

THE UNIVERSITY OF CHICAGO

REDACTION AND EXHIBITION: ENACTING AND COMMUNICATING PRIVACY
BY EXPLORING IDENTIFIABILITY AND ARTISTIC PRACTICE

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF COMPUTER SCIENCE

BY
EMMA ISABELLA CAROLINE PETERSON

CHICAGO, ILLINOIS

JUNE 2026

Copyright © 2026 by Emma Isabella Caroline Peterson
All Rights Reserved

*Allting finns om man söker och har tid, det vill säga råd att söka och under sökandet är
man fri och hittar saker som man inte alls hade tänkt sig.*

*[You can find anything if you search and have time, that is, if you can afford to look. And
while you're searching, you are free and find things that you never expected.]*

—Tove Jansson, *Sommarboken*

TABLE OF CONTENTS

LIST OF FIGURES	vii
LIST OF TABLES	viii
ACKNOWLEDGMENTS	ix
ABSTRACT	xi
1 INTRODUCTION	1
1.1 Dissertation Overview	3
1.1.1 Identifiability and Automated Redaction	4
1.1.2 Bringing Humans into the Loop in Automated Redaction	5
1.1.3 Exploring Privacy in Art and Applications to Privacy Communications	6
1.1.4 Designing HCI Privacy Provocations Using Artistic Mechanisms	7
2 EXPLORING IDENTIFIABILITY AND AUTOMATED REDACTION PROCESSES IN UNSTRUCTURED TEXT DATA	8
2.1 Dissertation Author’s Contributions	8
2.2 Motivation	8
2.3 Related Work	11
2.3.1 Key Redaction Techniques	12
2.3.2 Commercial Redaction Tools	13
2.3.3 Using LLMs For Redaction	13
2.4 Formative Interviews: Methods	14
2.4.1 Recruitment	15
2.4.2 Interview Structure	15
2.4.3 Data Analysis and Limitations	16
2.5 Formative Interviews: Results	17
2.5.1 Participants and Their Corpora	18
2.5.2 Views on Data Sharing and Redaction	18
2.5.3 Threat Models and Concerns	21
2.5.4 Logistics of Data Sharing and Redaction	23
2.6 Redaction Framework	24
2.7 Benchmarking Study: Methods	28
2.7.1 Datasets	29
2.7.2 Choice and Configuration of Legacy Tools	31
2.7.3 Choice and Configuration of LLMs	33
2.8 Results: Benchmarking Study	35
2.8.1 TalkBank Data	35
2.8.2 Synthetic Data	38
2.9 Discussion and Conclusions	39

3	MARI: A SEMI-AUTOMATED, HUMAN-IN-THE-LOOP GRAPHICAL REDAC-	
	TION WORKFLOW	44
3.1	Dissertation Author’s Contributions	44
3.2	Motivation	44
3.3	Related work	48
3.3.1	Automated Redaction	48
3.3.2	Redaction with LLMs	49
3.3.3	Graphical Workflows	50
3.4	MARI Design	51
3.4.1	Automated Redaction Pipeline	51
3.4.2	User-Facing Redaction Interface	53
3.5	Methodology: User study	55
3.5.1	Recruitment	58
3.5.2	Study Design	59
3.5.3	Data Collection	61
3.5.4	Datasets	62
3.5.5	Data Analysis	63
3.5.6	Preliminary Observations	63
3.6	Discussion and Conclusions	65
4	PRIVACY IMITATES ART: DRAWING LESSONS IN PRIVACY COMMUNICA-	
	TION FROM ARTISTS	67
4.1	Dissertation Author’s Contributions	67
4.2	Motivation	67
4.3	Related Work	70
4.3.1	Engagement Gaps in Usable Privacy	71
4.3.2	Surveillance and Privacy Art	72
4.3.3	HCI and the Arts	72
4.3.4	Interactive Design	73
4.4	Methods	74
4.4.1	Creating a Database of Privacy Art	74
4.4.2	Artwork Selection Criteria	75
4.4.3	Data Analysis	77
4.4.4	Study Team Composition and Positionality	79
4.4.5	Limitations	80
4.5	Results: Components in Privacy Art	82
4.5.1	Database Characteristics	82
4.5.2	Sensory Vocabulary of Privacy Art	83
4.5.3	Privacy Concepts Explored in Art	85
4.5.4	Subjects of Privacy Invasion	89
4.6	Results: Mechanisms to Present Components	91
4.6.1	Directing Viewer Attention	91
4.6.2	Viewer Perspective	93

4.6.3	Viewer Agency and Interactivity	95
4.6.4	Ambiance	98
4.7	Discussion	101
5	INTEGRATING ARTISTIC MECHANISMS INTO EXISTING PRIVACY COMMUNICATION METHODS	103
5.1	Motivation	103
5.2	Related Work	105
5.2.1	Persuasive Technology	105
5.2.2	Narrative and Storytelling in HCI	106
5.2.3	Creepiness by Design	106
5.2.4	Materiality and Tangible Interfaces	107
5.3	Methods	108
5.3.1	Study Design	108
5.3.2	Participant Data Collection	110
5.4	Discussion and Conclusions	111
5.4.1	Potential Limitations and Ethical Considerations	112
6	CONCLUSION	115
6.1	Future Work	116
6.1.1	Identifiability and Redaction	116
6.1.2	Privacy Communication	117
6.2	Final Thoughts	118
	REFERENCES	119
A	IDENTIFIABILITY INTERVIEW GUIDE	144
B	GENERAL PROMPT PROVIDED TO LLMS	148
C	FRAMEWORK PROMPT PROVIDED TO LLMS	149
D	PROMPT ENGINEERING STUDY	153
E	ADDITIONAL TABLES	156
F	REDACTION STUDY GUIDE	159
G	SAMPLED ARTWORKS	163
H	CODEBOOK	165
I	EXAMPLE CODED PIECE	168

LIST OF FIGURES

3.1	MARI Category Selection Page	55
3.2	MARI Category and Term Overview Pages	56
3.3	MARI Similar Terms Page Overview	57
3.4	MARI Final Preview Page	58
4.1	Privacy Artwork Structured Discovery Process	69
4.2	Standard Privacy Communications vs. Privacy Art	71
4.3	Audiovisual Components of Privacy Art	83
4.4	Surveillance Resistance in Privacy Art	87
4.5	Gaze Direction in Privacy Art	91
4.6	Viewer Perspective in Privacy Art	94
4.7	Refik Anadol's <i>Convergence LA</i>	96
4.8	Danja Vasiliev's <i>Netless2</i>	97
4.9	Ambiance in Privacy Art	98
5.1	Artistic Privacy Communication Method Mock-Up Comparison	114
I.1	Coded Artwork Example	169

LIST OF TABLES

2.1	Framework of Identifiable Data and Redaction Decisions	25
2.2	Word-Level Evaluation Results for TalkBank and Synthetic Datasets	33
2.3	Category-Level Evaluation Results for TalkBank and Synthetic Datasets	37
3.1	MARI User Study Categories for Redaction	60
D.1	Prompt Engineering Supplementary Results	155
E.1	Prompt Engineering Study: Claude Sonnet Missed Redactions	156
E.2	Prompt Engineering Study: GPT Missed Redactions	156
E.3	Interview Study Data Steward Demographics	157
E.4	Pilot Study TalkBank Results	157
E.5	Category Breakdown of Baseline Redactions	158
G.1	Privacy Art Sample, Part 1 of 2	163
G.2	Privacy Art Sample, Part 2 of 2	164
H.1	Privacy Art Codebook, Part 1 of 3	165
H.2	Privacy Art Codebook, Part 2 of 3	166
H.3	Privacy Art Codebook, Part 3 of 3	167

ACKNOWLEDGMENTS

Thank you, tack, kiitos, to my family and friends, who believed in me the whole time, even and especially when I did not. I couldn't have done it without you. I appreciate how you all made me smile through all the late nights, work on the go, and time glued to my desk; I am so incredibly fortunate to have such a huge support network that if I were to list everyone in it, this dissertation would be twice as long. Thank you to my parents, Tim and Caroline, for being in it for the long haul and for your unwavering belief in me; and to my siblings, Maja, Lina, Max, and Alex, for always being up for causing a distraction from my work with late-night pizza or spontaneous snack runs. Thank you to my friends for your forbearance in suffering through my long-winded vent sessions, making me take breaks, and travelling with me: Kathryn, Amanda, Brandon, Collin, Jackson, Maeve, Nick, and Ryan. You all know I have a propensity to err on the side of being gratuitously verbose, but even so, no words here can fully capture my gratitude for the steady encouragement, patience, and motivation from my friends and family throughout this journey.

Thank you to my advisor, Blase Ur, and to my committee members, Marshini Chetty and Alex Kale, for your thoughtful feedback and critical insights. Your guidance, questions, and critiques—whether smooth or challenging in the moment—ultimately strengthened this work. And to the other PhD students with whom my time here has overlapped in SUPER-group: Will, Weijia, Valerie, Kevin, Arthur, Madi, Tess, and Elleen; thank you for your support, feedback, and collaboration throughout this work. Thank you also to Dan and Pedro for your mentorship; your advice in my PhD experience was invaluable and helped me grow my research skills and foster my creativity.

MARI and redaction projects. This material is based on work supported by the National Science Foundation under Grant No. 2210193 and by a Meta Privacy-Enhancing Technologies Research Award. It also represents joint work. Authors on this project include Daniel Byrne, Arlie Jackson, Anwar Kader, Eliza Migdal, Valerie Zhao, and Blase Ur.

Thank you:

- Dan Byrne, for interview work, taxonomy and code contribution, data annotation, and psych-CS liaising;
- Valerie Zhao, for interview work, taxonomy contribution, and early UI design;
- Arlie Jackson and Eliza Migdal, for running tool analyses and data annotation;
- Anwar Kader, for data analysis and exceptional implementation work on MARI;
- Oliver Cai, Brianna Zeng, Ben Aizenberg, Ruotian Wang, and Will Brackenbury, for conceptual code contributions;
- Ruthe Foushee, Marisa Casillas, and Susan Goldin-Meadow, for social-behavioral insights and psychology department support.

Privacy art projects. This material is based upon work supported by the National Science Foundation under Grants No. CNS-2047827 and CNS-2149680. It also represents joint work. Authors on this project include Tess Eschebach, Nathaniel Kim, Bingning Jolin Liu, Marc Downie, Douglas Pancoast, and Blase Ur. Thank you to everyone who helped contribute to our database, especially Nick Briz and students in *CMSC 23218/SAIC 5007: Surveillance Aesthetics* (Fall 2024).

Thank you:

- Tess Eschebach, for expertise in qualitative coding work, project direction, data collection and analysis, and literature mastery;
- Nathaniel Kim and Jolin Liu, for qualitative coding work, data collection, and data analysis;
- Marc Downie and Douglas Pancoast, for expertise as artists and insight into the art world.

ABSTRACT

Human-centered privacy empowers individuals to understand and control their data. While often situated in human-computer interaction (HCI) and usable privacy and security, research in this field benefits from cross-disciplinary perspectives. This thesis develops a dialogue between computer science and the arts and humanities to examine both how privacy is enacted in practice and how it is communicated to users. We study privacy and identifiability in unstructured text data, such as conversation transcripts collected by social and behavioral science researchers, focusing on redaction as a key mechanism for protecting identity. We evaluate the capabilities of existing redaction tools and large language models (LLMs) for automating this process and, based on researchers' needs, design a semi-automated, human-in-the-loop workflow. This approach operationalizes context-dependent definitions of identifiability while leveraging LLMs to flag sensitive information in a graphical format, reducing the overhead of traditional methods while maintaining researcher control. In parallel, we investigate how privacy is made legible and engaging, drawing on a large corpus of privacy-related artworks we collected to analyze how artists communicate privacy concepts. We translate these insights into HCI by prototyping designs and outlining potential studies on effects of incorporating artistic mechanisms into privacy communication methods.

CHAPTER 1

INTRODUCTION

Privacy in computational systems is most commonly conceptualized as a problem of data collection and protection. This framing has led to extensive technical work spanning areas such as cryptography, systems security, differential privacy, and access control. Within these domains, privacy is often formalized into information flow, adversarial capabilities, or disclosure risks. In practice, legal protection or business goals are often prioritized over the humans behind the data, fulfilling institutional or regulatory requirements. However, while these approaches are necessary, they tend to treat users as data sources rather than as active agents and miss what the humans behind the data actually want or need. Thus, a purely data- and system-centric view of privacy is arguably incomplete. When individuals have control over their personal information, they can create and maintain boundaries with which data they choose to share, empowering them to act without fear that they will be manipulated or their information misused.

From a computer science perspective, privacy is not only a property of systems but also a property of interactions between users and these systems. Even strong, formalized methods of privacy modeling can fail to provide adequate privacy when misaligned with users' mental models or confronted with social and contextual dimensions of information flow. Research in human-computer interaction (HCI) and usable privacy explores ways in which to center privacy around humans, attempting to address this gap between formalized and experienced privacy. Works in these fields aim for usability and alignment with human behavior, demonstrating the need for user-centered design principles.

Lessons from HCI and usable privacy are broadly applicable—however, privacy-related challenges in computational systems cannot be fully understood or addressed through technical methods alone. Systems involving human subjects, sensitive data, or complex social contexts can strongly benefit from the arts and humanities, which have robust and extensive

background for empowering and centering the human element of computational systems. At the same time, computational tools offer powerful mechanisms for implementing, scaling, and evaluating privacy-preserving interventions in domains traditionally studied within the arts and humanities.

In this dissertation, we outline a conversation between computer science and humanities and the arts, using methods from computer science and technology to solve privacy problems in humanities spaces (among other human subjects research spaces), as well as using lessons from art and artists to apply to HCI privacy and security disciplines.

Methodologically, the dissertation employs a combination of qualitative and quantitative user studies, system-building, and design research. By grounding abstract concepts of privacy from theoretical frameworks into concrete artifacts, supported by practical findings with human subjects, this work extends existing approaches in HCI and privacy research. These works apply computational methods—including system design, prototyping, and user studies—to privacy challenges arising in humanities and human-subjects research contexts. We also draw on artistic and humanistic practices as a source of design and workflow insight, using them to inform the development of privacy-centered systems.

The contributions of this dissertation are grounded in the design, implementation, and evaluation of privacy-aware techniques across both computational and interpretive contexts. First, it contributes insights into how privacy is understood and enacted in practice, particularly in settings involving unstructured, qualitative data and reassessing definitions of identifiability. This work also evaluates and extends computational approaches to privacy protection in human subjects data, examining the limitations of existing automated redaction techniques and demonstrating the value of human-in-the-loop systems for improving both effectiveness and usability in real-world redaction workflows. In addition, this work introduces a cross-disciplinary perspective that draws on artistic practices to examine how privacy can be communicated to and reflected upon by end users. Through the design of

prototype systems, we explore how artistic representations of personal data can prompt reflection and influence how individuals perceive their own privacy.

The works that follow were conducted in an effort to understand how privacy can be seen not only as a technical property of data, but as a fundamentally human, perceptual, and interpretive challenge. The remainder of this dissertation develops this perspective through a series of studies, each examining privacy through the dual lens of computational mechanisms and human experience.

1.1 Dissertation Overview

In this section, we give a brief overview of each chapter. Chapter 2 explores *identifiability* and *redaction*, a method of removing or substituting information that might expose participants' identities to those who know them. In particular, we examine redaction and definition of identifiable information in unstructured text data, as well as a series of automated redaction methods including legacy tools and large language models (LLMs). We follow with Chapter 3, which concretizes the conceptual ideas of identifiability and the automated redaction research with a graphical interface designed for a human-in-the-loop redaction workflow. Chapter 4 inquires into privacy communication topics through a large-scale collection and study of gallery art and exhibitions, and Chapter 5 outlines a proof of concept and accompanying study outline implementing findings from this investigation. Specifically, we draw on mechanisms artists use in their privacy-oriented work and apply these in HCI methods to provoke reflection on privacy. We conclude with Chapter 6, discussing lessons learned from these studies.

1.1.1 Identifiability and Automated Redaction

Researchers across many fields often collect datasets of unstructured text, such as conversation transcripts. To benefit science, they often hope to share these datasets, yet first must redact identifiable information to protect participants’ privacy. Doing so at scale often requires partially automated redaction tools that retain a dataset’s utility by not over-redacting. To better understand requirements for redaction, we interviewed ten data stewards—primarily principal investigators from the social and behavioral sciences—about their re-identification threat models, current redaction practices, and types of information that make their participants identifiable. They articulated nuanced conceptions of re-identifiability and the need to redact more than just direct identifiers. We distilled their observations into a novel redaction framework. We then benchmarked current tools’ ability to flag such information. Specifically, we redacted text transcripts using three widely used legacy tools for flagging sensitive data, as well as recent large language models (LLMs). The LLMs correctly flagged many, but not all, potential identifiers redacted by trained humans. Notably, LLMs benefited when their prompts incorporated our framework. While LLMs outperformed legacy tools like Microsoft Presidio, their data flows may not be appropriate for all contexts.

Key insights:

- Human subjects researchers have diverse needs when it comes to redacting identifiable data, many of which are not covered by typical PII paradigms—especially for very sensitive social data or data collected from at-risk participants.
- Current commercial redaction tools miss a significant portion of identifiers in unstructured data compared to our baseline framework redactions.
- When given our redaction framework, LLMs correctly flag many classes of potential identifiers, indicating strong potential for automatically redacting unstructured text.
- Though LLMs perform well, they still miss more complex or niche identifiers, especially non-English vocabulary; there is still a need for human verification.

1.1.2 *Bringing Humans into the Loop in Automated Redaction*

We extend our identifiability and redaction studies by designing a workflow for partially-automated, human-in-the loop redaction of identifiers adhering to the prior redaction framework. The workflow is designed to automatically flag identifiable data, drastically reducing overhead for typical redaction workflows. Moreover, this workflow is implemented in a graphical user interface called MARI, which leverages existing offline redaction methods as well as LLMs to cover a broad scope of identifiable data; users are then presented with this data for verification, as well as other statistics and information pertaining to flagged terms. We evaluate this interface in a user study with real data, where participants are asked to perform redaction using various methods, comparing the application to manual and partially-assisted redaction scenarios. Preliminary observations show that though participants perceived a higher learning curve for the graphical method (MARI), participants performed better using MARI compared to the others, redacting identifiable information more efficiently and correctly. In particular, participants preferred our tool to manual or pre-filled redaction methods and perceived their redactions to be more efficient using our tool.

Key initial insights:

- The graphical method (MARI) most consistently achieves the best overall performance for participants, balancing throughput and accuracy in redaction more effectively than the manual and pre-filled methods.
- The pre-filled method increases throughput but reduces accuracy, while the manual method maximizes accuracy at the cost of throughput. The graphical method mitigates this trade-off by improving both correct redactions and overall speed.
- The graphical method encourages more active verification and engagement, whereas the other methods are more passive or linear in workflow.
- Participants find the graphical method reliable and efficient and the manual method slow but intuitive, but the pre-filled method to require higher cognitive effort.

1.1.3 *Exploring Privacy in Art and Applications to Privacy Communications*

User-facing mechanisms for communicating privacy, such as privacy policies and browser tools, are often overlooked and may fail to support meaningful understanding. In contrast, artworks emphasize engagement, provocation and reflection, and ambiguity and subjective experience. For decades, artists have explored surveillance, privacy, and identity through *privacy art*: gallery works exhibited across diverse media that focus on privacy or surveillance of individuals. To better understand *how* privacy topics are communicated in such works, we compiled a corpus of over 800 privacy artworks and conducted a qualitative analysis of a sample. We identified recurring patterns across media and time periods in how these artworks draw on familiar mechanisms to engage a variety of privacy topics, such as government or corporate surveillance. Many works ground their commentary in artists’ personal data, lived experience, or data collected from audiences and the public, situating privacy as both individually and collectively experienced. These artworks employ diverse representational and interactive mechanisms to communicate privacy concepts, spanning a broad affective range from fun to creepy. We synthesize these results into lessons for making common privacy communication methods more engaging and powerful through narrative and data ownership, materiality and tangible interactions, and setting an unexpected affective tone.

Key insights:

- Privacy artworks draw on a shared visual language of familiar surveillance objects and domestic metaphors to highlight topics like governmental and corporate surveillance, identity-surveillance dynamics, and AI-amplified privacy risks.
- Many works ground commentary in artists’ personal data or lived experiences, as well as data collected from the public and even artwork viewers.
- Artworks employ mechanisms such as tangible materials, data aggregation, and interactive experiences (sometimes with constrained user agency) to convey privacy concepts.
- Privacy artworks use varied tones from unsettling to playful to evoke affective responses, revealing opportunities for provocative privacy communication.

1.1.4 *Designing HCI Privacy Provocations Using Artistic Mechanisms*

We concretize our findings in the prior study by implementing design suggestions for integrating art with privacy communication strategies. This will act as a proof of concept, where we design a custom implementation using mechanisms from art and existing privacy communication methods. Our approach foregrounds materiality, data ownership, and unsettling affects as intentional qualities that shape how privacy concepts are perceived and experienced. We present a between-subjects study exploring how more conventional forms of privacy communication compare to those augmented with artistic mechanisms. Participants interact over time with either a digital pop-up banner containing descriptive text data or “watering” a physical artificial plant that exhibits data through an evolving form. The study examines how differences in modality (digital vs. physical representations of data) and perceived data ownership (self vs. other) influence interpretations of privacy, emotional responses, and perceptions of agency and intrusion. We reflect on the design process and outline initial insights into how material and affective qualities can expand privacy provocations beyond conventional approaches for communication. We position these reflections as early-stage directions for designing more experiential and provocative forms of privacy communication.

Key initial insights:

- Modality of privacy communication may shape how privacy-related information is interpreted, with physical artifacts potentially prompting more reflective and embodied engagement.
- Recontextualizing data between self and other may show how perceptions of data ownership and identifiability can be shaped by presentation and narrative framing rather than by the data itself.
- Introducing affective and provocative framing suggests opportunities for privacy communication that highlight reflection and critical engagement over just comprehension.

CHAPTER 2

EXPLORING IDENTIFIABILITY AND AUTOMATED REDACTION PROCESSES IN UNSTRUCTURED TEXT DATA

This chapter presents the findings of qualitative interview studies and benchmarking analysis of automated redaction tools for unstructured text. Our findings identify critical gaps in current automated redaction capabilities and establish a new framework for identifying subtle, context-dependent identifiers. The following sections detail how this framework was constructed and how it significantly improves the performance of LLMs in redaction tasks.

2.1 Dissertation Author’s Contributions

This chapter represents joint work with Daniel Byrne, Arlie Jackson, Anwar Kader, Eliza Migdal, Valerie Zhao, and Blase Ur. The author of this dissertation served as lead author and wrote substantial sections of the Introduction, Related Works, Framework, Benchmark Study, and Discussion sections in the version submitted for peer review. Co-authors contributed to the Introduction, Related Work, Formative Interview, Framework, and Discussion sections.

2.2 Motivation

Researchers who collect data from human subjects must balance the goal of collecting insightful scientific data with the need to protect study participants’ privacy. This trade-off occurs often for social and behavioral sciences (*SBS*) researchers. Through great effort and expense, many SBS researchers collect datasets of unstructured text data spoken by human subjects, recorded, and subsequently transcribed. For instance, human development researchers will transcribe conversations between caregivers and children, education researchers will do the

same for dialogues in classrooms, and linguistics researchers will do the same for speakers of regional dialects.

To protect participants, researchers must *redact* their corpora by removing or substituting information that might make human subjects identifiable. Conceptually, redaction requires removing *personal data*, which is itself a wide-reaching concept. The General Data Protection Regulation (*GDPR*) characterizes personal data as “any information relating to an identified or identifiable natural person... who can be identified, directly or indirectly” [81]. Personal data includes obvious examples like “an identification number” and “a name” , often referred to as personally identifiable information (PII). Longstanding techniques like named-entity recognition often flag such identifiers fairly accurately. However, the scope of personal data is far wider. For instance, the GDPR also highlights how “factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person” can also be personal data. Flagging these more ambiguous factors is far less straightforward.

Because redacting identifiers both reliably and at scale can be challenging and time-consuming, valuable datasets are rarely shared with other scientists, creating roadblocks to scholarly inquiry and requiring researchers to collect redundant datasets at significant cost. Our interdisciplinary team initially encountered redaction challenges when attempting—and failing—to redact a corpus of unstructured text data stewarded by some team members. We initially leveraged widely used tools like Microsoft Presidio [195], Google Sensitive Data Protection [109], and Amazon Macie [9], which aim to flag sensitive data in text. These tools missed both direct and subtle identifiers. They also *over-redacted* information unlikely to be identifying, potentially limiting the data’s research uses.

Spurred by this experience, we explore in this chapter the necessary requirements and current technical capabilities for redacting corpora of unstructured text transcripts. To understand the problem space, we first conducted qualitative interviews with ten *data stewards* who oversee corpora in fields like psychology, sociology, communication, and linguistics. We

asked participants how they redact their own data, their threat models in protecting human subjects, and information in their corpora that could make participants identifiable. The data stewards had nuanced conceptions of re-identifiability. Some worried about a participant being singled out; others worried about narrowing them down to dozens of candidates. All reported manually replacing or obfuscating potential identifiers before sharing data.

As part of this process, we worked with the data stewards iteratively to create and refine a novel framework of potential identifiers in the context of SBS corpora (Table 2.1). They identified seven major categories and twenty-three subcategories of information that may need to be redacted, ranging from direct identifiers to specific activities to linguistic patterns. While we focus on SBS researchers for concreteness, lessons from our results can be applied to other research fields with human subjects and unstructured text data.

To understand the degree to which current computational tools can automatically flag potential identifiers in unstructured text transcripts, enabling redaction at scale, we benchmarked these tools. Our first class of tools encompasses those specifically designed to flag potential identifiers in text. We tested three of the most commonly used [153, 234, 5, 17] tools of this class: Microsoft Presidio [195], Google Sensitive Data Protection [109], and Amazon Macie [9]. In response to the recent success of foundation large language models (*LLMs*) in processing text, we also tested eight state-of-the-art LLMs: Anthropic’s Claude Opus 4.5 [14], DeepSeek-V3.2 [66], Google’s Gemini 2.5 Pro [111], OpenAI’s GPT-5.1 [211], Meta’s Llama 4 Maverick and Llama 4 Scout [193], and Alibaba’s Qwen3-235B-A22B and Qwen3-30B-A3B [270]. Five are *open-weight* models, meaning they are publicly available for download and can be run locally given sufficient computational power. This would reduce chances of unredacted data being shared with organizations outside a research group’s IRB protocol. To avoid putting the onus on PIs to solve deployment issues, an institution or organization might deploy private LLMs [248] or shared local copies of these foundation LLMs that do not transmit data outside the organization.

For both legacy tools and LLMs, we compared items they redacted to potential identifiers redacted manually to consensus by three trained research assistants (*RAs*). Both RAs and the computational tools redacted a subset of data from TalkBank [174, 173, 172], a repository of human communication data. As TalkBank did not cover all framework categories, we augmented it with synthetic data. We also conducted a pre-study focused on prompt engineering to explore the impact of nine key prompts.

We found that, when given our redaction framework as part of their prompt, the eight LLMs correctly flagged many classes of potential identifiers on our main evaluation dataset, missing only between 1.8% (Gemini) and 12.8% (GPT) of the consensus redactions from our trained human annotators on TalkBank data. Notably, however, all LLMs we tested consistently failed to redact complex or nuanced identifiers, such as lengthy organization names or instances of code-switching. Furthermore, *any* missed redactions are a potential privacy leak. Without including our framework in the prompt, the LLMs on average completely missed between 5.3% (Gemini) and 19.7% (Claude) of the human RAs’ redactions. In contrast, the widely used legacy tools from Microsoft, Google, and Amazon missed between 22.2% (Presidio) and 94.0% (Macie) of redactions. We discuss the lessons for the security and privacy community, as well as for all researchers who collect text data.

2.3 Related Work

Contrasting with redaction’s importance in protecting the privacy of individuals, academic work on redaction is surprisingly small and split across sub-communities, with notable gaps in the security and privacy community. In fact, in a recent analysis of 38 de-identification guides, Guo et al. identified widespread inconsistencies in both the specific guidance given for de-identification and how key terms are defined [116].

2.3.1 Key Redaction Techniques

The academic literature presents a variety of techniques for flagging potential identifiers. Personal data that is highly structured (e.g., phone numbers, driver’s license numbers) is typically identified automatically using pattern-matching (e.g., regular expressions) or simple heuristics. Personal data that is less structured (e.g., names of people or places) instead relies on a combination of lists (e.g., of common names) and techniques derived from the natural language processing (*NLP*) sub-community. For instance, named-entity recognition (*NER*) encompasses a series of techniques that try to identify real-world objects with proper names (e.g., people, locations). Many NER approaches rely on grammar-based techniques alongside machine learning models trained on large amounts of annotated data.

A handful of different academic researchers have bundled these techniques into redaction tools. However, to our knowledge, none of these create a custom redaction workflow designed for human-in-the-loop verification. Bosch et al. used part-of-speech tags, lexical item indices, typed artifacts, and public records (e.g., census data) to redact data from online forums [36]. Others have proposed modeling tokens’ context and hierarchical relations to distinguish between identifiers and non-identifiers [236, 67, 84]. Many systems, including Textwash [151] and others [63], rely heavily on NER in concert with pattern matching to flag potential redactions. Beyond replacing identifiers with pseudonyms, researchers have also suggested generalizing words [12].

Researchers have also developed domain-specific methods [97] for redacting medical records [267] (and more specifically, electronic health records [194]), clinical text [28, 146], unstructured chat data [2], learning analytics [149]. Prior work has also discussed redacting everything from journalists’ files [187] to network traces [196, 59] and log files [87].

Any attempts to de-identify data can be filled with peril [204, 57]; redacting text is no exception [232]. Nicknames and typos can cause redactions to be missed [83]. References to locations [294] and correlations between terms [11] leak information. Overall, in a qualitative

study of 18 practitioners and 6 curators involved in redaction, Guo et al. [117] found that researchers often rely on manual and social processes to redact their data. Notably, they also reported that current “tools do not meet researchers’ [de-identification] needs” [117].

2.3.2 Commercial Redaction Tools

Commercial redaction tools similarly focus on removing highly structured identifiers (e.g., ID numbers, credit card numbers), as well as names and geographic locations [157]. The open-source Python library Microsoft Presidio [195], cloud-based Google Sensitive Data [109] (formerly Cloud DLP), and Amazon Macie [9] rely on pattern matching, keyword dictionaries, basic linguistic features, and straightforward heuristics [209] to suggest redactions, as do older tools like SPIRION [259] and CipherCloud [168].

While some commercial tools offer a user interface, these are often complex or difficult to navigate, and their shortcomings make them less effective for our intended workflow. From our initial attempts to use existing academic and commercial tools to redact an SBS corpus, we find they can fail to detect unique names, idiosyncratic phrases, or context-specific identifiers (e.g., a participant with fluid gender expression). On the opposite end, they can remove non-identifiers that provide crucial research context (e.g., the name of a movie character that influenced a young participant’s language development).

2.3.3 Using LLMs For Redaction

In parallel work over the last two years, a handful of other researchers have also explored the usefulness of LLMs for redaction, though not for redacting the same classes of information in our framework, nor for anonymizing data collected by SBS researchers. Asimopoulos et al. [17] proposed using LLMs for NER, evaluating their approach on the CoNLL-2003 dataset and finding it to outperform more traditional methods of NER. In contrast, our redaction framework and workflow consider aspects of identifiability far more broadly than NER. Pilán

et al. [222] use LLMs for generalization of uncommon words to protect privacy, replacing specific words with more general terms. Somewhat similarly, Frikha et al. [91] developed IncogniText, which uses an LLM specifically to remove attributes about the author of text (e.g., age, gender, income).

Other researchers have focused their redaction efforts specifically on the probability of words in text according to LLMs. For instance, Albanese et al. [5] estimate the in-context probability of each word in text, redacting the words with a probability below a specified threshold. Here, low-probability words are considered a proxy for potentially private information. Leveraging similar insights differently, Gusain and Leith [118] instead proposed analyzing which replacements language models deem likely to judge redaction success.

Dou et al. [72] focused on redaction specifically for limiting self disclosure in online social media posts. To do so, they fine-tuned RoBERTa and DeBERTa models, comparing their fine-tuned models’ performance to that of GPT-4. Finally, other researchers like Brahem et al. [38] and Gangadharaiyah et al. [96] have proposed the general idea of using LLMs for redaction, but do not provide specifics about the types of information they aim to redact, nor public benchmarks. Our prior work extended these definitions and LLM works, which we aim to implement in the backend in a novel interface.

2.4 Formative Interviews: Methods

To understand SBS researchers’ challenges and privacy concerns when redacting a corpus of unstructured text data they collected and wish to share with other scientists, we conducted qualitative interviews. The interviews aimed to capture researchers’ current redaction practices, threat models for re-identification, and requirements for assistive computational redaction tools. Interviews were conducted online via video conferencing and lasted 1–2 hours.

2.4.1 Recruitment

We targeted SBS PIs at universities in the United States who were responsible for stewarding at least one large corpus of unstructured text data. Additionally, we aimed for a diverse sample capturing various methodological approaches, disciplines, and intellectual traditions. We considered researchers in scope if they collected unstructured text data primarily in English as their primary research artifact through any means (e.g., observational studies, in-lab experiments, interviews). We prioritized researchers who worked with unstructured text data originating from spontaneously produced or spoken speech. Our team performed direct recruitment based on professional connections and targets identified from literature reviews in psychology, linguistics, anthropology, communication, political science, education, sociology, and cognitive science.

2.4.2 Interview Structure

Interviews with data stewards were semi-structured, containing four sections. Appendix A contains the full interview guide. We began by briefly discussing the data steward’s research and corpora of unstructured text data to contextualize subsequent responses. We then defined “redaction” as “removal or substitution of data that could be used to identify participants . . . so that data could be shared publicly or semi-publicly.” We asked data stewards to describe norms around data access/sharing in their fields, their own practices and experiences redacting unstructured text data for such purposes, and their abstract conceptions of (re-)identifiability.

In the second part of the interview, we assessed data stewards’ threat models. Specifically, we asked them to rate and justify their concern about publishing research data publicly in various redaction scenarios, as well as their specific privacy concerns about various individuals re-identifying that subject’s transcripts.

In the third part of the interview, data stewards iteratively refined a working framework of

classes of information that may need to be redacted from text corpora. Our inter-disciplinary research team created a first draft of this framework based on our own experiences attempting to automatically redact a large corpus, and we further expanded this initial draft through three pilot interviews. In this part, we shared an editable link to the working redaction framework from the end of the previous interview. Data stewards reviewed the framework in a “think-aloud” activity. We also asked data stewards to discuss each class of data in the framework relative to their own research, as well as harms of not redacting that data.

The final section of the interview delved into data stewards’ experiences redacting and sharing data. Questions included the cost to the data steward of manually redacting their own corpora, as well as any tools or software they had used.

Prior to finalizing this structure, we conducted pilot interviews to refine our questions with two linguists and one staff scientist. These ranged from 1.5-2.5 hours and included more questions than the final interviews. Over one-quarter of the framework changed substantially from our team’s initial draft to the final version, including adding linguistic patterns, code-switching, and personal interest categories, as well as refining nine others (e.g., events, regionalisms, slang). Additional smaller changes were process-focused (e.g., not redacting common diseases) or captured disagreements.

2.4.3 Data Analysis and Limitations

We audio-recorded and transcribed all interviews, deleting recordings after transcription. Following the thematic analysis and approach established by Guo et al. [117], we expanded and built on their framework in three key ways. First, we demonstrated that their codebook and documented themes generalize well to novel interview data focused on redaction. As an initial pass, we applied and augmented their codebook to our interviews. Although Guo et al. developed their codebook for researchers working primarily with quantitative data, we found that their codes closely aligned with the qualitative, unstructured text data de-

scribed by the data stewards we interviewed. However, their codebook alone did not capture many themes we observed. Through repeated readings of our interview data, we developed and added numerous codes, iteratively refining them until they stabilized, to ensure more robust coverage. Our OSF data release includes our final codebook, which contains 150 unique codes. Second, treating our finalized redaction framework classes and sub-classes as distinct—yet complementary—codes to Guo et al.’s codes and themes, facilitated a more precise analysis of the relationships among researchers’ approaches to redaction and different types of potentially identifying information. Third, through the cooperative and iterative co-construction of our redaction framework, we were able to systematically capture and code the extent to which our participants conditionally agreed, disagreed, or reasoned contextually about potentially identifying information across fields, processes, procedures, and threat models.

Our sample size ($N = 10$) is comparable to Guo et al.’s sample of 14 participants [117] and reached theoretical saturation after 8 interviews, especially in refining our redaction framework. Our sample was diverse in terms of data stewards’ research methods, scholarly fields, and universities (both public and private). Nonetheless, all data stewards worked in SBS fields in the United States, so our findings may not generalize to all other fields or countries.

2.5 Formative Interviews: Results

In this section, we first describe our participants and the text data they steward. We then discuss their general practices around sharing and redacting data, followed by their threat models and re-identification concerns. In Section 2.6, we then present the redaction framework we refined iteratively with participants.

2.5.1 Participants and Their Corpora

We interviewed ten data stewards, eight PIs (faculty members) and two postdoctoral researchers, all at universities in the United States. All reported major stewardship over at least one corpus of unstructured text data. Among data stewards (Table E.3 in the appendix), three were psychologists (cognitive, educational, and developmental), two were communication scholars (interpersonal and health), three were sociologists (medical, gender, and educational/rural), one was a medical anthropologist, and one was a linguist. The data stewards corpora ranged in size from 19 to over 2000 participants, and their participants ranged from infants to senior citizens. Their data was racially, geographically, and demographically diverse.

2.5.2 Views on Data Sharing and Redaction

Data Sharing: All ten data stewards reported that only trained members of their research staff listed on their institutions' IRB protocols had access to their unstructured text data. Two data stewards (P5, P6) also disclosed that some of their data was collected outside of their academic institutions and was subject to additional data use agreements. Moreover, P1, who works extensively with US military personnel, discussed additional requirements from military IRBs, policies, and laws.

Similarly, P3 and P4 reported that data collected in primary and secondary schools across different districts, even within the same state, may have varied access, sharing, and reporting requirements. P3 also discussed their legal requirement to upload all interview transcripts, but not their field notes, to the National Institute of Mental Health Data Archive. They went on to state that the density and mandated archival of their data necessitated using an external, IRB-approved transcription service. Five data stewards reported completing transcription work internally with trained RAs, without speech-to-text software or external services, while the other five reported using either software or services. All data stewards

who used software ran it locally and had a human review all outputs.

Defining Identifiability: When prompted to define “identifiable information,” data stewards’ responses coalesced around several themes relating to the number of possible data subjects. P1 said they think about identifiability in relation to “gray areas of innocuous things like mental health treatment” that may be used in conjunction to “narrow it down to a specific pool of people (between 15-30).” They further described mental health treatment as particularly sensitive due to its stigma in some communities. P5 echoed this sentiment, fearing the ability to “triangulate on multiple features. . . like a specific medical condition and a certain city.” Discussing a data subject, P10 noted, “there’s not that many biracial kids who are there. . . [people] will figure out who he is.” P2 and P5 said they hew closely to the CITI [56] definition of personally identifiable information. That said, P2 noted shortcomings of existing guidelines: “I sometimes think of it like if I were given this piece of information, could I find this person online with some heavy Googling and often you could, even if CITI tells you that’s okay to leave in. . . So even those kinds of details that are not considered individually identifiable but are very unique. . . I wouldn’t publish.”

