personal data collection to train GenAI becomes a *"necessary evil"* (P13).

> "But what do I do? Nothing, because I'm going to lose. In other words, imagine that you are in the middle of the sea in a storm and you find a small boat. The probability of the small boat failing is high, but if you don't get in, you'll drown. So, it's better to get in and try to save yourself in the small boat than not to get in and drown in the storm." (P17)

Equating Meta to a small boat in which to navigate a storm nicely captures the essence of surrendering to powerlessness. It involves people giving themselves and their data over to Meta's GenAI models – despite the uncertainty and discomfort. It highlights how people lack the capacity and the ability to act differently. Finally, it positions individuals and their relationships with Meta in a greater context that involves and is shaped by people's lives, livelihoods, and other people. In this greater context, maybe all together could find a way to navigate the storm and not drown.

5.2 Unknowns in GenAI: Acceptability of Information Flows

5.2.1 Expectations of Governance. All participants expected something different from (the processes around) personal data collection at Meta and elsewhere. Their expectations influenced what they deemed acceptable and unacceptable and what felt necessary or abusive. We elaborate on these expectations and their underlying feelings, specifically around consent. Participants expressed dissatisfaction with the opt-out mechanism implemented by Meta. They expected to have a voice before the fact and not the ability to oppose after the fact, *"they [Meta] didn't even give me the chance to make a decision at the time when they implemented whatever it was they implemented"* (P7). They, therefore, expected and would have preferred an opt-in mechanism where they would be adequately informed and give their consent. Still, some participants were critical of what an opt-in mechanism would have looked like based on their prior experience with Meta's products and their fatigue toward privacy policies:

> "I'm not familiar with [how Meta collects personal data]. I don't read the Privacy Policy, it sounds too harsh. I don't read them, and I might be selling my soul to the devil. I don't read anything. I say OK, I want to use it, next. I don't have the willingness or time to read all that stuff, which, by the way, they make it super long and with super small letters so that you get dizzy and say yes to everything." (P13)

Beyond their expectations of informed consent and dissatisfaction with the opt-out mechanism, participants highlighted two important elements regarding consent: *temporality* and *relationality*. Temporality and relationality are also characteristics of posts,

photos, captions, and comments on Instagram and other forms of personal data. Data is anchored to a temporal dimension in the form of timestamps. Personal data, by being related to individuals, might also contain and reflect their social relationships, which could be (further) exposed through GenAI.

The temporality of consent relates to when consent is given or withdrawn and when the policy update is enacted and applied to people's data. It is embedded in the temporality of using Meta's products, in this case, Instagram. Participants recognize how their usage of Instagram extends over time and question how the privacy update fits in this timeline: Does having consented to Instagram's terms of service in 2011 mean consenting in perpetuity to whatever update they come up with? Does the privacy update in 2024 apply to posts posted in 2011? Temporality, being central to people's timestamped digital lives, also influences the acceptability of some information flowing into and out of GenAI models (Section 5.2.3).

> "So I find it's also a bit weird. I don't feel very comfortable with [my posts] being used as AI training data. Especially considering that this [post] was quite a long time ago, and Instagram will just come back to it now and start using it." (P5)

The relationality of consent stems from recognizing the relational and interpersonal nature of posts and comments on Instagram – and, more broadly, of data – and how consent from an individual might include others. Consent in relation to others introduces tensions. Once participants were confronted with posts involving other people (e.g., group pictures, pictures made by participants of other people), they were doubtful about the need to ask for their consent to publish these posts and to allow Meta to use them to train GenAI models. P12 describes the relationality of consent when reflecting on the acceptability of a family photo of her with her mother and grandparents:

> "My grandparents, I mean, the face of my grandparents and my mother [make it unacceptable for this post to be used by Meta's GenAI models]. I mean, I think it's a question of consent. This one, where it's like a photo that has to do with me, well, it's something that you ask me, but it involves other people and they're not being asked. Even if I say yes, it brings other people in; it doesn't seem right to me." (P12)

5.2.2 Unknown(s) (in) GenAI. Participants discussed the (un)acceptability of (up to) ten different information flows, i.e., posts of/with them on Instagram and third parties used by Meta to train GenAI models. These discussions were permeated with uncertainties we delineate as *the unknown*. We choose this term to underline how GenAI is perceived as a completely different application of personal data, in contrast, for instance, to personalized ads – which are (sort of) known. One might (somewhat) expect and understand to see personalized ads for running shoes after spending a couple of hours on Google browsing pages about running shoes. Yet, with GenAI models, one can't really expect or understand how one's swimsuit photos will affect the algorithm or its outputs.

The unknown emphasizes the multiple uncertainties throughout the GenAI development process and how participants don't even know what they don't know about their data in relation to GenAI. It foregrounds how challenging it is to envision personal data flowing into and out of GenAI models and, even more so, governing these flows. Especially due to the generative nature of GenAI and the unpredictability of how participants' data would emerge in GenAI outputs. With questions and doubts from half of our participants, we want to show the many uncertainties that constitute the unknown:

> "Where can photos go? Also, other people's photos?" (P1). "What can they get from that photo? What information will they obtain?" (P2). " "I still don't know where the information is going to go. What are the AI models going to be doing?" (P3). "How can they train their models with my data?" (P4) "Actually, I have no idea which kind of data they could take" (P5). "How can my comments and those of others be used?" (P6). "They have access to everything?" (P7). "I don't know, I feel I'm not understanding, like, what will they do with this?" (P8). "Can this be bad? To us?" (P9). "Do you know what purpose it will have? What use does it have?" (P10).

The unknown also encompasses the temporality and relationality of people's data, the data in the system, and the system itself. It includes the potential and envisioned outputs of Meta's GenAI models. Participants were concerned and doubtful about the possible harms and exposition derived from these (e.g., deepfakes, identity theft). They discussed potential scenarios where the unknown could perpetuate their vulnerability: *"we go back to that vulnerable place because you don't know if someone made a video of you [with GenAI], which you didn't even know about"* (P7).

At the end of the interviews, we invited participants to envision alternative ways to collect personal data from individuals to train GenAI. These discussions circled back to *the unknown.* They centered on how to *know* better.

> "The thing is that everything about AI is quite new. And again, I think Meta should do a better job at explaining how exactly they're going to use it [the data] and not just having a legal document or just sending a notification saying, hey, we're going to use [your data to train] AI and that's it. It should be an actual awareness campaign." (P3)

One common strategy discussed by participants was to introduce friction, pushing people to know (or remember) constantly, being *"more in-your-face about it* (P6). For instance, by being asked *"for every single post, where you can say like, OK, this is whatever, you can go ahead and use it, or this is sensitive, too sensitive, and I don't feel like sharing it"* (P5). Nonetheless, friction can be misused and discourage people. Some participants identified friction from the need to explain how data processing by Meta impacts them when opting out, *"as a person with no knowledge on the subject, I find it a bit complex to find what to say [in the form]"* (P4).

5.2.3 Acceptable, Ambiguous, and Unacceptable Information Flows. The (un)acceptability of information flows is conditioned by participant's expectations of privacy and publicity. Participants with public-facing Instagram accounts, where they post public-facing posts, feel more comfortable with the flow of these posts, *"It's not*

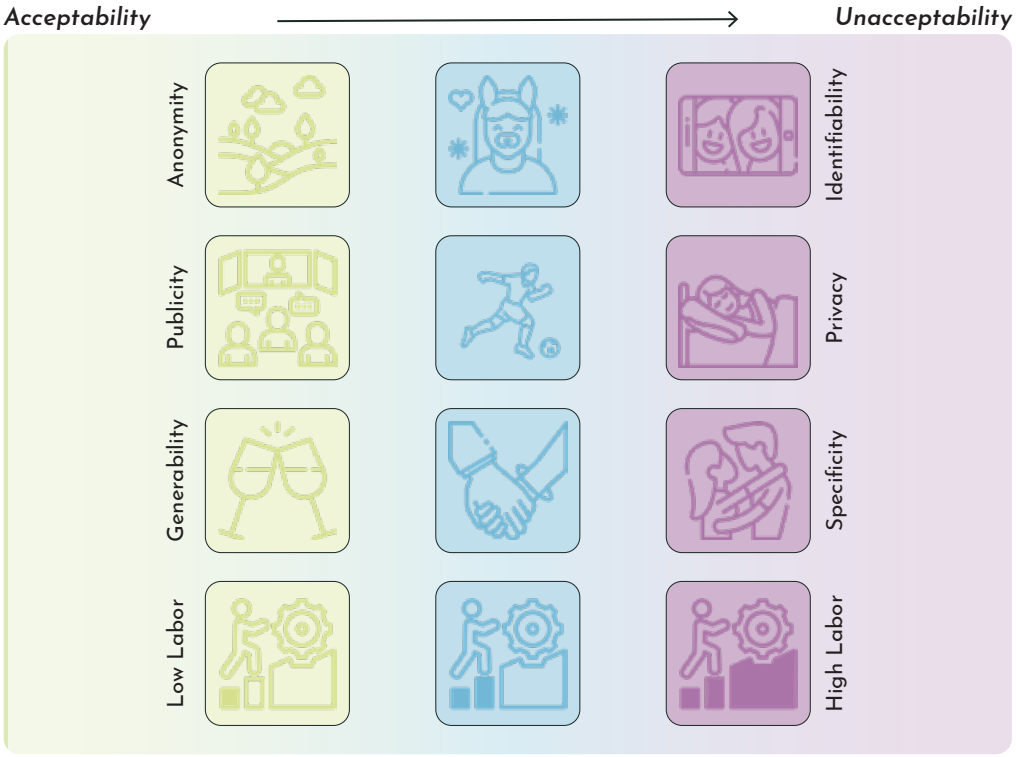

Figure 6: Spectrum of the (un)acceptability of personal data flowing into and out of GenAI models. It represents the multiple dimensions discussed by participants influencing (un)acceptability: identifiability, privacy, specificity, and labor.

like Instagram life is very private, right?" (P2). Although they are not necessarily more comfortable with these posts being used to train GenAI models. In contrast, participants with private Instagram accounts expect to control who can view and access their posts. They feel less comfortable, and often uncomfortable, with the flow of these posts outside of the boundaries they define by allowing people to follow them – including potential undesired exposure through Meta's GenAI tools. They recognize the exposure to publicity associated with using Meta's products.

> "I think this is, like, definitely my more private life. I don't like to share it. I don't want it to be public, although I know that it's in Meta's system and therefore it is public, let's say, some sort of public." (P3)

Information flows are considered **acceptable** in terms of the process and the content. Process-wise, participants envisioned several scenarios that could benefit them and others from Meta training GenAI models with their data. For instance, improving models through (more) diverse data or fine-tuning models to an individual based on increased access to their data, *"I use the tool [GenAI] a lot. In a certain way, it would benefit me if it were much more effective and efficient when it came to my interaction with it."* (P18). These potential benefits made it acceptable for different types of posts to be used to train GenAI, regardless of their content.

Content-wise, participants described different attributes (i.e., posts, photos, captions, comments) influencing the acceptability of information flows. Attributes that do not contain people or their information are considered acceptable as they are unrelated to any individual and do not expose them to any possible harm. Additionally, similar information can be found elsewhere, *"they are instruments, and let's say that you can find the same thing on Google or anywhere else"* (P1). These include simple text (e.g., nice) or emojis (e.g., 😉) in comments and captions, and photos of landscapes, food, objects, and pets, among others. P15 articulates this when reflecting on the difference in acceptability of a picture of his partner and a picture of his dog: *"It's not a human, I mean, no one is looking for a dog or trying to steal the identity of a dog, right?"* At the same time, this quote underlines potential assumed harms of personal data flowing out of GenAI, such as identity theft.

Attributes containing people or their information are considered acceptable when they contain: (1) generic information, (2) non-identifiable information, and (3) public-facing information. Generic information is decontextualized and commonplace. Its generic quality means that it is not considered personal or private and, therefore, is acceptable to be used to train GenAI models. It includes *"typical comments that people get that are not too personal or too private, like congratulations or just an emoji"* (P3), and photos of objects everyone has, things everyone does, or places everyone visits. Non-identifiable information does not relate to an identifiable person or the preferences and activities closely tied to people's identities. It includes images of people that are not recognizable, people with heavy filters, and large groups of people where individuals are not distinguishable. Again, it is acceptable as it is not considered personal or private, *"it [selfie with a filter] is me, but not so much me. Let's say I have a semi-identity. It doesn't feel very personal"* (P10).

Public-facing information corresponds to identifiable information online with the intention and expectation of publicity. It includes posts and photos made in a public context or with a public goal (e.g., in a professional context). It is meant to be shared with others and seen by others, and although others don't necessarily include GenAI models, its use is considered acceptable.

> "I think that it [public-facing post] is acceptable, I mean, that's a project I'm currently working on. It's fine because you give the information that you want to be out there, with a certain awareness or responsibility." (P7)

Information flows are **ambiguous** regarding the (potential) applications and the content. When discussing the ambiguity of some attributes, participants often questioned the value of said attribute being used to train GenAI models, *"at first glance, if I see this video, I say, what for? What purpose would it serve?"* (P16). Still, they envisioned potential (un)acceptable applications of their posts within the GenAI lifecycle. That is, specific applications where the information flows are acceptable or unacceptable. Yet, as there is no guarantee posts will be used only in a specific way or context, this leads to ambiguity.