All ten data stewards invoked the idea of attributes’ “Google-ability” as a heuristic for identifiability. P4 articulated this concern as “we all have the tools to be the FBI,” contrasting their concern with the need to be a “transparent researcher” and share precise details about participants to give context and ensure reproducibility.

Data stewards noted that identifiability depends heavily on which characteristics of their data are publicly known. P2 explained, “I live in the south. . . so everybody here is going to call a grocery cart a ‘buggy.’ So I’m not going to redact the word ‘buggy’—they’re all gonna say it that way. . . [it’s] public knowledge that I collected my data in a particular region. So to me that’s not identifiable.”

Redaction Practices: All ten data stewards reported replacing information in their text corpora with pseudonyms. Seven data stewards emphasized the need to substitute meaning-

ful tokens for redacted data, and six mentioned the use of pseudonyms that honor participants’ cultural and ethnic backgrounds. P5, P7, and P8 also spoke of selecting pseudonyms with the same number of syllables and letters as the original to preserve linguistic features.

P3, P5, P6, and P10 reported obfuscating unique participant features or attributes on the basis of commonality. More specifically, data stewards created abstractions to make participants less identifiable, such as transforming specific religious affiliations and cultural traditions into more generic descriptors. For example, P10 said, “I collected data in a children’s hospital in the Mountain West . . . a Santa was [video-conferenced] for a young, Muslim girl . . . I chose to simply record the incident simply as brown girl, does not celebrate Christmas.” Similarly, P3 said that much of their data came from a geographic location linked to a niche industry: “I simply described it as blue-collar work.” P10 explained, “I have one participant . . . from [specific small city], which I can’t say. So I would say like a particularly evangelical city or something because it is and that’s usually the important part.”

P3, P5, P6, and P10 employed “date-shifting”—changing the time or temporal sequence of events. P3 noted that doing so is not always straightforward: “my data on funerals is too close to my scientific research questions . . . I can’t lose it.” P10 also grappled with the practice, reflecting, “I don’t always de-identify the date . . . it obscures the social processes . . . for example, young siblings cannot visit the hospital during winter because of respiratory conditions that come through.” Notably, P10 also performed “member-checking”—asking the participant to confirm the redacted data’s fidelity.

Other data stewards employed more complex transformations they felt protected subjects while not misrepresenting their data. For instance, when discussing a potentially identifiable participant, P10 said, “We did composite her with two other similar cases so that her parents wouldn’t figure out like exactly what she said. And so like used like another girl who was also pulled out of school. So we used the reason that the other girl was pulled out of school.”

2.5.3 *Threat Models and Concerns*

To probe data stewards’ privacy threat models, we asked them to imagine they posted their unredacted data publicly on a website anyone can visit and to rate and justify their degree of concern for their participants’ privacy.

Eight data stewards said they would be “extremely concerned” about a stranger (i.e., unknown to the participant) re-identifying data. P8, a developmental psychologist, recoiled at the possibility “of someone looking at the conversations of little kids on the Internet . . . [and] trying to meet them in real life by using that knowledge.” Studying medical decision-making, P5 feared an HIV/AIDS diagnosis “could inadvertently ‘out’ someone as [queer] to an authoritarian government.” P5 further stated, “I study people with cancer . . . they could be forced out of their jobs if their employers found out.”

Regarding someone who personally knows a study participant re-identifying them, responses were split. Four data stewards said they would be “extremely concerned,” five reported they would be “moderately concerned,” and one reported that they would be only “somewhat concerned.” P1, an interpersonal communication scholar, was “deeply troubled” at the thought they could inadvertently ‘out’ someone as LGBTQ+ to a spouse or family member. P3 worried about a participant’s standing among high school peers and teachers if a particular transcript were attributed to them, explaining, “[the research participant] told me they saw the murder happen . . . only the school administration, me, and the police knew.”

Considering harms to participants more broadly, P1 recalled how collecting data during the US military’s “Don’t Ask, Don’t Tell” era created methodological and legal quagmires. They explained, “it was illegal for anybody to report that they were in the same-sex relationship because they would lose their spot in the military . . . we are more careful to redact than we otherwise would be.” On a longitudinal project about transitions from adolescence to adulthood, P6 related not knowing how to handle retrospective redaction for “one participant that identified as a queer woman in high school . . . [prior to] a religious conversion.” P3 dealt

with a similarly fraught question about a participant that transitioned genders, stating, “I would keep the truth as it emerged.” Capturing many data stewards’ broad approach, P1 explained, “We’re thinking about what is stigmatizing for that population... Cohort-specific or context-specific things that are especially stigmatizing we are more careful to redact.”

Notably, several data stewards expressed skepticism and concern about the real potential of re-identifying participants even when following redaction and de-identification guidelines created by external institutions, funders, organizations, and legal frameworks such as HIPPA. Concern regarding potential re-identification despite following “best-practice” redaction guidelines was frequently attributed to participants’ highly sensitive data traits (e.g. children with congenital, unilateral brain lesions or adolescent survivors of suicide attempts) and the systematic or incidental examination of seemingly innocuous indirect identifiers used in combination with each other to re-identify participants. P3 expressed this sentiment bluntly: “Kids dying [by suicide]. You’ve got to keep that private.” They go on to disclose that guidelines are “...not actually really protecting, it’s almost protecting reputation more than than privacy.” In contrast, P1 discussed the need to compensate for a gap in official DoD redaction guidance, noting that guidelines instruct researchers to remove any location smaller than a US state, however, “..in Wisconsin where there are no active duty, there’s only National Guard...there’s no military, no active duty military installations...we would take out the state there.” Indeed, most data stewards reported going beyond what was required by official guidance through the use of ad hoc heuristics or improvised techniques they themselves developed over time. Regarding redaction and participant privacy, P2 stated, “I think...my personal ethics were more stringent than what CITI or IRB would require.”

Interestingly, P2 also offered a meta perspective, suggesting that it would be relatively straightforward for an unspecified threat actor to infer participants’ identities through the data stewards. P2 reflects “people can look at my affiliation...guess that I probably collected data in [LOCATION] or that I got my PhD at [UNIVERSITY] and I say in a small

midwestern town, university town, and they would go she probably collected these data in [LOCATION].” Disturbingly, P2 also shared that the threats and harms extend beyond participants to the data stewards “...this group...has a database of...professors” that research certain social issues and processes. These real and present threats to not only participant re-identification but to the data stewards themselves is a significant and notable departure from what one of Guo et al.’s interviews dismiss as “sci-fi paranoia” [117]. Nevertheless, all ten data stewards felt their current redaction practices sufficiently protected participants, but all acknowledged that their methods have not been robustly assessed. None were aware of significant redaction failures to date in their field.

2.5.4 Logistics of Data Sharing and Redaction

We asked data stewards about the feasibility of redacting their data manually to a degree that they would feel comfortable sharing their data. All ten felt doing so was possible in theory, yet several emphasized that the time and financial costs required would make such efforts impractical for them in practice. Many data stewards remarked that most of their research budget was allocated for data transcription, annotation, and analysis, but not thorough redaction. P1 shared that “double checking the transcripts ... takes a really long time. And so it’s not in the kinds of datasets we have, they’re big but they’re not insurmountable.” P5 lamented that research grants tend to involve small sums of money, making redaction impossible. In explaining how their research assistants often redact data, P3 mentioned that those research assistants often have questions about what to redact: “Actually that’s the thing that my team will email me about the most where they’ll be like do we have to de-identify ‘Denver’ or like how do I de-identify this?”

None of the ten data stewards distributed their data to personnel not listed on their IRBs. There seemed to be two primary drivers: fear of running afoul any legal or ethical framework, and the time required to prepare, clean, and redact the data. Imagining sharing

data, P5 would want to “add [the recipients] to the IRB protocol . . . and sign a DUA” before emphasizing the effort required. Conversely, P4 bemoaned the fact that the general public is usually precluded from accessing data created from public grants. They explained, “we all say ‘data is available upon request,’ but I haven’t really thought about how to deal with that . . . I suppose I’d consult my IRB and enlist the help of an RA to redact it.”

Despite these barriers to sharing data, five participants (P3, P4, P7, P8, P9) aspired to make their data available on open-science platforms like Databrary, Open Science Foundation (OSF), CHILDES, or the National Institute of Mental Health Data Archive (NIMHDA). Additionally, all ten data stewards expressed an interest in using redaction software, though most required that it be run locally (i.e., not in the cloud) and released open-source.

Data stewards noted tensions between protecting subjects’ privacy and maintaining data’s research utility. P3 explained, “I don’t think I would redact the uncommon events because some of that is actually, that’s my research. Like last year when [a participant] was hospitalized for suicide. . . that’s too close to my scientific research questions.” P3 shared that these inherent tradeoffs between privacy and utility have culminated into larger debates within their field, noting that some scholars have “been really outspoken about how we should never hide the location where we’re doing work.”

2.6 Redaction Framework

Here, we present our finalized framework of information SBS researchers should consider redacting. We created this framework iteratively and cooperatively. We created an early first draft of this framework based on our interdisciplinary team’s experiences unsuccessfully trying to redact a large SBS corpus using existing commercial and academic tools. We then substantially revised the framework through three pilot interviews with data stewards (not explicitly reported) and the ten additional interviews from Section 2.5. Each interview participant reviewed the framework and suggested additions, deletions, and modifications.

Table 2.1: Our final framework of redaction decisions for SBS corpora and examples of common and uncommon cases. The final column shows which participants felt each category should be redacted always or *conditionally* (based on context).

Category	When to Redact	Examples (Common/General vs. Uncommon/Specific)	Participants Who Agreed (<i>Conditionally</i>)
Direct Identifiers			
→ Personal Names	Always	Jackson, Elizabeth, vs. Mordechai, Kynlee	P1, P2, P3, P4, P5, P6, P7, P8, P9, P10
→ Nicknames	If uncommon	Honey vs. Yuletidebug	P1, P2, P3, P4, P5, P6, P7, P8, P9, P10
→ Numerical Identifiers	Always	(Social Security Number, Phone Number)	P1, P2, P3, P4, P5, P6, P7, P8, P9, P10
Demographics			
→ Age	If specific	“She’ll be four in June” vs. 12/15/98	P1, P2, P3, P4, P5, P6, P7, P8, P9, P10
→ Sexual Orientation	If uncommon	Straight vs. Queer, Asexual	P2, P3, P5, P6, P7, P8, P9, P10
→ Gender	If uncommon	Male, Female vs. Trans, Gender-Fluid	P2, P3, P5, P6, P7, P8, P9, P10
→ Race / Ethnicity	If uncommon	Asian, Black vs. Multi-racial (Japanese + Nigerian)	P1, P2, P3, P4, P5, P6, P7, P8, P9, P10
→ Education	If specific	English class vs. seminar on “alien invasion films and sexuality”	P2, P3, P4, P5, P6, P7, P8, P9, P10
→ Profession	If uncommon	Teacher, Lawyer vs. Astronaut, VP at USENIX	P1, P2, P3, P4, P5, P6, P7, P8, P9, P10
→ Health Condition	If uncommon or sensitive	Cold, Allergies vs. HIV, Cancer, Ewing Sarcoma	P1, P2, P3, P4, P5, P6, P7, P8, P9
Locations			
→ Address	If private	30 Rockefeller Plaza vs. residential address	P1, P2, P3, P4, P5, P6, P7, P8, P9, P10
→ Landmark	If residential	London Eye, Times Square vs. Dakota Apartments, London Shard	P1, P2, P3, P4, P5, P6, P7, P8, P9, P10
→ Business / Organization	If uncommon	McDonald’s vs Chappie’s Cafe	P1, P2, P3, P4, P5, P6, P7, P8, P9, P10
Dates			
→ Dates	If specific and uncommon	Christmas Eve vs. “my father died on June 3rd”	P1, P2, P3, P4, P5, P6, P7, P8, P9, P10
Events			
→ Specific Events	If specific	Coachella Festival vs. St. John Bosco’s Parade of Flowers	P1, P2, P3, P4, P5, P6, P7, P8, P9
Linguistic Patterns			
→ Code-switching	If uncommon	Oy vey! vs. Caleta (Chilean slang for “a lot”)	P1, P2, P3, P4, P5, P6, P7, P8, P9, P10
→ Slang	If uncommon	“Groovy,” “No cap” vs. “Keerag”	P2, P3, P4, P5, P6, P7, P8, P9, P10
→ Regionalisms	If uncommon	“Bless your heart” (General Southern US) vs “Jawn” (Philadelphia)	P1, P3, P4, P5, P6, P7, P8, P9, P10
→ Familects	If uncommon	Bro vs. “choshimbo boy” (one family’s way to say “naughty boy”)	P1, P2, P3, P4, P5, P6, P7, P8, P9, P10
Personal Interests / Activities			
→ Hobbies	If uncommon	Pokémon card collector vs. Nokia phone collector	P2, P3, P4, P5, P6, P7, P8, P10
→ Traditions	If uncommon	Watching Superbowl vs. Celebrating (Druid) Winter Feast	P2, P3, P4, P5, P6, P7, P8, P10
→ Cultural References	If uncommon	Stormtrooper (Star Wars) vs. Skilling (local meteorologist)	P3, P4, P5, P6, P7, P8, P10
→ Personal Identity	If uncommon	Dungeons and Dragons player vs. member of the Kumeyaay tribe	P3, P4, P5, P6, P7, P8, P10

Overriding Considerations: A recurring finding was that redaction decisions are often context-dependent. Typically, the data stewards felt that a human responsible for the corpus would need to make a decision about whether to redact a term. As such, a key design requirement is that practical redaction tools efficiently integrate human judgment.

While prior redaction tools often focused on redacting text data written directly by the speaker, such as forum posts [36], the corpora controlled by our data stewards were typically transcribed either by a human research assistant or speech-to-text software. Each method poses challenges for effective redaction. For text produced directly by the speaker, characteristics like punctuation usage and capitalization can give hints about an otherwise anonymous author [1, 48, 39]. For transcriptions by humans, typos are common. For transcriptions by speech-to-text software, homonyms or common mispronunciations may cause transcription errors. Conversely, human transcribers may be susceptible to biases when transcribing, for example, transforming or normalizing phrases like “gonna” to “going to” and consequently eliding critical linguistic information.

Furthermore, overarching research questions can affect whether data should be redacted.

Research that investigates topics like human sexuality, intimate relationships, health, and politics raise expectations for protecting participant privacy. In contrast, answering certain research questions requires maintaining rich context clues. For instance, over-redacting cultural references (e.g., fictional characters) can hamstring research on language development.

In the rest of this section, we describe the broad categories of information data stewards felt might need to be redacted from their corpora, summarized in Table 2.1. The table also highlights which data stewards agreed if each type of information should be redacted either fully or conditionally (i.e., based on context). Our framework contains 23 classes of information across seven broad categories.

Direct Identifiers: These are singular pieces of information uniquely tied to a specific person. Such identifiers include names, phone numbers, ID numbers, and unique nicknames. All data stewards felt direct identifiers should be redacted.

Demographics: Attributes that link specific individuals to groups of differing sizes that share those attributes were often considered potentially identifying. Demographics include age, race, gender, sexual orientation, profession, and health status. Notably, data stewards expressed varied takes on how to redact health conditions like HIV/AIDs and other STDs. P1 and P5 suggested that because AIDs remains closely associated with gay and bisexual men, such a diagnosis could imply the data subject’s sexual orientation, socio-economic status, and age. Conversely, P2 and P10 took the opposite stance, arguing that redacting historically stigmatized diseases reinforces stigma and potentially harms participants. Data stewards also disagreed on whether to redact sexual orientation. P2, P6, and P10 all reported that members of younger generations were much less opposed, and even open, to leaving sexual orientation unredacted.

Locations: This category included street addresses, businesses, and landmarks. Businesses may be large multi-national chains like Starbucks, regional chains such as In-N-Out Burger, or local establishments, each with different impacts on identifiability. For instance, Starbucks

likely does not need to be redacted even though it is a named entity. That said, P9 noted that even combinations of common items (“that old apartment... was right next to a Cheesecake Factory and a Starbucks and a McDonald’s all on the same block”) often “becomes a little bit more localized and identifiable.”

Dates: In general, dates (events in time holding personal or cultural significance to a specific person or group) proved to be challenging for redaction. P10 recalled how their data collection began in late 2019 and the exact times of data collection became increasingly important because “a lot ... happens over Christmas” and then “we got to [COVID-19].” Similarly, P3 said they “don’t redact uncommon dates because ... that’s my research, like last year when she was hospitalized.” However, they mentioned shifting dates, yet “kept the distance between them the same ... it’s really tricky.”

Events: Participants discussed salient occurrences that recur or are marked culturally, personally, or symbolically. Examples include anniversaries, days of mourning, and memorials. Data stewards were split on redacting singular events. P3 and P10 noted changing dates to protect the privacy of surviving families and community members. Other data stewards felt an event would need both “publicly observable” and sensitive elements to warrant redaction. P3, P5, and P10 discussed “black-swan events,” such as surviving a mass shooting, witnessing a violent crime or death, or surviving a sexual assault or suicide attempt. P10 felt such events could be identifying, recalling “there was a girl that died by [disease] ... That doesn’t happen that often. Very identifying for the hospital.”

Linguistic Patterns: Any feature of language implying subpopulation membership could be identifying, including regionalisms, code-switching, famillects (family- or group-specific terminology), idiosyncratic speech, and personal narratives. P1 reported removing regionalisms related to military terminology while redacting. P8 also redacted code-switched “English-Yiddish” and “Spanglish” from their data. Regarding famillects, P9 noted that “with young children, there’s lots of weird special names that kids and families come up with” that

can be identifying to people who know them. Similarly, P8 noted that “a lot of family kind of have their own special word for a thing. . . they might have like a pet name for like a Binky or like a pack of butter.” P2 discussed how personal narratives are inherently identifying in their research: “somebody talked about their brother who had died” on a family vacation. P2 continued, “He died in a [PLACE] and. . . [was] considered an unnatural and very violent [death]. . . I didn’t publish that part because I knew somebody could Google this.”

Personal Interests / Activities: Data stewards felt hobbies or traditions uncommon in the general population often need to be redacted. Six data stewards acknowledged that decisions about redacting an interest or activity depended on its “Google-ability.” The remaining four left interests and activities unredacted as long as they did not represent a small group. P2 observed that “veterans of World War Two are now small [in number] . . . we’ve had to think about how to protect those members, but previously that wouldn’t be necessary to redact.”

2.7 Benchmarking Study: Methods

To understand the degree to which both existing purpose-built redaction tools and recent LLMs are capable of redacting potential identifiers from our framework in text transcript data, we conducted a benchmarking study. Specifically, we tested three widely used (legacy) redaction tools: Microsoft Presidio [195], Amazon Macie [9], and Google Sensitive Data Protection [109]. As LLMs are increasingly applied to language-processing tasks, we also benchmarked eight state-of-the-art LLMs on the same task: Anthropic’s Claude Opus 4.5 [14], DeepSeek-V3.2 [66], Google’s Gemini 2.5 Pro [111], OpenAI’s GPT-5.1 [211], Meta’s Llama4 Maverick and Llama4 Scout [193], and Alibaba’s Qwen3-235B-A22B and Qwen3-30B-A3B [270]. We evaluated redacting the same datasets using both the legacy tools and the LLMs. As described in Section 2.7.3, we conducted a pre-study focused on prompt engineering to configure the LLMs.

2.7.1 Datasets

In choosing the datasets for our benchmarking study, we aimed for both realism in the data we collected and full coverage of the categories in our conceptual framework (Section 2.6).

TalkBank Dataset: To this end, we sampled our primary dataset from TalkBank [174, 173, 172], a public, online repository of text corpora transcribed from naturalistic speech. While TalkBank is a multilingual repository, we used only data that was primarily in English. TalkBank is a vast collection of different corpora, most (but not all) of which lack much sensitive information, enabling them to have been released publicly. Thus, it would have been impractical to search for utterances that ought to have been redacted completely by hand. As a result, our process involved: (i) selecting TalkBank corpora that were potentially relevant based on the criteria above; (ii) using heuristics and NER to identify potential matches to framework categories; (iii) randomly sampling and then manually examining potential matches; and (iv) annotation with framework labels.

We sampled data from over 40 different TalkBank corpora, primarily from the CABank, CHILDES, and ClassBank datasets. We preferred conversational, rather than structured interview, data because conversational data encompasses a larger variety of topics and is more likely to contain identifiable information.

To span classes identified in our redaction framework within these corpora, we performed computationally assisted sampling. We employed a range of techniques to do so, including using spaCy’s NER system [130] to tag potential named entities, using similar language-detection functionality to highlight non-English text, and performing manual keyword searches for data we hypothesized might present challenges for automated redaction based on our interviews. Like many other classification problems, redaction ultimately requires an appropriate precision-recall tradeoff. Concretely, a system that overzealously flags potential redactions might protect data subjects’ privacy, yet would prove useless for the ultimate research purposes. Thus, to gauge the rate of false-positive redactions, we also

sampled *non-examples*, which are items that are *not* potential identifiers. We chose to focus on challenging non-examples that are likely to be misclassified, such as fictional characters or popular locations. While sometimes named entities, these sorts of non-examples are unlikely to make a data subject more identifiable.

Synthetic Dataset: As TalkBank data has all been released semi-publicly, it does not contain the full range of sensitive information covered by our redaction framework’s categories. To benchmark current redaction tools’ capabilities on these types of information as well, we created a companion synthetic dataset that captured the categories underrepresented in TalkBank. Our research team collaboratively created and assembled the utterances in our synthetic dataset based on everyday conversations, examples modified from existing corpora, and hypotheticals resulting from our interviews with SBS researchers. As such, while we designed these utterances to be realistic, none came directly from existing transcripts.

Our synthetic dataset includes underrepresented categories like niche hobbies (e.g., “goat yoga”) and sex or gender identifiers (e.g., “intersex”). It also includes uncommon phrases (e.g., “hot leaf juice” and “scissaurus”). As we are not evaluating the performance of a tool we built ourselves, but rather benchmarking existing third-party tools, we accepted the potential biases of this manual process because of the additional coverage it provides.

Human Redactions (Ground Truth): Three trained research assistants independently annotated both datasets by redacting potential identifiers and substituting them with labels corresponding to categories in our framework (Section 2.6). For all inconsistencies, the human annotators met and reached consensus on a ground truth label. In practice, many items in larger categories would benefit from more detailed labels; researchers may prefer different levels of granularity for semantic-, analytic-, or convention-related reasons. For the purposes of this study, we adhered to the framework as a more general structure. We distinguish between an *instance* of redaction and the number of words redacted. Each redaction instance contains one or more words. For example, one might redact “we catered from Ellie’s Diner

for the data science seminar” as “we catered from [BUSINESS_ORGANIZATION] for the [EVENT].” In this case, the single instance [BUSINESS_ORGANIZATION] contains two words, while [EVENT] contains three.

Our TalkBank dataset contains 4,466 manual redactions of potential identifiers (5,231 total words). Our synthetic dataset contains 136 manual redactions (253 total words). Table E.5 in the appendix maps the redactions for both datasets to our framework categories. The most frequent categories redacted were personal names, locations, demographics, and nicknames. Underrepresented categories include some direct identifiers, references to gender expression/sexuality, and regionalisms.

2.7.2 Choice and Configuration of Legacy Tools

We chose three of the most widely known legacy tools, configuring them in relatively standard ways. Notably, unlike LLMs, these tools cannot be customized in any straightforward way to recognize new classes of information (e.g., additional framework dimensions).

Microsoft Presidio [195] redacts sensitive information in text and images, including names, locations, phone numbers, and financial data, using identification and anonymization modules. It provides predefined recognizers for classes of identifiers, as well as support for custom patterns and models. A context-word recognizer uses surrounding text to increase detection confidence score. Finally, the anonymization module replaces identified terms with default or customized labels. Presidio is an open-source Python library that can be run locally, as is often a requirement of SBS data stewards. Microsoft Presidio can be installed using pip or Docker. If using pip, a user must install the analyzer and anonymizer packages with an NLP engine, such as spaCy. There are several options for recognizer customization, which we did not use. When run, the anonymizer engine returns a list of labels of redacted tokens, their locations, and a confidence score. It also returns the entire redacted text.

Google Sensitive Data Protection (SDP) [109], formerly known as Cloud Data Loss Prevention [108], uses 150 built-in InfoTypes to detect and redact sensitive information in structured data. The built-in InfoTypes are mostly designed to detect numerical data like phone numbers, country-specific identifiers like passport numbers, and secret data like encryption keys. The libraries also offer name, location, and ethnic group redaction, as well as custom-built redactions via dictionaries or regular expressions. Redaction is possible via a set replacement phrase or a random replacement of the same category. In addition, SDP can model risk using techniques like k -anonymity. Unlike Microsoft Presidio, Google SDP runs in the cloud. To use SDP, a user must have a Google Cloud account with billing enabled. The user can access the redaction capability directly through the Google Cloud console or through the Google Cloud API (requiring pip packages `google-auth` and `google-cloud`). We used the Google Cloud console to process our data.

Amazon Macie “discovers sensitive data using machine learning and pattern matching” [9]. We configured it with its default identifiers. Though Macie provides the option to create custom identifiers (using regular expressions or key words), we did not elect to use this option. Macie provides identifier options for personal information, financial information, and credentials. For our evaluation, we used all available identifiers provided at a sampling depth of 100%, run once. Unlike Microsoft Presidio, Amazon Macie runs in the cloud. To redact tokens with Macie, a user must have an AWS account with administrative access and a payment method for data scanning. After selecting a region, the user can create a bucket for data uploads and enable Macie through the user console. The user can create custom identifier rules, though we did not. A job must then be created with the correct bucket; there are several pages to click through to confirm settings, and an access key must be created. Macie gives a summary of results (e.g., a level of sensitivity and which documents contain sensitive data), but no specifics are provided in the dashboard if the user has not enabled data expansion.

Table 2.2: Evaluation results (word-level) on **TalkBank** (left) and the **synthetic data** (right). Trained RAs made 4,466 manual redactions totaling 5,231 words on TalkBank and 136 redactions totaling 253 words on the synthetic data.

		TalkBank Data					Synthetic Data				
		Correct		Missed		Excess	Correct		Missed		Excess
Tool (*open-weight model)		#	%	#	%	#	#	%	#	%	#
Legacy Tools	Amazon Macie	313	6.0%	4,918	94.0%	0	3	1.2%	250	98.8%	0
	Google Sensitive Data Protection	2,505	47.9%	2,726	52.1%	526	45	17.8%	208	82.2%	3
	Microsoft Presidio	4,072	77.8%	1159	22.2%	564	85	33.6%	168	66.4%	31
LLMs: General Prompt	Anthropic Claude Opus 4.5	4199	80.3%	1032	19.7%	204	161	63.6%	92	36.4%	10
	DeepSeek-V3.2*	4308	82.4%	923	17.6%	541	153	60.5%	100	39.5%	18
	Google Gemini 2.5 Pro	4952	94.7%	279	5.3%	622	202	79.8%	51	20.2%	32
	OpenAI GPT-5.1	4566	87.3%	665	12.7%	204	137	54.2%	116	45.8%	3
	Meta Llama4 Maverick*	4736	90.5%	495	9.5%	586	173	68.4%	80	31.6%	12
	Meta Llama4 Scout*	4822	92.2%	409	7.8%	839	188	74.3%	65	25.7%	17
	Alibaba Qwen3-235B-A22B*	4251	81.3%	980	18.7%	224	129	51.0%	124	49.0%	3
	Alibaba Qwen3-30B-A3B*	4424	84.6%	807	15.4%	603	168	66.4%	85	33.6%	12
LLMs: Framework Prompt	Anthropic Claude Opus 4.5	4977	95.1%	254	4.9%	372	244	96.4%	9	3.6%	18
	DeepSeek-V3.2*	4799	91.7%	432	8.3%	1249	243	96.0%	10	4.0%	44
	Google Gemini 2.5 Pro	5138	98.2%	93	1.8%	1398	245	96.8%	8	3.2%	38
	OpenAI GPT-5.1	4564	87.2%	667	12.8%	202	233	92.1%	20	7.9%	14
	Meta Llama4 Maverick*	4997	95.5%	234	4.5%	955	234	92.5%	19	7.5%	12
	Meta Llama4 Scout*	4880	93.3%	351	6.7%	741	236	93.3%	17	6.7%	16
	Alibaba Qwen3-235B-A22B*	4665	89.2%	566	10.8%	453	239	94.5%	14	5.5%	6
	Alibaba Qwen3-30B-A3B*	4837	92.5%	394	7.5%	791	240	94.9%	13	5.1%	21

2.7.3 Choice and Configuration of LLMs

We benchmarked recent versions of eight popular LLMs: Anthropic’s Claude Opus 4.5 [14], DeepSeek-V3.2 [66], Google’s Gemini 2.5 Pro [111], OpenAI’s GPT-5.1 [211], Meta’s Llama4 Maverick and Llama4 Scout [193], and Alibaba’s Qwen3-235B-A22B and Qwen3-30B-A3B [270]. Of these, DeepSeek, the Llama models, and the Qwen models are open-weight and can be run locally given sufficient equipment. We tested each model using API calls with default parameters, which emulates how a researcher may use an LLM to redact data without additional technical knowledge for fine-tuning.

We preceded our benchmarking study with a pre-study focused on the prompts used for LLMs. We initially designed two primary prompts. The first, the **general prompt** (reproduced verbatim in Appendix B), defined the task and provided formatting instructions as follows. To capture the novel insights in our redaction framework, we also tested an alternative prompt, the **framework prompt**. This prompt paralleled the structure of the general prompt, but adds an overview of our redaction framework alongside definitions/examples of

each class of information. Appendix C contains the full framework prompt, tweaked through iterative testing on additional sample data.

Because LLM performance is sensitive to the prompts used, our pre-study involved testing all four LLMs described above on the two full datasets described above. We then engaged in *prompt engineering*, or the exploration of different LLM prompts to maximize task accuracy. We focused our prompt engineering on Anthropic’s Claude Sonnet 3.5 [13] and OpenAI’s GPT-4o [210], the two LLMs that performed best in our pre-study (Table E.4 in the appendix). We built on the strategies for prompt improvement and error identification outlined by White et al. [286] and Marvin et al. [181]. Because our framework prompt outperformed the general prompt in this initial testing (Table E.4), we used the framework prompt as the basis for prompt engineering. Drawing on strategies from the aforementioned prior work [286, 181], we modified prompts by: (i) adding specific examples to each framework category; (ii) adding specific examples of items *not* to be redacted; (iii) asking the LLM to verify that no identifiable information remains; and (iv) specifying a persona or role for the LLM to emulate. We tried three key roles: (i) “redaction specialist”; (ii) “social-behavioral sciences researcher removing. . . information. . . to protect your participants’ privacy”; and (iii) “top-of-the-line redaction tool.” Appendix D details the prompts we tested and links to their full text in our OSF repository. We always provided LLMs the data one sentence at a time (following the prompt) without any accompanying metadata.

We tested all of these prompt variants on a subset of the TalkBank dataset focused on redaction errors from our initial tests, but also including text that was correctly redacted in our initial tests. Table D.1 in the appendix summarizes our results. Overall, we found that the prompt had at most a small impact on correctness, with one exception: the “counterexample” prompt performed poorest. Appendix D details these results.

Because the framework prompt on Claude Sonnet 3.5 performed best overall in terms of maximizing correct redactions and minimizing missed redactions (Table D.1 in the ap-

pendix), we chose to use the original framework prompt for our full benchmarking study. To understand the role of our novel framework in improving practical redaction performance by LLMs, we also tested our general prompt in the full benchmarking study.

2.8 Results: Benchmarking Study

Below, we detail our results, which are summarized in Table 2.2. Google’s Gemini 2.5 Pro performed best, making 96.8% of the ground truth redactions with our framework prompt and 79.8% with the more basic general prompt. The other LLMs made between 92.1% and 96.4% of the ground truth redactions with our framework prompt, compared to between 51.0% and 74.3% with the general prompt. Notably, however, all LLMs we tested failed to redact complex or nuanced identifiers, such as lengthy organization names or instances of code-switching, detailed in Table 2.3.

2.8.1 *TalkBank Data*

Overall, the legacy tools performed poorly, missing many of the consensus redactions made by our human RAs. Of these tools, Microsoft Presidio performed best, fully redacting 77.8% of the consensus redactions. However, it did not perform nearly as well as most LLMs. Microsoft Presidio and Google SDP accurately flagged many direct identifiers, basic demographics, and locations, yet they (and Amazon Macie) struggled with most nuanced categories from our framework. Surprisingly, Macie completely missed 94% of consensus redactions. Conceptually, all legacy tools focus on structured direct identifiers, missing large parts of our redaction framework.

The LLMs performed far better than legacy tools, especially when using our novel framework prompt (Appendix C) as opposed to the general prompt (Appendix B) that we expect more closely resembles the initial approach a researcher might use without knowledge of our framework. As shown in Table 2.2, Gemini 2.5 Pro performed best, making 98.2% of the

ground truth redactions with our framework prompt and 94.7% with the more basic general prompt. The other LLMs made between 87.2% and 95.5% of the ground truth redactions with our framework prompt, compared to between 80.3% and 92.2% with the general prompt. Claude Opus 4.5 improved the most from the generic to the framework prompt, going from 80.3% to 95.1% correct redactions, respectively.

Despite performing well for many categories, LLMs still missed more nuanced identifiers, as Table 2.3 details. For instance, longer terms or specific demographics (e.g., “six foot two”) were only partially redacted or completely missed. Most models often missed some ages, especially when written out alphabetically or referring to infants in terms of months rather than years. Dates were also often missed, especially days of the week or when written out.

Common improvements due to our framework prompt included specific locations, events, and linguistic categories. For example, more instances of African-American English and unique British English terms were flagged with the framework prompt compared to the general prompt. The LLMs were also more consistent in suggesting replacement labels for potential identifiers. The largest improvements occurred for demographics (697 new terms across models and datasets), events (446), and race or ethnicity (328). Surprisingly, some LLMs *missed* more items with the framework as opposed to the generic prompt (e.g., addresses). We suspect this is due to the density of information in the (relatively long) prompt. One improvement may consider running the prompts with one category at a time and consolidating results at the end, but this may be infeasible—or at the least, inefficient—for real applications.

Research transcriptions like those in the TalkBank dataset have many artifacts that do not appear in most written language, which is typically the source of LLM training data. Numerical values and linguistic markers (e.g., POSTCLITIC) were most commonly over-redacted. Family titles (e.g., “daddy,” “grandma”) were frequently over-redacted by all models despite not being especially identifying.

Table 2.3: Redaction performance by information category. The italicized “redaction” columns show the total number of redactions trained RAs made manually summed across both datasets. All other numbers indicate redactions *missed* by each tool using the generic (**G**) and framework (**F**) prompts for LLMs. Bolded numbers indicate when the framework prompt outperformed the generic prompt. We exclude Amazon Macie and Google SDP because of their poor performance overall.

Category	<i>Redactions</i>	Presidio	Claude		DeepSeek		Gemini		GPT		Llama-M		Llama-S		Qwen 235B		Qwen 30B	
			G	F	G	F	G	F	G	F	G	F	G	F	G	F	G	F
Personal Names	<i>2,433</i>	340	25	11	12	2	6	0	32	30	16	7	24	29	33	29	37	15
Nicknames	<i>432</i>	133	33	9	10	2	8	5	57	55	21	11	23	25	44	23	16	6
Demographics	<i>418</i>	29	245	24	241	27	22	18	96	96	74	23	39	26	206	86	127	53
Age	<i>25</i>	11	15	5	0	3	7	3	15	11	19	8	10	5	20	7	19	13
Sexual Orientation	<i>6</i>	6	6	0	5	0	2	0	5	0	6	0	0	0	6	0	6	0
Gender	<i>5</i>	3	3	0	3	0	0	0	2	0	2	0	1	0	3	0	1	0
Race/Ethnicity	<i>180</i>	28	109	6	105	4	12	2	11	10	26	3	10	5	61	9	42	9
Education	<i>63</i>	51	15	3	18	3	6	2	18	8	15	6	12	9	17	9	21	4
Profession	<i>53</i>	47	46	12	44	4	34	5	32	28	38	12	32	20	45	21	46	30
Health Condition	<i>25</i>	24	20	12	22	7	12	1	22	13	24	11	17	10	26	12	24	10
Locations	<i>825</i>	89	74	23	103	177	6	8	55	54	26	13	23	32	91	87	141	40
Address	<i>34</i>	21	1	0	0	0	0	0	4	4	4	4	5	4	1	0	0	4
Landmarks	<i>143</i>	48	25	9	21	5	7	2	30	25	17	11	14	12	28	11	30	12
Business/Organization	<i>81</i>	69	9	6	14	1	4	0	10	10	19	4	6	7	18	16	24	5
Dates	<i>116</i>	32	63	44	50	40	24	20	83	78	59	53	68	64	90	70	75	73
Events	<i>235</i>	148	185	10	141	111	42	4	86	86	64	16	45	33	169	90	118	54
Linguistic Patterns	<i>133</i>	96	83	47	71	19	66	18	82	60	79	39	70	53	91	53	80	36
Code-switching	<i>25</i>	14	17	4	15	4	13	4	17	8	13	5	12	3	17	2	10	5
Slang	<i>2</i>	2	2	0	2	0	2	0	2	2	2	0	0	0	0	0	2	0
Regionalisms	<i>24</i>	20	14	6	16	1	6	0	13	13	4	0	6	3	10	6	6	4
Personal Interests/Activities	<i>78</i>	47	37	5	32	7	13	0	31	31	9	6	18	8	32	20	13	8
Hobbies	<i>6</i>	5	6	0	6	0	3	0	6	2	6	1	5	0	6	3	5	0
Traditions	<i>6</i>	5	6	4	6	6	1	0	5	5	2	3	5	1	6	1	6	1
Cultural References	<i>30</i>	19	24	15	24	18	22	8	22	18	13	11	15	11	22	18	16	17
Personal Identity	<i>106</i>	38	61	8	62	1	12	1	45	40	17	6	14	8	62	7	27	8
TOTAL	<i>5,484</i>	1,327	1,124	263	1,023	442	330	101	781	687	575	253	474	368	1,104	580	892	407

In addition to providing redactions, LLMs occasionally gave justification for a redaction decision. These often referenced prevalence of items such as demographics or locations, assumptions about surrounding context, or comments about misspellings. Llama models also gave alternative categories for redaction when given the framework, resulting in several options for replacement terms. These explanations could perhaps be leveraged in future workflows. However, these models (in particular, Scout) also sometimes produced hallucinatory sample sentences *not* part of the dataset, which would require care when parsing results. Llama4 Scout also occasionally produced extraneous Python code for replacing terms in a sentence, despite instructions to limit additional text to JSON key-value pairs.

2.8.2 *Synthetic Data*

Since our synthetic examples are specifically geared towards edge cases and subtle (indirect) identifiers, legacy tools’ pattern-matching approaches fell short. For instance, Macie only flagged three potential identifiers (all names). Legacy tools often missed idiosyncratic language and specific data. Microsoft Presidio and Google SDP also over-redacted, often mislabeling mundane information as names.