> "I think I could make some exceptions for certain uses. Let's say the residents' association of the [place where the picture was taken] wants to use it. I wouldn't mind if they used it to generate content that promotes tourism in that area, right? But beyond that, I think I would need to know. I would need to be certain of the environment in which that photo could be used. I still wouldn't want this photo to end up somewhere else. As long as what the AI extracts is the information per se, that is, let's say, the landscape or the place itself rather than who is there, I think I wouldn't have any problem." (P16)

In terms of content, participants expressed ambiguity in relation to attributes containing people or their information where information is partly generic, *"I mean, I think if they were even photographs where there were no features, meaning your whole face wasn't visible or things like that, I wouldn't find it acceptable, but I wouldn't find it that serious"* (P14). That is, information is not fully generic – as in acceptable information flows – but generic enough that it could be (mis)used and contextualized in multiple ways. For instance, a picture of a group of friends with generic enough qualities that it might be considered a Stock image, but it is not; or the harmless preferences that can be inferred from content posted on Instagram:

> "When I heard the news, I thought, well, how can they train [GenAI] models with my information? What can they learn? I mean, that I like pasta more than hamburgers? I kind of imagined things that weren't serious. That I wouldn't mind."(P4)

Information flows are considered **unacceptable** as participants recognize attributes as a product of human labor and inherently human. First, people have a role in creating, editing, and posting their photos and captions, *"it is my photo, the aesthetics of the photo are mine, my intellectual property, my creativity"* (P4). It is not acknowledged by Meta. Training GenAI models with posts produced with care by individuals is perceived as exploitative, weird, and abusive. Second, people are captured in these posts. They expose individuals, their faces, bodies, and social relationships to publicity and potentially unknown harms. For instance, by containing or revealing sensitive or intimate information, *"everything seems very, very personal to me, very mine, very much like my feelings"* (P20); linking information related to an individual from multiple places and sources that is intentionally disconnected, *"many people didn't know that I had gone to that event, so it is very upsetting, I don't want to be connected to that."* (P7); or containing or revealing very clearly identifiable information about an individual, such as their whereabouts, their bodies, or their face:

> "It's my face, it's me, so there are no filters, you see it. I would be a little scared, I think there's nothing more to appreciate in that photo than my face, I mean, the background is blurred, I'm in the foreground. Again, it's exposing myself to an environment that can interpret this photo in 80,000 ways, and can use it in 80,000 ways." (P16)

A critical element leading to unacceptability was the presence of other people, both as exposing others and being exposed by others. These types of posts were always considered unacceptable and uncomfortable. Discussing these posts allowed participants to question their (digital) relationships with others and how they are and should be governed (Section 5.2.1). Unacceptability and discomfort were even greater when those other people were minors and individuals not active on social media – and were being exposed through the posts – or were dead – and could not even have a say. Moreover, as most participants have been using Instagram for more than ten years, the temporality of (inter)personal posts also contributes to their unacceptability:

> "I think this might be a bit weird because, like, there are pictures of us as teenagers kissing. Which I feel is quite weird. Well, the fact that it could be used, for example, to generate new images, you know, and then it would have like models of teenagers kissing, and maybe that would make AI generate images of younger people kissing, which might be a bit fucked up. You know, it can go to many levels, and well, yeah." (P5)

5.2.4 (Un)Acceptability Spectrum. We summarise acceptable, ambiguous, and unacceptable information flows into the (un)acceptability spectrum illustrated in Figure 6. This spectrum includes four dimensions[16]:

- **Identifiability:** How identifiable are individuals? To what extent could their likeness be (mis)used through GenAI? It is acceptable for data not associated with people to be used for GenAI. It is unacceptable for data about an identifiable person, or people, to flow into GenAI models – especially as it is unclear in which way it could flow out of these models. Unacceptability increases with the presence and potential exposure of other people.

[16]Note that a single data point could be both identifiable and low labor or general and low labor.

- **Privacy:** How private is the information being captured by the data? To what extent could GenAI expose it? It is acceptable for data generated in people's public spaces, already public-facing, to be used for GenAI. It is unacceptable for data that belongs to people's private spaces to flow into GenAI models.
- **Specificity:** How specific is the information being captured by the data? To what extent could it be enlarged and (mis)interpreted through GenAI? It is acceptable for generic data to be used for GenAI. It is unacceptable for data specific about a person or group of people to flow into and out of GenAI – as it could be potentially harmful.
- **Labor:** How much labor went into creating the data? To what extent could it be undermined through GenAI? It is acceptable for data generated with little human labor to be used for GenAI. It is unacceptable for data that was a product of human labor to flow into or out of GenAI models – especially without attribution and recognition.

5.3 Uncertainties about Data Governance: (Not) Opting-Out

> "I don't know what I expect [opting-out] because I don't know what is happening, to be honest." (P9)

Most participants (14/20) opted out of Meta's personal data collection to train their GenAI models. Thirteen of them during the interview. Notably, one participant in the EU opted out before the interview. She became aware of this possibility through her Instagram feed and followed the steps suggested in one post on her feed. Out of the 13 opt-out requests made by participants during the interviews, 12 were accepted. One participant, in the EU, received a response from Meta stating they were unable to process his request until he submitted evidence that his personal information appears in responses from Meta's GenAI – this aligns with the arbitrarity described by Knibbs [36]. Participants in LATAM sometimes could not access the form to opt out of their devices. In these cases, we filled out the form with them during the interview from the EU. None of the requests made by participants in LATAM were denied.

Similar to the potential applications of personal data flows in Meta's GenAI ecosystem, the opting-out process involved multiple *uncertainties*. The possibility to opt out was uncertain in itself – it was received as a nice surprise. Most participants (18/20) did not know it was possible to opt out – *"I don't think it's that clear that you have this option, I really was not aware of it"* (P9). They repeatedly highlighted how they would have never known how to find the form and follow the process by themselves. In this way, they also surrender to the status quo. They understand that pervasive data collection is the norm and product and service providers are not interested in promoting and supporting individuals to govern their data.

> "It seems to me that they are doing the bare minimum. You know, well, we [Meta] are going to send the notification, but most people are not going to pay attention to it; among the 800 notifications they have, we are going to be able to do what we want." (P13)

Participants expressed further uncertainties around the opting-out process. Especially around what it means for them, their social (and digital) relationships, and their data: *"so, what wouldn't they have access to, or what would they have access to?"* (P7). This echoed their overall uncertainty on how exactly Meta would use their data to train GenAI models (Section 5.2). Similarly, participants were uncertain about the potential short- and long-term implications of opting out as they continued using Meta's products: Would it be the same? Would it change how they can use Meta's products? Would it restrict their access to Meta's AI tools? Moreover, participants were uncertain about the uncertainties of the opting-out process as described by Meta on their Privacy Policy and opt-out form (e.g., What if their data was already being used to train GenAI models? How would opting-out change that?) They found the information available insufficient and ambiguous, and the process arbitrary.

> "I don't know; for example, it's not clear to me if once you decide to opt-out, then that only applies to posts after [opting-out] or if it's retroactive. It's very ambiguous, and [Meta] doesn't guarantee that they won't use them [posts]. They really do review it [opt-out request], and they might say no, it doesn't seem like that to us. It seems very arbitrary to me." (P14)

Whether to opt out was also an uncertain decision – participants expressed doubt and concern about it. Amongst participants who decided not to opt out, there were three distinct motivations: (1) wanting to contribute to GenAI models, (2) wanting to wait and see how GenAI models are deployed by Meta, and (3) not trusting Meta. The first one stems from recognizing the role (personal) data plays in the development and deployment of GenAI models and wanting better models (for themselves and other people), *"the bigger the model, the better it works. The more people opt out, the less it works"* (P6). Note that even participants who did not want to opt-out, for this reason, would have wanted some posts (e.g., family pictures) not to be used to train GenAI models. The second one derived from further uncertainty about Meta's GenAI models and wanting more certainty before making a decision, *"I don't want to [opt-out] yet, but I like that the option is there"* (P10). The third one further illustrates how participants *surrender to powerlessness* (Section 5.1), they acknowledge the power imbalance between them and Meta and don't trust the mechanisms set in place by Meta to (slightly) tumble it.

> "They [Meta] will continue to use it and not care what people say. They only have that [opt-out] form because they have to comply with the rule that «we are giving people the ability to do this» but they are not actually doing it." (P3)

The motivation of participants who decided to opt out stemmed from them considering Meta's processing of their data to train GenAI models unacceptable, uncomfortable, and creepy (Section 5.2). Note that even participants who wanted to opt out would have accepted some of their posts (e.g., objects, landscapes) being used to train GenAI models. Participants perceived the opt-out process as relatively easy, with the caveat that we supported them in filling the required text box describing how data processing impacts them, and we directed them to the form – *"the difficult thing is knowing where it [the form] is to do it"* (P8). Nonetheless, participants were baffled by Meta having to review and accept their

Table 2: Participants self-reported prior knowledge of Meta's data collection practices, Meta's GenAI policy update, and their opting-out requests before and during the interviews.

		Participants in LATAM	*Participants in the EU*
Prior Knowledge	*Meta's Data Collection Practices*	2/10	2/10
	Meta's Privacy Policy Update	1/10	3/10
Opting-Out	*Prior to the Interview*	0/10	1/10
	During the Interview	7/10	6/10

opt-out request: *"Why do they have to approve something that I am asking for and that I as a user am refusing?"* (P7). It made it hard for participants to envision the potential impact of opting out as they were expectant of Meta's decision. It also intensified participants' feelings of powerlessness: even when given a choice, that choice is not entirely theirs but mediated by Meta. In most cases, this decision came within minutes, during the interview. Even when it was a positive one, it introduced further uncertainties:

> "They have already accepted it. What is not clear, for example, is what it means that they have accepted it. It says they will not process my information, but what does that mean? How do I know that they are not processing it?" (P8)

Overall, the decision to opt-out was permeated by multiple uncertainties; the opt-out process, its practicalities, and implications were uncertain, and the outcomes of the process led to further uncertainties and unanswered questions.

5.4 Different Starting Points: Perspectives from EU and LATAM

Participants in the EU and LATAM expressed similar opinions about their data flowing into and out of GenAI models and had similar prior knowledge of Meta's data collection practices and policy update (Table 2). One important difference was their awareness of local legislation and the rights these grant them over their personal data. Most participants in the EU (7/10) were familiar with the GDPR and mentioned it during the interview, for instance, concerning how Meta was required to inform them of their policy update and allow them to opt-out.

> "The good thing is that we are in Europe, we are legally more protected than in other countries because of GDPR, they [Meta] need to communicate with us." (P6.)

Awareness of the GDPR carried the expectation of data protection it being enforceable on their data even when it was created and posted from abroad (e.g., tagged posts where the primary user was based in LATAM) – *"This is a bit fucked-up, because if you, I mean, if you were within European territory, I would expect that you would be under European laws."* (P5). In contrast, none of the participants in LATAM (0/10) were familiar with local legislation or mentioned it during the interviews. For them, the interview – and later the Zine – represented a first encounter with actionable information about data governance and the specific rights they have over their data. This does not illustrate divergent perspectives but a different starting point. It demonstrates the importance of leveling the field when conducting research across continents and regulations – creating a space for participants to start from the same place, even when it might lead to redundancy for some.

6 Discussion

We synthesize our findings into three provocations on personal data flowing into and out of GenAI models. They are meant to be generative and open up discussion. With our provocations, we argue for a position to highlight shortcomings in the consensus, raise new questions, and challenge the status quo [6, 13].

6.1 Should We Surrender to Powerlessness?

With the theme of **surrendering to powerlessness**, we describe how participants feel they lack the capacity to stop Meta from exploiting their data and to react to policy updates with which they might disagree. Powerlessness is multifaceted: insufficient awareness, normalization, and acceptance of the status quo, oversimplified categories of who is a user and therefore gets to opt-out, and non-enforceable regulations all contribute to powerlessness. Powerlessness increases through the pivotal role Meta's products play in how (most) people relate with others and their livelihoods – making it almost impossible to stop using them. Our findings on powerlessness mirror reactions to policy updates by other service providers such as WhatsApp [29] and Reddit [54]. People are outraged and confused; they are eager to take action, and although in some cases successfully do so (e.g., [54]), in most cases, they fail to account for the relational implications of doing so (e.g., stopping using a service requires convincing others to do so [29, 56]) and revert to acceptance. Legislative differences between continents accentuate this tendency towards acceptance [7, 46], as people's ability to exercise their rights depends on where they are based and continue to perpetuate colonial hierarchies [35, 52]. What alternatives are then to shift the power imbalance between individuals and tech giants like Meta across continents?

The power imbalance between individuals and service providers and the extractivist and exploitative nature of personal data processing have been problematized across disciplines (e.g., [2, 35, 64, 74, 78]). Previous work has explored how people can shift this power imbalance by leveraging their rights [4, 53, 78]. We echo these perspectives, framing tech giants' reliance on personal data as an opportunity to reverse the power imbalance. Based on insights from participants in our study, we propose three generative spaces where greater data leverage can be fostered:

- **Opt-Out Strikes,** emphasizing the insufficiency of policy updates without affirmative and informative consent practices by actively withdrawing consent and access to data as a response. They require rights to object or restrict processing and delete data. On a small scale, they were enacted by participants who opted-out of Meta's GenAI data processing. Still, questions remained about Meta's honoring their decision.
- **Relational Data Poisoning,** responding to the discomfort expressed by participants concerning other people's privacy by *"poisoning"* (e.g., applying a filter, cropping, editing) data when it is shared and relational to prevent the non-consensual use and exploitation of others. It requires familiarity with digital tools and rights to delete data. On a small scale, it was enacted by participants who cropped or deleted posts of/with minors, friends, and family after the interview. Participants suggested tech companies could facilitate relational poisoning by identifying other people in the data and proactively excluding them from algorithmic processing.
- **Conscious Data Labor,** foregrounding individuals as creators, sources, and ultimate beneficiaries of data by transferring it to platforms where they can derive value. It requires rights to access or data portability. It was discussed by P15 and enacted by him on a small scale. P15 accounts for the effort and labor he puts into generating his data. He has uploaded it to platforms other than Instagram, where he can monetize it or at least actively and explicitly consent to its use in GenAI.