In contrast, LLMs performed fairly well on the synthetic data. Without framework prompting, Gemini performed the best, missing 20.2% of ground truth redactions. Over-redactions included non-specific, common terms (e.g., “aquarium” and “rap”). With the general framework, LLMs also often missed education-related terms, linguistic anomalies, and instances of code-switching. For example, the word “metastasized” was often missed, though it holds a strong connotation with cancer and indicates a sensitive medical condition. Interestingly, however, Llama4 models and Qwen3-30B redacted some idiosyncratic phrases with the general framework prompt, suggesting that some LLMs already inherently associate uncommon language patterns as potentially identifying.

With our framework prompt, LLMs generally improved their detection rate. Qwen3-235B improved the most, reducing missed redactions from 49.0% to 5.5%. Over-redactions were similar in nature to the generic prompt. Fewer linguistic anomalies—in particular, code-switching—were missed, though education-related terms still often went unflagged. Category-by-category analysis (Table 2.3) shows that LLMs improved significantly in most categories, though they still struggled somewhat with dates, events, and health statuses. While these data may not be directly identifying, they can be used in conjunction with other terms to identify a participant.

2.9 Discussion and Conclusions

The catalyst for our work was to understand researchers’ needs for redacting potentially identifiable information in their unstructured text research data, as well as the effectiveness of current computational tools in automatically flagging these at scale. This investigation ultimately aims to help researchers protect the privacy of their human subjects while simultaneously enabling sharing of useful (i.e., not over-redacted) data with other researchers.

In our formative interviews of ten data stewards (eight PIs and two postdocs), we observed frequent worry about protecting the privacy of human subjects and nuanced, yet mostly informal, practices for trying to protect data subjects through manual redaction. At the same time, the data stewards noted the onerous time requirements of manual redaction, as well as concerns about outsourcing such a critical task to potentially inexperienced RAs. Echoing findings from Guo et al. [117], none of the data stewards knew of any computational tools they could trust for automating redaction.

Through these interviews, we co-developed and iteratively refined a redaction framework outlining different categories of information that may need to be redacted from SBS text transcripts depending on the research context and risk to human subjects. Our framework encompassed some categories of data that have received substantial attention in the academic literature and in existing tools. For instance, our framework’s direct identifiers, such as personal names and numerical identifiers, are a major focus of legacy tools. Our novel framework also highlighted aspects that have been discussed far less in the literature and operationalized far less in practice. For instance, data stewards highlighted in their threat models how transcripts containing linguistic code-switching and famillects, or personal interests and aspects of identity, could make participants identifiable to people who know them.

To gauge the ability of current computational tools to aid researchers in redacting text transcripts, we then benchmarked both purpose-built legacy tools and emerging LLMs with

various prompting strategies. While Microsoft Presidio is widely used for redaction in the privacy engineering community, it performed poorly in our benchmarks, completely missing 22.2% of the human redactions on our TalkBank dataset and 66.4% on our synthetic dataset. The tools from Google and Amazon did even worse. While Presidio did well at flagging structured direct identifiers, it fell far short when considering the more expansive view of re-identifiability in the novel framework co-created and refined with our data stewards.

In contrast, the eight state-of-the-art LLMs we tested were far more effective at flagging potential identifiers. Notably, we found that including our redaction framework as part of the prompt substantially improved the LLMs’ effectiveness at redaction, whereas following typical strategies for prompt engineering [286, 181], such as providing examples, counterexamples, or a defined role for the LLM, had minimal impact. That said, even the best-performing LLM, Gemini 2.5 Pro, completely missed 1.8% of redactions made by human RAs, likely necessitating manual correction.

Anticipated Redaction Workflows: The data stewards we interviewed understandably felt highly responsible for protecting their human subjects from re-identification or other privacy leaks. As such, even the fairly high accuracy of Gemini or other LLMs would likely be insufficient for fully automated redaction. In our interviews, the data stewards shared their own mental models for what makes their particular participants identifiable, as well as the risks to specific participants based on information shared in their specific transcripts. As such, we find it very hard to imagine any redaction workflow that would not involve manual verification of redactions (and non-redactions) by a human researcher before data release.

That said, there remain open questions about how best to incorporate humans into redaction workflows. How will a human manually review transcripts that have been redacted by a tool before sharing them? How should data dependence (e.g., a concept appearing repeatedly throughout one transcript) be handled? How can this workflow leverage data stewards’ unique knowledge of potential identifiers that may appear in the data, risks to

participants, and aspects of the data that must be preserved because of their centrality to the research? As our framework prompt contains human-intelligible definitions of each facet of our redaction framework, one could imagine a data steward directly modifying the framework prompt before using it. The data steward might articulate additional categories that should be redacted and remove categories that do not make their subjects identifiable.

Pitfalls of LLMs in Redaction Workflows: We note that even open-weight LLMs like Llama, DeepSeek, and Qwen can be difficult or impossible to run on commodity hardware (e.g., researchers’ laptops). Sending unredacted data to cloud services can raise both ethical and legal questions. For instance, IRB agreements or promises made to human subjects might prevent sending data to the cloud. Part of Microsoft Presidio’s appeal has been its open-source nature and ability to be run locally on minimal hardware. However, we anticipate that a research group with sufficient resources would be able to run open-weight models locally without needing to access the internet, especially smaller models like Qwen3-30B.

We ran additional benchmarks on smaller (typically older) LLMs that are more feasible to run locally, such as OpenAI’s GPT-2, Meta’s Llama2-7B, and Meta’s Llama3-8B-Instruct. Unfortunately, we found that their results were unusable. Notably, GPT-2 and Llama2-7B often generated recommendations for manual redaction, rather than redacting the input data. Similarly, over half the time Llama3-8B-Instruct generated different text samples instead of redacting, hallucinating text samples not provided as input. Furthermore, even these smaller models require substantial resources (e.g., a high-end GPU) to run locally. That said, future small LLMs, particularly those trained to follow instructions via reinforcement learning from human feedback (RLHF), may hold promise, and models like Qwen3-30B-A3B show that it is indeed possible for smaller, open-weight models to vastly outperform legacy tools.

In the meantime, researchers may need to rely on data use agreements and focus on models with appropriate terms of service. Alternatively, some organizations (including our own) have deployed private LLMs [248], or local copies of foundation LLMs that do not

share data outside the organization. Rather than requiring each individual PI to solve computational or regulatory problems on their own, the deployment of private LLMs could be a way forward.

Generalizability: While we focused on SBS research data for concreteness, our work presents lessons we expect to apply broadly to any kind of unstructured text transcripts, whether related to research or not. These lessons include types of information that make speakers identifiable in text transcripts, measurements of the category-by-category redaction effectiveness of both widely used legacy tools and emerging LLMs, and how including our framework in an LLM redaction strategy improves redaction effectiveness. That said, redaction strategies will necessarily depend on context. Moreover, redaction strategies may differ between short-form data (e.g., social media posts), medium-form data (e.g., a full interview transcript), and long-form data (longitudinal data from one individual).

Our redaction framework is intended as a starting point for researchers in considering types of data to redact. We anticipate that data stewards—who know their data best—would tweak the framework before using it in an LLM prompt. Our framework can encourage researchers to consider the appropriateness of redacting each category relative to their subjects’ re-identifiability and risks, as well as the data’s research purposes. All prompts we tested, even in our pre-study, can be characterized as either zero-shot or few-shot. Future work may consider changing how information is conveyed to the LLMs via interactive uses of LLMs, providing different amounts of text at a time to the LLM, or fine-tuning LLMs.

Conclusions: In short, while LLMs show promise for redaction beyond legacy tools, they do not fully solve the problem of redaction for SBS researchers. Nonetheless, we expect that the use of computational tools as part of a redaction workflow would likely save researchers substantial time and could even flag identifiers humans miss. It may also be possible to leverage the unique strengths of different tools in concert. For instance, whereas LLMs

sometimes struggled with identifying dates, legacy tools like Presidio fared better at flagging such structured data. Our redaction framework and benchmark data provide a starting point for researchers hoping to redact their data, yet researchers ultimately know their participants, their data, and applicable regulations best. We therefore envision that any redaction process be at most *semi*-automated, with a human in the loop, to protect the privacy of human subjects.

CHAPTER 3

MARI: A SEMI-AUTOMATED, HUMAN-IN-THE-LOOP GRAPHICAL REDACTION WORKFLOW

This chapter presents the design and preliminary evaluation of MARI (Mostly Automated Redaction Interface), a novel graphical workflow tool developed to address the limitations identified in Chapter 2. This work represents an ongoing research project currently in the final stages of data collection and analysis. While final statistical results are pending, this chapter provides a rigorous account of the system architecture, the study protocol, and the preliminary qualitative insights that validate the need for a semi-automated, graphical approach to privacy redaction.

3.1 Dissertation Author’s Contributions

This chapter represents joint work with Daniel Byrne, Anwar Kader, and Blase Ur. The author of this dissertation served as lead author on the version to be submitted for peer review and is sole author of all sections presented in this chapter, with the exception of Related Work, which had coauthoring on initial versions of Sections 3.3.1 and 3.3.2.

3.2 Motivation

Human subjects data is often expensive and labor-intensive to collect, making data sharing a valuable practice that allows researchers to access and analyze datasets they did not gather themselves. Much of this data takes the form of *unstructured text*—natural language that lacks a standardized format, such as conversational transcripts, diary entries, or social media posts. While these datasets offer insights across multiple areas of social-behavioral science, their inconsistent structure complicates efforts to identify and protect sensitive information.

Sharing such data therefore requires privacy practices like *redaction*, which involves removing or substituting *identifiable data*—information that could potentially re-identify study participants. Unfortunately, the very unstructured nature of these texts makes locating and redacting identifiable information a time-intensive and challenging task.

Furthermore, existing frameworks of identifiable information often miss items that can identify a participant to someone who knows them (Chapter 2). Most frameworks focus on terms that can be found using pattern-matching or search, such as personal identification numbers, credit card numbers, and addresses. However, these structured identifiers are relatively rare in natural language, where information can be expressed in myriad ways. Certain regionalisms or idiosyncratic speech patterns, for instance, can uniquely identify participants, yet it is impractical to search for every possible dialect variation or unusual word combination.

This fundamental mismatch between the fluidity of human language and the rigidity of rule-based detection creates a problem for automated redaction. Current tools [9, 108, 195] exist for automatically flagging potential identifiers, though they are equipped mostly for pattern-matching. Tools with additional natural language processing (NLP) capabilities can be more successful, though they may still miss categories of identifiable information that are not readily expressed with consistent linguistic features (Chapter 2). Large language models (LLMs) also have potential for success in redaction tasks (Chapter 2). However, they risk under-redacting by missing subtle identifiers—a direct privacy breach—or over-redacting by obscuring crucial research terms, diminishing data utility. Since any missed identifier compromises privacy, and researchers must preserve data integrity for potential use, human verification of automated outputs remains an essential final step.

Moreover, a significant barrier to widespread adoption lies in the user experience: none of these tools offer an intuitive, graphical workflow. While some have dashboards with data statistics, the primary workflow for the redaction process entails running a program in the

background on data provided by the user. In many cases, this demands technical proficiency, requiring researchers to write code, manage dependencies, or navigate complex configuration setups before a single piece of data is redacted. This lack of transparency makes it difficult for non-technical researchers to inspect why a specific term was flagged or to manually intervene in real-time, effectively decoupling the researcher from the decision-making process.

To that end, we introduce *MARI* (Mostly Automated Redaction Interface), a graphical user interface designed to help data stewards efficiently review proposed redactions while incorporating their human judgment about research context and re-identifiability risks. We implemented MARI’s models and graphical user interface in Python and will be open-sourcing its code.

MARI transforms redaction from a passive, opaque operation into an active, iterative workflow. Rather than simply outputting a static, pre-redacted file, MARI performs initial automated flagging and then presents the flagged segments directly to the user for immediate verification. This interface empowers researchers to correct false positives (reverting erroneously flagged data) and catch false negatives (manually redacting missed identifiers) within the same environment, ensuring that the final dataset balances privacy protection with data utility.

MARI’s effectiveness stems from a hybrid redaction pipeline informed by our prior automated tool study on unstructured text. To concretize this workflow, MARI integrates multiple layers of identifier detection: First, building on an augmented version of Microsoft Presidio, MARI handles structured identifiers (e.g., phone numbers, IDs) and other standard identifiers like names with pattern-matching and basic NLP techniques. To address transcription errors, such as misspellings or homophone substitutions that often cause standard tools to miss identifiers, MARI incorporates additional linguistic features that capture words semantically related to potential redactions. Finally, for complex semantic and cultural nuances, MARI leverages large language models selected for their superior performance

in our initial evaluation. This includes Meta’s Llama 4 Maverick [193] for local, offline execution (ensuring data protection) and Google’s Gemini 2.5 Pro [111] for optional cloud-based computation (offering advanced contextual understanding).

To validate MARI’s efficacy, we conduct a within-subjects usability study involving participants with prior human subjects data experience. Participants experience three workflows: fully manual redaction, a backend-only baseline where we used MARI’s underlying models (Presidio, Llama, and Gemini) to generate a pre-redacted file for participants to correct, and the MARI workflow featuring the interactive graphical interface. Participants are asked to redact three datasets of similar composition, one with each process, with the same amount of time for each. We evaluate effectiveness by number of terms redacted, corrected, and verified in the timeframe and user surveys with Likert-scale questions on a modified System Usability Scale (SUS). After all tasks, participants answer Likert-scale questions about graphical redaction methods, rank each of the redaction tasks on different metrics, and participate in a structured interview. We also conduct qualitative analysis on screen recordings of participants’ redaction process and open-ended responses to interview questions.

Preliminary observations suggest that the MARI workflow enables significantly higher task completeness within the time constraint compared to both baselines. Although some participants noted an initial learning curve, the data shows that MARI consistently outperformed other conditions in actuality, achieving the highest proportion of correctly redacted terms across all participants. Furthermore, MARI’s approach of collecting and displaying all identifiers per category upfront helped participants identify a broader range of potential redactions compared to linear methods, suggesting the tool effectively balances automation with user oversight. Notably, participants preferred MARI to the other two methods and perceived it as more efficient.

3.3 Related work

Automated redaction has evolved from simple pattern-matching scripts to AI- and LLM-driven systems, yet significant gaps remain in handling the nuanced, context-dependent identifiers common in human subjects research. This section reviews the landscape of automated redaction tools, the emerging role of large language models (LLMs), and the dearth of accessible graphical interfaces that would enable non-technical researchers to effectively participate in the redaction workflow.

3.3.1 *Automated Redaction*

Commercial Redaction Tools. Commercial redaction tools focus on removing highly structured identifiers (e.g., ID numbers, credit card numbers), as well as names and geographic locations [157]. The open-source Python library Microsoft Presidio [195], cloud-based Google Sensitive Data [109] (formerly Cloud DLP), and Amazon Macie [9]

Tools like Microsoft Presidio [195], Google Cloud DLP [109], and Amazon Macie [9] offer automated detection of personally identifiable information (PII). These systems rely primarily on pattern matching, keyword dictionaries, and basic linguistic features and heuristics to identify structured data such as phone numbers, credit card numbers, and standard ID formats [157, 209]. Newer “AI-powered” platforms like Redactable [228], iDox [135], and AI-Redact [258] have introduced more robust detection capabilities, yet they remain focused on high-confidence, easily-detectable entities rather than the subtle, context-dependent identifiers that characterize human subjects data. Furthermore, these tools are designed to completely remove and “black out” information rather than mask it, reducing data utility.

Redacting Unstructured Text. While early redaction efforts focused on structured domains like medical records or legal logs, applying these techniques to natural language reveals a need for significant paradigm shifts. Traditional approaches rely heavily on NER and

pattern matching, which excel at identifying rigid entities but struggle with the semantic ambiguity of natural language. For instance, standard NER models often fail to recognize that a specific linguistic pattern (e.g., a regional dialect) serves as a stronger identifier in a small cohort, a limitation echoed in domain-specific studies ranging from clinical text [28, 146] to chat data [2].

Identifiers are often accompanied by contextual clues, which many traditional systems miss. Bosch et al. [36] attempted to mitigate this by incorporating public records and part-of-speech tags, while others proposed modeling token context and hierarchical relations to distinguish identifiers from non-identifiers [236, 67, 84]. However, these approaches are still insufficient for broader applications—Guo et al. report that researchers still often rely on manual and social processes to redact their data and that current tools are insufficient for de-identification [117].

3.3.2 *Redaction with LLMs*

Large language models (LLMs) have emerged as powerful tools for redaction, offering contextual understanding that surpasses traditional pattern-matching and NER systems. Studies demonstrate their superiority in detecting named entities [17], generalizing specific terms to broader categories [222], and removing author-attributable traits [91]. Other approaches further leverage word likelihoods to flag potential sensitive content [5, 118]. However, despite these advances, current LLM-based approaches remain narrow in scope.

Common limitations include context fragmentation, small category pools, and a lack of human oversight. Benchmarks reveal that even advanced models struggle with context-dependent identifiers, resulting in low security scores when sensitive information relies on reasoning across multiple segments [140, 55]. Furthermore, most systems target narrow, predefined categories (e.g., standard PII) rather than the full spectrum of identifiable information we defined. Existing work treats redaction as a black-box operation [72, 38], failing

to provide mechanisms for researchers to verify, correct, or override model decisions. This approach ignores the essential trade-off between privacy and data utility, highlighting the need for a hybrid system that combines LLM detection with human verification.

While tools like RedacBench [140] allow users to specify policies and redact data using LLMs, the user must create their own redaction parameters and operate in browser with API LLM access. PRvL [99] uses a suite of LLMs to redact PII in unstructured text and similar datasets, yet the tool only focuses on typical PII, with no capabilities for flagging cultural or linguistic identifiers. Many of these tools also by default miss demographic information other than which can be found in typical ID documentation.

3.3.3 Graphical Workflows

The majority of redaction tools are command-line libraries (Presidio), cloud APIs (Google DLP, AWS Macie), or desktop applications designed for legal or compliance workflows rather than research settings (Adobe Acrobat Pro, Redactable, iDox). None of the major tools offer an intuitive, graphical workflow that allows non-technical data stewards to verify or correct redactions before they are finalized. Existing graphical solutions, such as Adobe Acrobat Pro [3] or Redactable [228], focus on permanent “black-out” redaction for legal compliance rather than substitution for preserving context in a research setting. Tools with dashboards prioritize metrics and statistics, focusing on risk assessment rather than editorial review [9, 108, 135]. This creates a “black box” effect where researchers cannot easily understand why a specific term was flagged, nor can they efficiently override false positives (such as fictional character names) or catch false negatives without resorting to manual text editing. To our knowledge, there are currently no graphical tools with masking capabilities (i.e., redaction with context-preserving replacements) that include aggregated results and allow users to modify redactions before finalization.

3.4 MARI Design

In this section, we outline the design of our graphical user interface, MARI. Typical redaction methods are time-consuming and miss important identifiers or redact items that do not need to be redacted. To that end, MARI combines automated detection methods with an interactive review process that allows users to verify, modify, and contextualize flagged content. Our tool also integrates existing offline redaction techniques with large language models to broaden coverage of identifiable data types in our framework.

3.4.1 *Automated Redaction Pipeline*

The backend of our automated redaction system is designed as a modular framework, providing users with three processing options: a customized iteration of Microsoft Presidio, Llama 4 Maverick, and Google Gemini 2.5 Pro. Users can select a combination of these models that best fits their needs, allowing for a balance between computational efficiency, data privacy, and linguistic nuance. For the LLMs, we pass in the framework prompt from the previous study (see Appendix C).

Microsoft Presidio is an open-source PII detection toolkit, integrated here as a local Python library via the pip package manager. While Presidio traditionally relies on a combination of NER models (e.g., spaCy or transformers) and regular expressions, our implementation features an additional layer of analysis. We retained the standard Presidio models for baseline detection but augmented them with a dedicated step that performs a secondary pass over the text, leveraging conversational context and additional semantic characteristics to flag potential identifiers that the primary NER models missed. This approach avoids the overhead of model retraining to prevent introducing dataset-specific biases that could limit the system’s generalizability across different domains. The augmented Presidio option ensures that MARI can perform reasonably correct, high-speed redaction for users prioritizing local execution and data residency.

To address the limitations of legacy tools—which often struggle with contextual ambiguity in natural language—we integrated two LLMs selected from our preliminary benchmarking tests. We include LLMs based on their ability to process language at a more complex level and follow our redaction framework.

Gemini is a larger, more powerful model that can be accessed through API calls given an API key. This model performed the best among all models in our benchmark tests, missing just 1.8% of identifiers in the TalkBank dataset. However, Gemini is closed-weight, meaning researchers would need to use cloud-based computation on their data.

Maverick offers a middle ground between Presidio and Gemini. We chose this model as a small-scale, open-weight architecture capable of being deployed locally, provided sufficient infrastructure for computation. In our benchmarks, Maverick demonstrated the best performance among open-weight models of comparable parameter counts, making it the best choice for secure, local LLM-based redaction.

While there may be residency issues with sending research data to an LLM, some steps have been made to ensure user privacy. We note that DuckDuckGo’s Duck.ai [73] allows for ostensibly private LLM chats with SotA LLMs, where prompts are stripped of metadata and chat content is not retained. Proton’s Lumo AI assistant [225] also provides confidential chats, with the option for automatic chat deletion. However, since there is no API access for either system, it is infeasible to integrate them into our current local application workflow. Duck.ai also does not provide access to Gemini 2.5 Pro, and Maverick is locked behind a paywall; since Lumo is non-API, we have not been able to perform rigorous testing on its redaction capabilities. Furthermore, though there is a privacy guarantee stated by these services, many researchers may not feel comfortable redacting their data under any online connection. Thus, we do not use either of these services in our application.

3.4.2 *User-Facing Redaction Interface*

The user interface is structured as a multi-stage workflow designed to guide users through the process of modifying and applying redactions. Our interface was implemented in Python, using the Tkinter GUI toolkit. This method was chosen for its flexibility and ability to run locally, as opposed to a web application.

MARI’s interface design is grounded in Ben Shneiderman’s “Eight Golden Rules of Interface Design” [250] to mitigate the cognitive load of redaction. The application maintains uniform terminology, color coding, and interaction patterns across all stages leading up to the final file preview, offering immediate visual confirmation of every action. The interface also prioritizes “[reducing] short-term memory load” by presenting flagged terms in categorized groups rather than forcing users to scan the entire transcript linearly. This prevents the uncertainty that is common with batch-processing tools, allowing researchers to verify their decisions before finalizing the dataset. MARI treats every redaction decision as a reversible toggle rather than a permanent decision, and users can always go back to a previous step to modify a prior decision. Where possible, we reduced clickable items and guided the user with text to provide context to possible actions.

MARI Workflow. The user starts the process by loading their data into the application and selecting which combination of redaction architectures to use. After automated redaction has run, users receive a summary of terms redacted and next steps, beginning the editing phase. Settings, such as active categories and redactions, are saved per project and persist between application instances. A persistent timeline navigation bar at the top of the interface enables users to move between different stages after reaching the preview phase, though it proceeds linearly until then to ensure users do not miss parts of the process. Users receive instructional pages before each stage to highlight their key features and functionality.

Editing begins with a category selection page, where users determine which categories of

terms will be subject to redaction (Figure 3.1). These categories adhere to our framework with an additional category as a catch-all for potential terms flagged by LLMs whose label do not fully match a category. All categories are active by default to ensure the highest coverage of potential identifiers; users select categories to ignore based on which are essential to their data utility or are not sensitive enough for redaction.

For each active category, users are directed to a sequence of category-specific pages that display all terms pre-labeled under that category, regardless of labeling accuracy (Figure 3.2a). Within these pages, users can inspect individual terms and access detailed views that present their occurrences in the original document alongside their corresponding replacements (Figure 3.2b). This functionality supports granular control, allowing users to selectively redact or retain specific instances. Additionally, users can examine terms in full context by navigating to its position in the preview phase.

Following the category-specific pages, there is a page highlighting terms that are similar to those already redacted, helping identify potential errors such as typos, mistranscriptions, or other inconsistencies (Figure 3.3). Similarly, users can view specific terms with chosen granularity and toggle which terms should be redacted. To avoid oversaturating the page, we highlight only terms that are most sensitive and have high agreement for redaction between researchers (see 2.5).

The redaction workflow concludes with a preview page, where users can review the fully redacted document and make final adjustments, including adding missed redactions or modifying existing ones (Figure 3.4). At any time before export, the user can go back and make changes to any items, categories, or term collections. After changes are approved by the user, they can then export the file to a chosen location.

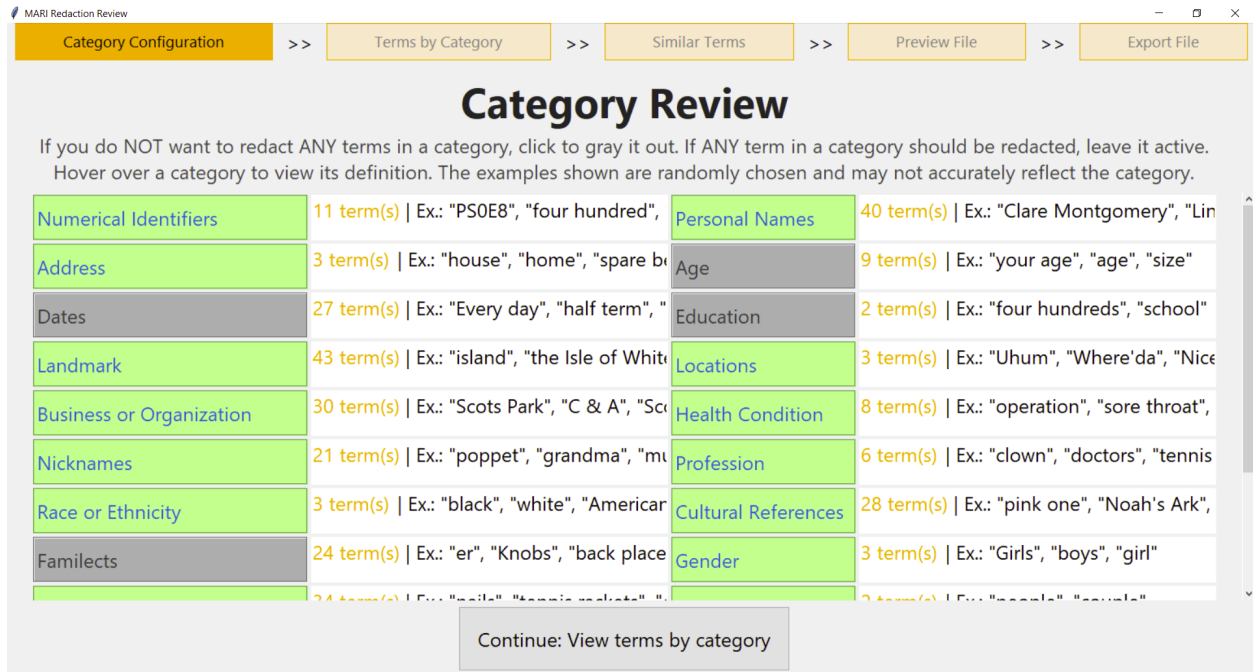


Figure 3.1: Category selection page for MARI. All categories are active by default; the user can select or deselect any categories to customize redaction to their privacy requirements.

3.5 Methodology: User study

To test the efficacy and efficiency of MARI, we are currently conducting a within-subjects usability study, for which four participants have completed the full protocol. Study participants completed three different tasks related to data redaction: manual redaction in a Google spreadsheet, correcting MARI redaction output in a Google spreadsheet, and using MARI's graphical interface for redaction. A within-subjects design was selected to control for individual differences in redaction ability and familiarity, as well as interpretation of identifiable information.

Participants completed all three tasks while engaging in a think-aloud protocol, enabling us to capture their reasoning, decision-making strategies, and points of confusion during redaction. The order of the tasks and datasets for each participant were pseudorandomized to mitigate ordering effects and learning biases. After each task, participants responded to modified SUS statements about the relevant redaction method; after completing all tasks,

MARI Redaction Review

Category Configuration >> Terms by Category >> Similar Terms >> Preview File >> Export File

Personal Names, category (2/24)

Below, clicking on a checked box will STOP redacting all instances of that term. To selectively choose which instances to redact (or not), click on the term to view all appearances in context. Hover over a term to view one example in context.

40 terms flagged in this category.

Term	Redacted	Not Redacted	Term	Redacted	Not Redacted	Term	Redacted	Not Redacted
☒ Amy (35 0)	☒ Clare (20 0)	☐ She (0 62)	☒ Lindy (12 0)	☒ Jenny (9 0)				
☐ he (0 44)	☐ him (0 12)	☒ Bill (7 0)	☒ Lucy (3 0)	☒ Mr Staton (3 0)				
☒ Eve (2 0)	☐ [...] (0 348)	☒ Charlie (2 0)	☐ you (0 427)	☒ Helen (2 0)				
☒ Deborah (2 0)	☒ Caroline (2 0)	☒ Christine (1 0)	☒ Taff (1 0)	☒ Peter (1 0)				
☒ Mrs. (1 0)	☒ Mr Statton (1 0)	☒ Clare Montgomery (1 0)	☒ Mr. Statton (1 0)	☒ Bill Boyd (1 0)				
☒ David (1 0)	☐ friend (0 1)	☒ Lindsay (1 0)	☐ I (0 474)	☒ Miss MacDonald (1 0)				
☒ Orangey (1 1)	☒ Bronwen (1 0)	☒ White (3 8)	☐ you (0 2)	☒ Rhionna (1 0)				
☒ Shhhh (1 0)	☒ Mum (8 0)	☒ will (1 15)	☐ myself (0 2)	☐ he'll (0 2)				

Previous Next category

(a) An overview of all terms flagged in the “Personal Name” category. The user can choose to redact or not redact each term globally. If a term is clicked, the user views each term on a per-instance basis (below).

MARI Redaction Review

Category Configuration >> Terms by Category >> Similar Terms >> Preview File >> Export File

will - Personal_Names

Redact	Sentence	Replacement	Preview
☒	Well, will , put it, put it into	Personal_Names	See in context
☐	tip it carefully ... that will do for the time being	(No Replacement)	See in context
☐	Will I listen?	(No Replacement)	See in context
☐	don't think Marks and Spencers will give them free.	(No Replacement)	See in context
☐	I will take the yellow one I	(No Replacement)	See in context
☐	Yes it will .	(No Replacement)	See in context

Previous page 1 of 3 Next

Back to Personal_Names

(b) The per-instance view for the term “will”. This page allows for fine-grained toggling of instances where the term occurs; for example, “will” occurs as a name once, but other instances are verbs.

Figure 3.2: With MARI, the viewer first views terms sorted by category. They may make changes to replacement types, toggle redaction, and view instances where the term occurs, both redacted and not.

Category Configuration
>>
Terms by Category
>>
Similar Terms
>>
Preview File
>>
Export File

Similar Terms Overview

Here, choose any similar terms (typos/mistranscriptions) you wish to redact. Click on the term to view all appearances in context. Hover over a term to view one example in context.

41 highly identifiable terms with similarities found.

Redacted Terms	Similar Terms	# Similar Terms
Amy	Ex. Amt, Ams	2
Clare	Ex. clear, Ckare, Flare	3
Lindy	Ex. Lindsay	1
Bill	Ex. ball, Bull, big, till, build	10
Lucy	Ex. Lice	1
Eve	Ex. leave, e, eat	3
Taff	Ex. tac, staff	2
Lindsay	Ex. Lindy	1
will	Ex. wild, Bill, wish, till, wall	7
nic	Ex. nice, thick, pick, Pic	4

Previous
Next Step

(a) The “similar terms” overview page highlights specific terms the user has selected for redaction, indicating whether they have terms linguistically similar that have not been redacted. The red rectangles were added to highlight terms that were typos or mistranscriptions.

MARI Redaction Review
Category Configuration
>>
Terms by Category
>>
Similar Terms
>>
Preview File
>>
Export File

Terms linguistically similar to "Bill"

Below, clicking on a checked box will STOP redacting all instances of that term. To selectively choose which instances to redact (or not), click on the term to view all appearances in context.

10 term(s) similar to "Bill" found.

☐ ball (0 1)	☐ bell (0 1)	☐ big (0 7)	☐ bin (0 2)	☐ bit (0 28)
☐ build (0 1)	☐ Bull (0 1)	☐ ill (0 2)	☐ sill (0 1)	☐ till (0 3)

Back

(b) Selecting a term to view its similarities presents a page structured like a category overview. Users can select a term to see a view like that of Figure 3.2b.

Figure 3.3: The “similar terms” page allows viewers to find terms that might otherwise be missed: typos, mistranscriptions, and other terms similar to those already redacted.

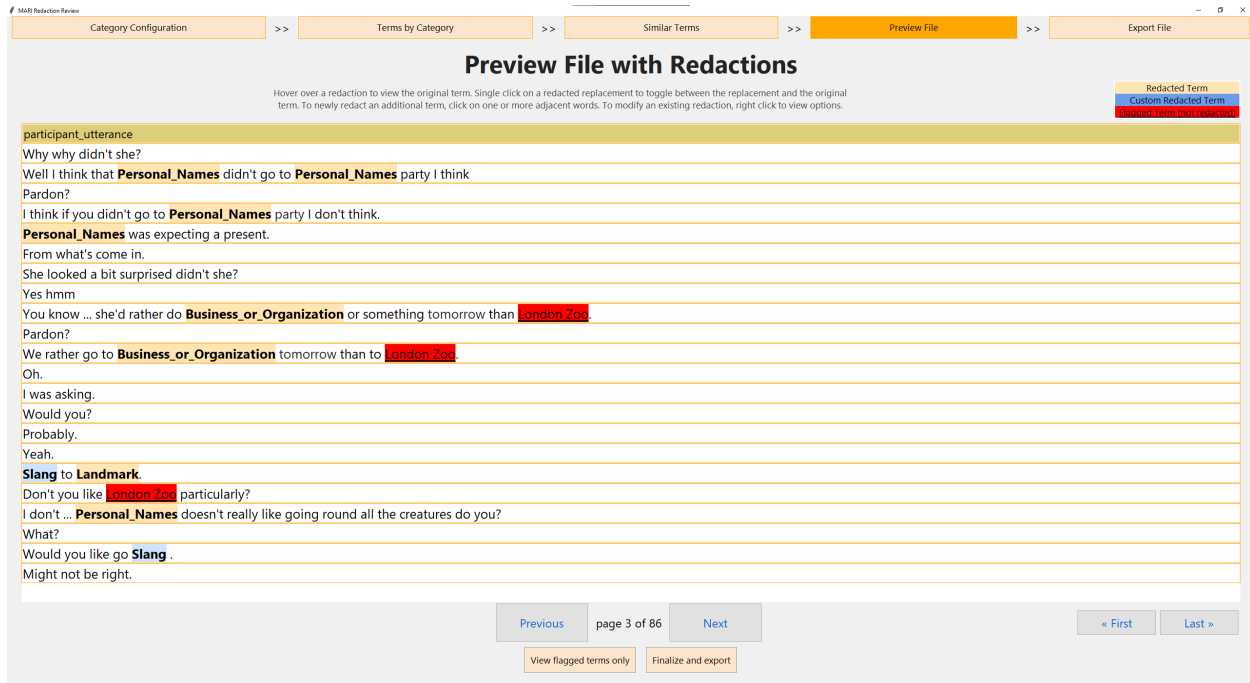


Figure 3.4: The overview page of our application shows the user a preview of their file with redactions applied. Here, the user can edit the sentence (if there are typos), change redaction status for any term or combination of terms, and modify replacements.

participants were asked to rank each method relative to each other and respond to various statements on redaction. Participants also completed a demographics survey and participated in a brief semi-structured interview regarding their experience with the redaction process and prior knowledge of redaction.

3.5.1 Recruitment

We recruited adults (age 18+) with at least some college education and familiarity with navigating computer programs. These criteria were chosen to ensure a baseline level of technical literacy, as participants were required to interact with a novel graphical interface and work with spreadsheets during the redaction process. Furthermore, we required that participants have some past experience analyzing data collected from human subjects. Eligible participants include someone who has worked with data collected from human sub-

jects via (i) working as a research assistant or lab manager, (ii) completing an academic thesis/dissertation, or (iii) co-authoring an academic publication, or others with equivalent experience. Individuals with human subjects experience are more likely to be familiar with common forms of identifiable information, existing redaction practices, and ethical considerations surrounding participant privacy. This requirement reflects the applied context of our system, which is designed for use in research where human subjects data is handled. By targeting this population, we aimed to evaluate the system with users who reflect its intended audience, while ensuring that participants could provide meaningful feedback on both the usability and effectiveness of the redaction workflows.

3.5.2 *Study Design*

In each task, participants were asked to redact one of three sets of data (of approximately equal length) containing potentially sensitive and personally identifiable information; however, not all data in these sets were necessarily considered sensitive or personally identifiable. This ensured that participants had to actively interpret and apply redaction criteria rather than redacting all content.

Prior to the tasks, participants were provided with standardized definitions of sensitive and personally identifiable information based on our framework and instructions how to redact data with each method. Positive (correct redactions) and false positive (non-identifiable) examples were also provided for informational contrast and to further highlight task goals. To simulate a real-world scenario for redaction, we asked participants to redact only a subset of categories according to our framework. This selection can be seen in Table 3.1; categories were chosen based on the total number of baseline redactions in each category to attain approximately the same proportion of identifying terms in each dataset.

For each task, participants were assigned one of three modes for performing redaction: (i) manual inline redaction using an unannotated Google Sheet, (ii) inline correction of pre-

Table 3.1: Categories chosen for redaction in our user study. Note that Demographics as a category is not included for redaction, but is a contextual header for Health Condition. Cultural References were also disincluded from the Personal Interests / Activities category.

Direct Identifiers
↪ Personal Names
↪ Nicknames
↪ Numerical Identifiers
Demographics
↪ Health Condition
Locations
↪ Address
↪ Landmark
↪ Business / Organization
Events
↪ Specific Events
Personal Interests / Activities
↪ Hobbies
↪ Traditions
↪ Personal Identity

redacted content in a Google Sheet, and (iii) redaction using the MARI graphical interface. Each participant used all three interfaces and datasets in pseudorandomized order.

In each 90-minute session, 60 minutes were allocated for the data-redaction tasks split into 20-minute intervals for each task (15 minutes for a redaction task, and 5 minutes for a short survey and break). The remaining time was allocated for briefing participants, redaction method tutorials, and short interviews.

In each task, participants were instructed to redact identifiable information according to a predefined subset of our framework. Participants were asked to review datasets and replace sensitive or identifying elements with the appropriate category labels. They were instructed to use their best judgment and apply the most general label when uncertain. Conversely, when a clear match existed, they selected the most specific category or subcategory. For multi-token entities, a single redaction label replaced a span of text for consistency (e.g., “John Smith” was replaced with “Personal_Name”, not “Personal_Name Personal_Name”).

Each task used the same datasets either unredacted (in their original state) or pre-redacted by Presidio, Maverick, and Gemini. For the manual task, participants worked directly on raw text and were instructed to ignore items that did not fall within the designated redaction categories. In the prefilled condition, participants corrected existing redactions

using a separate spreadsheet containing the original text for reference. In contrast, the MARI interface provided integrated access to contextual information (e.g., hover actions to display the original text).

3.5.3 Data Collection

We collected both quantitative and qualitative data to assess usability, performance, and user perception.