We recognize that the impact of these strategies is limited and transfer the responsibility to individuals. Yet, we argue some of the power should fall on individuals and outline these as pragmatic ways of achieving that. Although these might be ineffective on a small scale, prior research demonstrates that even a small data strike can substantially reduce the utility of a recommender system without sacrificing access to the underlying services [78].

Provocation 1: We are not powerless data subjects, but powerful data contributors. Individuals contribute greatly through the data they generate and can leverage their contributions to shift the power imbalance between them and service providers, whether through (1) opt-out strikes, (2) relational data poisoning, or (3) conscious data labor.

6.2 From Personal to (Inter)Personal and Relational Data

Data is often portrayed as a resource to be refined and exploited [63], "the new gold" [68] or " the new oil" [26]. Thus justifying its processing across various applications, including GenAI. These perspectives fail to account for the entanglements between data and people and how "data are people" [85]. Loukissas [45] argues "all data are local" and calls for acknowledging how data(sets) are created by people in specific configurations (e.g., times, places, bodies, devices). Desjardins et al. [23] highlight how data are lively and dynamic. We continue to expand on these ways of describing and challenging data by emphasizing how data and its algorithmic derivatives are (inter)personal and relational. They capture, contain, and reveal information about a person and other people in their lives and physical environments. Although these aspects are more evident in social media, social and relational by design, they expand to other forms of data and other settings, such as the (smart) home, where families share a physical space and the digital technologies within (e.g., [10, 30, 40, 79]).

Throughout our study, we discuss concepts and protections across legislation that emphasize the personal nature of personal data, that is, related to an identified or identifiable person. These overlap with how service providers conceptualize their users: a person. Yet, users and their data – much like people – are (inter)personal: shared, social, and relational. Data's (inter)personal nature first emerges by analyzing Meta's privacy update. They describe how data from a person – whether or not they are users of Meta – might leak or spread to other people's accounts and data. Thus, people's social relationships permeate their digital lives and their data. The (inter)personal nature of data is further discussed by participants as they encounter other people in their data and/or find themselves in other people's data. Recognizing data as (inter)personal invited participants to reconsider critical aspects around ownership (e.g., Whose data is it?), agency (e.g., Who should have a say?), and consent (e.g., Should I have asked others?). These considerations highlight one of the shortcomings of design, policy, and current forms of data governance: they are centered around individuals. Who should have rights over shared data? How could these rights be enforced across individual and legislative boundaries?

Provocation 2: Data are not personal, but (inter)personal, shared, and relational. We should establish design, policy, and data governance approaches that go beyond the individual and address data's (inter)personal nature. Especially as (inter)personal data flows into and out of GenAI models mediated by individuals with different preferences and sensibilities.

6.3 Towards Encountering the Unknown through Design

> "The activity we did of looking at my posts, it makes you think a lot about what you share and what you don't. At this point, it [data] is not even directed at other people, but literally at the entire world, and at a system that will know more about you than you do yourself." (P19)

Our results underline the many uncertainties, concerns, and questions surrounding how personal data flows into and out of GenAI. These uncertainties contribute to difficulties in articulating and addressing algorithmic aspects of privacy and privacy risks [30, 41, 70]. Privacy can be violated in non-obvious ways that are difficult to articulate and need to be disentangled from data, and algorithms. We delimited the multiple uncertainties around data flowing into and out of GenAI models as "the unknown." It is partly due to processes of collecting, tracking, aggregating, and exploiting personal data being *invisible* to individuals [79] and, in some cases, manipulative [73]. It is aggravated by information on privacy policies and consent notices not containing clear information [22] nor the information users wish to know [41]. Further, it expands

with how (Gen)AI technologies create new types of privacy risks or exacerbate known privacy risks [42].

As GenAI and AI technologies pose new – and unknown – privacy risks; it is paramount to support individuals in grappling with these. We argue that overcoming the unknown(s) requires design efforts that invite people to encounter data and their algorithmic derivatives in different ways grounded in their data and lived experiences. This should be done across user groups, considering conditions like education levels, environment, and socioeconomic conditions. We support our argument with observations from participants, such as the quote above, where P19 described the value of thinking about the algorithmic aspects of privacy through her own Instagram posts. Privacy research has largely focused on (re)imagining and (re)designing online consent processes, privacy notices, and privacy labels aiming to better inform or educate people (e.g., [16, 22, 76, 82, 84]). We recognize the value of these efforts, yet they don't fully grapple with the fact that frequently underlying uncertainties remain uncertain, abstract, and opaque. Discussing personal data, an abstract construct, is not the same as discussing about people's Instagram posts, a concrete construct. Through concrete discussions, we found that people would prefer to actively delimit how their data flows into GenAI. Some would find it acceptable to have parts of their data flowing into and out of GenAI models. We see this as an opportunity to balance data contributions to GenAI models with people's preferences and expectations that can be enabled through design. Design can contribute to creating concrete encounters that invite people to overcome the unknown, by knowing and feeling their personal data and algorithmic derivatives (e.g., [32, 66]). These can promote conversations beyond identifying privacy concerns towards determining how to respond to them (or not).

Provocation 3: We should not design to provide better or more information about personal data and GenAI, but rather ways that invite people to encounter <u>their</u> data and algorithmic derivatives. The governance of (inter)personal data into and out of GenAI models is mediated by the unknown. How data flows is difficult to understand and articulate. It requires ways of knowing, beyond receiving information, that promote a felt understanding of data collection and exploitation.

7 Conclusion

This paper contributes with a qualitative interview study and analysis of how people experience personal data collection in GenAI ecosystems. We focused on personal data collection by Meta, specifically Instagram, in line with their recent policy update on processing user data to train GenAI models. We conducted semi-structured interviews with 20 Latin American Instagram users in the EU and LATAM. We discussed the acceptability of their data flowing into and out of GenAI through different scenarios. We unpack power dynamics in data collection, the (inter)personal nature of data, and the multiple unknowns concerning GenAI. We delimit an (un)acceptability spectrum describing the acceptability of personal information being processed for GenAI according to four dimensions: identifiability, privacy, specificity, and labor. We propose and discuss a set of generative provocations highlighting the shortcomings of processing personal data to train GenAI models and open further discussion.

Acknowledgments

We thank the participants for their generous and insightful contribution to this paper and the many exciting discussions that followed. Special thanks to Gabi, Willy, Clau, Pauli, Aloe, and Berto for supporting this work. Thank you to the anonymous reviewers for their thoughtful and thorough feedback.

References

[1] Irwin Altman. 1975. *The Environment and Social Behavior: Privacy, Personal Space, Territory, Crowding.*
[2] Mark Andrejevic. 2014. The Big Data Divide. (2014).
[3] Noah Apthorpe, Yan Shvartzshnaider, Arunesh Mathur, Dillon Reisman, and Nick Feamster. 2018. Discovering Smart Home Internet of Things Privacy Norms Using Contextual Integrity. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 2 (July 2018), 59:1–59:23. https://doi.org/10.1145/3214262
[4] Stefan Baack. 2015. Datafication and empowerment: How the open data movement re-articulates notions of democracy, participation, and journalism. *Big Data & Society* 2, 2 (Dec. 2015), 2053951715594634. https://doi.org/10.1177/2053951715594634 Publisher: SAGE Publications Ltd.
[5] Andy Baio. 2022. Exploring 12 Million of the 2.3 Billion Images Used to Train Stable Diffusion's Image Generator. https://waxy.org/2022/08/exploring-12-million-of-the-images-used-to-train-stable-diffusions-image-generator/
[6] Jeffrey Bardzell and Shaowen Bardzell. 2015. *Humanistic HCI.* Springer International Publishing, Cham. https://doi.org/10.1007/978-3-031-02214-2
[7] Benedicta Ehimuan, Ogugua Chimezie, Ob, Onyinyechi Vivian Akagha, Oluwatosin Reis, and Bisola Beatrice Oguejiofor. 2024. Global data privacy laws: A critical review of technology's impact on user rights. *World Journal of Advanced Research and Reviews* 21, 2 (Feb. 2024), 1058–1070. https://doi.org/10.30574/wjarr.2024.21.2.0369
[8] Anne Bowser, Katie Shilton, Jenny Preece, and Elizabeth Warrick. 2017. Accounting for Privacy in Citizen Science: Ethical Research in a Context of Openness. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17).* Association for Computing Machinery, New York, NY, USA, 2124–2136. https://doi.org/10.1145/2998181.2998305
[9] Alex Bowyer, Jack Holt, Josephine Go Jefferies, Rob Wilson, David Kirk, and Jan David Smeddinck. 2022. Human-GDPR Interaction: Practical Experiences of Accessing Personal Data. In *CHI Conference on Human Factors in Computing Systems.* ACM, New Orleans LA USA, 1–19. https://doi.org/10.1145/3491102.3501947
[10] Alex Bowyer, Kyle Montague, Stuart Wheater, Ruth McGovern, Raghu Lingam, and Madeline Balaam. 2018. Understanding the Family Perspective on the Storage, Sharing and Handling of Family Civic Data. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems.* ACM, Montreal QC Canada, 1–13. https://doi.org/10.1145/3173574.3173710
[11] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (Jan. 2006), 77–101. https://doi.org/10.1191/1478088706qp063oa
[12] Virginia Braun and Victoria Clarke. 2013. *Successful Qualitative Research: A Practical Guide for Beginners.* SAGE Publications LTD. https://uk.sagepub.com/en-gb/eur/successful-qualitative-research/book233059
[13] Barry Brown, Alexandra Weilenmann, Donald McMillan, and Airi Lampinen. 2016. Five Provocations for Ethical HCI Research. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems.* ACM, San Jose California USA, 852–863. https://doi.org/10.1145/2858036.2858313
[14] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS '20).* Curran Associates Inc., Red Hook, NY, USA, 1877–1901.
[15] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting Training Data from Large Language Models. https://doi.org/10.48550/arXiv.2012.07805 arXiv:2012.07805 [cs].

[16] Claire C Chen, Dillon Shu, Hamsini Ravishankar, Xinran Li, Yuvraj Agarwal, and Lorrie Faith Cranor. 2024. Is a Trustmark and QR Code Enough? The Effect of IoT Security and Privacy Label Information Complexity on Consumer Comprehension and Behavior. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–32. https://doi.org/10.1145/3613904.3642011

[17] Andy Crabtree, Peter Tolmie, and Will Knight. 2017. Repacking 'Privacy' for a Networked World. *Computer Supported Cooperative Work (CSCW)* 26, 4-6 (Dec. 2017), 453–488. https://doi.org/10.1007/s10606-017-9276-y

[18] Congreso de Colombia. 2012. Ley 1581 de 2012 - Gestor Normativo. https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=49981

[19] Camara de Diputados del Honorable Congreso de la Union. 2010. Ley Federal de Protección de Datos Personales en Posesión de los Particulares. http://www.gob.mx/indesol/documentos/ley-federal-de-proteccion-de-datos-personales-en-posesion-de-los-particulares

[20] Honorable Congreso de la Nacion Argentina. 2000. Ley de Proteccion de Datos Personales, Ley 25326. https://www.argentina.gob.ar/justicia/derechofacil/leysimple/datos-personales

[21] Congreso de la Republica del Peru. 2011. Ley N.° 29733 - Derecho Fundamental a la Proteccion de los Datos Personales. https://www.gob.pe/institucion/congreso-de-la-republica/normas-legales/243470-29733

[22] Martin Degeling, Christine Utz, Christopher Lentzsch, Henry Hosseini, Florian Schaub, and Thorsten Holz. 2019. We Value Your Privacy ... Now Take Some Cookies: Measuring the GDPR's Impact on Web Privacy. In *Proceedings 2019 Network and Distributed System Security Symposium*. https://doi.org/10.14722/ndss.2019.23378 arXiv:1808.05096 [cs].

[23] Audrey Desjardins, Heidi R. Biggs, Cayla Key, and Jeremy E. Viny. 2020. IoT Data in the Home: Observing Entanglements and Drawing New Encounters. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–13. https://doi.org/10.1145/3313831.3376342

[24] Audrey Desjardins, Jena McWhirter, Justin Petelka, Chandler Simon, Yuna Shin, Ruby K Peven, and Philbert Widjaja. 2023. On the Making of Alternative Data Encounters: The Odd Interpreters. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–20. https://doi.org/10.1145/3544548.3581323

[25] Mindy Nunez Duffourc, Sara Gerke, and Konrad Kollnig. 2024. Privacy of Personal Data in the Generative AI Data Lifecycle. *New York University Journal of Intellectual Property & Entertainment Law* 13, 2 (2024), 219–268.

[26] The Economist. 2017. The world's most valuable resource is no longer oil, but data. *The Economist* (2017). https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data

[27] European Parliament. Directorate General for Parliamentary Research Services. 2022. *Governing data and artificial intelligence for all: models for sustainable and just data governance*. Publications Office, LU. https://data.europa.eu/doi/10.2861/915401

[28] Abenezer Golda, Kidus Mekonen, Amit Pandey, Anushka Singh, Vikas Hassija, Vinay Chamola, and Biplab Sikdar. 2024. Privacy and Security Concerns in Generative AI: A Comprehensive Survey. *IEEE Access* 12 (2024), 48126–48144. https://doi.org/10.1109/ACCESS.2024.3381611 Conference Name: IEEE Access.