After completing each task, participants completed a short survey of Likert scale responses regarding the task they just completed. The questions were the same for each task, based on modified SUS statements to reflect task-specific metrics such as perceived efficiency, confidence in redaction accuracy, and ease of use. Using consistent questions across conditions enabled direct comparison of subjective usability metrics.

Following the completion of all tasks, participants answered a short demographics survey, verbally answered interview questions, and completed additional Likert survey questions related to graphical redaction methods. Participants were also asked to rank each of the tasks relative to each other for preference, efficiency, accuracy, and overall progress.

Participants' actions were screen-recorded as they completed the redaction tasks on a provided laptop to capture editing patterns, navigation, and interaction with interface features. Participants' think-aloud processes during task completion were audio-recorded and transcribed for further analysis, providing insight into participant reasoning, error identification, and strategies for resolving ambiguity. These were used to contextualize quantitative findings and identify usability challenges.

Finally, we examined the redacted documents for each task to evaluate participants' actual correctness and completeness scores. We recorded number of potential identifiers seen in the timespan (according to our baseline redactions), the number of these that were correctly redacted, the number missed, and if applicable, the number replaced with a new

category label. As a point of comparison, we also examined the fraction of terms seen and redacted to the count in the full document; for the prefilled and graphical tasks, however, this includes terms that the participants did not see or correct, so these are treated as supplementary completeness metrics.

3.5.4 *Datasets*

Datasets were derived from publicly available data from CABNC [6], an English-language corpus containing transcriptions of British natural conversation. We truncated three transcripts, one per task, to approximately 10,000 words, assuming participants would not be able to fully read and redact the entire subset in the allotted time based on an average reading speed of 200-300 words per minute.

Though these datasets are publicly available, identifying contextual information was publicly released under ethical guidelines and participant consent. Thus, we were able to find many instances of identifiable information corresponding to the framework (Table 2.1). Two members of our team manually redacted these datasets according to the framework, then met to discuss and resolve conflicting redactions to establish a ground truth baseline.

For tasks involving pre-redacted datasets, we used a basic form of the backend of MARI to redact each of the datasets. To maintain consistency across participants, we created a baseline project save including identical redactions with no modifications. All categories in the framework were used in pre-redaction to properly evaluate how MARI’s redaction modification workflow compares with that of a spreadsheet redacted with the same backend process. We also injected synthetic utterances with deliberate typos and mistranscriptions at a rate of approximately 1% of total words and 3% of total identifiers per dataset, distributing these randomly within an even interval across each dataset. These were added to test users’ ability to detect and redact terms that were not flagged originally through the redaction system but would be flagged by MARI’s similarity detection system.

3.5.5 *Data Analysis*

We quantitatively analyzed Likert-scale responses to SUS statements to determine usability scores for each of the methods, as well as for Likert-scale questions about using graphical interfaces for redaction. As an additional metric, participants' rankings of the methods on different metrics were taken into account. Finally, we conducted analysis for each method on number of terms modified (redacted or unredacted) per number of potential redactions seen, overall percentage of identifiers redacted (according to our baseline), and correctness of these redactions (redacted or not).

Qualitative analysis was conducted in two portions. For the screen recordings, we took audio transcriptions of the think-aloud process and segmented these into sentences related to on-screen actions. We coded action-statement pairs in-vivo with terms considered for redaction, the relevant category and replacement (if applicable), and action taken (redacted or not), as well as with a binary confidence metric for (a) whether or not a term should be redacted, and (b) if the redaction label applied to the term was correct (if applicable). For the interviews, we performed iterative inductive coding; codes were drawn from participant responses and grouped into higher-level, discrete categories following repeated readings of the interview transcripts.

3.5.6 *Preliminary Observations*

Data collection for the MARI usability study is currently ongoing, with the first four participants having completed the full study protocol. These initial observations provide early validation of the experimental design and highlight consistent trends across the three redaction conditions.

Early trends suggest a difference in workflow efficiency between conditions and indicate a strong preference for the graphical workflow. Three of the four participants ranked MARI as their most efficient method, while one ranked it second to the prefilled method due to the

learning curve, stating “if [they] had more time to ... figure out how to use it, then it would be more efficient.” The same three participants also ranked MARI as their preferred method overall, expressing that it was useful once they got used to it.

Results were more mixed on perceived accuracy and completeness, with only half ranking MARI as the most accurate; however, all participants had the most correct terms redacted with MARI than the other two methods, though there were more false positives in the MARI and prefilled methods by virtue of the datasets’ construction. Participants typically accepted categorical labels provided by MARI and the prefilled method when the term was ultimately redacted. This was more common when the replacement was a name, suggesting the tool successfully builds trust for obvious cases while prompting review for nuanced ones. Furthermore, since participants all went through the manual and prefilled methods linearly, they saw the highest number of potential identifiers with MARI, which collects all identifiers per category and displays these first. This may also account for the discrepancy in perceived completeness and actual completeness, since no participants got through the full step-by-step process and all saw only a subset of all categories and select portions of the file preview rather than the full set of selected categories.

Due to the small sample size, quantitative metrics such as mean completeness scores and standard deviations are not yet reported, as they would not be statistically representative. The primary value of this initial report is to confirm that the study protocol and MARI workflow are viable for the target audience. Quantitative comparisons of completeness, correctness, and SUS scores will be aggregated and analyzed once the full sample is collected. These preliminary observations ensure that data collected in subsequent sessions will be comparable and that the study design effectively captures the usability metrics.

3.6 Discussion and Conclusions

The design of MARI and the accompanying usability study represent a response to the gap between the nuanced, context-dependent nature of human identifiability and the rigid, pattern-matching limitations of current automated tools (Chapter 2). While Chapter 2 demonstrates that LLMs significantly outperform legacy tools and perform even better when guided by a robust framework, it also shows that even the best models miss subtle identifiers and overredact, highlighting the need for human oversight. MARI operationalizes the “human-in-the-loop” paradigm, designed to bridge the divide between high-speed automation and the irreplaceable judgment of a data steward. Our workflow transforms the redaction process from a one-and-done, black-box operation into an interactive, transparent workflow.

Crucially, this work disrupts the typical choice researchers must make to either rely on imperfect automated tools that risk privacy leaks, or spend excessive effort and cost on manual redaction. MARI proposes a middle ground where automation handles the bulk of the work, surfacing potential risks for efficient human review. The qualitative data from the think-aloud protocols will be instrumental in understanding how researchers interpret and correct model suggestions. We expect to find that the ability to view terms in their original context and instantly toggle their redaction status in bulk is an important component of user preference, as it restores agency to the user without sacrificing the speed of automation.

We further anticipate that the results of the evaluation study will validate the efficacy of the human-in-the-loop approach. Specifically, we expect that the graphical interface will significantly improve task completeness and efficiency compared to both manual redaction and the backend-only baseline. By organizing flagged terms into categories and allowing for global toggles and per-instance verification, MARI aims to reduce the cognitive load associated with scanning long transcripts. We hypothesize that participants will demonstrate higher accuracy in redacting subtle identifiers (such as specific identity details) with MARI.

However, the deployment of such a system faces practical challenges. The computational

requirements for running local LLMs (like Llama 4 Maverick) may be prohibitive for smaller research groups. While MARI offers a cloud-based alternative via Gemini, this introduces data privacy concerns that must be navigated carefully to comply with IRB protocols. Future iterations of the system will need to balance these trade-offs, potentially through institutional LLM deployments.

Overall, MARI represents a significant step toward making high-quality privacy redaction accessible to the broader scientific community. By grounding the design in data stewards' real workflow needs and the technical capabilities of modern LLMs, we aim to create a tool that is both effective and trustworthy. The pending results of the usability study will provide evidence that a graphical, semi-automated approach can successfully scale redaction of unstructured text, enabling researchers to share their valuable datasets with greater confidence and reduced burden.

CHAPTER 4

PRIVACY IMITATES ART: DRAWING LESSONS IN PRIVACY COMMUNICATION FROM ARTISTS

This chapter presents the findings of a large-scale qualitative analysis of privacy artworks. While the full paper includes a broader literature review and discussion, this dissertation chapter focuses specifically on the methodology for constructing the privacy art database and the results of the qualitative coding process. The following sections detail how these mechanisms were discovered and characterized, providing the theoretical framework for integrating artistic strategies into functional privacy communication.

4.1 Dissertation Author’s Contributions

This chapter represents joint work with Tess Eschebach, Nathaniel Kim, Jolin Liu, Marc Downie, Douglas Pancoast, and Blase Ur. The author of this dissertation is a lead author of the work and wrote substantial sections of the Methods and Results sections in the version published at CHI 2026, similarly reflecting contributions to the data analysis process. The author was a minor contributor to the Related Work and Discussion sections for the version published at CHI 2026, so these sections have been significantly rewritten for this dissertation.

4.2 Motivation

Communicating with users about privacy decisions is a major focus of usable privacy research so users can make informed choices about their personal data [86]. Users are informed about privacy through **privacy communications**: user-facing information about privacy such as privacy policies or cookie banners, as well as alternative methods like privacy dashboards, educational materials, and experimental research prototypes. However, privacy policies are often long and obtuse, and some user interfaces are similarly difficult for users to understand.

While researchers have proposed some alternative communication mechanisms to improve engagement (4.3.1), for the most part, these methods have not been widely adopted. Users most commonly receive privacy communications from companies, either those disclosing to users that their data is being collected [256, 19] or those advertising pro-privacy features [246]. More generally, one might argue that privacy concepts are usually communicated to users in ways that are staid, boring, and corporate.

For decades, artists working largely outside the usable privacy research community have been creating works that grapple with privacy. Art is a “distinctive way to illuminate key technical and social issues” in computing [124], as well as a way to spur action in viewers and cover important and difficult topics. Within this tradition, privacy-oriented art has become an increasingly visible part of the artistic canon, with artists aiming to both reveal and resist privacy intrusions. More broadly, artists continue to use a wide range of media and techniques to probe complex questions in computing [262]. Freed from the constraints of corporate communication, they are able to engage more directly with political and personal dimensions of the topic.

Artistic work around privacy highlights how artists express and interpret privacy concerns, as well as the diverse methods they use to do so. This body of work also offers valuable insights for interaction designers, suggesting techniques that may help them communicate privacy topics to users, even though their design contexts and constraints often differ significantly from those of artists. To this end, we brought together an interdisciplinary team of researchers with backgrounds in usable privacy, media studies, and fine art.

There have been a few small-scale efforts to catalog art about privacy and surveillance, but at the time of our study, no existing collection included more than a few dozen pieces. We assembled a database of 859 artworks that we characterize as *privacy art*, or artworks that relate to the privacy or surveillance of individuals and that have been exhibited in art spaces (e.g., galleries, museums). In total, 332 *privacy artists* created these works. Creating this

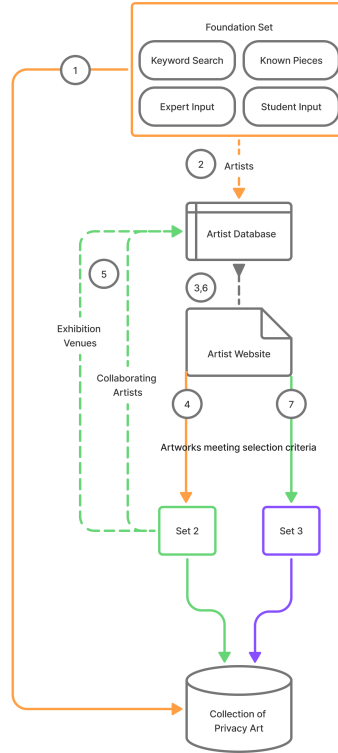


Figure 4.1: A graphical summary of the process our team followed to discover privacy artworks for our database. We first add artworks to our Foundation Set with various search methods (1). Artists are added to a database (2) and we visit their websites to analyze their works (3), adding the first round of works meeting selection criteria to Set 2 (4). We then search exhibition venues and collaborating artists associated with Set 2 for more artists, adding them to the artist database (5) and repeating the process to add to Set 3 (6,7). All works from Foundation Set, Set 2, and Set 3 are collected in our full artwork database.

largest-of-its-kind database required a structured discovery process (Figure 4.1) combining numerous keyword searches, snowball sampling, and feedback from practicing privacy artists.

To ascertain what aspects of privacy these artworks covered, we qualitatively coded digitally archived descriptions (e.g., images, videos, artist statements, media coverage) on a weighted random sample until we reached qualitative stability at 73 works. We found that privacy art commonly highlighted existence and impact of the surveillance of individuals by governments and/or corporations. Some works focused on the bi-directional relationship between surveillance and identity, as well as privacy risks heightened by artificial intelligence. Artworks additionally presented means of resisting surveillance, encompassing specific real

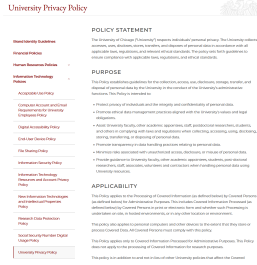
and speculative approaches, as well as general techniques (e.g., camouflage). Finally, privacy art commented on, and provocatively collapsed, distinctions between public and private.

We observed a number of mechanisms that conveyed artworks’ key messages about privacy. First, we observed a visual language that featured familiar forms like surveillance objects (e.g., cameras, microphones) and homes, similar to imagery and metaphors in privacy communication research and design (Figure 4.3). Artworks employed unique mechanisms to frame these forms. Many privacy artworks leaned into a lack of objectivity to critique surveillance practices. Artworks differed widely in tone, ranging from familiar to unsettling (Figure 4.9). Broadly, subjects observed in privacy art included the artist, the viewer, specific individuals, groups in aggregate, and even surveillance devices themselves. Differently, some privacy artworks widely expanded scope by speculating on the future of society.

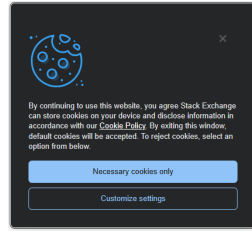
4.3 Related Work

Privacy communication in digital systems faces a fundamental challenge: the gap between abstract data practices and human experience. While surveillance has become perceptibly ubiquitous across public and private domains [180, 18, 298], and its psychological effects—including self-censorship and behavioral modification—are well-documented [180], the mechanisms for communicating these realities to end users remain an open problem. Existing approaches prioritize cognitive clarity over emotional resonance, leaving users informed but disengaged.

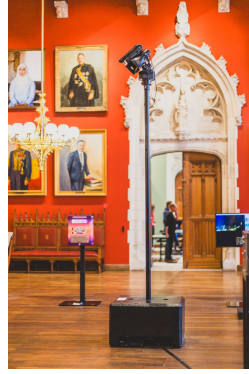
Usable privacy research has developed increasingly sophisticated mechanisms for communicating privacy topics, yet struggles with sustained engagement. In parallel, surveillance and privacy art has evolved strategies for making data collection tangible and emotionally salient. HCI design methodologies, including embodied, affective, and speculative approaches, offer conceptual bridges between these domains. This suggests that effective privacy communication requires more than information: it requires material presence and affective engagement.



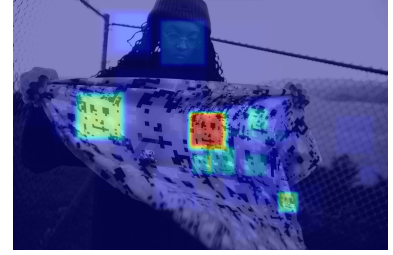
(a) The start of UChicago's online privacy policy [273].
©University of Chicago



(b) Cookie banner notification for Stack Overflow [260].
©Stack Exchange Inc



(c) Dries Depoorter's *Surveillance Speaker* [68].
©Dries Depoorter



(d) Adam Harvey's *Hyperface* quilt, tricking facial recognition systems [122].
©Hyphen-Labs and Adam Harvey



(e) Francine Leclercq's *tapestry Embroidered Surveillance* [161].
©Francine LeClercq



(f) James Bridle's *collage Every CCTV Camera* [40].
©James Bridle



(g) Eva and Franco Mattes's *My Little Big Data* [182]. ©Eva and Franco Mattes

Figure 4.2: An informal comparison of standard privacy communications (Figures 4.2a–4.2b) and privacy art (Figures 4.2c–4.2g).

4.3.1 Engagement Gaps in Usable Privacy

Privacy policies, cookie consent banners, and data dashboards are common in usable privacy systems [185, 184, 82, 245], yet users routinely ignore them, click through without reading, or feel overwhelmed [150, 147, 191]. Efforts to improve comprehension, like privacy “nutrition labels” [147, 229] and icon-based systems [65, 129, 23, 238], increase clarity but introduce new challenges, such as interpretive ambiguity and reliance on supporting text [119, 129]. More dynamic formats (e.g., comics, video, interactive policies) improve attention [268, 164, 255], and immersive approaches like gamification and VR attempt to increase transparency through engagement and play [166, 141, 29], yet these remain largely informative and abstracted from privacy scenarios. By reframing privacy communication

through the lens of art, we propose that the path forward for usable privacy lies not in better icons or clearer text, but in designs that are physically embodied and emotionally resonant.

4.3.2 *Surveillance and Privacy Art*

Parallel to HCI research, artists have been exploring surveillance and privacy as creative subjects for decades. What began as individual works has coalesced into what Brighenti terms “artveillance”—a collection of artwork that addresses topics, concerns, and procedures within surveillance and privacy studies [41]. While this body of work has received attention in art criticism and media coverage [127, 177, 42], it has remained largely peripheral to usable privacy research.

Unlike prior works that focus on the artist’s personal ethics or the use of technology for interactivity [262], ours investigates the specific *design mechanisms* used to communicate privacy concepts. We build on existing scholarship regarding resistance, spatial dynamics, and the concept of “gaze” [285, 199, 41], but expand the analysis to include mechanisms like specific data sources or tonal variation (e.g., unsettling or humorous).

Existing academic literature and media typically analyzes small cohorts [198, 199, 41, 52, 285, 75, 189, 121, 49] or catalogs limited mostly to contemporary art [41]. The most comprehensive prior effort to our knowledge cataloged 43 artists [31]. In contrast, this study compiles 859 artworks across 332 artists and analyzes 73, providing a significantly larger dataset for identifying design mechanisms.

4.3.3 *HCI and the Arts*

Research through Design (RtD) positions design itself as a research method, producing knowledge through the creation of artifacts [297, 220]. While HCI has a strong RtD tradition, privacy-specific applications have focused on individual prototypes rather than synthesizing

broader design patterns [89, 272, 290]. Focus on the “everyday” risks narrowing the scope of inquiry; thus, there is also some push to move beyond the scope of plausibility in design fiction beyond the constraints of utility [35].

But is design art? Some argue that design’s ease of integration offers a unique pathway for critique that art alone may lack [20]. Others posit that critical design is inherently distinct from art, existing in a realm separate from the mundane [74].

The intersection of art and HCI has a long history, calling on humanities- and arts-based methods in computing research [21, 188]. CHI’s Art.CHI gallery [80, 171] and interdisciplinary collaborations [79, 223, 124] demonstrate this trend. Consequently, we posit that the value of interaction design extends beyond the confines of the everyday, embracing artistic modes that exist both within and outside the realm of critical design. Our work seeks to capture the benefits of this broader spectrum, moving past the binary of whether design is or is not art, to instead explore how art can inform design practices in ostensibly utilitarian spaces like organizational privacy communications.

4.3.4 *Interactive Design*

The concept of *materiality* examines how physical properties shape user experience, while *Tangible User Interfaces* (TUIs) use physical objects to manipulate digital information [137, 113]. Rendering digital activity into physical forms—such as printing receipts of screen time—can effectively disrupt addictive cycles and empower users [237]. Beyond screen-based interactions, researchers have employed VR, theater, dance, and haptic feedback to render data collection and privacy topics visceral and immediate [141, 254, 253, 212, 54, 69]. By extracting mechanisms from physically interactive artworks, we aim to enhance privacy communication through immersive, tactile experiences that transcend traditional digital interfaces.

4.4 Methods

We began with a structured discovery process to assemble a database of privacy art. We then qualitatively analyzed a sample of works from this database to characterize communication methods in privacy art using iterative open coding. This let us draw connections between different works and explore privacy artists’ mechanisms, topics, and techniques.

4.4.1 *Creating a Database of Privacy Art*

Our discovery process (Figure 4.1) began with creating a *Foundation Set* sourced from keyword searches, research team members’ prior knowledge (see Section 4.4.4), suggestions from experts (e.g., practicing artists), and dozens of students in an interdisciplinary course on privacy art. We conducted initial keyword searches between Fall 2024 and Spring 2025 using search engines (primarily Google and DuckDuckGo) and arts databases including the MoMA Collections [202], Ars Electronica’s Prix Archive [16], and Smithsonian Libraries Art and Artist Files [257]. Keywords included common surveillance and privacy terms from media, research, and literature. These ranged from simpler searches like “privacy art” to more specific searches like “facial recognition artworks.” Specifically, our searches combined each of four art terms (*art*, *artwork(s)*, *artist(s)*, and *exhibition*) with one or more privacy or topic terms:

- *Privacy terms*: privacy, (mass-/counter-) surveillance
- *Topic terms*: (location) tracking, circumvention, phone, facial recognition, (personal) data, social media, camera

To augment keyword searches, we relied on expert knowledge to ensure our database included key instances of privacy art. For instance, we received input from three practicing artists with backgrounds in media, arts, and architecture, as well as library subject specialist in the arts. These experts particularly provided insight on historic artworks where organized digital documentation may be more limited. Additionally, we crowdsourced 61 submissions

from 24 different students in an interdisciplinary course on privacy art taught by members of our study team. For each artwork, we applied the inclusion/exclusion criteria detailed in Section 4.4.2.

For all artworks, we learned about the piece by visiting the contributing artists' personal websites to review descriptions of the work. We investigated all other works by that artist, again applying our selection criteria. For artists without a personal website, we relied on exhibition pages, media coverage, or academic literature.

To expand our database, we further investigated contributing artists and exhibition venues from the Foundation Set in a process similar to snowball sampling. First, we collected the names of artists with artworks exhibited in shows alongside pieces in the Foundation Set. Works found through this broadened search were placed into *Set 2*. For *Set 3*, we repeated this process with venues and artists from Set 2.

4.4.2 *Artwork Selection Criteria*

Similar to how previous work has examined privacy representations in specific media, we focus on *publicly exhibited artworks in formal settings* like galleries, museums, and arts festivals. These have unique interaction techniques, constraints, and tropes, allowing for more specific and feasible mappings of the spaces. To most readily translate our findings to usable privacy research, we chose to focus on the privacy and surveillance of *individuals*, as opposed to larger groups or organizations. We also focus predominantly on *visual* art for more direct applications to interactive HCI privacy communications. We set formal inclusion and exclusion criteria via iterative team discussions while collecting initial artworks. Potential edge cases were flagged for discussion with our larger team and included or excluded based on team consensus. Our full selection criteria follow:

1. We only include pieces that were at some point displayed in galleries, as public art (e.g., installations in public buildings/parks), or publicly accessible digitally in an art context.
2. We only include works typically displayed as an “artwork.” Works in literature, film, or long-form media format were excluded unless displayed or exhibited as an artwork in a more formal setting. The ephemeral experience of formal settings is more similar to privacy communications in the wild, providing immediate visual or short-form experiences. We similarly considered works exploring privacy in theater [141, 254, 253, 212] out of scope. Works in formal settings have consistent interaction modes, where viewers can enter and exit the space (e.g., gallery) at will, unlike singular, self-contained experiences of theater or film.
 - *Example:* Hito Steyerl’s short film *How Not to Be Seen* [265] was exhibited in art galleries, so it is included.
 - *Edge case:* George Orwell’s novel *1984* [213], though widely known for commentary on surveillance, is a book not primarily exhibited as an artwork, so it is excluded.
3. Since we care about privacy communication with end users, not the inner workings of surveillance entities, we only include works that focus primarily on the privacy of an individual or the general public.
 - *Example:* Kyle McDonald’s *ICESPY* [186], concerning citizen surveillance (i.e., identification) of ICE agents, is in scope since it involves privacy invasions of individuals (specific government workers).
 - *Edge case:* Trevor Paglen’s *Black Sites* [214], centered on exposing CIA secret prisons, is not in scope, since it exposes the practices of an organization without commenting directly on the surveillance of individuals.

4. We included pieces that artists did not themselves present as privacy art if they prompted significant critical discussion relating to privacy (e.g., in media discussing the piece). Art is subject to interpretation past intent.

- *Example:* Arne Svenson’s *The Neighbors* [266] consists of photographs taken through windows, prompting outrage from (unwitting) subjects who felt their privacy had been invaded. While the artist said the piece is about life’s unnoticed moments, most critical discussion has mentioned privacy, so we include it.
- *Edge case:* Jason Bruges’s *Platform 5* [45] could be interpreted as a piece involving data persistence and anonymity, but we did not find such commentary from the artist or media, so we exclude it.

5. Many historical oil paintings depict private or intimate moments. We included those for which ensuing critical/scholarly discussion focused on privacy invasion. The voyeuristic facet of viewers’ intrusion into private moments connects these works to others more overtly focused on privacy invasion.

- *Example:* We include Edgar Degas’s *Woman Bathing in a Shallow Tub*, as critical analysis often highlights the voyeuristic invasion of privacy of looking through a keyhole at a seemingly oblivious subject [30, 43].
- *Edge case:* Mary Cassatt’s *The Bath* [51], while portraying a private moment in the home, is typically discussed as tender and innocent, not invasive [295], so we exclude it.

4.4.3 Data Analysis

Because of the size of our database and time-intensive qualitative coding process, we coded a weighted sample of artworks from our database. Because pieces in our Foundation Set were the first suggested by experts or identified by keyword searches, we wanted to oversample

those; we did so by weighting them $1.5\times$ works in Sets 2–3. To ensure a breadth of styles and perspectives, we ensured that each artist appeared only once in our sample.

We analyzed 73 sampled artworks using iterative open coding [235]. We employed a rigorous codebook development and consensus-coding process. Each piece was coded in MAXQDA by at least two researchers and all disagreements were resolved within the broader team. All artworks, independent of media, were coded as compiled PDFs containing information from the artist’s website (e.g., images, artist statement). If an artist statement was not available, we included text about the work from art criticism or other scholarly literature. If a video or interactive digital element was embedded, coders experienced the element in a web browser and applied codes (e.g., perspective, interactivity) to applicable screenshots [235]. However, most codes were already present in text.

Though there is inherent subjectivity in qualitative and artistic analysis, we applied codes as rigorously as possible to our compilations of images and text about the work. For example, audiovisual component codes (Section 4.5.2) were derived from observable or audible elements. Beyond cataloging formal elements relevant to privacy communication, we also identified affective aspects aligned with interaction design practices (Section 4.3.4). More subjective themes, such as topics (Section 4.5.3) and ambiance (Section 4.6.4), emerged through textual analysis. Where possible, we used in-vivo-style coding [235]. While an artist may not explicitly outline *why* they used particular mechanisms (e.g., color scheme, composition), artist statements can be informative as to which values may be ascribed to a work as a whole.

Initially, two researchers independently coded four artworks selected for their distinct concepts and execution, generating a draft codebook that they then applied to three additional pieces, discussing the state of the codebook with the larger research team. At this point, two additional coders joined the team, and all four coders reached consensus on codes for seven additional artworks. The four researchers then independently coded 15 artworks

from our sample, meeting as a full team after every five artworks to discuss new codes and themes. At that point, we moved to a split-coding approach where these four researchers worked in rotating pairs, independently coding each artwork. Each pair met to discuss and resolve disagreements after every five pieces, while the full research team (including faculty in both the arts and computer science) also met regularly to review the state of the codebook.

We continued this process until no new codes were introduced for five artworks in a row per pair of coders. This occurred after 73 artworks. Our full codebook can be found in Appendix H, and an example of a coded work can be found in Appendix I. Due to the subjectivity of and vastness of creative works, we focus on reaching a stable understanding of components, mechanisms, and their affective aspects, rather than conventional saturation of knowledge. Though artworks with unique mechanisms may exist beyond our sample, our aim is not to enumerate all possible mechanisms; rather, we examine those common among multiple artworks, so we may compare and contrast how these are used.

4.4.4 Study Team Composition and Positionality

Our interdisciplinary team’s expertise spans both the art and computer science domains. All four coders are students at various levels focusing on computer science, but with additional academic experience (e.g., double majors/minors in media arts) or prior experiences publicly exhibiting art. Our broader study team includes faculty members in computer science (specializing in security and privacy), the arts, and media studies. All faculty members have taught academic courses that discuss surveillance art. This combination of artistic understanding and technical privacy knowledge enabled our team to interpret artworks through both lenses, facilitating nuanced analysis that bridges the gap between artistic expression and applications in privacy research.

4.4.5 Limitations

Our work has a number of limitations, many of which are common in qualitative analysis.

Scope Though our database of privacy art is by far the largest to date, it is likely impossible to create one that is comprehensive. While we relied on web searches, suggestions from practicing privacy artists, and multiple other methods, it is near certain that we missed works. For example, our keyword searches included terms related to *surveillance art*, which was characterized in the 2000s and 2010s alongside the rise in surveillance technologies [199, 41]. This may bias our database towards more recent works and those that include these technologies (e.g., surveillance cameras). Some privacy artworks are likely to remain inaccessible—not archived online, lost to time, or lacking documentation. Furthermore, while our research team is multilingual and includes multiple cultural backgrounds, we are all affiliated with universities in a single country. There are likely works that we cannot access based on either language barriers or different cultural interpretations of privacy. Due to our methods, though, it is likely that we covered many of the most notable privacy artworks.

Weighted Sampling Since we intentionally biased our sampling process toward the Foundation Set, we potentially overrepresent well-known artists and more recent works in our selection. While this may limit our analysis of less-well-known artists and older works, we seek to map common techniques in the space in our analysis rather than evaluating every topic or mechanism. Additionally, these works have expectation to be public and to be analyzed.

Formal Arts Institutions A comprehensive study of privacy communications must navigate a vast ecosystem of media [238], ranging from grassroots movements and memes to digital creations where creators rarely anticipate academic scrutiny [85]. Analyzing these independent, often non-consensual contexts requires deep community collaboration [53, 201] and presents unique methodological challenges. Our large-scale analysis concentrates on the

visual elements of formally exhibited artworks (4.4.2). This choice is driven by the practical advantages of formal venues: they offer the consistent documentation and accessibility necessary for broad study, alongside an inherent expectation of public viewing and critical debate. However, this pragmatic selection comes with a known trade-off. By relying on institutional archives, we inevitably inherit the Western-centric and colonial biases embedded in those spaces [120, 22], which systematically exclude the very independent practices—such as protest art, graffiti, and street-level creativity—that warrant their own dedicated future exploration. By establishing a robust foundation through formal archives, this study creates a necessary baseline from which future research can expand to illuminate the rich, diverse landscapes of independent and marginalized artistic expression.

Subjectivity Art is subjective, creating interpretive challenges. Our team viewed works through a technical privacy lens. While we used consensus coding on available statements by artists themselves and/or art critics, we cannot capture all possible interpretations. Similarly, selection criteria boundaries are inherently debatable. Our team debated selection criteria over months and spent extra time deliberating over edge cases. Though our criteria are certainly not the only basis for evaluating privacy artworks, we chose them based on applicability to usable privacy research.

Perspective We characterize our interpretations and those suggested by the artists (or critics), not viewers of the artwork, differing from actual experiences of art. It is unreasonable to assume that an artwork will have the same effect on every viewer. Our team viewed most artworks outside their original context (i.e., online or in photographs, rather than in a gallery or in-person exhibition). We also relied heavily on artist statements, which are often a secondary or tertiary artifact in a gallery setting. We attempted to standardize formats and consulted direct sources (e.g., quotes or artist statements), if possible. Nonetheless, our approach may have missed notable nuances of these artworks.

4.5 Results: Components in Privacy Art

In this section, we provide an overarching look at what privacy artists are saying in their art. Section 4.5.1 describes our database and sample set composition and Section 4.5.2 details audiovisual components (e.g., symbols, objects). After discussing topics present in the works (Section 4.5.3), we discuss how artworks emphasize different subjects of privacy invasion and data collection (Section 4.5.4).

4.5.1 Database Characteristics

As of this writing, our database includes 859 artworks by 332 unique artists, 78 of whom collaborated with others. Our database spans over 350 years of privacy art: the earliest work is *Las Meninas*, painted by Diego Velázquez in 1656; the latest pieces are either ongoing or are currently exhibited at an artistic venue.

Sample Set Composition. Our sample set includes 73 works (see Appendix G). While many works are collaborative, all artists appear only once in the sample set. Artists came from national backgrounds spanning 32 countries; most artists are Europe-based (49%), followed by the Americas (37%, 31% US-based), and Asia (11%).

The earliest work our team coded was painted in 1657-1659, and the most recent works were pieces ongoing or exhibited in 2024. The most common decade (N=36) in our sample are pieces from 2010-2019. The least represented (N=5) era is pre-2000s. Half those from 2010-2019 (N=18) deal with data access and agency, mirroring increases in consumer technology use and rise of social media. Beyond this connection, works from different time frames use diverse methods and explore myriad topics.

Privacy artworks in our sample set employ a wide spectrum of media, some even employing multiple media (see Appendix G). There are 27 digital pieces, including 23 software-controlled systems (i.e., using code to drive or manipulate the installation), a virtual reality



Figure 4.3: Examples of the visual and auditory components of privacy art.

experience (A25), and a video game (A30). Ten pieces use electronic hardware, such as computers. We also saw 11 video-based artworks. Six others included animated images (e.g., GIFs). Still images are in 19 pieces, including eight photographs, and two (A41, A55) paintings. Seven pieces take on tangible forms through textiles and wearables. Another seven pieces are three-dimensional sculptures using raw materials or manufactured products. Beyond discrete media, nine pieces use immersive exhibits that collapse the distance between artwork and audience. Seven live performances further engage real-time dynamics through human or machine agents. Finally, eight architectural installations use preexisting or constructed spaces (e.g., buildings or rooms) to explore privacy in physical environments.

4.5.2 Sensory Vocabulary of Privacy Art

Artworks following our criteria are constructed from audiovisual components, which set tone, create a visual shorthand, or indicate symbolism. *Visual components* are items or symbols appearing visually in an artwork.

Familiarity as an Anchor in Reality. Thirty-three artworks turn to familiar settings (Section 4.5.3), human forms, and objects to develop comfort or conduct satirical critique.

Thirteen pieces use imagery from familiar settings—for example, offices (A6, A40, A59) or airports (A50)—showing how innocuous locations may be part of the surveillance apparatus (see also Section 4.5.3).

Fifteen pieces include human forms both real (N=11) or digital (N=5), inviting the viewer to relate to the subject or face the uncanny. Commonplace objects also play a role in centering privacy experiences in daily life. Five pieces feature smartphones; three (A11, A13, A30) display text messages on a phone in the piece and two (A6, A44) provide interaction on a viewer's device. Two pieces comment on the security from (A32) or embedded into (A20) clothing. For example, Doring's *The Hoodie* (A32) uses AI-generated images of a hoodie as commentary on how it might be the only counter-surveillance tactic an individual has at their disposal (see Figure 4.3b).

Transforming the Abstract Digital into the Tactile Physical. Some artists translate digital information into familiar tangible forms using materials that emulate touch, comfort, and daily life. Five works use paper to ground digitized information in a more traditional format, including printed screenshots of text messages (A6) and NSA documents (A55). Two pieces (A33, A53) transform fleeting, intangible data into a permanent print form. For example, a simple robotic arm in MSCHF's *Eavesdropper* (A53) writes out overheard words of viewers in Sharpie (see Figure 4.3e), turning casual speech into an enduring artifact. Three pieces (A7, A59, A70) embed digital information into textiles, such as cross-stitch depicting surveillance footage (A70), a word cloud of common passwords sewn into a quilt (A7), and carpet tiles with color patterns correlated with the frequency of email exchanges (A59). Through these tactile materials, digital surveillance becomes something viewers can literally feel—a reminder of how deeply it threads through daily life.

Cameras and Microphones. Twenty pieces include visuals of surveillance devices, drawing attention to objects viewers may typically ignore due to familiarity. Cameras were visible

in 17 pieces, types including eight bullet, five dome, two (A39, A49) web, two (A60, A67) video, and a smartphone (A13). While some cameras are functional (see Section 4.5.4), cameras in eight pieces are nonfunctional or are even presented in a novel medium—in photos (A9), VR (A25), or marble sculpture (A12) (see Figure 4.3c). Though nonfunctional cameras do not capture data, they imply observation and act as a proxy for surveillance. Some pieces contain visible audio equipment to record (A22, A53) or play audio (N=5), including microphones (A22, A53), speakers (N=5), and megaphones (A3, A47). Together, these visual cues of monitoring equipment evoke an atmosphere where observation feels ubiquitous.

Linguistic and auditory content. Many artists turn to language and sound to shape how viewers engage with surveillance, with 35 works incorporating written or auditory communication. Plaintext is an important component in 26 pieces, filling a role beyond supplementary statements or descriptive text for the piece. For example, ten works use labels for data visualizations, six provide additional context, and two (A33, A42) give instructions (e.g., prompts in an app (A42)). In 13 pieces, plaintext even features as a central subject: social media posts (A39, A45, A59), text messages (A6, A11), transcripts (A16, A53), and browser history (A33, A59). Fifteen pieces include music (N=9) or narration (N=7). Music (N=9) may help set a mood (see Section 4.6.4). Robotic voices lend technology a voice in four pieces for communication with viewers (A3, A5, A43) or for narration (A1). In contrast, three pieces (A8, A62, A73) use human voices to highlight humanity behind technology. In Coupe’s *Sanctum* (A73), a human voice narrates the text of an online post, mapping the individual being viewed to a specific identity and age.

4.5.3 *Privacy Concepts Explored in Art*

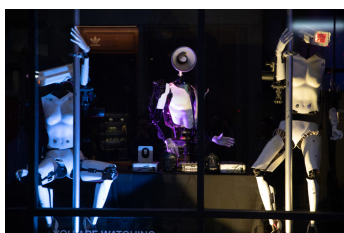
We determined topics in privacy artworks by consulting artists’ statements and documentation concerning the works. The most common topic (N=33) in our sampled dataset was

organized surveillance of individuals, commonly in conjunction with resisting surveillance (N=23). Other common topics include identity in context of privacy invasions (N=27), relationship between public and private space (N=15), and exposure of personal data (N=9).

Organized Surveillance and Data Collection. Thirty-three pieces comment on organized surveillance by entities like governments or corporations, often portraying these forces as intrusive, inequitable, or quietly omnipresent. Twenty-nine comment on pervasive government and large-scale surveillance of citizens and the resulting power disparity. For example, Luo’s *If This Is a Global Surveillance Center* (A23) accesses video feeds of public surveillance cameras, equating control of data to “dominance over physical space” [170]. Other works address government surveillance even more directly, such as the US National Security Agency (NSA) (e.g., A55). Ten artworks focus on corporate-driven privacy invasions, with three (A33, A71, A73) exploring data collection within specific companies like Facebook, illustrating how private information can be mined within everyday platforms on which people routinely rely.

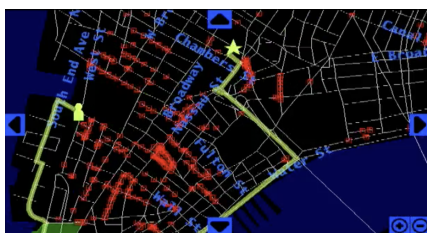
Resisting surveillance. Twenty-three pieces additionally focus on resisting surveillance, with 15 exposing surveillance and four offering direct interventions. Four pieces (A1, A17, A24, A32) also explore this concept abstractly, such as repurposing surveillance equipment to monitor air pollution instead of people (A24).

Fifteen pieces expose surveillance practices, which can point out omnipresent observation and encourage reflection on surveillance apathy. By revealing practices that often fade into the background of daily life, artists encourage viewers to confront the extent to which constant observation has become normalized. Six glamorize surveillance, drawing attention to the practice by ornamentation or idolization (see Section 4.6.4, Figure 4.4a). Five pieces expose existing CCTV cameras, showcasing the prevalence of surveillance in public settings, with three (A35, A42, A46) exposing specific locations. In *The Institute for Applied Auton-*



(a) Giles Walker's *Peepshow* installation **exposes surveillance with glamorous pole dancing CCTV cameras** [203, 287, 284].

©Wieden+Kennedy



(b) The IAA's interactive map, *iSee*, **exposes surveillance camera locations**, indicated by red squares [283, 218].

©Rich Pell



(c) Hasan Elahi's *Citizen obfuscates* surveillance by **collaging** mundane photographs (e.g., meals, toilets) [77, 76].

©Hasan Elahi

Figure 4.4: Examples of resisting surveillance in privacy art (see Section 4.5.3).

omy's *iSee* (A35), CCTV cameras in New York City are collected on a map, so viewers can identify and avoid their specific locations (see Figure 4.4b). Two pieces (A9, A23) explore the prevalence of cameras ambiguous from location, such as repeated images of different CCTV cameras in London (A9) (see Figure 4.2f). Five pieces expose government surveillance tactics (A31, A36, A55) and the web of individuals behind these systems of control (A36, A37, A58). Lavigne's *3 Degrees of Separation* (A37) uses a network graph to expose individuals with surveillance skills publicly posted on LinkedIn. One piece (A66) exposes the environmental cost of cookie tracking—an unseen expense of surveillance.

Beyond exposure, several artists imagine direct countermeasures. Four pieces (A2, A31, A44, A71) show direct interventions that could be used to hide from surveillance. Two (A31, A71) use obfuscation. In Elahi's *Citizen* (A31), part of *Tracking Transience* (2003-Ongoing), Elahi floods the FBI with mundane photos of his life (e.g., meals, toilets) to combat profiling with radical transparency (see Figure 4.4c) [144, 78]. Differently, Grosser's downloadable tool *GoRando* (A71) obfuscates users' reactions on Facebook by randomization. Two pieces serve as proof of concept, showing inventive possibilities of directly combating surveillance: Burtch's *Mic Jammer* (A44) mutes phone microphones, and Harvey's *HyperFace* (A2) cloak tricks a specific facial recognition system (see Figure 4.2d).

Privacy-Invasive Technology, Algorithms, and Identity. Twenty-seven works cover the interplay between personal identity and privacy invasions. Eleven pieces comment on how algorithmic systems construct, distort, or conceal assumptions about who a person is, revealing the gap between lived identity and computational shortcuts used to define it. For example, Onuoha’s *Classification.01* (A21) hides an algorithm and its calculations from viewers, flagging viewers as “similar” through hidden sensors, pointing to opaque judgments embedded in machine classification. Three pieces (A1, A31, A36) highlight profiling and the use of assumptions about identity as an excuse to surveil marginalized groups, such as Boundaoui’s documentary *The Feeling of Being Watched* (A36) exposing FBI tactics used to surveil her Arab American community.

Relationship Between Public and Private Space. Similar to architectural literature [125, 138], several works probe the shifting boundary between public and private spaces. Eight use public infrastructure to spotlight or draw attention to surveillance of public spaces. Three of these (A45, A46, A73) seamlessly integrate with existing public infrastructure, such as glamorously decorating public surveillance cameras (A46) or projecting videos onto a building (A73). Four pieces simulate more personal locations like homes (A41) or bedrooms (A8, A29, A69) (see Figure 4.3a) to emphasize how normally safe, private spaces may still be invaded. Although many works focus on physical spaces, four works (A39, A45, A52, A59) explore how increased data collection blurs the line between public and private in the digital space. Lowe’s *You Saw Me?* (A45) projects Facebook status updates on the side of a building in a busy public area, highlighting how seemingly private or personal online activity may be seen by unexpected viewers.

4.5.4 *Subjects of Privacy Invasion*

Artists play with different collection sources and subjects of privacy invasions (the public, the artist, and the viewer). Pieces collect data through a variety of channels, including cameras (N=17), public datasets (N=10), tech-mediated communication¹ (N=10), microphones (A8, A16, A53), and a viewer’s personal device (A6).

Invading Privacy of the Public (Aggregated and Individual). Seventeen pieces focus on invading the privacy of the aggregated public, showing the limitations of privacy in any number of private spaces (both physically and digitally). Twelve pieces draw on public datasets, including publicly available city datasets (A24, A34, A54, A66) (e.g., traffic), location data (A24, A35, A54), social media posts (A39, A73), and leaked passwords (A7). Similarly, four (A23, A38, A68, A70) works aggregate and obscure (such as through generative AI (A38)) camera footage of the unknowing public. Giasson’s *VOX* (A16) features secretly-overheard conversations transcribed onto transparent panels; this aggregated data gives little insight into the subjects’ identities. These works illuminate how anonymity in masses can be less secure than it appears, revealing structural vulnerabilities that affect populations rather than isolated individuals.

While 17 pieces focus on the public in aggregated format, six focus on public individuals. Three (A15, A39, A67) highlight specific individuals, which may show how publicly accessible information can be used against any person. For example, Riekstins’s *Back to the Light* (A15) restores deleted files from secondhand hard drives, showing that data persists on physical devices. Two works (A37, A58) focus on individuals’ part of the surveillance apparatus (see Section 4.5.3). Hovers’s *The Right to be Forgotten* (A18) uniquely focuses on a very specific subject, the first man to successfully claim his “Right to be Forgotten” in Europe. Hovers creates a series of multimedia images based on a digitally accessible image of the

1. Tech-mediated communication may include SMS (A6, A11, A30), browsing history (A33, A59), or social media (A39, A71, A73).

man, illustrating that one can have a digital footprint despite attempting to erase it. Using specific individuals' data demonstrates that breaches of privacy are not abstract phenomena, but direct incursions into real lives and identities.

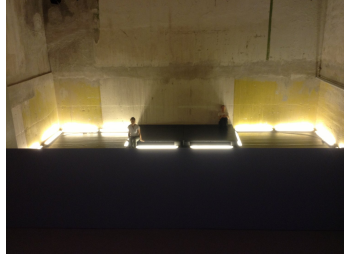
Invading the Personal Privacy of Artists and Performers. Artists may invade the privacy of themselves (N=11), performers (A4, A10, A27), or study participants (A44) to act as a viewer proxy, showcasing intimate privacy invasions.

In 11 pieces, artists invade their own privacy in a more intrusive way than they could that of a viewer. Six pieces display artists' personal data. Two (A59, A64) contrast dense and specific data displayed in different formats, showing how small segments of data can be more personal. Mattes and Mattes's *My Little Big Data* (A59) shares over 13 years of the artists' digital information in a video essay and installation with data visualizations and schematics of their apartments (see Figure 4.2g). In contrast, Goni's *Deletion Process* (A33) presents the artist's search queries one at a time, which may showcase the information derived from even a small piece of data. Three pieces (A4, A10, A44) transform a viewer's perception of "normal" observational contexts (e.g., performances) into surveillance with computer vision (A10) and other means.

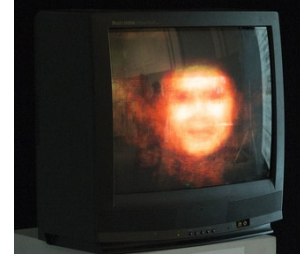
Invading the Privacy of the Viewer. Sixteen pieces invade the viewer's privacy *non-consensually*. By viewing or experiencing the artwork, viewers inherently become part of the piece. This can potentially draw a stronger, more personal reaction, as the act of experiencing the piece becomes inseparable from the experience of being observed. Viewers' data is typically collected through cameras (N=13); two works employ (A8, A53) microphones (see Figure 4.3e). Six pieces process video data before it is displayed to the viewer, three (A21, A57, A67) use computer vision visuals (see Section 4.6.2), and one converts video data into a different medium (A3). For example, in Depoorter's *Surveillance Speaker* (A3), a camera captures gallery visitors' actions, which are run through a computer vision algorithm and



(a) Viktoria Binschok's *Suspicious Minds* uses **collage** to show multiple cropped photographs [32]. ©Viktoria Binschok



(b) Anna Ådahl's *The Exhibited* employs **lighting** to show more private and less private spaces [299] ©Anna Ådahl



(c) Peters Riekstins's *Back to the Light* layers different faces into a **composite**, unidentifiable face [231, 175]. ©Kristine Madjare

Figure 4.5: Different mechanisms for directing gaze including collage, lighting, and composite (see Section 4.6.1).

narrated out loud with a robotic voice (see Fig 4.2c). Two (A49, A73) pieces simply present the viewer with a video of themselves without modification, emphasizing the immediate feeling of being observed. Using the viewer as a data source forces them to confront how easily their personal behaviors, movements, and voices can be captured, analyzed, and displayed, transforming passive observation into a deeply personal experience.

4.6 Results: Mechanisms to Present Components

In this section, we provide an overview of mechanisms used to showcase the topics presented in Section 4.5. We examine how artworks influence the viewer's gaze (Section 4.6.1) and control the viewer's perspective (Section 4.6.2), concluding with viewer interaction (Section 4.6.3). Finally, we examine various tonal portrayals of privacy violations in Section 4.6.4.

4.6.1 Directing Viewer Attention

Artworks can direct attention toward important elements through techniques of emphasis but also reveal elements through techniques of obscurement. Common techniques include lighting, collages, or even obscuring information.

Alerting Viewers at the Moment of Privacy Invasion. Three pieces (A39, A42, A56) alert viewers to an invasion of privacy the moment it happens. By providing real-time notifications, these works may show the frequency of surveillance and allow viewers the opportunity to recognize and potentially avoid future intrusions. Blinder’s *Dark Side of the Prism* (A56), a downloadable tool, plays a snippet of Pink Floyd’s *Dark Side of the Moon*² each time the NSA’s PRISM³ invades a user’s privacy. Wunderling’s *Audience* (A39) displays a public image from Flickr and takes a photo of the viewer, then immediately notifies the owner of the public image that it has been viewed by commenting a blurred version of the viewer’s photo on the image. The viewer in turn sees the publicly posted comment. In-the-moment alerts expose the immediacy of privacy invasions while prompting viewers to consider their own vulnerability.

Exposing Those Surveilled and Surveillance with Lighting. Eight pieces use lighting to indicate surveillance and direct viewer attention. Lighting can be an analogy for visibility, paralleling the implied spotlight of being under surveillance. Five pieces use lights to direct viewer attention to a specific area of a piece, where two (A4, A25) highlight surveilled subjects. In Huang’s *The Eye and I* (A25), a central lighthouse emits a rotating spotlight, illuminating cells within a VR panopticon. Two pieces use light to indicate that surveillance devices are active (A21, A57), such as a subtle white bounding box framing the viewer’s face in a mirror (A57). Instead of using light to explicitly direct a viewer’s attention, Ådahl’s *The Exhibited* (A4) uses contrasting bright and dark spaces where performers can move between to “escape” the spotlight, though they remain visible in every location (see Figure 4.5b).

2. The iconic album cover of Pink Floyd’s *Dark Side of the Moon* (1973) prominently features a triangular prism on a black background.

3. PRISM is an online surveillance program run by the US NSA that collects user data from websites and services like YouTube, Facebook, and Yahoo, among others.

Showing Connections and Scope with Collages of Information. Twenty-seven pieces use a collage effect for display. In these pieces, viewer attention is directed to *multiple* visual components (e.g., images, videos), which may show the vastness of surveillance and data collection. Six pieces showcase still images in grids (see Figure 4.5a), ranging from photos of surveillance cameras (A9) to objects (e.g., food, toilets) from an artist’s personal life (A31), both exposing and obfuscating surveillance. Five pieces display video surveillance footage in grids, showcasing the vast amount of data collected by CCTVs. Two pieces use grids to display non-photographic content such as identification (A48) and leaked government documents (A55). One (A7) uses overlapping collage to display a colorful word cloud of common passwords.

Obscuring Information with Composite. Six pieces used composites i.e., combinations of multiple overlaid images that create a new image. Using composites could illustrate how a specific individual can be lost when part of an aggregate. Over 1,000 individuals’ faces are obscured in Riekstins’s *Back to the Light* (A15) as they are overlaid on top of one another, making a single hazy and unidentifiable face (see Figure 4.5c). Similarly, Tarar’s *Landsat.Earth* (A26) showcases low-resolution images of earth for monitoring of the atmosphere stitched together. Though this technology is part of the surveillance apparatus, in this form it is unable to be used for the direct surveillance of individuals.

4.6.2 Viewer Perspective

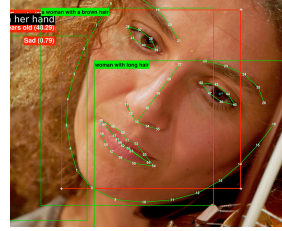
The relationship between observer and subject is a common theme in privacy art [285, 199, 41]. Attention drawn to surveillance may be fetishistic in nature (A46, A47). In Vermeer’s *A Girl Reading a Letter by an Open Window*, the viewer peers into a young woman’s private moment (featuring symbolic elements such as window blinds thrown open, suggesting an affair) [274] (see Figure 4.6a). In other cases, the viewer is the subject of a voyeuristic



(a) Johannes Vermeer's *A Girl Reading a Letter* [cropped] shows a **voyeuristic** perspective [282, 274, 62].
©Hans Peter Klut/Elke Estel⁴



(b) James Coupe's *Sanc-tum* shows the viewers looking up at the **camera POV** from the devices that surveil them [60].
©James Coupe



(c) Trevor Paglen's video *Image Operations* visualizes **computer vision** through bounding boxes with correct and incorrect labels [215]. ©Trevor Paglen



(d) Björn Schülke's *Drone #4 looms* above viewers in the gallery, who must look up to see the hanging sculpture [241]. ©Björn Schülke

Figure 4.6: Examples of different viewer perspectives in privacy art (see Section 4.6.2).

gaze—watched by a giant eye (A25) or being in a space with large windows leaving nowhere to hide (A69). In two (A4, A8) pieces, viewers are both the observer and the observed. In Ådahl's *The Exhibited* (A4), the performers watch back, inspired by historical “Human Zoo[s]” [299]. In Ghani's *Security Blanket* (A8), initial viewers whisper secrets into a wall; future viewers listen.

Camera POV. Thirteen pieces used a camera point of view (POV), immersing viewers in the perspective of surveillance (see Figures 4.6b– 4.6c). Ten pieces specifically focus on surveillance systems which look down on subjects from a bird's-eye-view, allowing viewers to explore what surveillance cameras “see” and feel the inescapability of being in view (see Figure 4.6b). This POV is commonly (N=11) displayed as videos on monitors such as flat screen televisions (A39, A50, A73), cathode ray-style televisions (A14, A15 A60), CCTVs (A29, A67), or computer monitors (A49). Uniquely, Leclercq's *Embroidered Surveillance* (A70) displays “1/24 second” stills of surveillance footage in cross-stitch, creating a frozen, permanent contrast to the fleeting nature of a moment. Six pieces make use of livestream video, with three (A14, A49, A73) directly surveilling the viewer (see Section 4.5.4) and one (A67) surveilling members of the public nearby. Viewers hold the objectifying gaze of the surveillant and may question their role in systems of observation.

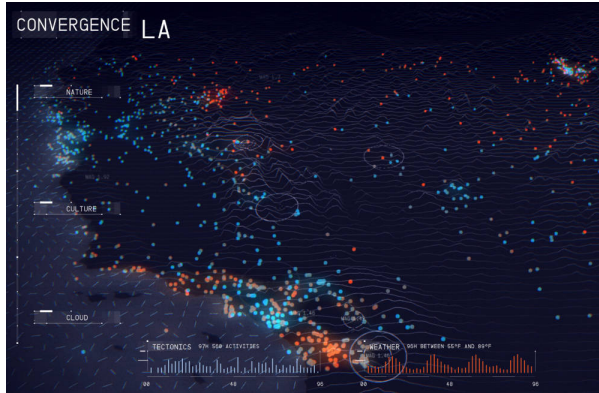
Visualizing AI. Seven pieces use a camera POV in conjunction with bounding boxes and image distortion to display how cameras “see” with AI and computer vision. Five works use bounding boxes to outline faces (A2, A10, A57, A61) or bodies (A67) and two (A2, A10) comment on possible facial recognition errors (see Figure 4.6b), drawing attention to the assumptions embedded in algorithms. Two pieces (A32, A38) feature morphed bodies and environments as a result of small datasets being fed into a GAN (generative adversarial network) (see Figure 4.3b), creating an eerie stitching of images. These visualizations reveal the interpretive power and the potential misjudgments inherent in AI-driven surveillance.

Surveillance Infrastructure Looms Above Viewers. Seven works situate surveillance tools looming over viewers, such that they must physically look up to experience the work (N=6). A looming perspective can imply observation and allude to tools of surveillance being beyond reach or control. Three pieces (A3, A5, A22) are displayed above a viewer in a gallery setting, two of which hang from the ceiling (A5, A22). Schülke’s *Drone #4* (A22) is a monochrome conglomerate of surveillance equipment hung by a string above viewers, with parts swooping downwards (see Figure 4.6d). Two pieces (A45, A54) are projected onto skyscrapers, potentially reinforcing the idea that surveillance towers over everyday life. Differently, Bridle’s *Every CCTV Camera* (A9) features photos of surveillance cameras shot from a worm’s-eye-view (see Figure 4.2f).

4.6.3 Viewer Agency and Interactivity

In 16 pieces, viewers are active agents rather than passive subjects, which may create a more visceral experience of surveilling and being surveilled. Viewers can interact without being surveilled (N=10), as a countersurveillance intervention (A35, A42, A56, A71) (see also Section 4.5.3), or under surveillance (A6, A8, A14, A60).

4. Gemäldegalerie Alte Meister, Staatliche Kunstsammlungen Dresden



(a) Zoomed-out view showing a map of LA and weather data [10].



(b) Zoomed in with coordinates and digital data transport [10].

Figure 4.7: Refik Anadol’s *Convergence LA* (web-version) viewers **explore** a map of LA by zooming from natural (Figure 4.7a) to digital (Figure 4.7b) data [10]. ©Refik Anadol

Viewer Exploration in Absence of Being Surveilled. Eight works let viewers directly explore the piece with the agency of the surveillant, making decisions on interpreting and scaling the presented data. This shifts the viewer from passive subject to active participant of the surveillance apparatus, highlighting the mechanics of surveillance and the potential for control or resistance. Five include web-based interactive data visualizations. Anadol’s web version of *Convergence LA* (A54) allows viewers to directly interact with scale by zooming in and out from a typographic map of Los Angeles into a “molecular” level of data transfer (see Figure 4.7). Differently, other works offer tools designed to counteract surveillance rather than replicate it: two pieces (A56, A71) are downloadable tools to combat surveillance online through alerts and obfuscation. In Vasiliev’s *Netless2* (A17), viewers may engage with the work (which details a new data transfer system) by sending messages through a network integrated into a model train system (see Figure 4.8).

Compulsory Participation. Viewers who are observed (N=9) (see Section 4.5.4) have some agency to move out of view of the camera or remain in silence to avoid recording, but they have still been surveilled. Works that force viewer participation highlight the tension between agency and surveillance, showing that seemingly innocuous participation



Figure 4.8: Danja Vasiliev’s *Netless2* allows viewers to **interactively** send secure messages through a model train [280]. ©Danja Vasiliev

can serve the surveillance apparatus. Two pieces directly ask for viewer participation—like telling a secret (A8)—or texting a viewer’s device with an action—like “I left you something on the printer” (A6). In these cases, information is taken from a viewer (e.g., secret, cell information). Sometimes, the interaction control is more subtle, e.g., viewer distance (A14, A39, A60), where a video of a dog barks louder as the viewer approaches (A60).

Info Withheld About Surveillance. Six pieces have information deliberately censored, which could make a statement on the limited information available about the everyday surveillance apparatus. Viewers are directed to consider what is *absent as much as what is present, prompting reflection on the unseen forces that shape observation*. Two pieces (A21, A57) use hidden cameras and indirectly show the viewer the outcome of the observation. Information may also be withheld by censoring or caricaturing dataset contents. Treister’s *Camouflage* (A55) obscures segments of leaked NSA documents concerning surveillance and data mining with geometric paintings mimicking shapes in the documents, warping their original meanings. These works reveal the gaps of information between the subject and



(a) Ann-Sofi Sidén's *Sticky Floors* frames a **mundane** tone by showing daily surveillance footage at a pub [251].
©Ann-Sofi Sidén

(b) Egor Kraft's *Twelve Nodes* presents proposed Fair Data rules on marble blocks in a serious, **formal** tone [154].
©Egor Kraft

(c) Hsin-Chien Huang's *The Eye and I* creates **unsettling** VR panopticon with the eye and surrounding cells [134].
©Hsin-Chien Huang

(d) Fabiola Larios's *Surveillance Cutie* adds **fun** to surveillance with pink bedazzled vintage TVs [159].
©Fabiola Larios

Figure 4.9: Examples of pieces that feature different ambiances such as mundane, formal, unsettling, and fun (see Section 4.6.4).

surveillant inherent in surveillance. What is purposefully hidden can be as revealing about power and control as the data that is openly displayed.

4.6.4 Ambiance

What is shown, and *how* it is shown, contribute to the atmosphere of a piece and the viewer's emotional response to it. In our sample set, we found four common tones to privacy invasions: *mundane*, *overly formal*, *unsettling*, or *fun*.

Mundanity of Surveillance. In seven pieces, surveillance blends into ordinary settings such that it becomes seemingly uneventful, normalized, or overlooked. Ghani's *Security Blanket* (A8) recreates the setting of a cozy childhood bedroom in soft, neutral colors where visitors whisper secrets into a hidden wall, recorded without their knowledge. Its success collecting intimate, private answers from viewers demonstrates how the nostalgic comfort and security of childhood may lower people's defenses against surveillance and privacy invasion. Three pieces (A68, A69, A70) comment on the existing surveillance infrastructure. Three (A4, A31, 68) use mundane contents as the data subject. For example, Sidén's *Sticky Floors* (A68) documents the full daily cycle of a pub through its own security cameras, making constant observation feel like part of the venue's natural rhythm (Figure 4.9a).

Ostentatious Formality. Ten artworks are framed as serious or formal to emulate and critique formal structures of power. Three pieces (A11, A18, A55) feature framed images of text messages (A11), images (A18), and government documents (A11, A18) to emphasize their importance and subvert expectations about what belongs in a frame. Two pieces (A12, A51) use marble to mimic the formality of government regulation. For example, Kraft’s *Twelve Nodes* (A51) uses marble as a backdrop to display imagined rules for protecting privacy; the choice draws on marble’s legacy in the first written legal code of ancient Rome, imbuing the proposed rules with stately significance [154] (see Figure 4.9b). Ai’s *Surveillance Camera* (A12) monumentalizes surveillance: a marble CCTV camera as an anti-monument, the material drawing on cultural traditions of commemoration and past political symbols to emphasize the longevity of government surveillance [4] (see Figure 4.3c).

Unsettling Invasions of Privacy. Eighteen pieces induce viewer anxiety about privacy invasions. Six pieces convey the constant, inescapable sense of surveillance, similar to the “Panopticon” [27] (see Figure 4.9c). Four pieces (A25, A38, A43, A72) describe eerie and dystopian futures of technological control and surveillance. Other works may induce disorientation (A5, A14, A38, A49). Nauman’s *Live-Taped Video Corridor* (A14) confronts visitors with an unsettling split image: the viewer’s live reflection shrinking away in one monitor and complete absence in the other.

Thirteen pieces feature exposed wires or raw components, and seven pieces feature industrial styling (e.g., concrete, exposed pipes). These visual styles may emphasize intrusions of technology into private life. For example, Scher’s *Papa Bed* (A29) has a mass of wiring and technology sprawling around a plain bedframe. Eleven pieces distort familiar objects into uncanny new forms. Six pieces include modified human visuals such as digital avatars (A1, A25), AI-distorted figures (A32, A38), or blocky facial features to trick AI (A2). Digital representations draw on both familiarity and a sense of detachment, creating a sense of tension for the viewer.

Making Privacy Violations Fun. Sixteen artworks created fun or playful atmospheres for the viewer, potentially for engagement or as juxtaposition to the underlying privacy violation messaging. For example, in Larios’s *Surveillance Cutie* (A49), pink reflective foil and warm colors create a bubbly, cutesy tone, in contrast to the stark message of control under surveillance (see Figure 4.9d). Five pieces also use familiar music. Some music acts as jokes: for example, Stevie Wonder’s *I Just Called to Say I Love You* plays during a mysterious phone call to viewers’ phones (A6). Six pieces make surveillance glamorous. This can sometimes be literal, with four pieces (A34, A46, A47, A49) decorating or idealizing surveillance infrastructure. Walker both glamorizes and fetishizes surveillance in *Peepshow* (A47), where anthropomorphic robots with security cameras for heads perform a pole dancing routine in a club setting (see Figure 4.4a). The viewer is enticed to stare at the display; they then confront the visual of the surveillance camera as it stares back.

Ten works are framed as humorous or satirical. In five pieces, the surveillance apparatus itself is parodied. For example, Bunting’s *The Status Project* (A72) does a satirical take of the class system to comment on surveillance and the power of corporations (classified as artificial humans). Five pieces use exaggerated, over-the-top humor. While this technique may underscore absurdities of surveillance systems (A6, A27, A42), some pieces highlighted fatuous measures to protect personal security in a highly surveilled world (A1, A20). For example, Mann’s *EXISTech* (A20) promotes merchandise such as bags and bras retrofitted with large security cameras designed to deter crime. Two pieces create parodies of existing media (A1, A28). For example, Steyerl’s *How Not to Be Seen* (A1) is a satirical documentary based on a Monty Python skit⁵ describing silly ways to avoid detection, such as “being female and over fifty” [265].

5. Monty Python is a comedy troupe formed in Britain in the 1960s that was known for their absurdist television sketches.

4.7 Discussion

In this chapter, we have presented a robust analysis of 73 privacy artworks, revealing a vast and diverse ecosystem of communication strategies that stand in stark contrast to the standardized, often ineffective interfaces of usable privacy research. By moving beyond the narrow scope of individual case studies to examine a large-scale corpus, we have identified a rich “sensory vocabulary” that artists have developed over centuries to make the invisible mechanics of surveillance tangible, emotional, and immediate. Privacy art transforms abstract data flows into physical objects, turns passive viewers into active subjects, and leverages a spectrum of human affect—from the mundane to the unsettling—to force a confrontation with the reality of privacy invasions.

The analysis underscores a critical gap in current privacy communication: while digital interfaces strive for cognitive clarity and regulatory compliance, they often fail to generate the emotional resonance necessary for genuine engagement. In contrast, the artworks in our database succeed by embracing ambiguity, materiality, and narrative. They do not shy away from “creepiness” or discomfort; instead, they leverage these feelings to break through viewer apathy. Whether through the tactile weight of a data-weaved quilt (A7), the disorienting perspective of a looming drone (A22), or the shock of seeing one’s conversations displayed publicly and permanently (A53), these works prove that the most effective way to communicate privacy is not through better text, but through better experience.

From this expansive landscape, we have distilled four specific mechanisms that hold the highest potential for translation into functional privacy systems: materiality, creepiness (affect), storytelling, and data source (the viewer’s or another individual’s). These were not selected for their frequency, but for their capacity to address gaps in current interfaces. Materiality offers a way to ground the ephemeral nature of digital consent in physical reality; creepiness provides a mechanism to disrupt the banality of surveillance through productive discomfort; storytelling contextualizes raw data within human narratives to make it intelli-

gible and memorable; and the choice of data source bridges the gap between abstract policy and lived experience by making the user the immediate subject of the visualization.

While the full scope of privacy art is too broad to implement directly in mass-market systems, these specific vectors offer a focused framework for designing a new generation of privacy communication. They suggest a path where the goal is not to turn privacy policies into art, but to borrow the disruptive strategies of the art world to make privacy communication truly human. Now, we can explore how these mechanisms can be operationalized in a functional artifact, testing whether the provocation of the gallery can be successfully translated into the utility of the interface. With lessons learned from artists, we can then design systems in computer science contexts that do not just tell users what is happening, but make them feel it and reflect on privacy in their own lives.

CHAPTER 5

INTEGRATING ARTISTIC MECHANISMS INTO EXISTING PRIVACY COMMUNICATION METHODS

This chapter presents a workshop-style position paper that proposes a novel research study and artifact design rather than reporting on completed experimental results. Building directly on the findings of Chapter 4, this work outlines a theoretical framework and a concrete study protocol for integrating artistic mechanisms into functional privacy communication. While the study described herein has not yet been executed, the chapter serves as a rigorous proposal for future work, inviting critical feedback on the feasibility, ethics, and potential impact of translating artistic provocation into usable privacy tools.

5.1 Motivation

Privacy communication in digital systems is predominantly handled through standardized interfaces, such as banner-based notices and modal dialogs. While these mechanisms provide transparency and regulatory compliance, they are frequently routine and low-engagement interactions, encouraging habituation and limiting users' attention to how their data is used. In contrast, artistic practice has long explored how to communicate complex, sensitive topics through provocation. In Chapter 4, we examined this space through a large-scale qualitative analysis of artworks engaging with privacy and surveillance. However, while this work surfaces a robust set of design implications, it remains unclear how such strategies compare to conventional privacy communication mechanisms in shaping user perception and experience.

To investigate this avenue further, we take a selection of mechanisms and methods artists use and design a privacy communication method using these practices. For example, we look at transforming digital abstractions into physical material, invading the privacy of the public and the viewer, and setting an unsettling tone through aesthetic and data interpretation.

In this chapter, we propose a between-subjects study to examine the effects of augmenting a traditional privacy communication method, a digital pop-up banner, with artistic mechanisms drawn from our prior study (Chapter 4). Participants will either encounter a text-based pop-up containing personal data and inferences drawn from it, or they will interact with a plant-like material artifact that displays the data in unsettling ways. Materiality and affective tone (in particular, creepiness) are the primary mechanisms in the presentation of these two modes of communication.

Further, we explore the perception of data ownership in a simulated privacy invasion: participants will experience the communications over time starting with either interpretations of their own data or of synthetic data, intending to compare how participants react to violation of their own privacy versus that of others. This again draws on ideas from our previous analysis, where artists often use data from the public or even artwork viewers to tell stories about privacy.

We anticipate that interacting with a physical artifact will persuade users to reflect deeper on privacy as opposed to the experience of viewing a banner pop-up. In particular, the material plant-like design will spur interaction, and the nature of the inferences and creepy aesthetics will provoke a stronger affective response. Using personal data will center the user in the simulated privacy invasion; by juxtaposing the discomfort of “seeing something forbidden” with the realization of “this is my data,” this may provoke a stronger response to the perceived invasion of privacy. By centering the user in the simulated invasion through the use of personal data, we aim to demonstrate that artistic mechanisms can effectively disrupt the apathy often associated with standard privacy notices, turning passive observers into active participants within their own privacy paradigm.

5.2 Related Work

Drawing from our prior work, we investigate concepts relevant to enhancing privacy communications and employing artistic mechanisms to communicate privacy concepts. More specifically, we pull on the mechanisms of unsettling tones, visitor- or user-centric roles, and tactile media. Related research has been conducted in *persuasive technology* and storytelling in design, discomfort as a mechanism for provoking speculative reasoning, and physical media as a vessel for persuasive technology.

5.2.1 *Persuasive Technology*

Persuasive technology is technology designed to change a user’s attitudes or behavior through interactivity [88]. Persuasive technology can significantly enhance users’ security and privacy awareness, and it can be used to develop proactive, customized security awareness systems following persuasive strategies, such as including facilitated steps for enhancing security, short knowledge tests, or real-world security stories [25]. The space of persuasive technology applied to privacy and security topics is mostly application-based, though other privacy-enhancing methods use similar persuasive tactics, such as gamification. For example, Ganesh et al. use games designed to encourage privacy-enhancing behavior in the context of mobile phone usage [95, 94, 93]. They aim to improve users’ awareness of mobile security issues and engage them actively with material that increases users’ self-efficacy in privacy behaviors. While persuasive technology can be used for good, there are ethical concerns for more potential deception and distortion of user agency [88, 26, 112]. *Dark patterns*, manipulative design practices intended to make users take specific actions without full knowledge or autonomy [112], can lead to privacy issues, addiction, or financial and emotional harm [26]. In contrast to intentionally creepy design, these often deliberately obfuscate information and creepy features, removing user agency in favor of the designer. We aim to provoke participants into reflecting on privacy topics, perhaps exposing dark patterns or highlighting

data practices previously unknown to viewers. In contrast to other methods, we will examine the role of artistic mechanisms specifically in communicating messages encouraging privacy-preserving behaviors.

5.2.2 *Narrative and Storytelling in HCI*

Several studies have explored narrative- or story-based design methodologies to enhance user engagement and address broader societal issues in a privacy context. Some works highlight how speculative narrative approaches and scenarios can help designers and users understand and address larger societal concerns [64, 105]. For example, de Almeida and Rafael use speculative design and storytelling to critique data colonialism in Brazil, highlighting the sociopolitical implications of datafication [64]. There are also efforts to use visual storytelling to improve communication between users and designers [163, 115]. With users playing an essential role in real-world applications of technology, researchers must explore the narrative role of users and their hypothetical representations during design [206, 34]. User-centric approaches advocate for more nuanced user characterization in scenarios and humanizing scenario characters with more realistic backgrounds [206]. We will draw on the concept of user-centric storytelling to enhance our implementation with the context of data ownership, giving the participant a more active and immersive role.

5.2.3 *Creepiness by Design*

Researchers have leveraged feelings of discomfort to examine the phenomenon of “creepiness” in design, exploring social norms and their violations in conjunction with emergent technology [271, 158]. Technologies with creepiness as an intentional design or storytelling feature can help researchers uncover unseen impacts of innovative technologies through a societal and ethical lens [226, 233]. Researchers use creepy features to foster discussion and investigate nuanced interpretations about privacy and security [249, 230, 233], a springboard for

ethical design [264], and giving users a more active role in the design process [249, 264, 190]. Differently, Pierce uses speculation on how smart home cameras may be used *outside* of the design process as a creepiness factor to define design concepts for understanding both current and future social and ethical issues with smart technologies [221]. The presented frameworks include data leakage, obtuse or opaque data usage information, and sneaky new features that users did not originally ask for. Considering these concepts abstractly as creepy can help inform design and critical thinking about other similar technologies, making users more aware of potential privacy violations on other platforms. While we also examine the design potential of discomfort and unsettling features as mechanisms, we focus on the specific mechanisms that contribute to an unsettling affect and how they may be employed specifically for communication rather than speculation.

5.2.4 *Materiality and Tangible Interfaces*

Materiality in HCI examines tangible properties of interactive systems and how they shape user experiences. *Tangible user interfaces* (TUIs) operationalize materiality by using physical objects to represent and manipulate digital information [137, 278]. Rather than viewing interfaces as purely digital, materiality emphasizes hybrid contexts, where physical and digital elements intertwine to create new forms of expressive interaction and ground abstract digital concepts in reality [113, 142, 143]. A number of HCI researchers have explored translating digital activity into physical material [132, 100, 165, 139], sometimes in the context of persuasive technology. For example, Sathya and Nakagaki create a system that prints physical paper receipts detailing YouTube activity and watchtime duration to encourage users to take control of their time management and break addictive watch cycles [237]. Together, materiality and TUIs frame physical form as an active participant in interaction design, where digital meaning emerges through engagement with the tangible world. With many of the artworks we encountered in our study, interaction played a role in user immersion and

artist expression; we aim to extract the mechanisms used in physically interactive or tactile works to enhance communication in a privacy-specific context.

5.3 Methods

We present a between-subjects study comparing a conventional privacy communication interface with an artistic, materially grounded alternative. The study investigates how representational modality and data ownership affects users’ perceptions of personal data use, agency, and affective response, particularly perceived “creepiness.”

5.3.1 Study Design

Participants are randomly assigned to one of two conditions: (1) a pop-up-banner-style interface (baseline condition), or (2) a plant-emulating physical artifact system (artistic condition). Participants complete short interaction sessions over time in which behavioral data is collected unobtrusively. In the baseline condition, participants immediately view and respond to the privacy banner. In the artistic condition, participants observe the accumulation of data being translated into material change and interact with the system via a watering action that triggers state updates in the plant artifact.

Within each condition, participants are further assigned to one of two sequences designed to probe expectations related to data ownership. In sequence (a), participants first encounter their own data, unexpectedly followed by synthetic data presented as belonging to another individual. In sequence (b), participants are initially exposed to synthetic data framed as another person’s, followed by an unexpected reveal of their own data.

Data Collection. The system captures highly specific behavioral and contextual data, recognizable to participants as their own, but collected locally to ensure that raw data remains inaccessible to researchers. All data is collected with participant consent.

During each session, the system records fine-grained interaction signals, including:

- cursor movement and click coordinates,
- scrolling behavior,
- keystroke timing (excluding content), and
- action sequences and dwell times.

To increase personalization, the system also incorporates coarse-grained browsing activity, such as:

- domain-level website visits,
- visit frequency and recency within the session, and
- derived groupings (e.g., news, social media, shopping).

To limit sensitivity, page content, query parameters, and user-entered data are neither captured nor stored, and all raw data is processed locally on the participant’s device.

Data Representation. A mock-up of each condition can be seen in Figure 5.1.

In the baseline condition, participants interact with a conventional privacy banner interface modeled on standard consent dialogs. Data is presented starting with general statistics like click counts or activity categories, becoming increasingly fine-grained and specific with items like domain names, cursor paths, and timestamps. This allows participants to begin recognizing their own data, as well as hone their awareness of data collection. Interface controls (e.g., accept/decline) mirror familiar interaction patterns to ground the experience in existing privacy communication conventions.

In the artistic condition, participants engage with a physical, plant-like artifact that materially manifests data over time. To create an uncanny or unsettling affect, the plant would be gnarled and twisted, with eyes instead of flowers to evoke the feeling of being watched. Winding roots, audio, and LED cues as peripherals also contribute to its form.

The artifact translates data into changes in form, movement, and structure. Data dimensions can be mapped to artifact properties, such as:

- Activity level: eye aperture
- Browsing categories: different iris colors
- Sites visited: text-to-speech whispers
- Keystroke timing: LED veins pulsing at typing rate
- Cursor path: approximated by a winding root

To avoid misinterpretation of changes in data profile as system error, all outputs are framed as belonging to a specific data profile (e.g., “Alice: user 1” or “unknown user 1”). Synthetic datasets are designed to include distinct but coherent behavioral patterns, including consistent category distributions and data structure. Transitions between datasets are implemented as legible state changes with profile labels, reinforcing the interpretation of a change in data source rather than a system malfunction. While we would ideally use real, donated behavioral data to maximize ecological validity, this presents significant logistical and ethical barriers. Furthermore, managing real, potentially sensitive, data introduces a substantial liability and privacy risk. Thus, we will aim to use rigorously constructed synthetic data, balancing realism with user safety.