[29] Carla F. Griggio, Midas Nouwens, and Clemens Nylandsted Klokmose. 2022. Caught in the Network: The Impact of WhatsApp's 2021 Privacy Policy Update on Users' Messaging App Ecosystems. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–23. https://doi.org/10.1145/3491102.3502032

[30] Alejandra Gómez Ortega, Jacky Bourgeois, and Gerd Kortuem. 2023. What is Sensitive About (Sensitive) Data? Characterizing Sensitivity and Intimacy with Google Assistant Users. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3544548.3581164

[31] Alejandra Gómez Ortega, Jacky Bourgeois, and Gerd Kortuem. 2024. Participation in Data Donation: Co-Creative, Collaborative, and Contributory Engagements with Athletes and their Intimate Data. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference (DIS '24)*. Association for Computing Machinery, New York, NY, USA, 2388–2402. https://doi.org/10.1145/3643834.3661503

[32] Alejandra Gómez Ortega, Renee Noortman, Jacky Bourgeois, and Gerd Kortuem. 2024. Dataslip: Into the Present and Future(s) of Personal Data. In *Proceedings of the Eighteenth International Conference on Tangible, Embedded, and Embodied Interaction (TEI '24)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3623509.3633388

[33] Elisa Harlan and Katharina Brunner. 2024. We Are All Raw Material for AI. https://interaktiv.br.de/ki-trainingsdaten/en/

[34] Yuheng Hu, Lydia Manikonda, and Subbarao Kambhampati. 2014. What We Instagram: A First Analysis of Instagram Photo Content and User Types. *Proceedings of the International AAAI Conference on Web and Social Media* 8, 1 (May 2014), 595–598. https://doi.org/10.1609/icwsm.v8i1.14578 Number: 1.

[35] Lauren Klein and Catherine D'Ignazio. 2024. Data Feminism for AI. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Rio de Janeiro Brazil, 100–112. https://doi.org/10.1145/3630106.3658543

[36] Knibbs. 2023. Artists Allege Meta's AI Data Deletion Request Process Is a 'Fake PR Stunt' | WIRED. https://www.wired.com/story/meta-artificial-intelligence-data-deletion/

[37] Priya C. Kumar, Mega Subramaniam, Jessica Vitak, Tamara L. Clegg, and Marshini Chetty. 2020. Strengthening Children's Privacy Literacy through Contextual Integrity. *Media and Communication* 8, 4 (Nov. 2020), 175–184. https://www.cogitatiopress.com/mediaandcommunication/article/view/3236

[38] Priya C. Kumar, Michael Zimmer, and Jessica Vitak. 2024. A Roadmap for Applying the Contextual Integrity Framework in Qualitative Privacy Research. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (April 2024), 1–29. https://doi.org/10.1145/3653710

[39] Albrecht Kurze, Andreas Bischof, Sören Totzauer, Michael Storz, Maximilian Eibl, Margot Brereton, and Arne Berger. 2020. Guess the Data: Data Work to Understand How People Make Sense of and Use Simple Sensor Data from Homes. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–12. https://doi.org/10.1145/3313831.3376273

[40] Hyosun Kwon, Joel E. Fischer, Martin Flintham, and James Colley. 2018. The Connected Shower: Studying Intimate Data in Everyday Life. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 4 (Dec. 2018), 1–22. https://doi.org/10.1145/3287054

[41] Lin Kyi, Abraham Mhaidli, Cristiana Teixeira Santos, Franziska Roesner, and Asia J. Biega. 2024. "It doesn't tell me anything about how my data is used": User Perceptions of Data Collection Purposes. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–12. https://doi.org/10.1145/3613904.3642260

[42] Hao-Ping (Hank) Lee, Yu-Ju Yang, Thomas Serban Von Davier, Jodi Forlizzi, and Sauvik Das. 2024. 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)*. Association for Computing Machinery, New York, NY, USA, 1–19. https://doi.org/10.1145/3613904.3642116

[43] Lauren Leffer. 2023. Your Personal Information Is Probably Being Used to Train Generative AI Models. https://www-scientificamerican-com.tudelft.idm.oclc.org/article/your-personal-information-is-probably-being-used-to-train-generative-ai-models/

[44] Ámbito Jurídico Legis. [n. d.]. Protegen derechos a la intimidad y buen nombre frente a la difusión no consentida de imágenes en redes sociales. https://ambitojuridico.com/noticias/tecnologia/constitucional-y-derechos-humanos/protegen-derechos-la-intimidad-y-buen-nombre

[45] Yanni Alexander Loukissas. 2019. *All Data Are Local: Thinking Critically in a Data-Driven Society*. The MIT Press. https://doi.org/10.7551/mitpress/11543.001.0001

[46] Aofei Lv and Ting Luo. 2018. Authoritarian Practices in the Digital Age| Asymmetrical Power Between Internet Giants and Users in China. *International Journal of Communication* 12, 0 (Sept. 2018), 19. https://ijoc.org/index.php/ijoc/article/view/8543 Number: 0.

[47] Meta. 2024. Generative AI at Meta. https://www.facebook.com/privacy/guide/genai/

[48] Meta. 2024. How Meta uses information for generative AI models and features. https://www.facebook.com/privacy/genai

[49] Meta. 2024. Llama 2: open source, free for research and commercial use. https://llama.meta.com/llama2/

[50] Meta. 2024. Meta AI: Expand your world with Meta AI. https://ai.meta.com/meta-ai

[51] Meta. 2024. Object to Your Information Being Used for AI at Meta. https://help.instagram.com/contact/233964459562201

[52] Milagros Miceli and Julian Posada. 2022. The Data-Production Dispositif. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (Nov. 2022), 1–37. https://doi.org/10.1145/3555561

[53] Stefania Milan and Lonneke Van Der Velden. 2016. The Alternative Epistemologies of Data Activism. *Digital Culture & Society* 2, 2 (Dec. 2016), 57–74. https://doi.org/10.14361/dcs-2016-0205

[54] Sara Morrison. 2023. The ongoing and increasingly weird Reddit blackout, explained. https://www.vox.com/technology/2023/6/14/23760738/reddit-blackout-explained-subreddit-apollo-third-party-apps

[55] Sara Morrison. 2023. The tricky truth about how generative AI uses your data. https://www.vox.com/technology/2023/7/27/23808499/ai-openai-google-meta-data-privacy-nope

[56] Casey Newton. 2021. Warning Signal: the messaging app's new features are causing internal turmoil. https://www.theverge.com/22249391/signal-app-abuse-messaging-employees-violence-misinformation

[57] Helen Nissenbaum. 2004. PRIVACY AS CONTEXTUAL INTEGRITY. *Washington Law Review* 79 (2004).

[58] Helen Nissenbaum. 2019. Contextual Integrity Up and Down the Data Food Chain. *Theoretical Inquiries in Law* 20, 1 (March 2019), 221–256. https://doi.org/10.1515/til-2019-0008

[59] Commissioner Office of the. 2018. Payment and Reimbursement to Research Subjects. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/payment-and-reimbursement-research-subjects Publisher: FDA.

[60] Leysia Palen and Paul Dourish. 2003. Unpacking "Privacy" for a Networked World. *NEW HORIZONS* 5 (2003).