5.3.2 *Participant Data Collection*

To evaluate how effective the artistic intervention is relative to the baseline digital pop-up, we capture participants’ cognitive, affective, and behavioral responses.

Following each interaction session, participants complete a short survey on a Likert scale, adapted from privacy scales such as the Internet Users’ Information Privacy Concerns (IUIPC) [178] or Value of Other People’s Privacy (VOPP)[123]. This includes statements on perceived creepiness and intrusiveness, alongside a measure of sense of agency to determine if the user felt in control of the data. Complementing these self-reports, we log behavioral metrics as indicators of engagement, such as interaction latency (time spent viewing the interface before acting) and interaction frequency (e.g., “watering” action count). These quantitative data provide insight as to which dimensions of the experience users interact with.

Following the final session, participants engage in semi-structured interviews focused on their interpretation of the artifact’s material changes and the moment of transition between synthetic and personal data. We ask participants to reflect on their experience, if any portions of the design impact their perception of privacy or data control. These qualitative insights can give more information on what specifically influenced participants’ behaviors, as well as which mechanisms impacted them the most.

Since our study contains potential stressors (5.4.1), we include a debriefing session to ensure participant well-being after the experience. We explicitly clarify the nature of the data (e.g., regarding synthetic data), contextualize the study’s research goals, and provide resources for any lingering discomfort.

5.4 Discussion and Conclusions

As we move toward the execution of this study, we anticipate shifting the user’s engagement from passive consent to active, affective reflection. We hypothesize that the baseline condition (the digital banner) will likely result in “banner blindness” observed in prior literature, where users rapidly dismiss the notification without processing the data. In contrast, the artistic condition leverages materiality to force reflection. The physical presence of the plant-like artifact, coupled with the deliberate “watering” interaction required to trigger state updates, creates an interruption that breaks the habitual flow of browsing.

We further anticipate that mapping users’ behavioral data to the plant’s behavior will generate a stronger affective response, specifically the feeling of “creepiness” or discomfort. By translating abstract metrics like cursor paths and keystroke timing into things like eye distortions, pulsing LED veins, and whispered audio cues, the system makes the invisible machinery of surveillance tangible and visceral. We expect participants in the artistic condition to report higher levels of privacy reflection, not necessarily because they understood the data better, but because they felt the intrusion more acutely. The gnarled aesthetic and

the substitution of flowers with watching eyes are intended to evoke a strong discomfort that lingers after the interaction ends, challenging the standard of bland, informational interfaces.

We hypothesize that the transition between synthetic data (framed as another person’s) and real, personal data will serve as a pivotal moment of revelation. This unexpected reveal is designed to center the user in the simulated invasion, transforming them from a detached observer of a stranger’s privacy violation into the direct subject of the surveillance, or from the subject of surveillance to the surveillant in power. This shift will demonstrate that artistic mechanisms concerning data ownership and subjects can effectively disrupt the apathy often associated with standard privacy notices, turning passive observers into active participants in their own privacy paradigm.

As we finalize the prototype, our immediate focus will be on refining the data-to-form mapping. Artists thrive on open-ended interpretation, but users in a research setting require clarity to avoid confusion. A major risk is that the “plant” metaphor becomes too obscure, leading to confusion rather than reflection. If the artifact becomes too ambiguous or the data representation too opaque, participants may disengage entirely or assume the system is broken. We need to ensure that visual changes in the plant are legible enough to be interpreted as data-driven rather than random noise, as well as clear enough to be recognizable as a specific individual’s. We plan to conduct a small-scale pilot study to test data readability and how best to integrate the “creepy” aesthetic.

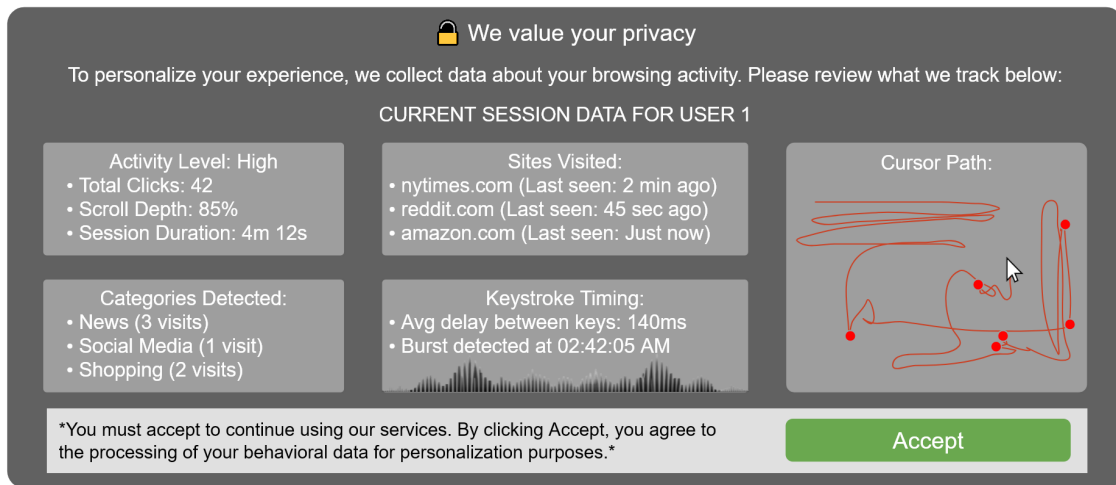
5.4.1 Potential Limitations and Ethical Considerations

The ethical dimensions of this study are equally complex. While artists can freely invade the privacy of viewers for commentary, a system designed for research must navigate consent carefully. Our decision to process behavioral data locally is a necessary compromise, but it raises the question whether the system simulates intrusion effectively if the participant knows their data is not distributed or seen. Conversely, a participant may experience severe

discomfort when confronted with the new data condition (e.g., if they are suddenly confronted with their own data after only seeing others'). Careful consideration must be taken to ensure that participants trust that their data are truly local once this shift is revealed.

Ideally, this study would also use real, donated data to provide participants with an authentic experience of their own digital footprint in the “other” scenario. However, obtaining such data involves complex consent protocols where donors must agree to have their private behaviors used in such a study designed to simulate a privacy violation; additionally, there is risk of accidental re-identification or leaking sensitive data, which is not acceptable. By using synthetic data, we ensure that the “violation” remains a controlled experience rather than a genuine breach of trust.

Furthermore, by simulating a privacy invasion against a non-existent person, we must navigate a delicate boundary between provocation and psychological distress. If the synthetic data is too obviously fake, users might treat the entire exercise as a game. On the other hand, we must carefully balance realism with participant psychological impact. Even though the data is synthetic, the behavioral patterns (e.g., “visited a suicide hotline,” “searched for divorce lawyers”) might be realistic enough to trigger a genuine feeling of having violated someone’s dignity. Users might leave the study feeling guilty or anxious, believing they have harmed a real person. Thus, it is important to rigorously debrief users after their participation in the study.



(a) A mock-up of the baseline condition. Participants view a pop-up banner with text summaries of their data, as well as charts showing interaction frequency and a cursor trace map.



(b) A prototype of the artistic condition. Participants experience their data in a physical format through mappings of the same data to different aspects of the plant (movement, aesthetic, audio). Muscle wire is the primary actuator, and LEDs and speakers provide additional sensory output.

Figure 5.1: A comparison of the privacy communication methods outlined in this study. Each condition uses the same data, but the user experiences the data in significantly different ways.

CHAPTER 6

CONCLUSION

This dissertation has explored the multifaceted nature of privacy in computational systems, arguing that a purely technical or data-centric view is insufficient to address the complex realities of human experience. Whether protecting the sensitive transcripts of social-behavioral scientists or communicating data practices to end-users, the core challenge remains the same: bridging the gap between abstract data flows and the lived human experience. By building this bridge between computer science, humanities, and the arts, this work has demonstrated that effective privacy protection requires not only robust algorithms but also intuitive workflows and strong communication strategies.

In Chapter 2, we challenged the limited definitions of identifiability that dominate current systems. By engaging directly with data stewards, we revealed that identifiability is not a fixed property of data, but a fluid, context-dependent paradigm shaped by language, culture, and social nuance. This insight drove the development of a new redaction framework that captures the subtle, often invisible components of identity that automated tools miss. However, recognizing these is only the first step; another challenge lies in acting on them. This led to the design of MARI in Chapter 3, a human-in-the-loop interface that operationalizes the framework. MARI does not just automate redaction; it creates a space where computational power and human judgment converge. This ensures that data protection remains a human-driven act of agency rather than solely an algorithmic process.

However, once data is collected and processed, how do we make its implications felt by the people it concerns? This question propelled the second track of this research, where we turned to the arts for strategies to make the invisible visible. In Chapter 4, our privacy artwork analysis revealed how artists have mastered the art of disrupting apathy around privacy using a myriad of mechanisms. These mechanisms do not just inform; they provoke, unsettle, and embody privacy concepts. In Chapter 5, we proposed translating these artistic mechanisms

into functional privacy communication—in particular, through materiality, affective tone, storytelling, and centering humans behind the data.

Both of these tracks make the black box of privacy transparent, ensuring privacy is an active practice rather than a static state. Whether a researcher is deciding whether to redact a specific phrase in a transcript, or a user is deciding whether to act on data usage policies, the goal is to restore the human to the center of the decision. This suggests that just as *protecting data* requires human judgment to interpret nuance and context, *communicating privacy* requires human engagement to interpret risk and consequence. In both cases, the solution lies not in better automation, but in better human-centered design. The redaction framework provides the language to define what is sensitive; the artistic mechanisms provide the medium to make that sensitivity felt.

6.1 Future Work

6.1.1 Identifiability and Redaction

One immediate proposition for future work in redaction and identifiability would be to perform case studies on real data with real potential users of the application, i.e., researchers. This line of research combines the interview study with the usability study, where we would ask participants to redact their own research data with our tool. However, this could potentially raise ethical questions in participant consent for data that has already been collected; it may be worth conducting this work in collaboration with other researchers *currently* conducting research resulting in unstructured text. These researchers may be able to notify participants about using our application for the data practices after collection, reducing potential concern. Though we have run a study on real-world data and conduct our user study on another subset of real-world data, we are limited to publicly available transcripts with narrow parameters. This future study would give us a more diverse set of data with which

to test our application and give more insight as to how MARI works in real-world contexts.

A similar line of research might include expanding the scope of the study to encompass either more diverse (e.g., non-English language transcripts) or more niche (e.g., therapists' notes) data than we were able to access. Large data sources are required for good performance from LLMs, which is lacking for certain languages, leading to potential bias. For example, an in-depth study on code-switching corpora in particular may provide interesting insight as to which identifiers LLMs recognize when confronted with multiple languages. Furthermore, for niche or very specific datasets, there may be different requirements for the balance between utility and privacy, or differences in LLM analysis; it may be worth conducting a study on hyperspecific datasets to see if there is a noticeable difference in performance.

6.1.2 Privacy Communication

The path forward from the work presented in this dissertation is paved with opportunities to deepen and broaden the application of artistic mechanisms in privacy. Future work may consider performing a longitudinal or large-scale study of custom artistic privacy communications. If conducted over a longer period of time, one could see how users' perceptions and privacy-preserving behaviors change over time relative to artistic privacy communications they receive. There may also be more opportunity to explore diverse reactions to privacy art mechanisms—in a larger-scale study, one can examine how different cultural backgrounds, age brackets, or other demographics impact participants' experiences with privacy communications. This would also give more generalizable insight as to the effectiveness of a certain artistic privacy communication. A co-design workshop or “game jam”-style event with non-artists can also produce unique ideas and uses of artistic mechanisms for common privacy communications, expanding possibilities for communication study.

Another potential line of research might examine different artistic contexts of privacy art and see how these might apply to privacy communications. For instance, popular art,

street art, and art created by non-artists may be more relatable or accessible to the everyman, though the creatives may not have as many art resources as professional artists. This may provide a more in-depth understanding of public concerns regarding privacy, reflecting individuals' experiences with privacy and further humanizing the message. However, given that our inclusion criteria for the main study were somewhat arbitrary, the boundary for inclusion on these types of art may be even more nebulous.

6.2 Final Thoughts

As we navigate an increasingly data-driven world, the stakes of privacy research have never been higher. Ultimately, advancing human-centered privacy requires a genuine dialogue between computer science and the arts and humanities: by operationalizing context-dependent definitions of identifiability to *enact privacy* alongside legible, artist-inspired *privacy communication*, we can truly empower individuals to understand and control their data. This dissertation advances the field of usable privacy by redefining how we might approach identifiability, workflow design, and cross-disciplinary collaboration. We move beyond rigid definitions of identifiability toward context-aware frameworks that utilize the fluidity of human language and social nuance. We demonstrate that the most effective privacy tools are not fully automated systems, but hybrid workflows that integrate computational scalability with human judgment, ensuring that automation serves as an assistant rather than overtaking human agency. Finally, we bridge the gap between protection and communication by showing that the strategies used to protect data are linked to the mechanisms needed to communicate about it effectively. This perspective suggests that the future of privacy lies not in choosing *between* technical precision and human experience, but in *synthesizing* them to create systems that are both secure and deeply human.

REFERENCES

- [1] Ahmed Abbasi and Hsinchun Chen. Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems (TOIS)*, 26(2):1–29, 2008.
- [2] Allison Adams, Eric Aili, Daniel Aioanei, Rebecca Jonsson, Lina Mickelsson, Dagmar Mikmekova, Fred Roberts, Javier Fernandez Valencia, and Roger Wechsler. Anonymate: A toolkit for anonymizing unstructured chat data. In *Proceedings of the Workshop on NLP and Pseudonymisation*, pages 1–7, 2019.
- [3] Adobe Inc. Adobe acrobat pro, 2026. URL <https://adobe.com>.
- [4] Weiwei Ai. Surveillance camera, 2010. URL <https://www.phillips.com/detail/ai-weiwei/NY010620/46>.
- [5] Federico Albanese, Daniel Ciolek, and Nicolas D’Ippolito. Text sanitization beyond specific domains: Zero-shot redaction & substitution with large language models. *arXiv:2311.10785*, 2023.
- [6] Saul Albert, Laura E. de Ruiter, and J.P. de Ruiter. CABNC: the Jeffersonian transcription of the Spoken British National Corpus. 2015. URL <https://saulalbert.github.io/CABNC/>.
- [7] Amy Alexander. Sven, 2006/2007. URL <https://amy-alexander.com/live-performance/sven/>.
- [8] Meriç Algün Ringborg. Which no one will ever see, 2012. URL <https://www.mericalgun.com/works.html#Which%20No%20One%20Will%20Ever%20See>.
- [9] Amazon Web Services. Macie. <https://aws.amazon.com/macie/>, 2024.
- [10] Refik Anadol. Convergence la, 2017. URL <https://refikanadol.com/works/convergence-la/>.
- [11] Balamurugan Anandan and Chris Clifton. Significance of term relationships on anonymization. In *2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, volume 3, pages 253–256. IEEE, 2011.
- [12] Balamurugan Anandan, Chris Clifton, Wei Jiang, Mummoorthy Murugesan, Pedro Pastrana-Camacho, and Luo Si. t-Plausibility: Generalizing words to desensitize text. *Trans. Data Priv.*, 5(3):505–534, 2012.
- [13] Anthropic. Claude 3.5 sonnet. <https://www.anthropic.com/claude/sonnet>, 2025.
- [14] Anthropic. Claude Opus 4.5. <https://www.anthropic.com/claude/opus>, 2025.
- [15] Burak Arikan. Mypocket, 2008. URL <https://burak-arikan.com/mypocket/>.

- [16] Ars Electronica Linz GmbH & Co. ARS ELECTRONICA ARCHIVE - PRIX. <https://archive.aec.at/prix/>. Accessed: 2025-08-27.
- [17] Dimitris Asimopoulos, Ilias Siniosoglou, Vasileios Argyriou, Thomai Karamitsou, Eleftherios Fountoukidis, Sotirios K Goudos, Ioannis D Moscholios, Konstantinos E Psannis, and Panagiotis Sarigiannidis. Benchmarking advanced text anonymisation methods: A comparative study on novel and traditional approaches. *arXiv:2404.14465*, 2024.
- [18] Larry Catá Backer. Global panopticism: states, corporations, and the governance effects of monitoring regimes. *Indiana Journal of Global Legal Studies*, 15(1):101–148, 2008.
- [19] Kenneth A Bamberger and Deirdre K. Mulligan. *Privacy on the ground: driving corporate behavior in the United States and Europe*. MIT Press, 2015.
- [20] Jeffrey Bardzell and Shaowen Bardzell. What is “critical” about critical design? In *Proceedings of the 2013 SIGCHI Conference on Human Factors in Computing Systems*, 2013. ISBN 9781450318990. doi:10.1145/2470654.2466451. URL <https://doi.org/10.1145/2470654.2466451>.
- [21] Jeffrey Bardzell, Shaowen Bardzell, Carl DiSalvo, William Gaver, and Phoebe Sengers. The humanities and/in hci. In *Proceedings of the 2012 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 2012. ISBN 9781450310161. doi:10.1145/2212776.2212405. URL <https://doi.org/10.1145/2212776.2212405>.
- [22] Tim Barringer and Tom Flynn. *Colonialism and the object: Empire, material culture and the museum*. Routledge, 2012.
- [23] Susanne Barth, Dan Ionita, and Pieter Hartel. Understanding online privacy—a systematic review of privacy visualizations and privacy by design guidelines. *ACM Comput. Surv.*, 55(3), February 2022. ISSN 0360-0300. doi:10.1145/3502288. URL <https://doi.org.edu/10.1145/3502288>.
- [24] Aram Bartholl. Choir of missed connections, 2024. URL <https://arambartholl.com/choir-of-missed-connections/>.
- [25] Mohammed Abdullah Bawazir, Murni Mahmud, Nurul Nuha Abdul Molok, and Akram Zeki. The effects of persuasive technology for influencing end-users’ information security awareness. *Journal of Information Systems and Digital Technologies*, 4(1):56–81, 2022.
- [26] Dennis Benner, Sofia Schöbel, Andreas Janson, and Jan Marco Leimeister. How to achieve ethical persuasive design: A review and theoretical propositions for information systems. *AIS Transactions on Human-Computer Interaction (THCI)*, 14(4):548–577, 2022.
- [27] Jeremy Bentham. *Panopticon; or, The Inspection- House*. T. Payne, 1791.

- [28] Hanna Berg, Aron Henriksson, and Hercules Dalianis. The impact of de-identification on downstream named entity recognition in clinical text. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, pages 1–11, 2020.
- [29] Jenny Berkholz, Anika Rahman, and Gunnar Stevens. Playing with privacy: Exploring the social construction of privacy norms through a card game. *Proc. ACM Hum.-Comput. Interact.*, 9(1), January 2025. doi:10.1145/3701202. URL <https://doi.org/10.1145/3701202>.
- [30] Charles Bernheimer. Degas’s brothels: Voyeurism and ideology. *Representations, Special Issue: Misogyny, Misandry, and Misanthropy*, 20:158–186, 1987. doi:10.2307/2928506.
- [31] Ann-Christin Bertrand and James Bridle. *WATCHED! Surveillance, art and photography*. Verlag der Buchhandlung Walther König, 2016.
- [32] Viktoria Binschtok. Suspicious minds, 2009. URL <https://viktoriabinschtok.wordpress.com/work-3/suspicious-minds/>.
- [33] Justin Blinder. Dark side of the prism, 2013. URL <https://justin.work/dark-side-of-the-prism/>.
- [34] Stefan Blomkvist. Persona—an overview. *Retrieved November, 22(2004):15*, 2002.
- [35] Mark Blythe, Kristina Andersen, Rachel Clarke, and Peter Wright. Anti-solutionist strategies: Seriously silly design fiction. In *Proceedings of the 2016 SIGCHI Conference on Human Factors in Computing Systems*, 2016. ISBN 9781450333627. doi:10.1145/2858036.2858482. URL <https://doi.org/10.1145/2858036.2858482>.
- [36] Nigel Bosch, R Crues, Najmuddin Shaik, and Luc Paquette. “hello,[redacted]”: Protecting student privacy in analyses of online discussion forums. In *Proceedings of the 13th International Conference on Educational Data Mining*, 2020.
- [37] Assia Boundaoui. The feeling of being watched, 2018. URL <https://www.feelingofbeingwatched.com/>.
- [38] Mariem Brahem, Jasmine Watissee, Cédric Eichler, Adrien Boiret, Nicolas Anciaux, and Jose Maria de Fuentes. retellme: Design rules for using large language models to protect the privacy of individuals in their textual contributions. In *International Workshop on Data Privacy Management @ ESORICS*, 2024.
- [39] Michael Brennan, Sadia Afroz, and Rachel Greenstadt. Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity. *ACM Transactions on Information and System Security (TISSEC)*, 15(3):1–22, 2012.
- [40] James Bridle. Every cctv camera (cc), 2017. URL <https://jamesbridle.com/works/every-cctv-camera-cc>.

- [41] Andrea Mubi Brighenti. Artveillance: At the crossroad of art and surveillance. *Surveillance & Society*, 7(2):137–148, 2010.
- [42] Jeffery Brown. In a world full of surveillance, artist trevor paglen stares back, 2018. URL <https://www.pbs.org/newshour/show/in-a-world-full-of-surveillance-e-artist-trevor-paglen-stares-back>.
- [43] Kathryn Brown. The aesthetics of presence: Looking at degas’s bathers. *The Journal of Aesthetics and Art Criticism*, 68:331–341, 2010. doi:10.1111/j.1540-6245.2010.01428.x.
- [44] Jonah Brucker-Cohen. Human error: Demand ware (2022), 2022. URL <https://www.coin-operated.com/2022/07/27/human-error-demand-ware-2022/>.
- [45] Jason Bruges. Platform 5, 2012. URL <https://www.jasonbruges.com/platform-5>.
- [46] Heath Bunting. The status project, c. 2012. URL <https://irational.org/cgi-bin/cv2/temp.pl>.
- [47] Allison Burtch. Mic jammer, 2017. URL <https://github.com/allisonburtch/micjammer>.
- [48] José Cabrero-Holgueras and Sergio Pastrana. A methodology for large-scale identification of related accounts in underground forums. *Computers & Security*, 111:102489, 2021.
- [49] Susan Cahill, Morgan Campbell, and Stéphanie McKnight. Art and surveillance project. Website, 2025. URL <http://www.artandsurveillance.com/>.
- [50] Daniel Canogar. Cmyk, 2014. URL <https://www.danielcanogar.com/work/cmyk>.
- [51] Mary Cassatt. The bath, 1891.
- [52] Claudio Celis. Critical surveillance art in the age of machine vision and algorithmic governmentality: Three case studies. *Surveillance & Society*, 18(3):295–311, 2020. doi:10.24908/ss.v18i3.13410.
- [53] Janet X. Chen, Allison McDonald, Yixin Zou, Emily Tseng, Kevin A. Roundy, Acar Tamersoy, Florian Schaub, Thomas Ristenpart, and Nicola Dell. Trauma-informed computing: Towards safer technology experiences for all. In *Proceedings of the 2022 SIGCHI Conference on Human Factors in Computing Systems*, 2022. ISBN 9781450391573. doi:10.1145/3491102.3517475. URL <https://doi.org/10.1145/3491102.3517475>.
- [54] Joana Chicau, Sarah Fdili Alaoui, Anne Lee Steele, Hazel Ryan, Yadira Sánchez, Romaine Gadelrab, Caroline Sindors, Mukul Patel, Gavin Starks, John Fass, and Rebecca Fiebrink. Designing counter-choreographies: Embodied choreographic approaches for critical examination of online tracking. In *Proceedings of the 2025 Conference on Creativity and Cognition*, C&C ’25, page 872–885, New York, NY, USA, 2025. Association

- for Computing Machinery. ISBN 9798400712890. doi:10.1145/3698061.3726928. URL <https://doi.org/10.1145/3698061.3726928>.
- [55] Jenny Chim, Jonathan Pearson, Dan Schofield, and Maria Liakata. Evaluating privacy leakages in llm-driven ambient clinical documentation. *Frontiers in Digital Health*, 8, 2026. URL <https://www.frontiersin.org/journals/digital-health/articles/10.3389/fdgth.2026.1761624/full>.
 - [56] CITI Program. Website. <https://about.citiprogram.org/>, 2024.
 - [57] Aloni Cohen. Attacks on deidentification’s defenses. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 1469–1486, 2022.
 - [58] Beatriz da Costa. Pigeonblog, 2006-2008. URL <https://nideffer.net/shaniweb/pigeonblog.php>.
 - [59] Scott E. Coull, Michael P. Collins, Charles V. Wright, Fabian Monroe, and Michael K. Reiter. On web browsing privacy in anonymized netflows. In *USENIX Security Symposium*, pages 339–352, 2007.
 - [60] James Coupe. Sanctum, 2015. URL <http://jamescoupe.com/?p=1740>.
 - [61] Lorrie Cranor. Security blanket, 2013. URL <https://lorrie.cranor.org/blog/2013/08/12/security-blanket/>.
 - [62] Google Arts & Culture. Girl reading a letter by an open window. URL <https://artsandculture.google.com/asset/girl-reading-a-letter-by-an-open-window/3wFQaidzxA5mqg?hl=en>.
 - [63] Chad Cumby and Rayid Ghani. A machine learning based system for semi-automatically redacting documents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2011.
 - [64] Fabio de Almeida and Sónia Rafael. Data sins: Exploring data colonialism through storytelling-based speculative design practices. In *Proceedings of the 2025 Conference on Creativity and Cognition, C&C ’25*, page 141–155, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400712890. doi:10.1145/3698061.3726961. URL <https://doi.org/10.1145/3698061.3726961>.
 - [65] Sander de Jong and Dayana Spagnuolo. Iconified representations of privacy policies: A gdpr perspective. In *World Conference on Information Systems and Technologies*, pages 796–806. Springer, 2020.
 - [66] DeepSeek. DeepSeek-V3.2 Release | DeepSeek API Docs. <https://api-docs.deepseek.com/news/news251201>, 2025.

- [67] Soteris Demetriou, Nathaniel D. Kaufman, Jonah Baim, Adam J. Goldsher, and Carl A. Gunter. Toward an extensible framework for redaction. In *Proceedings of the 1st International Workshop on Security and Privacy for the Internet-of-Things*, 2018.
- [68] Dries Depoorter. Surveillance speaker, 2018-2024. URL <https://driesdepoorter.be/surveillancespeaker/>.
- [69] Rod Dickinson, Nathan Semertzidis, and Florian Floyd Mueller. Machine in the middle: Exploring dark patterns of emotional human-computer integration through media art. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI EA '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391566. doi:10.1145/3491101.3503555. URL <https://doi.org/10.1145/3491101.3503555>.
- [70] Mimi Onuoha. Classification.01, 2017. URL <https://mimionuoha.com/classification01>.
- [71] Bogomir Doring. The hoodie, 2008. URL <https://nieuweinstituut.nl/en/projects/the-hoodie/bogomir-doring>.
- [72] Yao Dou, Isadora Krsek, Tarek Naous, Anubha Kabra, Sauvik Das, Alan Ritter, and Wei Xu. Reducing privacy risks in online self-disclosures with language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024.
- [73] DuckDuckGo. Duck.ai - intelligent search and generative ai, 2026. URL <https://duck.ai>.
- [74] Anthony Dunne and Fiona Raby. *Speculative everything: Design, fiction, and social dreaming*. MIT press, 2013.
- [75] Myk Eff. The top 10 works of surveillance art. Medium, 2025. URL <https://medium.com/spy-novel-research/the-top-10-works-of-surveillance-art-15a39bcd2ce>.
- [76] Hasan Elahi. Citizen, 2006. URL <https://elahi.wayne.edu/citizen.php>.
- [77] Hasan Elahi. Recap: Hiding in plain sight by hasan elahi, 2006. URL <https://www.objectifs.com.sg/recap-hasan-elahi/>.
- [78] Hasan Elahi. Fbi, here i am!, 2011. URL https://www.ted.com/talks/hasan_elahi_fbi_here_i_am.
- [79] Chris Elsdén, David Chatting, Michael Duggan, Andrew Carl Dwyer, and Pip Thornton. Zoom obscura: Counterfunctional design for video-conferencing. In *Proceedings of the 2022 SIGCHI Conference on Human Factors in Computing Systems*, 2022. ISBN

9781450391573. doi:10.1145/3491102.3501973. URL <https://doi.org/10.1145/3491102.3501973>.
- [80] David England, Celine Latulipe, Nick Bryan-Kinns, Ernest Edmonds, and Sean Clark. Art.chi ii: Digital art in a post-digital world. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 2016. ISBN 9781450340823. doi:10.1145/2851581.2856474. URL <https://doi.org/10.1145/2851581.2856474>.
 - [81] European Parliament and Council of the European Union. Regulation (eu) 2016/679: General data protection regulation article 4: Definitions, April 27, 2016. URL <https://eur-lex.europa.eu/eli/reg/2016/679/oj>.
 - [82] Florian M. Farke, David G. Balash, Maximilian Golla, Markus Dürmuth, and Adam J. Aviv. Are privacy dashboards good for end users? evaluating user perceptions and reactions to google’s my activity. In *Proceedings of the 30th USENIX Security Symposium*, 2021.
 - [83] Elaine Farrow, Johanna D Moore, and Dragan Gasevic. Names, nicknames, and spelling errors: Protecting participant identity in learning analytics of online discussions. In *LAK23: 13th International Learning Analytics and Knowledge Conference*, pages 145–155, 2023.
 - [84] Oluwaseyi Feyisetan, Tom Diethe, and Thomas Drake. Leveraging hierarchical representations for preserving privacy and utility in text. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 210–219, 2019. doi:10.1109/ICDM.2019.00031.
 - [85] Casey Fiesler and Nicholas Proferes. “participant” perceptions of twitter research ethics. *Social Media + Society*, 4(1), 2018. doi:10.1177/2056305118763366.
 - [86] Simone Fischer-Hübner and Farzaneh Karegar. Overview of usable privacy research: Major themes and research directions. *The Curious Case of Usable Privacy: Challenges, Solutions, and Prospects*, pages 43–102, 2024.
 - [87] Ulrich Flegel. Pseudonymizing UNIX log files. In *International Conference on Infrastructure Security*, pages 162–179. Springer, 2002.
 - [88] BJ Fogg. Persuasive computers: perspectives and research directions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI ’98, page 225–232, USA, 1998. ACM Press/Addison-Wesley Publishing Co. ISBN 0201309874. doi:10.1145/274644.274677. URL <https://doi.org/10.1145/274644.274677>.
 - [89] Sarah Fox, Noura Howell, Richmond Wong, and Franchesca Spektor. Vivewell: Speculating near-future menstrual tracking through current data practices. In *Proceedings of the 2019 Designing Interactive Systems Conference*, 2019. ISBN 9781450358507. doi:10.1145/3322276.3323695. URL <https://doi.org/10.1145/3322276.3323695>.

- [90] Batya Friedman, Peter H Kahn Jr, Alan Borning, and Alina Huldtgren. Value sensitive design and information systems. In *Early engagement and new technologies: Opening up the laboratory*, pages 55–95. Springer, 2013.
- [91] Ahmed Frikha, Nassim Walha, Krishna Kanth Nakka, Ricardo Mendes, Xue Jiang, and Xuebing Zhou. Incognitext: Privacy-enhancing conditional text anonymization via llm-based private attribute randomization. *arXiv:2407.02956*, 2024.
- [92] Kelsey R. Fulton, Rebecca Gelles, Alexandra McKay, Yasmin Abdi, Richard Roberts, and Michelle L. Mazurek. The effect of entertainment media on mental models of computer security. In *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*, pages 79–95, Santa Clara, CA, August 2019. USENIX Association. ISBN 978-1-939133-05-2. URL <https://www.usenix.org/conference/soups2019/presentation/fulton>.
- [93] Anirudh Ganesh, Chinenye Ndulue, and Rita Orji. Permarun- a persuasive game to improve user awareness and self-efficacy towards secure smartphone behaviour. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380959. doi:10.1145/3411763.3451781. URL <https://doi.org/10.1145/3411763.3451781>.
- [94] Anirudh Ganesh, Chinenye Ndulue, and Rita Orji. Smartphone security and privacy—a gamified persuasive approach with protection motivation theory. In *International Conference on Persuasive Technology*, pages 89–100. Springer, 2022.
- [95] Anirudh Ganesh, Chinenye Ndulue, and Rita Orji. Tailoring a persuasive game to promote secure smartphone behaviour. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2023.
- [96] Akshatha Gangadharaiah, Bijeeta Pal, and Sameera Ghayyur. Transforming data through deidentification. In *Proceedings of the 2024 USENIX Conference on Privacy Engineering Practice and Respect*, 2024.
- [97] Simson Garfinkel et al. *De-identification of Personal Information*. US Department of Commerce, National Institute of Standards and Technology, 2015.
- [98] Paul Garrin. Yuppie ghetto with watchdog, 1989. URL <https://zkm.de/en/artwork/yuppie-ghetto-with-watchdog>.
- [99] Leon Garza, Anantaa Kotal, Aritran Piplai, Lavanya Elluri, Prajit Kumar Das, and Aman Chadha. Prvl: Quantifying the capabilities and risks of large language models for pii redaction. In *2025 IEEE 7th International Conference on Trust, Privacy and Security in Intelligent Systems, and Applications (TPS-ISA)*, pages 148–159, 2025. doi:10.1109/TPS-ISA67132.2025.00025.

- [100] William W Gaver, John Bowers, Kirsten Boehner, Andy Boucher, David WT Cameron, Mark Hauenstein, Nadine Jarvis, and Sarah Pennington. Indoor weather stations: investigating a ludic approach to environmental HCI through batch prototyping. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3451–3460, 2013.
- [101] Jakub Geltner. Nest, 2015. URL <http://www.geltner.cz/2009/nest-05/>.
- [102] Mariam Ghani. Security blanket, 2005. URL <https://www.mariamghani.com/work/195>.
- [103] Ghostery. Free tracker & ad blocker extension. Version 4.4.25, 2025. URL <https://www.ghostery.com/ghostery-ad-blocker>.
- [104] Steve Giasson. Vox, 2015-2016. URL <http://www.artandsurveillance.com/artists-artworks-exhibits/>.
- [105] Kentaro Go and John M. Carroll. The blind men and the elephant: views of scenario-based system design. *Interactions*, 11(6):44–53, November 2004. ISSN 1072-5520. doi:10.1145/1029036.1029037. URL <https://doi.org/10.1145/1029036.1029037>.
- [106] Kyriaki Goni. Deletion process, 2013, 2015, 2019. URL <https://kyriakigoni.com/projects/deletion-process>.
- [107] Kyriaki Goni. Deletion process_only you can see my history: investigating digital privacy, digital oblivion, and control on personal data through an interactive art installation. In *ACM SIGGRAPH 2016 Art Gallery*, SIGGRAPH ’16, page 324–333, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342803. doi:10.1145/2897843.2915187. URL <https://doi.org/10.1145/2897843.2915187>.
- [108] Google. Cloud Data Loss Prevention (DLP). <https://cloud.google.com/dlp/>, 2021.
- [109] Google. Sensitive Data Protection. <https://cloud.google.com/security/products/sensitive-data-protection>, 2024.
- [110] Google. Gemini models. <https://ai.google.dev/gemini-api/docs/models/gemini>, 2025.
- [111] Google. Gemini 2.5 Pro. <https://docs.cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-pro>, 2025.
- [112] Colin M. Gray, Cristiana Teixeira Santos, Nataliia Bielova, and Thomas Mildner. An ontology of dark patterns knowledge: Foundations, definitions, and a pathway for shared knowledge-building. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI ’24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703300. doi:10.1145/3613904.3642436. URL <https://doi.org/10.1145/3613904.3642436>.

- [113] Shad Gross, Jeffrey Bardzell, and Shaowen Bardzell. Structures, forms, and stuff: the materiality and medium of interaction. *Personal and Ubiquitous Computing*, 18(3): 637–649, 2014.
- [114] Ben Grosser. Go rando, 2021. URL <https://bengrosser.com/projects/go-rando/>.
- [115] Mengyao Guo, Kexin Nie, Jinda Han, Guan Wang, Xin Wang, zhishun Chi, Jie Fu, and Ze Gao. Visual storytelling in hci: A workshop on narrative development through sequential art. In *Proceedings of the 2025 Conference on Creativity and Cognition, C&C '25*, page 3–12, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400712890. doi:10.1145/3698061.3728391. URL <https://doi.org/10.1145/3698061.3728391>.
- [116] Wentao Guo, Aditya Kishore, Adam J. Aviv, and Michelle L. Mazurek. A qualitative analysis of practical de-identification guides. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, 2024.
- [117] Wentao Guo, Paige Pepitone, Adam J. Aviv, and Michelle L. Mazurek. How researchers de-identify data in practice. In *Proceedings of the 34th USENIX Security Symposium*, 2025.
- [118] Vaibhav Gusain and Douglas Leith. Towards quantifying the privacy of redacted text. In *European Conference on Information Retrieval*, pages 423–429. Springer, 2023.
- [119] Hana Habib, Yixin Zou, Yaxing Yao, Alessandro Acquisti, Lorrie Cranor, Joel Reidenberg, Norman Sadeh, and Florian Schaub. Toggles, dollar signs, and triangles: How to (in)effectively convey privacy choices with icons and link texts. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi:10.1145/3411764.3445387. URL <https://doi.org/10.1145/3411764.3445387>.
- [120] Elizabeth Harney and Ruth B. Phillips. *Mapping modernisms: Art, indigeneity, colonialism*. Duke University Press, 2018.
- [121] Hugh Hart. The art of surveillance. URL <https://www.wired.com/2007/11/the-art-of-surveillance/>.
- [122] Adam Harvey. Hyperface, 2017. URL <https://adam.harvey.studio/hyperface/>.
- [123] Rakibul Hasan, Rebecca Weil, Rudolf Siegel, and Katharina Krombholz. A Psychometric Scale to Measure Individuals’ Value of Other People’s Privacy (VOPP). 4 2023. doi:10.60882/cispa.24614742.v2. URL https://publications.cispa.de/articles/conference_contribution/A_Psychometric_Scale_to_Measure_Individuals_Value_of_Other_People_s_Privacy_VOPP_/24614742.

- [124] Drew Hemment, Morgan Currie, SJ Bennett, Jake Elwes, Anna Ridler, Caroline Sindlers, Matjaz Vidmar, Robin Hill, and Holly Warner. Ai in the public eye: Investigating public ai literacy through ai art. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 931–942, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701924. doi:10.1145/3593013.3594052. URL <https://doi.org/10.1145/3593013.3594052>.
- [125] Lefebvre Henri and Nicholson-Smith Donald. *The production of space*. Blackwell, Oxford, 1991.
- [126] Dan Hett. Sorry to bother you, 2018. URL <https://danhett.itch.io/sorry>.
- [127] Kashmir Hill. This surveillance artist knows how you got that perfect instagram photo, 2022. URL <https://www.nytimes.com/2022/09/24/technology/surveillance-footage-instagram.html>.
- [128] Steven Hill, Zhimin Zhou, Lawrence K Saul, and Hovav Shacham. On the (in)effectiveness of mosaicing and blurring as tools for document redaction. *Proc. Priv. Enhancing Technol.*, 2016(4):403–417, 2016.
- [129] Leif-Erik Holtz, Katharina Nocun, and Marit Hansen. Towards displaying privacy information with icons. In *IFIP primelife international summer school on privacy and identity management for life*, pages 338–348. Springer, 2010.
- [130] Matthew Honnibal. spaCy: Industrial-strength natural language processing. <https://spacy.io>, 2024.
- [131] Esther Hovers. The right to be forgotten, 2021-2024. URL <https://estherhovers.com/works-main>.
- [132] Yuan-Yao Hsu, Wenn-Chieh Tsai, Wan-Chen Lee, and Rung-Huei Liang. Botanical printer: An exploration on interaction design with plantness. In *Proceedings of the 2018 Designing Interactive Systems Conference*, pages 1055–1068, 2018.
- [133] Yuhan Hu, Zhengnan Zhao, Abheek Vimal, and Guy Hoffman. Soft skin texture modulation for social robotics. In *2018 IEEE International Conference on Soft Robotics (RoboSoft)*, pages 182–187, 2018. doi:10.1109/ROBOSOFT.2018.8404917.
- [134] Hsin-Chien Huang. The eye and i, 2023. URL https://hsinchienhuang.com/pix/_3artworks/i_theEyeAndI/p0.php?lang=en.
- [135] iDox.ai. idox ai - intelligent information management, 2026. URL <https://idox.ai/>.
- [136] Aki Ishida and Noritaka Minami. Obsolescent masculinity: Nakagin capsule tower demolition (tokyo), 2022. URL <https://disegnojournal.com/newsfeed/obsolescent-masculinity-nakagin-capsule-tower-demolition-tokyo>.

- [137] Hiroshi Ishii. Tangible user interfaces. In *The human-computer interaction handbook*, pages 495–514. CRC Press, 2007.
- [138] Jane Jacobs. *Death and Life of Great American Cities, 1961*. Knopf Doubleday Publishing Group, 1992.
- [139] Yvonne Jansen, Pierre Dragicevic, and Jean-Daniel Fekete. Evaluating the efficiency of physical visualizations. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 2593–2602, 2013.
- [140] Hyunjun Jeon, Kyuyoung Kim, and Jinwoo Shin. Redacbench: Can ai erase your secrets? *arXiv preprint arXiv:2603.20208*, 2026. URL <https://arxiv.org/abs/2603.20208>.
- [141] Frederike Jung, Jonah-Noël Kaiser, Kai Von Holdt, Wilko Heuten, and Jochen Meyer. The art of privacy – a theatrical privacy installation in virtual reality. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI EA '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394222. doi:10.1145/3544549.3583893. URL <https://doi.org/10.1145/3544549.3583893>.
- [142] Heekyoung Jung and Erik Stolterman. Form and materiality in interaction design: a new approach to hci. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '11, page 399–408, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450302685. doi:10.1145/1979742.1979619. URL <https://doi.org/10.1145/1979742.1979619>.
- [143] Heekyoung Jung and Erik Stolterman. Digital form and materiality: propositions for a new approach to interaction design research. In *Proceedings of the 7th Nordic Conference on Human-Computer Interaction: Making Sense Through Design*, pages 645–654, 2012.
- [144] Gary Kafer. Reimagining resistance: Performing transparency and anonymity in surveillance art. *Surveillance & Society*, 14:227–239, 09 2016. doi:10.24908/ss.v14i2.6005.
- [145] KairUs. Found footage stalkers, 2018. URL <https://kairus.org/portfolio/forensic-fantasies-trilogy/>.
- [146] Mehmet Kayaalp, Allen C. Browne, Fiona M. Callaghan, Zeyno A. Dodd, Guy Divita, Selcuk Ozturk, and Clement J. McDonald. The pattern of name tokens in narrative clinical text and a comparison of five systems for redacting them. *Journal of the American Medical Informatics Association*, 21(3):423–431, 2014.
- [147] Patrick Gage Kelley, Joanna Bresee, Lorrie Faith Cranor, and Robert W. Reeder. A "nutrition label" for privacy. In *Proceedings of the 5th Symposium on Usable Privacy and Security*, SOUPS '09, New York, NY, USA, 2009. Association for Computing

- Machinery. ISBN 9781605587363. doi:10.1145/1572532.1572538. URL <https://doi.org/10.1145/1572532.1572538>.
- [148] Dave Kemp. Data collection, 2010. URL <https://davekemp.ca/projects/data-collection/>.
 - [149] Mohammad Khalil and Martin Ebner. De-identification in learning analytics. *Journal of Learning Analytics*, 3(1):129–138, 2016.
 - [150] Agnieszka Kitkowska, Mark Warner, Yefim Shulman, Erik Wästlund, and Leonardo A Martucci. Enhancing privacy through the visual design of privacy notices: Exploring the interplay of curiosity, control and affect. In *Sixteenth Symposium on Usable Privacy and Security (SOUPS 2020)*, pages 437–456, 2020.
 - [151] Bennett Kleinberg, Toby Davies, and Maximilian Mozes. Textwash—automated open-source text anonymisation. *arXiv preprint arXiv:2208.13081*, 2022.
 - [152] Cory Knobel and Geoffrey C. Bowker. Values in design. *Commun. ACM*, 54(7):26–28, July 2011. ISSN 0001-0782. doi:10.1145/1965724.1965735. URL <https://doi.org/10.1145/1965724.1965735>.
 - [153] Damian P Kotevski, Robert I Smee, Matthew Field, Yvonne N Nemes, Kathryn Broadley, and Claire M Vajdic. Evaluation of an automated presidio anonymisation model for unstructured radiation oncology electronic medical records in an australian setting. *International Journal of Medical Informatics*, 168, 2022.
 - [154] Egor Kraft. Twelve nodes, 2019. URL <https://kraft.studio/twelve-nodes/>.
 - [155] Isadora Krsek, Kimi Wenzel, Sauvik Das, Jason I. Hong, and Laura Dabbish. To self-persuade or be persuaded: Examining interventions for users’ privacy setting selection. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI ’22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391573. doi:10.1145/3491102.3502009. URL <https://doi.org/10.1145/3491102.3502009>.
 - [156] Kisho Kurokawa. Nakagin capsule tower, 1972.
 - [157] Clete A Kushida, Deborah A Nichols, Rik Jadrnicek, Ric Miller, James K Walsh, and Kara Griffin. Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies. *Medical care*, 50(Suppl):S82, 2012.
 - [158] Markus Langer and Cornelius J König. Introducing and testing the creepiness of situation scale (cross). *Frontiers in psychology*, 9:2220, 2018.
 - [159] Fabiola Larios. Surveillance cutie, 2024. URL https://fabiola.io/portfolio/works/surveillance_cutie.html.

- [160] Sam Lavigne. 3 degrees of separation from the military-industrial-prison-data-surveillance state, 2015. URL <https://lav.io/projects/3-degrees-of-separation/>.
- [161] Francine Leclercq. Embroidered surveillance, 2024. URL <https://www.designboom.com/art/hand-embroidered-artworks-francine-leclercq-surveillance-cameras-footage-03-20-2024/>.
- [162] Mosh Levy, Alon Jacoby, and Yoav Goldberg. Same task, more tokens: the impact of input length on the reasoning performance of large language models, 2024. URL <https://arxiv.org/abs/2402.14848>.
- [163] Makayla Lewis and Lizzie Coles-Kemp. A tactile visual library to support user experience storytelling. *DS 81: Proceedings of NordDesign 2014, Espoo, Finland 27-29th August 2014*, pages 386–395, 2014.
- [164] Makayla M. Lewis and Lizzie Coles-Kemp. Who says personas can’t dance? the use of comic strips to design information security personas. In *Proceedings of the 2014 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 2014. ISBN 9781450324748. doi:10.1145/2559206.2581323. URL <https://doi.org/10.1145/2559206.2581323>.
- [165] Rung-Huei Liang, Wei-Ming Chung, Hsin-Liu Kao, and Tsen-Ying Lin. Intouch: Crossing social interaction with perception. In *International Conference of Design, User Experience, and Usability*, pages 306–315. Springer, 2013.
- [166] Junsu Lim, Hyeonggeun Yun, Auejin Ham, and Sunjun Kim. Mine yourself!: A role-playing privacy tutorial in virtual reality environment. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI EA ’22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391566. doi:10.1145/3491101.3519773. URL <https://doi.org/10.1145/3491101.3519773>.
- [167] Machine Listening. Always learning, 2018. URL <https://machinelisting.exposed/site-map/works/always-learning>.
- [168] LookOut. Integrated endpoint-to-cloud security. <https://www.lookout.com/>, 2020.
- [169] Melanie Lowe. You saw me?, 2008. URL <http://www.artandsurveillance.com/artists-artworks-exhibits/>.
- [170] He-Lin Luo. If this is a global surveillance center, 2019. URL https://www.digiarts.org.tw/DigiArts/DataBasePage/1_147254841059061/En.
- [171] Nic Lupfer, Bill Hamilton, Andrew Webb, Rhema Linder, Ernest Edmonds, and Andruid Kerne. The art.chi gallery: An embodied iterative curation experience. In *Proceedings of the 2015 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 2015. ISBN 9781450331463. doi:10.1145/2702613.2725457. URL <https://doi.org/10.1145/2702613.2725457>.

- [172] B. Macwhinney. The CHILDES project: Tools for analyzing talk. note: Grant support for CHILDES – NICHD HD082736. Lawrence Erlbaum Associates, Mahwah, NJ, 2000.
- [173] B. Macwhinney. A transcript-video database for collaborative commentary in the learning sciences. In R. Goldman, R. Pea, B. Barron, and S. Derry, editors, *Video research in the learning sciences*, pages 537–546. Lawrence Erlbaum Associates, Mahwah, NJ, 2007.
- [174] B. Macwhinney and J. Wagner. Transcribing, searching and data sharing: The CLAN software and the TalkBank data repository. *Gesprachsforschung*, 11:154–173, 2010.
- [175] Kristine Madjare and rixcriga. [rixc gallery] trihars exhibition “ram” opening, 2016. URL <https://www.flickr.com/photos/rixcriga/albums/72157669540894514/with/29083597595>.
- [176] Jill Magid. System azure security ornamentation, 2002. URL <https://www.jillmagid.com/projects/system-azure-security-ornamentation>.
- [177] Emanuel Maiberg. ‘fucklapd.com’ lets anyone use facial recognition to instantly identify cops, 2025. URL <https://www.404media.co/fucklapd-com-lets-anyone-use-facial-recognition-to-instantly-identify-cops/>.
- [178] Naresh K. Malhotra, Sung S. Kim, and James Agarwal. Internet users’ information privacy concerns (iuipc): The construct, the scale, and a causal model. *Info. Sys. Research*, 15(4):336–355, December 2004. ISSN 1526-5536. doi:10.1287/isre.1040.0032. URL <https://doi.org/10.1287/isre.1040.0032>.
- [179] Steve Mann. Existech corp., 2001. URL <http://wearcam.org/domewear/>.
- [180] Ivan Manokha. Surveillance, panopticism, and self-discipline in the digital age. *Surveillance and Society*, 16(2), 2018.
- [181] Ggaliwango Marvin, Nakayiza Hellen, Daudi Jjingo, and Joyce Nakatumba-Nabende. Prompt engineering in large language models. In I. Jeena Jacob, Selwyn Piramuthu, and Przemyslaw Falkowski-Gilski, editors, *Data Intelligence and Cognitive Informatics*, pages 387–402, Singapore, 2024. Springer Nature Singapore. ISBN 978-981-99-7962-2.
- [182] Eva Mattes and Franco Mattes. My little big data, 2019. URL <https://0100101110101101.org/my-little-big-data/>.
- [183] Lauren Lee McCarthy. Mwitm, 2017. URL <https://get-lauren.net/MWITM-Man-Woman-In-The-Middle>.
- [184] Aleecia M. McDonald and Lorrie Faith Cranor. The cost of reading privacy policies. *I/S: A Journal of Law and Policy for the Information Society*, 4:543, 2008.

- [185] Aleecia M. McDonald, Robert W. Reeder, Patrick Gage Kelley, and Lorrie Faith Cranor. A comparative study of online privacy policies and formats. In *Proceedings of the International Symposium on Privacy Enhancing Technologies Symposium*, 2009.
- [186] Kyle McDonald. Icespy, 2018, 2025. URL <https://icespy.org/>.
- [187] Susan E. McGregor, Polina Charters, Tobin Holliday, and Franziska Roesner. Investigating the computer security practices and needs of journalists. In *24th USENIX Security Symposium (USENIX Security 15)*, pages 399–414, 2015.
- [188] Shaun McNiff. *Art-based research*. Jessica Kingsley Publishers, 1998.
- [189] Vice Media. Surveillance art. URL <https://www.vice.com/en/tag/surveillance-art/>.
- [190] Jessica Megarry, Peta Mitchell, Markus Rittenbruch, Yu Kao, Bryce Christensen, and Marcus Foth. Probing for privacy: A digital design method to support reflection of situated geoprivacy and trust. *Digital Society*, 2(3):55, 2023.
- [191] Vikram Mehta, Daniel Gooch, Arosha Bandara, Blaine Price, and Bashar Nuseibeh. Privacy care: A tangible interaction framework for privacy management. *ACM Trans. Internet Technol.*, 21(1), February 2021. ISSN 1533-5399. doi:10.1145/3430506. URL <https://doi.org/10.1145/3430506>.
- [192] Meta. Introducing llama 3.1: Our most capable models to date. <https://ai.meta.com/blog/meta-llama-3-1/>, July 23, 2024.
- [193] Meta. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation. <https://ai.meta.com/blog/llama-4-multimodal-intelligence/>, 2025.
- [194] Stephane M Meystre, F Jeffrey Friedlin, Brett R South, Shuying Shen, and Matthew H Samore. Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC medical research methodology*, 10(1):70, 2010.
- [195] Microsoft. Presidio: Data Protection and De-identification SDK. <https://microsoft.github.io/presidio/>, 2024.
- [196] Meisam Mohammady, Lingyu Wang, Yuan Hong, Habib Louafi, Makan Pourzandi, and Mourad Debbabi. Preserving both privacy and utility in network trace anonymization. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pages 459–474, 2018.
- [197] Joana Moll. Carbolytics, 2022. URL <https://www.janavirgin.com/carbolytics.html>.

- [198] Torin Monahan. The right to hide? anti-surveillance camouflage and the aestheticization of resistance. *Communication and Critical/Cultural Studies*, 12(2):159–178, 2015. doi:10.1080/14791420.2015.1006646.
- [199] Torin Monahan. Ways of being seen: surveillance art and the interpellation of viewing subjects. *Cultural Studies*, 32(4):560–581, 2018.
- [200] MSCHF. Eavesdropper 1, 2022. URL <https://leaflet.perrotin.com/view/338/no-more-tears-im-lovin-it>.
- [201] Maria Murray, Nadia Pantidi, and John McCarthy. Agency, power and confrontation: The role for socially engaged art in cscw with rural communities in support of inclusion. *Computer Supported Cooperative Work*, 33(3):435–472, 2023. ISSN 0925-9724. doi:10.1007/s10606-023-09482-7. URL <https://doi.org/10.1007/s10606-023-09482-7>.
- [202] Museum of Modern Art. MoMA Archives, Library, and Research Collections. <https://research.moma.org/home>. Accessed: 2025-08-27.
- [203] mwvisser. Peepshow - art & electronic media online companion. URL <https://artelectronicmedia.com/en/artwork/peepshow/>.
- [204] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pages 111–125. IEEE, 2008.
- [205] Bruce Nauman. Live-taped video corridor, 1967, 1968. URL <https://www.guggenheim.org/artwork/3153>.
- [206] Lene Nielsen. From user to character: an investigation into user-descriptions in scenarios. In *Proceedings of the 4th Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques*, DIS '02, page 99–104, New York, NY, USA, 2002. Association for Computing Machinery. ISBN 1581135157. doi:10.1145/778712.778729. URL <https://doi.org/10.1145/778712.778729>.
- [207] Maggie Oates, Yama Ahmadullah, Abigail Marsh, Chelse Swoopes, Shikun Zhang, Rebecca Balebako, and Lorrie Faith Cranor. Turtles, locks, and bathrooms: Understanding mental models of privacy through illustration. *Proceedings on Privacy Enhancing Technologies*, 2018.
- [208] Julian Oliver. Stealth cell tower, 2015. URL <https://julianoliver.com/projects/stealth-cell-tower/>.
- [209] Yuya Jeremy Ong, Mu Qiao, Ramani Routray, and Roger Raphael. Context-aware data loss prevention for cloud storage services. In *2017 IEEE 10th International Conference on Cloud Computing (CLOUD)*, pages 399–406. IEEE, 2017.
- [210] OpenAI. Hello gpt-4o. <https://openai.com/index/hello-gpt-4o/>, 2025.

- [211] OpenAI. GPT-5.1: A smarter, more conversational ChatGPT. <https://openai.com/index/gpt-5-1/>, 2025.
- [212] Opera on Tap. Looking at You: A New Techno-Noir Opera. <https://operaontap.org/looking-at-you/>. Accessed: 2025-08-27.
- [213] George Orwell. *1984 (Updated ed.)*. Harold Bloom, 2007.
- [214] Trevor Paglen. Black sites, 2006. URL <https://paglen.studio/2020/04/24/black-sites/>.
- [215] Trevor Paglen. Image operations, 2018. URL <https://paglen.studio/2020/04/23/image-operations/>.
- [216] Nam June Paik. Good morning mr. orwell, 1984. URL <https://www.eai.org/titles/good-morning-mr-orwell>.
- [217] Yuri Pattison. Yuri pattison at kunst halle sankt gallen, 2017. URL <https://artviewer.org/yuri-pattison-at-kunst-halle-sankt-gallen/>.
- [218] Rich Pell and The Institute for Applied Autonomy. isee - the institute for applied autonomy (2004), 2009. URL <https://vimeo.com/6163268>.
- [219] Chanda Phelan, Cliff Lampe, and Paul Resnick. It’s creepy, but it doesn’t bother me. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI ’16, page 5240–5251, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450333627. doi:10.1145/2858036.2858381. URL <https://doi.org/10.1145/2858036.2858381>.
- [220] James Pierce. On the presentation and production of design research artifacts in hci. In *Proceedings of the 2014 Designing Interactive Systems Conference*, 2014. ISBN 9781450329026. doi:10.1145/2598510.2598525. URL <https://doi.org/10.1145/2598510.2598525>.
- [221] James Pierce. Smart home security cameras and shifting lines of creepiness: A designed inquiry. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI ’19, page 1–14, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359702. doi:10.1145/3290605.3300275. URL <https://doi.org/10.1145/3290605.3300275>.
- [222] Ildikó Pilán, Benet Manzanares-Salor, David Sánchez, and Pierre Lison. Truthful text sanitization guided by inference attacks. *arXiv:2412.12928*, 2024.
- [223] Ethan Plaut, Kiri West, Fabio Morreale, Maya Gibson, Grace Thompson, Kara Woodward, and Danielle Lottridge. Surveillance on exhibit: Using problematic technology to teach about problematic technology. In *Proceedings of the 2025 CHI Conference on*

- Human Factors in Computing Systems*, CHI '25, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400713941. doi:10.1145/3706598.3713710. URL <https://doi.org/10.1145/3706598.3713710>.
- [224] Surveillance Camera Players. The mass psychology of fascism, 1999, 2000. URL <https://notbored.org/reich.html>.
 - [225] Proton. Lumo: Privacy-first AI assistant, aug 2025. URL <https://lumo.proton.me/>.
 - [226] Argenis Ramirez Gomez and Katarzyna Stawarz. Exploring creepy futures: Reflecting on the value of creepiness as design fiction. 2025.
 - [227] Argenis Ramirez Gomez, Carolina Fuentes Toro, Samuelson Atiba, Nervo Verdezoto Dias, and Katarzyna Stawarz. Understanding# creepytech: Exploring the context of creepiness of emerging technology. 2025.
 - [228] Redactable, Inc. Redactable. <https://www.redactable.com/>, 2026.
 - [229] Daniel Reinhardt, Johannes Borchard, and Jörn Hurtienne. Visual interactive privacy policy: The better choice? In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi:10.1145/3411764.3445465. URL <https://doi.org/10.1145/3411764.3445465>.
 - [230] Nathan Reitering, Bruce Wen, Michelle L Mazurek, and Blase Ur. What does it mean to be creepy? responses to visualizations of personal browsing activity, online tracking, and targeted ads. *Proceedings on Privacy Enhancing Technologies*, 2024(3). doi:10.56553/popets-2024-0101. URL <https://par.nsf.gov/biblio/10578101>.
 - [231] Peters Riekstins. Back to the light, 2016. URL <https://we-make-money-not-art.com/back-to-the-light/>.
 - [232] Luc Rocher. *Anonymization in the 21st Century: A Critical Examination of De-identification for Modern Large-Scale Data*. PhD thesis, UCL-Université Catholique de Louvain, 2019.
 - [233] Daniel Romm, Hongyu Zhang, Priyanka Verma, Grant McKenzie, and Emily Chen. "data horror": Mapping (spatial) data privacy violations onto a cognitive account of horror. 2021.
 - [234] Eidan J Rosado. Pii-codex: a python library for pii detection, categorization, and severity assessment. *The Journal of Open Source Software*, 8(86):5402, 2023.
 - [235] Johnny Saldaña. *The Coding Manual for Qualitative Researchers*. SAGE publications Ltd, Los Angeles, CA, 2021.

- [236] David Sánchez and Montserrat Batet. C-sanitized: A privacy model for document redaction and sanitization. *Journal of the Association for Information Science and Technology*, 67(1):148–163, 2016.
- [237] Anup Sathya and Ken Nakagaki. Attention receipts: Utilizing the materiality of receipts to improve screen-time reflection on youtube. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2024.
- [238] Florian Schaub, Rebecca Balebako, Adam L Durity, and Lorrie Faith Cranor. A design space for effective privacy notices. In *Eleventh symposium on usable privacy and security (SOUPS 2015)*, pages 1–17, 2015.
- [239] Florian Schaub, Aditya Marella, Pranshu Kalvani, Blase Ur, Chao Pan, Emily Forney, and Lorrie Faith Cranor. Watching them watching me: Browser extensions’ impact on user privacy awareness and concern. In *NDSS workshop on usable security*, volume 10, 2016.
- [240] Julia Scher. Papa bed, 2003. URL <https://www.estherschipper.com/artists/51-julia-scher/works/1292-julia-scher-papa-bed-2003/>.
- [241] Björn Schülke. Drone #4, 2005. URL <http://www.schuelke.org/drone-4.html>.
- [242] John S. Seberger, Irina Shklovski, Emily Swiatek, and Sameer Patil. Still creepy after all these years:the normalization of affective discomfort in app use. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI ’22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391573. doi:10.1145/3491102.3502112. URL <https://doi.org/10.1145/3491102.3502112>.
- [243] John S. Seberger, Hyesun Choung, Jaime Snyder, and Prabu David. Better living through creepy technology? exploring tensions between a novel class of well-being apps and affective discomfort in app culture. *Proc. ACM Hum.-Comput. Interact.*, 8 (CSCW1), April 2024. doi:10.1145/3637299. URL <https://doi.org/10.1145/3637299>.
- [244] Phoebe Sengers, Kirsten Boehner, Shay David, and Joseph ’Jofish’ Kaye. Reflective design. In *Proceedings of the 4th Decennial Conference on Critical Computing: Between Sense and Sensibility*, CC ’05, page 49–58, New York, NY, USA, 2005. Association for Computing Machinery. ISBN 1595932038. doi:10.1145/1094562.1094569. URL <https://doi.org/10.1145/1094562.1094569>.
- [245] Vandit Sharma and Mainack Mondal. Understanding and improving usability of data dashboards for simplified privacy control of voice assistant data. In *Proceedings of the 31st USENIX Security Symposium*, 2022.
- [246] Xiaoxin Shen, Eman Alashwali, and Lorrie Faith Cranor. What do privacy advertisements communicate to consumers? *arXiv:2405.13857*, 2024.

- [247] Mark Shepard. Minor urbanism, 2012. URL <https://www.andinc.org/site/minor-urbanism/>.
- [248] Abhishek Shivanna. A guide to integrating large language models in your organizations. Forbes, November 5, 2024. URL <https://www.forbes.com/councils/forbestechcouncil/2024/11/05/a-guide-to-integrating-large-language-models-in-your-organizations/>.
- [249] Irina Shklovski and Erik Grönvall. Creepyleaks: Participatory speculation through demos. In *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*, NordiCHI '20, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450375795. doi:10.1145/3419249.3420168. URL <https://doi.org/10.1145/3419249.3420168>.
- [250] Ben Shneiderman. *Designing the user interface: strategies for effective human-computer interaction*. Pearson Education India, 2010.
- [251] Ann-Sofi Sidén. Sticky floors, 2014. URL <https://www.annsosisiden.com/index.html#detail-sticky>.
- [252] SIGCHI. Sigchi policies. Version 4.4.25, 2025. URL <https://programs.sigchi.org/policies#sigchi-acm-privacy-policy>.
- [253] Michael Skirpan, Maggie Oates, Daragh Byrne, Robert Cunningham, and Lorrie Faith Cranor. Is a Privacy Crisis Experienced, a Privacy Crisis Avoided? <https://archive.aec.at/prix/>. Accessed: 2025-08-27.
- [254] Michael Warren Skirpan, Jacqueline Cameron, and Tom Yeh. More than a show: Using personalized immersive theater to educate and engage the public in technology ethics. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–13, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356206. doi:10.1145/3173574.3174038. URL <https://doi.org/10.1145/3173574.3174038>.
- [255] Garrett Smith, Sarah Carson, Rhea G. Vengurlekar, Stephanie Morales, Yun-Chieh Tsai, Rachel George, Josh Bedwell, Trevor Jones, Mainack Mondal, Brian Smith, Norman Makoto Su, Bart Knijnenburg, and Xinru Page. "i know i'm being observed:" video interventions to educate users about targeted advertising on facebook. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703300. doi:10.1145/3613904.3642885. URL <https://doi.org/10.1145/3613904.3642885>.
- [256] Stephanie A. Smith and Steven R. Brunner. To reveal or conceal: Using communication privacy management theory to understand disclosures in the workplace. *Management Communication Quarterly*, 31(3):429–446, 2017.

- [257] Smithsonian Libraries and Archives. Art and Artist Files. <https://library.si.edu/art-and-artist-files>. Accessed: 2025-08-27.
- [258] SOKT Technologies Private Limited. Ai-redact. <https://ai-redact.com/>, 2026.
- [259] Spirion. Building a data loss prevention (dlp) solution. <https://www.spirion.com/blog/building-dlp-solution/>, 2021.
- [260] Stack Exchange. Stack Overflow Meta, 2026. URL <https://meta.stackoverflow.com/questions/266557/why-was-stack-overflow-chosen-as-a-name-for-this-site>. Accessed: 2026-04-21.
- [261] Stanza. The nemesis machine: Manifestation, 2023. URL https://www.stanza.co.uk/Nemesis_Manifestation/index.html.
- [262] Luke Stark and Kate Crawford. The work of art in the age of artificial intelligence: What artists can teach us about the ethics of data practice. *Surveillance & Society*, 17(3/4):442–455, 2019.
- [263] Dan Stavy. Selfie stick, 2017. URL <https://stavdan.com/selfie-stick/>.
- [264] Katarzyna Stawarz, Alison Burrows, and Argenis Ramirez Gomez. Leveraging creepiness to facilitate ethical design: Lessons learned from a design workshop. In *38th International BCS Human-Computer Interaction Conference*, pages 432–440. BCS Learning & Development, 2025.
- [265] Hito Steyerl. How not to be seen: A fucking didactic educational .mov file, 2013. URL https://www.moma.org/collection/works/181784?artist_id=43752&page=1&so_v_referrer=artist.
- [266] Arne Svenson. The neighbors, 2023. URL <https://www.danzigergallery.com/exhibitions/arne-svenson-1>.
- [267] Latanya Sweeney. Replacing personally-identifying information in medical records, the Scrub system. In *Proceedings of the AMIA annual fall symposium*, page 333. American Medical Informatics Association, 1996.
- [268] Madiha Tabassum, Abdulmajeed Alqhatani, Marran Aldossari, and Heather Richter Lipford. Increasing user attention with a comic-based policy. In *Proceedings of the 2018 chi conference on human factors in computing systems*, pages 1–6, 2018.
- [269] Shaheer Tarar. Landsat.earth, 2023. URL <https://shaheer.info/>.
- [270] Qwen Team. Qwen3: Think Deeper, Act Faster. <https://qwenlm.github.io/blog/qwen3/>, 2025.

- [271] Omer Tene and Jules Polonetsky. A theory of creepy: technology, privacy and shifting social norms. *Yale JL & Tech.*, 16:59, 2013.
- [272] Marc Teyssier, Marion Koelle, Paul Strohmeier, Bruno Fruchard, and Jürgen Steimle. Eyecam: Revealing relations between humans and sensing devices through an anthropomorphic webcam. In *Proceedings of the 2021 SIGCHI Conference on Human Factors in Computing Systems*, 2021. ISBN 9781450380966. doi:10.1145/3411764.3445491. URL <https://doi.org/10.1145/3411764.3445491>.
- [273] The University of Chicago. University Privacy Policy, 2026. URL <https://intranet.uchicago.edu/policies/information-technology-policies/university-privacy-policy>. Accessed: 2026-04-21.
- [274] Lucy Thompson. Vermeer’s curtain: Privacy, slut-shaming and surveillance in ‘a girl reading a letter’. *Surveillance and Society*, 15(2):326, 2017.
- [275] Nye Thompson. Uninvited, 2021. URL <https://nyethompson.net/works/UNINVITED.html>.
- [276] Timo Toots. Reality tv, 2017, 2023. URL <https://www.timo.ee/reality-tv/>.
- [277] Suzanne Treister. Camouflage, 2013. URL <https://www.suzannetreister.net/Camouflage/Camouflage.html>.
- [278] Brygg Ullmer and Hiroshi Ishii. The metadesk: models and prototypes for tangible user interfaces. In *Proceedings of the 10th annual ACM symposium on User interface software and technology*, pages 223–232, 1997.
- [279] Ben Van den Berghe. Backspace, 2022. URL <https://benvandenbergh.com/>.
- [280] Danja Vasiliev. Netless2, 2012. URL <https://k0a1a.net/danja/netless2>.
- [281] Danja Vasiliev. Netless, 2012. URL <https://k0a1a.net/netless/netless-abstract.pdf>.
- [282] Johannes Vermeer. A girl reading a letter by an open window, c. 1657—1659.
- [283] Ian Wagner. isee, 2016. URL <https://www.digiart21.org/art/isee>.
- [284] Giles Walker. Kinetic sculpture, 2022. URL <https://www.gileswalker.org/kinetic-sculpture>.
- [285] Zhixuan Wang. Watch me watching: Surveillance art and the politics of observation. *Aspectus*, (3), 2021.
- [286] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*, 2023.

- [287] Wieden+Kennedy. Giles walker brings peepshow to w+k london, 2019.
- [288] Lauren Wissot. “do not be daunted by the magnitude of the challenge in front of you”: Assia boundaoui on her surveillance doc, the feeling of being watched, 2018. URL <https://filmmakermagazine.com/105231-do-not-be-daunted-by-the-magnitude-of-the-challenge-in-front-of-you-assia-boudaoui-on-her-surveillance-doc-the-feeling-of-being-watched/>.
- [289] Richmond Y Wong, Deirdre K Mulligan, Ellen Van Wyk, James Pierce, and John Chuang. Eliciting values reflections by engaging privacy futures using design workbooks. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):1–26, 2017.
- [290] Richmond Y. Wong, Ellen Van Wyk, and James Pierce. Real-fictional entanglements: Using science fiction and design fiction to interrogate sensing technologies. In *Proceedings of the 2017 Designing Interactive Systems Conference*, 2017. ISBN 9781450349222. doi:10.1145/3064663.3064682. URL <https://doi.org/10.1145/3064663.3064682>.
- [291] Paweł W. Woźniak, Jakob Karolus, Florian Lang, Caroline Eckerth, Johannes Schöning, Yvonne Rogers, and Jasmin Niess. Creepy technology: what is it and how do you measure it? In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI ’21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi:10.1145/3411764.3445299. URL <https://doi.org/10.1145/3411764.3445299>.
- [292] WTTDOTM. Traffic cam photobooth, 2024. URL <https://trafficcamphotobooth.com/>.
- [293] Jens Wunderling. Audience, 2010. URL <http://jenswunderling.com/works/audience/>.
- [294] Elad Yacobson, Orly Fuhrman, Sara HersHKovitz, and Giora Alexandron. De-identification is insufficient to protect student privacy, or—what can a field trip reveal? *Journal of Learning Analytics*, 8(2):83–92, 2021.
- [295] Susan Fillin Yeh. Mary cassatt’s images of women. *Art Journal*, 35:359–363, 1976. doi:10.2307/776228.
- [296] Leah Zhang-Kennedy and Sonia Chiasson. A systematic review of multimedia tools for cybersecurity awareness and education. *ACM Comput. Surv.*, 54(1), January 2021. ISSN 0360-0300. doi:10.1145/3427920. URL <https://doi.org/10.1145/3427920>.
- [297] John Zimmerman, Jodi Forlizzi, and Shelley Evenson. Research through design as a method for interaction design research in hci. In *Proceedings of the 2007 SIGCHI Conference on Human Factors in Computing Systems*, 2007. ISBN 9781595935939. doi:10.1145/1240624.1240704. URL <https://doi.org/10.1145/1240624.1240704>.

- [298] Shoshana Zuboff. The age of surveillance capitalism. In *Social theory re-wired*, pages 203–213. Routledge, 2023.
- [299] Anna Ådahl. The exhibited, 2013. URL <https://www.annaadahl.com/works/the-exhibited-2013>.

APPENDIX A

IDENTIFIABILITY INTERVIEW GUIDE

As described in our recruitment email, we are researchers designing software that protects participant privacy in unstructured, text transcript data by automatically redacting identifiable information. By “identifiable information,” we mean any parts of the data that someone outside of the research staff might use to infer the participant’s identity. We define redaction to be the removal or substitution of such identifiable information so that data could be shared publicly or semi-publicly. To this end, we are studying how other researchers think about redaction and about what kinds of data make participants identifiable.

[Moderator re-shares the consent form—previously emailed—and asks for verbal consent.]

I will start the audio recording now. Thank you for consenting to the audio recording. If there are any questions you do not wish to answer or do not know the answer to, please feel free to say “I don’t know.”

1. Could you quickly state for us your institutional and departmental affiliation, as well as research specialty?

In this section, we would like to dig deeper into how you think about the tradeoffs between redacting or not redacting identifiable information. As a reminder, we define “redaction” to be any removal or substitution of data that could be used to identify participants. I will give you an example about the kinds of tradeoffs we’re talking about. We work with the Language Development Project dataset, which is longitudinal collection of text transcripts recorded from children speaking with their caregivers at home. The goal of the research was to study these children’s language development. But, something in the transcript that pertains to this research goal might also identify the children, so there’s a decision we need to make about whether to redact this. As a hypothetical example, if the child refers to its imaginary friend as “parallelepiped” or some very advanced word, people who know this family might be able to recognize them from this, but it’s also interesting in the context of the research project.

We’re interested in learning about your perspective regarding similar situations. While we are focusing on the project you shared with us, also feel free to mention relevant information about any other projects.

2. Who would have access to these text transcripts?

3. In your field, what is the general or standard practice for redacting text transcripts?
4. In your opinion, how well do these practices protect participants?
5. Have there been any well-known incidents in your field related to participant privacy?
6. What do you think makes a piece of information identifiable?
7. Is there a threshold of specificity that separates identifiable information from non-identifiable information?

Imagine you publish your dataset without any redactions publicly on a website anyone can visit. To what extent would you be worried about the following outcomes on the following scale? *1–not at all concerned; 2–slightly concerned; 3–somewhat concerned; 4–moderately concerned; 5–extremely concerned*

8. Someone who does **not personally know** study participants looking through the data and being able to identify specific participants? Why?
9. Someone who **personally knows** a study participant (and knows they participated in the study) looking through the data and being able to identify them? Why?
10. Someone who personally **participated** in the study being able to identify their own transcripts? Why?
11. Is there anyone else you can think of who may be able to identify your participants based on unredacted transcript data?

Putting aside the issue of identifiability for one second. Are there any categories of information you can think of that are not necessarily identifiable, but can *cause harm* to the participants themselves? In other words, categories that may still be advisable to redact, even though they are not necessarily identifiable?

Now, given our discussion of identifiability and harm, we will send you a link containing a taxonomy of potential information to redact that we are iteratively co-creating with participants in this study. As you read over the taxonomy, please narrate aloud your thoughts about each point in the taxonomy. Some of your thoughts we're interested in hearing about are:

12. Has the item ever shown up in your corpus?
13. Have you thought about redacting it? Does it depend on the context?
14. Beyond identifiability, do you think this is a category that may cause harm to the participant by either redacting or leaving it in the dataset? Please highlight them in light orange if you could.
15. Are there any additional categories or examples you have come across that we did not include?

Next, we'd like you to think about data collected by others that you may have worked with, or at least heard about.

16. While looking at data collected by other researchers, have you encountered information that could have made the participant identifiable?

17. What kinds of information?

18. How did you respond? Did you, for example, omit the participant's data or stop analyzing the dataset altogether? How did you decide whether or not to continue using this data?

19. From your discussions with others in your field, regardless of whether you've looked at their data, can you think of information they may collect that feels potentially identifiable?

20. Have you ever encountered text transcripts from other research groups that seemed too limited in information due to redactions (in other words "over-redacted")?

21. How did you respond?

In this section, we want to learn more about your experience redacting and sharing data. This helps us contextualize how our software may be helpful to you.

22. For this project, would it have been feasible to redact those text transcripts **by hand** based on our taxonomy? What about based on your redaction practices (if you have one)?

23. (If yes) How was the redaction process organized? Who did the redaction? What was the resource cost in terms of time and money?

24. (If no) Why not?

25. What challenges have you encountered when redacting text transcripts?

26. Has another researcher or research group ever requested access to your data?

27. (If yes) How did you handle this? How often does this occur? How often do you accept/reject these requests? Why?

28. (If no) How would you handle this?

29. Do you ever share the **raw** text transcripts that you collect with anyone outside of those listed on your IRB? For example, researchers outside of your lab or collaborations, the scientific community, or the general public?

If so, answer the following questions in terms of what you have done. If not, answer hypothetically.

30. Do you share all of your data, or only part?
31. When you shared these transcripts, were you concerned about someone identifying your participants?
32. Did you encounter any logistical obstacles in sharing these data, given that they were raw text transcripts?
33. What are your other key considerations related to data sharing?
34. Through what platforms or repositories do you share data?
35. Is there a data use agreement or other restrictions on use?

We have completed the interview! Do you have any final thoughts or questions?

APPENDIX B

GENERAL PROMPT PROVIDED TO LLMS

You are a linguistics researcher, removing potentially identifiable information (PotII) from transcripts.

Please flag all instances of potentially identifiable information in each sentence in the following text (enclosed by <t></t>). Flagged PotII should be replaced with a mask according to its subcategory, like [Personal_Name]. For example, in the utterance “I wish Anna would come pick up her packages”, “Anna” should be replaced with [Personal_Name]. Your response should be a single JSON object and should include the original_sentence, the modified_sentence, a list of replaced_items, and a list of PotII_categories. Any additional text should only go under an additional notes key.