[61] Luisa Parraguez Kobek and Erick Caldera. 2016. Cyber Security and Habeas Data: The Latin American response to information security and data protection. (Nov. 2016). https://bdigital.uexternado.edu.co/handle/001/8397 Publisher: Facultad de Finanzas, Gobierno y Relaciones Internacionales.

[62] Jessica Pater, Amanda Coupe, Rachel Pfafman, Chanda Phelan, Tammy Toscos, and Maia Jacobs. 2021. Standardizing Reporting of Participant Compensation in HCI: A Systematic Literature Review and Recommendations for the Field. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–16. https://doi.org/10.1145/3411764.3445734

[63] Barbara Prainsack. 2019. Logged out: Ownership, exclusion and public value in the digital data and information commons. *Big Data & Society* 6, 1 (Jan. 2019), 205395171982977. https://doi.org/10.1177/2053951719829773

[64] Barbara Prainsack, Seliem El-Sayed, Nikolaus Forgó, Łukasz Szoszkiewicz, and Philipp Baumer. 2022. Data solidarity: a blueprint for governing health futures. *The Lancet Digital Health* 4, 11 (Nov. 2022), e773–e774. https://doi.org/10.1016/S2589-7500(22)00189-3

[65] Presidência da República. 2018. L13709. https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm

[66] Irina Shklovski and Erik Grönvall. 2020. CreepyLeaks: Participatory Speculation Through Demos. In *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*. ACM, Tallinn Estonia, 1–12. https://doi.org/10.1145/3419249.3420168

[67] Irina Shklovski, Scott D. Mainwaring, Halla Hrund Skúladóttir, and Höskuldur Borgthorsson. 2014. Leakiness and creepiness in app space: perceptions of privacy and mobile app use. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Toronto Ontario Canada, 2347–2356. https://doi.org/10.1145/2556288.2557421

[68] Sandro Shubladze. 2023. Council Post: How To Make Use Of The New Gold: Data. https://www.forbes.com/councils/forbestechcouncil/2023/03/27/how-to-make-use-of-the-new-gold-data/ Section: Innovation.

[69] Yan Shvartzshnaider, Schrasing Tong, Thomas Wies, Paula Kift, Helen Nissenbaum, Lakshminarayanan Subramanian, and Prateek Mittal. 2016. Learning Privacy Expectations by Crowdsourcing Contextual Informational Norms. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 4 (Sept. 2016), 209–218. https://doi.org/10.1609/hcomp.v4i1.13271

[70] Patrick Skeba and Eric P. S. Baumer. 2020. Informational Friction as a Lens for Studying Algorithmic Aspects of Privacy. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2 (Oct. 2020), 101:1–101:22. https://doi.org/10.1145/3415172

[71] Solid. 2024. Community - Solid. https://solidproject.org/community

[72] Jhoandry Suarez. 2024. ¿Qué significa que Meta entrene su IA con mis datos y cómo puedo oponerme desde Latinoamérica? https://colombiacheck.com/investigaciones/que-significa-que-meta-entrene-su-ia-con-mis-datos-y-como-puedo-oponerme-desde

[73] Lorena Sánchez Chamorro, Romain Toebosch, and Carine Lallemand. 2024. Manipulative Design and Older Adults: Co-Creating Magic Machines to Understand Experiences of Online Manipulation. In *Designing Interactive Systems Conference*. ACM, IT University of Copenhagen Denmark, 668–684. https://doi.org/10.1145/3643834.3661513

[74] Linnet Taylor. 2021. Public Actors Without Public Values: Legitimacy, Domination and the Regulation of the Technology Sector. *Philosophy & Technology* 34, 4 (Dec. 2021), 897–922. https://doi.org/10.1007/s13347-020-00441-4

[75] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. LaMDA: Language Models for Dialog Applications. https://doi.org/10.48550/arXiv.2201.08239 arXiv:2201.08239 [cs].

[76] Christine Utz, Martin Degeling, Sascha Fahl, Florian Schaub, and Thorsten Holz. 2019. (Un)informed Consent: Studying GDPR Consent Notices in the Field. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*. ACM, London United Kingdom, 973–990. https://doi.org/10.1145/3319535.3354212

[77] Nicholas Vincent, Brent Hecht, and Shilad Sen. 2019. "Data Strikes": Evaluating the Effectiveness of a New Form of Collective Action Against Technology Companies. In *The World Wide Web Conference (WWW '19)*. Association for Computing Machinery, New York, NY, USA, 1931–1943. https://doi.org/10.1145/3308558.3313742

[78] Nicholas Vincent, Hanlin Li, Nicole Tilly, Stevie Chancellor, and Brent Hecht. 2021. Data Leverage: A Framework for Empowering the Public in its Relationship with Technology Companies. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 215–227. https://doi.org/10.1145/3442188.3445885

[79] Zezhong Wang, Shunming Wang, Matteo Farinella, Dave Murray-Rust, Nathalie Henry Riche, and Benjamin Bach. 2019. Comparing Effectiveness and Engagement of Data Comics and Infographics. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–12. https://doi.org/10.1145/3290605.3300483

[80] Mark Warner, Agnieszka Kitkowska, Jo Gibbs, Juan F. Maestre, and Ann Blandford. 2020. Evaluating 'Prefer not to say' Around Sensitive Disclosures. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–13. https://doi.org/10.1145/3313831.3376150

[81] Jason Wiese, Sauvik Das, Jason I. Hong, and John Zimmerman. 2017. Evolving the Ecosystem of Personal Behavioral Data. *Human–Computer Interaction* 32, 5-6 (Nov. 2017), 447–510. https://doi.org/10.1080/07370024.2017.1295857

[82] Maximiliane Windl and Sebastian S. Feger. 2024. Designing Interactive Privacy Labels for Advanced Smart Home Device Configuration Options. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference (DIS '24)*. Association for Computing Machinery, New York, NY, USA, 3372–3388. https://doi.org/10.1145/3643834.3661527

[83] Amy Winograd. 2023. LOOSE-LIPPED LARGE LANGUAGE MODELS SPILL YOUR SECRETS: THE PRIVACY IMPLICATIONS OF LARGE LANGUAGE MODELS. 36, 2 (2023).

[84] Yaqing Yang, Tony W Li, and Haojian Jin. 2024. On the Feasibility of Predicting Users' Privacy Concerns using Contextual Labels and Personal Preferences. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–20. https://doi.org/10.1145/3613904.3642500

[85] Matthew Zook, Solon Barocas, Danah Boyd, Kate Crawford, Emily Keller, Seeta Peña Gangadharan, Alyssa Goodman, Rachelle Hollander, Barbara A. Koenig, Jacob Metcalf, Arvind Narayanan, Alondra Nelson, and Frank Pasquale. 2017. Ten simple rules for responsible big data research. *PLOS Computational Biology* 13, 3 (March 2017), e1005399. https://doi.org/10.1371/journal.pcbi.1005399