APPENDIX C

FRAMEWORK PROMPT PROVIDED TO LLMS

You are a linguistics researcher, removing potentially identifiable information (PotII) from transcripts.

We format the PotII taxonomy of potentially identifiable information as follows:

```
[
Overarching category 1
  » description
> subcategory a
  » description
> subcategory b
  » description
]
```

POTII TAXONOMY

```
[
Direct Identifiers
  » personal information that can identify an individual
> Personal Names
  » first names, last names, or similar
> Nicknames
  » a familiar or silly moniker between people
> Numerical Identifiers
  » traditional PII such as phone numbers, emails, SSN, etc.
```

Demographics

```
  » terms describing characteristics of people
> Age
  » a person's age, either in relation to a specific date or isolated
```

- > Sexual Orientation
 - » words relating to the romantic or sexual orientation of a person
- > Gender
 - » words relating to the gender identity or sex of a person
- > Race or Ethnicity
 - » social, cultural, or biological trait group a person belongs to
- > Education
 - » reference to educational experience or degree, institution, or specific courses
- > Profession
 - » title of a person's job or occupation
- > Health Condition
 - » references to a personal disease, disorder, or other health-related condition, whether physical or mental

Locations

- » physical locations or regions
- > Address
 - » residential or occupational address
- > Landmark
 - » notable or recognizable feature of some place, whether formal (mapped) or informal (local reference)
- > Business or Organization
 - » franchises, chains, other businesses, or formally established groups

Dates

- » date associated with publicly-known or personal events

Events

- » holidays, social occasions, and other common noteworthy happenings
- > Specific Events
 - » event that is niche, experienced by very few people, or made-up

Linguistic Patterns

- » unusual linguistic habits like familects, portmanteaus, malapropisms, or unlikely n-grams and similar
- > Code-switching
 - » alternating between two or more languages or dialects within a segment of text
- > Slang
 - » informal language often used by specific groups
- > Regionalisms
 - » terms common to a specific region or local group
- > Familects
 - » terms or phrases used as in-jokes or similar amongst family members

Personal Interests or Activities

- » pursuits and interests outside of work and school
- > Hobbies
 - » leisure activities and topics or subjects that people enjoy
- > Traditions
 - » customs or practices that can be cultural, familial, or between people who know each other
- > Cultural References
 - » activities or practices specifically related to a culture; pop culture references, fictional characters, religious figures, or other non-material beings
- > Personal Identity
 - » non-inherent identities associated with some sufficiently large in-group

]

Please flag all instances of potentially identifiable information in each sentence in the following text (enclosed by <t></t>), according to the above PotII taxonomy. Flagged PotII should be replaced with a mask according to its subcategory, like [Personal_Name], [Health_Condition], or [Profession]. For example, in the utterances “I wish Anna would come pick up her packages”, “he just came out as demisexual”, and “I think I have symptoms of type 1 diabetes”, “Anna”, “demisexual”, and “type 1 diabetes” should

be replaced with [Personal_Name], [Orientation], and [Health_Condition] respectively. Your response should be a single JSON object and should include the original_sentence, the modified_sentence, a list of replaced_items, and a list of PotII_categories. Any additional text should only go under an additional notes key.

APPENDIX D

PROMPT ENGINEERING STUDY

Many prompt guidelines emphasize chat experience, language capabilities (e.g., tone, writing assistance), or utility (e.g., code). Because our study is run with a “one-and-done” use of the LLM, we do not consider strategies that improve subsequent responses.

We use the framework prompt (Appendix C) as a base for each prompt style, then modify it based on the specified alternative approach. In particular, we modify the prompt in the following ways, with the names corresponding to the filenames for prompts in our OSF repository¹. Items are all preceded by `prompt_FRAMEWORK`.

- `*_examples`: adding specific examples to each framework category to be redacted
- `*_counterexamples`: adding specific examples of items **not** to be redacted
- `*_verification`: specifically asking the LLM to verify that no identifiable information remains in the redacted text
- `*_role{num}`: modifying or specifying a persona or role for the LLM to emulate

For the role prompts, we modified the first line in each prompt to give the LLM context for its role:

- `*_role1`: “You are a redaction specialist, removing potentially identifiable information (PotII) from transcripts.”
- `*_role2`: “You are a social-behavioral sciences researcher, removing potentially identifiable information (PotII) from your transcripts to protect your participants’ privacy.”
- `*_role3`: “You are a top-of-the-line redaction tool, removing potentially identifiable information (PotII) from unstructured conversational data.”

1. https://osf.io/29ejq/overview?view_only=e0119096f9ca4dd0b885ed1b6581e5e7

The results of our prompt engineering study can be found in Table D.1. For these results, “partially correct” refers to either slightly alternative redactions that still obscure potential identifiers (e.g., “Mrs. Miller,” as opposed to our annotation of “Miller,” is replaced) or redactions in an utterance that were partially correct (e.g., “music theory” and “astrophysics” as opposed to “masters,” “music theory,” “PhD,” and “astrophysics”). Redactions were deemed “missed” if items were left completely or partially unobscured (e.g., missing a full phrase, potentially identifiable information left in the utterance).

Overall, the prompts did not have a significantly different impact on correctness, with one exception. The counterexample prompt performed poorest. With Claude, this prompt had near triple the amount of missed redactions compared to prompts that missed the least (two role-specific prompts: the redaction specialist and the redaction tool), differing by an average of about 4 percentage points. GPT similarly had more errors occur with this prompt (also differing by about 4 percentage points, with nearly twice as many missed redactions).

Counter to our hypotheses, the “examples” prompt did not perform as well as the “role-playing” prompts or the “verification” prompt. There are several potential reasons for this. We speculate that this could induce an oversaturation of information, where key information is lost in the prompt, or potentially overfitting, where the LLM prefers utterances that are most similar to the examples. LLMs are prone to degradation in performance with larger inputs [162], which appears to align with our results for this prompt. The “verification” prompt did not markedly impact numerical results, though this strategy did increase justification and “note” outputs, especially with Claude. For researchers interested in more context around flagged redactions, this method has potential. The “roleplaying” prompts and verification had similar performance for both models. These are perhaps the closest to the original framework prompt, since there is only one additional line at most. No role significantly outperformed the others, though the role specifically referencing privacy had a slight advantage in just the GPT results.

Table D.1: Results for the prompt engineering study. Roles are redaction specialist (**RS**), social-behavioral sciences researcher (**SBS**), and a top-of-the-line tool (**RT**). The Claude Sonnet dataset contained 1,431 items (1,148 redactions) and the GPT dataset 2,160 items (1,700 redactions).

Prompt	Correct Redactions		Partial Redactions		Missed Redactions		Excess Redactions
	#	%	#	%	#	%	#
Anthropic Claude Sonnet 3.5							
Framework	1,136	79.4%	185	12.9%	110	7.7%	426
Specific Examples	1,098	76.7%	194	13.6%	139	9.7%	371
Counterexamples	882	61.6%	196	13.7%	353	24.7%	282
Verification	1,117	78.1%	190	13.3%	124	8.7%	445
Role A: RS	1,123	78.5%	194	13.6%	114	8.0%	421
Role B: SBS	1,109	77.5%	195	13.6%	127	8.9%	421
Role C: RT	1,125	78.6%	192	13.4%	114	8.0%	456
Role B + Verification	1,126	78.7%	193	13.6%	112	7.8%	440
OpenAI GPT-4o							
Framework	1,376	63.7%	455	21.1%	329	15.2%	240
Specific Examples	1,406	65.1%	401	18.6%	353	16.3%	223
Counterexamples	1,122	51.9%	473	21.9%	565	26.2%	154
Verification	1,411	65.3%	429	19.9%	320	14.8%	251
Role A: RS	1,381	63.9%	441	20.4%	338	15.6%	273
Role B: SBS	1,412	65.4%	438	20.3%	310	14.4%	262
Role C: RT	1,332	61.7%	459	21.3%	369	17.1%	250
Role B + Verification	1,415	65.5%	428	19.8%	317	14.7%	239

APPENDIX E

ADDITIONAL TABLES

Table E.1: Category breakdown results for missed redactions in the **Claude Sonnet** dataset used in the prompt engineering study. The dataset contained 1,431 items (1,148 redactions).

Framework category	Items Redacted	Original	Framework	Specific Examples	Counterexamples	Verification	Role A	Role B	Role C	Verification + Role B
Personal Names	550	35	5	6	5	8	4	4	4	5
Nicknames	104	12	3	6	15	4	4	6	4	1
Demographics	102	14	7	9	39	7	7	8	7	8
Age	3				1					
Race / Ethnicity	26	1	1	1	6	1				1
Education	7									
Profession	10	6	6	6	6	6	6	6	5	6
Health Condition	7	4		3	5	3	3	3	3	3
Locations	247	9		2	23	2			1	
Landmark	27	2	2	4	8	3	2	3	3	3
Business or Organization	12	2								
Dates	75	28	34	35	72	31	36	36	33	36
Events	108	9	6	8	83	8	6	6	7	7
Linguistic Patterns	72	44	28	38	45	36	32	36	34	27
Code-switching	6	3	2	2	3	2	2	2	2	2
Regionalisms	7	2		1	5		2	2		
Personal Interests / Activities	5	2		2	3			2		4
Traditions	6	3	3	5	4	2	2	2	2	4
Cultural References	11	5	5	7	9	7	4	6	5	2
Personal Identity	46	4	3	4	21	4	4	5	4	3
TOTAL	1431	185	110	139	353	124	114	127	114	112

Table E.2: Category breakdown results for missed redactions in the **GPT** dataset used in the prompt engineering study. The dataset contained 2,160 items (1,700 redactions).

Framework category	Items Redacted	Original	Framework	Specific Examples	Counterexamples	Verification	Role A	Role B	Role C	Verification + Role B
Personal Names	782	80	37	22	55	23	31	23	37	23
Nicknames	153	30	11	14	17	9	11	9	9	9
Demographics	184	57	31	45	69	36	32	29	34	31
Age	16	7	6	7	5	5	5	5	6	5
Race / Ethnicity	64	9	5	3	8	5	3	3	9	5
Education	11									
Profession	35	24	22	29	28	26	22	25	23	27
Health Condition	14	14	8	9	6	7	8	7	9	9
Locations	342	33	12	17	35	20	23	19	17	18
Address	6				4	4	4		4	
Landmark	77	20	20	19	21	16	15	16	20	18
Business or Organization	28	6	8	6	13	5	9	6	8	5
Dates	90	50	53	53	62	48	45	48	57	49
Events	143	46	26	30	111	30	37	24	38	32
Linguistic Patterns	80	50	32	41	49	37	40	38	38	29
Code-switching	10	5	4	6	4	4	5	5	5	3
Slang	2	2	2	2	2	2	2		2	2
Regionalisms	13	7	7	7	9	3	4	5	5	6
Personal Interests / Activities	24	9	8	7	11	7	8	8	8	8
Traditions	6	6	2	3	5	2	2	3	2	5
Cultural References	24	17	14	15	19	13	18	18	15	13
Personal Identity	56	30	21	18	32	18	14	19	23	20
TOTAL	2160	502	329	353	565	320	338	310	369	317

Table E.3: Overview of data stewards who participated in our formative interviews. The table shows their field, role, the number of participants in their primary dataset, and whether that dataset is longitudinal.

ID	Field	Position	# Participants	Longitudinal?
P1	Communication	Faculty	500-1500	Yes
P2	Communication	Faculty	10-50	No
P3	Sociology	Faculty	1000-2000	Yes
P4	Psychology	Faculty	100-500	No
P5	Anthropology	Faculty	50-100	No
P6	Sociology	Faculty	50-100	Yes
P7	Linguistics	Faculty	50-100	No
P8	Psychology	Faculty	50-100	Yes
P9	Psychology	Post-doc	50-100	No
P10	Sociology	Post-doc	10-50	Yes

Table E.4: **TalkBank** results in our **pilot study**. Our trained RAs made 4,466 manual redactions (5,231 words total).

Tool	Correct Redactions		Partial Redactions		Missed Redactions		Excess Redactions
	#	%	#	%	#	%	#
LLM: General Prompt							
Anthropic Claude Sonnet 3.5	3,397	64.9%	1,195	22.9%	639	12.2%	301
Google Gemini 1.5 Pro	4,151	79.4%	453	8.7%	627	12.0%	607
Meta Llama 3.1 405B Instruct	2,779	53.1%	305	5.8%	2,147	41.0%	361
OpenAI GPT-4o	3,502	66.9%	579	11.1%	1,150	22.0%	646
LLM: Framework							
Anthropic Claude Sonnet 3.5	4,879	93.3%	167	3.2%	185	3.5%	506
Google Gemini 1.5 Pro	3,859	73.8%	691	13.2%	681	13.0%	272
Meta Llama 3.1 405B Instruct	3,052	58.3%	1,224	23.4%	955	18.3%	446
OpenAI GPT-4o	3,511	67.1%	1,218	23.3%	502	9.6%	549

Table E.5: Redactions by category in the TalkBank and Synthetic datasets. We report the number of manual redactions (of words or phrases), as well as the total number of words included in these redactions. Because the data labels and the framework were developed concurrently and finalized separately, they may not exhibit perfect alignment. Data labels were aligned with the closest matching framework category after the framework was finalized, which can result in partial mismatches (i.e., coarser labels) between the final redacted items in the data and the stabilized framework.

Framework category	<u>TalkBank</u>		<u>Synthetic</u>	
	Redactions	Total Words	Redactions	Total Words
Direct Identifiers	–	–	–	–
Personal Names	2,323	2,405	26	28
Nicknames	371	420	11	12
Numerical Identifiers	–	–	–	–
Demographics	398	418	–	–
Age	10	19	6	6
Sexual Orientation	1	1	4	5
Gender	–	–	3	5
Race / Ethnicity	143	170	4	10
Education	8	23	13	40
Profession	28	43	5	10
Health Condition	7	14	4	11
Locations	672	823	2	2
Address	1	6	5	28
Landmarks	72	125	11	18
Business / Organization	41	71	5	10
Dates	47	105	5	11
Events	119	235	–	–
Specific Events	–	–	–	–
Linguistic Patterns	56	110	13	23
Code-switching	14	16	3	9
Slang	2	2	–	–
Regionalisms	19	24	–	–
Familects	–	–	–	–
Personal Interests / Activities	45	78	–	–
Hobbies	–	–	3	6
Traditions	3	6	–	–
Cultural References	13	25	4	5
Personal Identity	73	92	9	14

APPENDIX F

REDACTION STUDY GUIDE

Introduction

We are researchers at the University of Chicago. We are studying how to protect participant privacy in social-behavioral science research (SBS research) through redaction—a process to remove or obscure information for the privacy of data subjects—and we are looking at the effectiveness of different redaction systems. If you choose to participate, we will have you complete three redaction tasks, and we will conduct an interview with you to ask about your experiences with these tasks at the end. With your permission, we will screen-record your completion of the redaction tasks and record audio. We will not record any video of you, and we will not ask you for any personal information. Recordings will only be handled by the research team; audio will be transcribed and subsequently deleted.

Redaction tasks

For the purposes of this study, we define redaction as the removal of identifying information. Identifying information will be defined according to this framework, which gives us a shared understanding of what types of information to redact. Any redactions for each task should be done according to this framework. *[Moderator presents framework here and walk through with participant.]*

Participant consent

In a moment, I will hand you an information sheet about our study. Please review this sheet, after which I will ask for your verbal consent to proceed with the study. *[Moderator shares the consent information sheet with the participant.]*

Do you have any questions about the study?

Are you at least 18 years old and willing to volunteer to participate in the research? *[Wait for verbal yes. If no, do not proceed.]*

Do we have your permission to screen-record your completion of the redaction tasks and record audio? *[Wait for verbal yes. If no, do not proceed.]*

Framing

We've been asked to evaluate three different redaction methods for potential use in a future research project. Imagine you're a research assistant who has been asked to redact identifying information from transcripts of natural conversations. Participants have consented to release some, but not all, forms of identifying information to protect their privacy. You want to make sure that anyone who knows the participant personally cannot re-identify them from the text, but you still want other researchers to be able to study language development. You have decided to redact these types of identifying information. *[Moderator indicates types of identifiers to redact from framework.]*

As you go through each task, we ask you to think aloud. Think of this like guiding us through your process. For example, if I were to ask you how many windows are in your lab, a think-aloud process might walk me through each room and describe the size and shape of each window on each wall. For these tasks, that might mean why you decided to redact—or not redact—a term, as well as what actions you are taking to do so.

Please get through as much as you can in the allotted time; focus on both correctness and speed.

METHODS:

- *manually redacting in a Google sheet*
- *correcting pre-redacted data in a Google sheet*
- *using a graphical redaction interface*

Task 1 (15 min)

In this task, you will be asked to redact this dataset by [METHOD 1]. *[Moderator presents participant with tutorial slides for specific method.]* You will have 15 minutes to complete this task. Feel free to ask any clarifying questions about the process, but use your best judgment for specific items.

Survey/Break 1 (5 min)

Okay, task 1 is completed. Please fill out this survey about the task you just completed, then we will take a short break to prepare for the next task.

Task 2 (15 min)

In this task, you will be asked to redact this dataset by [METHOD 2]. *[Moderator presents participant with*

tutorial slides for specific method.] You will have 15 minutes to complete this task. Feel free to ask any clarifying questions about the process, but use your best judgment for specific items.

Survey/Break 2 (5 min)

Okay, task 2 is completed. Please fill out this survey about the task you just completed, then we will take a short break to prepare for the next task.

Task 3 (15 min)

In this task, you will be asked to redact this dataset by [METHOD 3]. *[Moderator presents participant with tutorial slides for specific method.]* You will have 15 minutes to complete this task. Feel free to ask any clarifying questions about the process, but use your best judgment for specific items.

Survey/Break 3 (5 min)

Okay, the final task is completed. Please fill out this survey about the task you just completed, then we will ask you to complete an interview.

Interview (15 min)

This survey will ask you about all the task methods you used. *[Moderator presents exit survey.]*

Thank you for participating in our study. In this final portion, we will ask you some questions regarding all the redaction tasks you just completed.

1. What are your thoughts on the redaction processes for each task? Were there any particular tasks that felt more intuitive or easier to work with than others?
2. Were there any difficulties you encountered while redacting information during any of the tasks? If so, how did you overcome them?
3. Did you feel that one redaction task was quicker than the others?
4. Did you feel that one redaction task was more effective than the others?
5. How confident were you in the accuracy of your redactions in each task? Did any of the tasks make you feel more or less confident in the correctness of the redacted information?
6. How did the layout or setup of any of the tasks affect your experience? Were there any features that made a particular task easier or harder to complete?

7. How was your decision-making process influenced when it came to determining what should or should not be redacted?
8. Based on your experience, how would you improve any of the redaction tasks?

Conclusion and debrief

Please fill out this short demographics survey. [*Moderator presents demographics survey.*]

Thank you for your time. If you have any questions, please feel free to ask or contact us at redactionproject@uchicago.edu.

APPENDIX G

SAMPLED ARTWORKS

Table G.1: The sample of artworks from our larger database that we qualitatively coded and the numbers (*A-#*) that reference them in the text (part 1 of 2). Nationality was determined based on public information on where an artist lives/works and their citizenship.

A-#	Title	Artist	Year	Medium	Venue	Artist's Nationality
1	How Not to Be Seen: A Fucking Didactic Educational .MOV File [265]	Hito Steyerl	2013	Video	Gallery	Germany
2	Hyperface [122]	Adam Harvey	2017	Textile, Wearable	Festival	Germany, USA
3	Surveillance Speaker [68]	Dries Depoorter	2018-2024	Hardware	Gallery, Public Art	Belgium
4	The Exhibited [299]	Anna Adahl	2013	Live Performance, Multimedia Installation	Gallery	Sweden
5	Choir of Missed Connections [24]	Aram Bartholl	2024	Sculpture	Festival/Pop-up	Germany
6	Stealth Cell Tower [208]	Julian Oliver	2015	Sculpture	Public Art	New Zealand
7	Security Blanket [61]	Lorrie Cranor	2013	Textile	Festival/Pop-up	USA
8	Security Blanket [102]	Mariam Ghani	2005	Architectural Space, Multimedia Installation	Festival/Pop-up	Afghanistan, Lebanon, USA
9	Every CCTV Camera (CC) [40]	James Bridle	2017	Photographs	Gallery	Greece, UK
10	Image Operations [215]	Trevor Paglen	2018	Video	Gallery	USA, Germany
11	MWITM [183]	Lauren Lee McCarthy, Kyle McDonald	2017	Software, Still Image	Gallery	China, USA
12	Surveillance Camera [4]	Ai Weiwei	2010	Sculpture	Gallery	China, Portugal
13	Selfie Stick [263]	Dan Stavy	2017	Sculpture, Software	Public Art	Israel, USA
14	Live-Taped Video Corridor [205]	Bruce Nauman	1967, 1968	Architectural Space, Multimedia Installation	Gallery	USA
15	Back to the Light [231]	Peters Riekstins	2016	Multimedia Installation	Gallery	Latvia
16	VOX [104]	Steve Giasson	2015-2016	Still Image	Festival/Pop-up	Canada
17	Netless2 [280, 281]	Danja Vasiliev	2012	Software	Public Art	Germany, Russia
18	The Right to be Forgotten [131]	Esther Hovers	2021-2024	Photographs	Gallery	Netherlands
19	Found Footage Stalkers [145]	KairUs	2018	Photographs	Gallery	Austria, Finland
20	EXISTech Corp. [179]	Steve Mann	c. 2001	Hardware, Wearable	Public Art	Canada
21	Classification.01 [70]	Mimi Onuoha	2017	Sculpture	Gallery	Nigeria, USA
22	Drone #4 [241]	Björn Schülke	2005	Sculpture	Gallery	Germany
23	If This is a Global Surveillance Center [170]	He-Lin Luo	2019	Multimedia Installation	Gallery	Taiwan
24	Pigeonblog [58]	Beatriz da Costa, Cina Hazegh, Kevin Ponto	2006-2008	Wearable, Multimedia Installation	Festival/Pop-up	Germany
25	The Eye and I [134]	Hsin-Chien Huang	2023	Multimedia Installation	Festival/Pop-up	Taiwan
26	LANDSAT.EARTH [269]	Shaheer Tarar	2023	Digital	Online	UK
27	The Mass Psychology of Fascism [224]	Surveillance Camera Players	1999, 2000	Live Performance	Public Art	USA
28	Good Morning Mr. Orwell [216]	Nam June Paik	1984	Live Performance, Livestream	Mass Media	South Korea, USA
29	Papa Bed [240]	Julia Scher	2003	Sculpture	Gallery	USA
30	Sorry to Bother You [126]	Dan Hett	2018	Video Game	Online	Turkey, UK
31	Citizen [77, 76]	Hasan Elahi	2006	Photographs	Online	Bangladesh, USA
32	The Hoodie_ [71]	Bogomir Doring	2008	Still Image	Gallery	Serbia, Netherlands
33	Deletion Process [107, 106]	Kyriaki Goni	2013, 2015, 2019	Hardware, Multimedia Installation	Gallery	Greece
34	The Nemesis Machine: Manifestation [261]	Stanza	2023	Hardware, Multimedia Installation	Gallery	UK, USA
35	iSee [283, 218]	The Institute for Applied Autonomy	2004	Digital, Software	Online	USA
36	The Feeling of Being Watched [288, 37]	Assia Boundaoui	2018	Video	Festival/Pop-up	Algeria, USA
37	3 Degrees of Separation from the Military-Industrial-Prison-Data-Surveillance State [160]	Sam Lavigne	2015	Digital	Online	USA
38	Uninvited [275]	Nye Thompson	2021	Video	Gallery	Austria, Switzerland, UK, USA
39	Audience [293]	Jens Wunderling	2010	Digital	Gallery	Germany
40	CMYK [50]	Daniel Canogar	2014	Sculpture, Video	Gallery	Spain, USA
41	A Girl Reading a Letter by an Open Window [282, 274]	Johannes Vermeer	c. 1657-1659	Painting	Gallery	Netherlands
42	Traffic Cam Photobooth [292]	WTTDOTM	2024	Digital, Software	Online	USA
43	Always Learning [167]	Machine Listening	2018	Video	Gallery, Festival/Pop-up	Australia
44	Mic Jammer [47]	Allison Burtch	2015	Hardware, Software	Gallery	USA
45	You Saw Me? [169]	Melanie Lowe	2008	Digital, Multimedia Installation	Public Art, Gallery	Canada

Table G.2: The sample of artworks from our larger database that we qualitatively coded and the numbers (*A-#*) that reference them in the text (part 2 of 2). Nationality was determined based on public information on where an artist lives/works and their citizenship.

A-#	Title	Artist	Year	Medium	Venue	Artist's Nationality
46	System Azure Security Ornamentation [176]	Jill Magid	2002	Live Performance	Public Art	USA
47	Peepshow [203, 284]	Giles Walker	2007	Live Performance, Multimedia Installation	Festival/Pop-up	UK, Canada
48	Data Collection [148]	Dave Kemp	2010	Photographs, Video, Multimedia Installation	Gallery	Canada
49	Surveillance Cutie [159]	Fabiola Larios	2024	Multimedia Installation	Festival/Pop-up	Mexico, USA
50	Trusted Traveller [217]	Yuri Pattison	2017	Architectural Space, Digital, Multimedia Installation	Gallery	Ireland, UK
51	Twelve Nodes [154]	Egor Kraft	2019	Hardware, Software, Multimedia Installation	Gallery	Austria, Germany, Japan, Russia
52	Backspace [279]	Ben Van den Berghe	2022	Architectural Space	Gallery	Belgium
53	Eavesdropper 1 [200]	Mschf	2022	Hardware	Gallery	USA
54	Convergence LA [10]	Refik Anadol	2017	Digital, Multimedia Installation	Public Art	Turkey, USA
55	Camouflage [277]	Suzanne Treister	2013	Painting, Still Image	Gallery	UK
56	Dark Side of the Prism [33]	Justin Blinder	2013	Digital, Software	Online	USA
57	Reality TV [276]	Timo Toots	2017, 2023	Hardware, Software	Gallery, Public Art	Estonia
58	Suspicious Minds [32]	Viktoria Binschok	2009	Photographs	Gallery	Russia, Germany
59	My Little Big Data [182]	Eva and Franco Mattes	2019	Video, Architectural Space, Multimedia Installation	Gallery	Italy, USA
60	Yuppie Ghetto with Watchdog [98]	Paul Garrin	1989	Architectural Space, Video, Multimedia Installation	Gallery	USA
61	Minor Urbanism [247]	Mark Shepard	2012	Video	Online	USA
62	Which no one will ever see [8]	Meriç Algün	2012	Multimedia Installation	Gallery	Turkey, Sweden
63	Nest [101]	Ringborg				
64	MyPocket [15]	Jakub Geltner	2015	Sculpture	Public Art	Czechia
		Burak Arikan	2008	Software, Multimedia Installation	Gallery	Turkey
65	Demand Ware [44]	Jonah Brucker-Cohen	2022	Digital, Software	Festival/Pop-up	Ireland, USA
66	Carbolytics [197]	Joana Moll	2022	Digital, Software, Multimedia Installation	Gallery, Online	Spain, Germany
67	SVEN [7]	Amy Alexander	2006-2007	Digital, Multimedia Installation	Gallery, Public Art	USA
68	Sticky Floors [251]	Ann-Sofi Sidén	2014	Video, Multimedia Installation	Festival/Pop-up	Sweden
69	Nakagin Capsule Tower [156, 136]	Kisho Kurokawa	1972	Architectural Space	Building	Japan
70	Embroidered Surveillance [161]	Francine Leclercq	2024	Textile	Festival/Pop-up	France, USA
71	Go Rando [114]	Ben Grosser	2021	Digital, Software	Gallery	USA
72	The Status Project [46]	Heath Bunting	c. 2012	Digital	Gallery, Festival/Pop-up	UK
73	Sanctum [60]	James Coupe	2015	Hardware, Software	Gallery, Public Art	UK, USA

APPENDIX H

CODEBOOK

Table H.1: Our codebook and the artworks in our sample to which we applied each code (part 1 of 3).

Code	Definition	Artworks (A-#)
Elements & Principles of Design	Arrangement, shape, and composition	
Colors		
↔ Bright Colors	Whites, yellows, hot pink, etc.	7, 10, 13, 15, 21, 25, 28, 31, 32, 34, 35, 37, 38, 40, 42, 44, 45, 47, 49, 52, 55, 56, 58, 59, 60, 64, 67, 72
↔ Dark Colors	Blacks, dark greys, etc.	3, 4, 10, 11, 15, 18, 20, 23, 25, 26, 28, 29, 32, 33, 35, 38, 39, 40, 42, 44, 48, 51, 54, 59, 60, 64, 67
↔ Monochrome	One color with different tones, or black and white	1, 6, 12, 14, 15, 18, 20, 22, 25, 26, 27, 38, 47, 53, 55, 58, 59, 62, 64, 68, 70
↔ Contrasting Colors	Uses strongly opposing colors, such as red and green	1, 2, 4, 11, 17, 26, 29, 33, 34, 51, 52, 54, 55
Distortion	Alters the perception of visual or auditory components	2, 4, 14, 15, 16, 18, 25, 26, 28, 32, 38, 39, 55, 59, 71
↔ Tactile/Analog Representation of Digital	Digital object is represented through physical media	6, 7, 33, 53, 55, 57, 59, 62, 70, 73
↔ Digitized	Pixelated, synthetic, or screen-mediated aesthetics	1, 10, 18, 25, 26, 50, 52, 66, 70
Scale	Emphasizes size in respect to space	4, 7, 9, 11, 13, 14, 15, 16, 17, 19, 20, 21, 25, 26, 29, 31, 32, 34, 35, 37, 48, 51, 52, 53, 54, 55, 57, 58, 59, 64, 66, 72, 73
Collage	Piecing together many smaller works or elements, often using repetition and variation	4, 5, 7, 9, 13, 15, 18, 19, 23, 26, 28, 31, 32, 37, 38, 42, 48, 49, 52, 55, 58, 61, 63, 67, 68, 69, 73
↔ Composite	Overlapping or merging multiple elements into the same space	15, 26, 28, 32, 38
Perspective	Position the artwork is viewed from	
↔ Voyeuristic	Viewer in the watcher, camera POV	4, 14, 23, 25, 27, 33, 38, 39, 41, 42, 47, 61, 67, 68, 69, 70, 73
↔ Surveillance is Looming	Situated above the viewer or subject	3, 9, 22, 26, 54, 67
↔ Shifts	Literal viewpoint shifts across the artwork	4, 8, 10, 38, 50, 58
Motion	Visible movement, physically or visually	1, 3, 36, 38, 49, 45, 47, 49, 50, 53, 54, 59, 61, 62, 66, 67, 68
Lighting	Lights are used to direct attention	4, 5, 13, 21, 25, 28, 34, 43, 45, 50, 54, 57
Display	The modes in which the artwork is presented or output to viewers	
Data Visualization	Network graphs, bar charts, interactive interfaces	24, 33, 34, 35, 37, 54, 59, 61, 64, 66, 72
↔ Bounding Boxes	Rectangles displayed around objects (e.g., faces) to suggest detection	2, 10, 54, 57, 61, 67
Livestream	Video plays in real time	4, 14, 23, 49, 67, 73
Framed Image	Parts of the artwork are framed for display	11, 18, 52, 55

Table H.2: Our codebook and the artworks in our sample to which we applied each code (part 2 of 3).

Code	Definition	Artworks (A-#)
Visual and auditory lexicon	The sensory vocabulary that the artist uses as components to build from	
Plaintext	Words, source code	1, 6, 7, 10, 11, 15, 16, 18, 27, 33, 34, 37, 38, 39, 45, 48, 51, 53, 54, 59, 61, 62, 64, 70, 72, 73
Everyday Object	Common objects in daily life (e.g., smartphones, bedding)	5, 6, 8, 11, 13, 17, 20, 25, 29, 30, 32, 39, 40, 41, 44, 49, 50, 57, 59, 60, 67, 68, 69, 73
Human	Human form (e.g., bodies, faces)	1, 2, 4, 10, 13, 15, 18, 25, 27, 28, 32, 36, 37, 38, 39, 41, 57, 58, 59, 60, 61, 67, 70
Camera (object)	Visible camera	3, 5, 9, 12, 14, 20, 22, 25, 29, 39, 46, 47, 49, 50, 60, 62, 63, 73
Audio	Audio or sound plays	38, 54, 60
↔ Music	Music is used within the artwork or is the artwork itself	6, 10, 25, 28, 47, 56, 62, 67
↔ Robotic Voice	A robotic or synthetic voice can be heard in the artwork	1, 3, 5, 43
↔ Human Voice	A human voice can be heard in the artwork	8, 62, 73
Public Infrastructure	Human-built public infrastructure (e.g., buildings, streets)	9, 17, 23, 24, 34, 36, 45, 46, 50, 52, 59, 60, 61, 70, 73
Common Symbols	A shape or object that is representative of another object or idea	1, 27, 41, 51, 62, 66, 71
Microphone (object)	Visible microphone	22, 44, 53
Techniques	Strategies and methods artist utilizes to construct, critique, or reframe surveillance and privacy invasion	
Interactivity	The viewer can interact directly with the artwork	3, 4, 6, 8, 14, 17, 21, 25, 26, 30, 33, 34, 35, 37, 42, 48, 53, 57, 60, 66, 71, 73
Speculative	Imagines future possibilities	17, 20, 21, 23, 24, 25, 28, 38, 43, 51, 65, 67, 72
↔ Proof of Concept	Demonstrates feasibility through a functional prototype	2, 6, 17, 24, 35, 44, 64, 65, 66, 67
Counter-Surveillance Interventions	Direct methods to resist or evade surveillance (e.g., camouflaging, obfuscation)	1, 2, 11, 17, 23, 24, 26, 31, 32, 35, 37, 44, 46, 55, 56, 59, 62, 71
Exposing Surveillance	Reveals hidden or often ignored surveillance	9, 12, 20, 22, 23, 24, 25, 29, 35, 36, 39, 42, 43, 46, 55, 56, 58, 66
Imitation	Mimics or exaggerates an existing object, topic, or cultural form, often as satire or parody	5, 6, 12, 13, 20, 22, 31, 34, 40, 47, 50, 55, 60, 62, 3, 67, 69
Information Withheld From Viewer	Restricts what the viewer knows	11, 14, 21, 31, 48, 55, 57, 59, 65
Tone	Atmosphere that frames how the artwork is experienced	
Unsettling	Discomfort or eeriness	5, 6, 12, 14, 15, 18, 19, 23, 25, 29, 30, 32, 36, 38, 39, 48, 49, 60, 65, 66, 69, 73
↔ Industrial	Utilitarian or cyberpunk (e.g., pipes, wires, metal)	3, 4, 13, 15, 20, 21, 25, 29, 34, 38, 39, 40, 47, 49, 50, 51, 53, 54, 59, 60, 69
↔ Panopticon	Constant sense of overwhelming surveillance	12, 20, 23, 34, 41, 69
Fun and Playful Glamourized / Fetishized Mundane Ornate	Light-hearted and whimsical in tone Glamorous or pleasurable to view Ordinary, everyday life Refined, elegant, or high-class	3, 8, 28, 31, 42, 46, 47, 49, 67 9, 34, 37, 41, 46, 47, 49 4, 8, 31, 49, 68, 69, 70 12, 46, 52

Table H.3: Our codebook and the artworks in our sample to which we applied each code (part 3 of 3).

Code	Definition	Artworks (A-#)
Observation	Dynamics of watching and being watched (visual and data collection)	
Data Source/Subject	Whose data is collected/privacy invaded	
↪ Specific Individual (public)	Specific individual from the public is highlighted	1, 6, 8, 15, 16, 18, 19, 23, 30, 37, 39, 41, 45, 48, 58, 67, 68, 69, 70
↪ Viewer	The viewer is the subject of surveillance/privacy invasion	3, 4, 6, 8, 13, 21, 25, 34, 39, 42, 49, 53, 57, 60, 67, 69, 71, 73
↪ The Public	The general public in an aggregate form	3, 6, 16, 17, 19, 23, 24, 34, 54, 61, 66, 67, 68, 70, 73
↪ Artist Self-Surveillance	Artist uses their own data or personal experiences	11, 14, 27, 28, 30, 31, 33, 37, 59, 64, 72
↪ Performers	Those hired by an artist or part of a collective art group	4, 10
Data Collection Method	How data used in the artwork is collected	
↪ Camera	Uses cameras	3, 10, 13, 14, 21, 23, 25, 38, 39, 42, 49, 57, 60, 67, 68, 70, 73
↪ Text-mediated Communications	Uses text or social media data	6, 11, 15, 30, 39, 59, 71, 72, 73
↪ Microphone	Uses microphones	8, 16, 53
↪ Viewer Device	Uses the viewer's personal device	42, 45
Duration of Data Collection	The artwork emphasizes the length of time data is collected, or comments on the persistence of data traces	6, 8, 13, 15, 17, 18, 19, 21, 33, 35, 39, 45, 48, 54, 59, 61, 64, 68, 73
Implied Observation	Suggests that the viewer is being watched, e.g. non-functioning cameras	4, 5, 20, 21, 22, 47, 50, 56, 57, 63, 69
Topics	The ideas and themes the artwork engages or critiques	
Privacy-invading Technology, Algorithms, and Identity	Technological systems that invade privacy and their implications	
↪ Personal Identity	How a person is defined by themselves or others	2, 4, 11, 13, 15, 18, 19, 21, 25, 32, 33, 36, 40, 41, 45, 50, 59, 62, 64, 67, 71, 72
↪ Artificial Intelligence	Classification algorithms, computer vision	2, 3, 10, 13, 21, 32, 34, 38, 39, 43, 54, 57, 61, 64, 67, 72, 73
↪ Algorithmic Assumption	Inferences made by algorithms	10, 21, 32, 64, 71
Organized Surveillance / Data Collection	Surveillance and data collection by an organization	
↪ Government Surveillance	Law enforcement, intelligence agencies	1, 6, 9, 12, 17, 18, 23, 24, 25, 26, 28, 33, 34, 35, 36, 37, 42, 46, 47, 50, 54, 55, 56, 58, 61, 62, 66, 67, 70, 71, 72
↪ Corporate Surveillance	Social media companies, browsers	33, 35, 43, 44, 55, 56, 66, 71, 72, 73
Public vs. Private Space	Questions which digital / physical spaces are private or public	1, 4, 5, 16, 23, 31, 40, 41, 45, 52, 54, 59, 60, 62, 69

APPENDIX I

EXAMPLE CODED PIECE

To begin coding, the coder reads through the document to become familiar with overall content. Codes pertaining to the visual content of the document are applied to a bounding box drawn around images (see the fourth panel in Figure I.1). In this example, the consistent images on the left are ignored, since these refer to other artworks by the artist. The fourth panel contains an image of the artwork; other images contain information about the process. Process images typically contain information about mechanisms or more subjective, interpretive elements. In the artwork image, our sample contains the codes “Data Visualization” (display mode), “Human” (visual object), and “Collage” (design element), among others. Codes pertaining to the textual content of the document are applied to a highlighted paragraph (see the first panel in Figure I.1). Each paragraph is considered one unit of text for consistency in coder analysis. Codes applied to text have a broader scope, since text may contain both information about the mechanisms of the artwork and themes or topics the artist addresses.

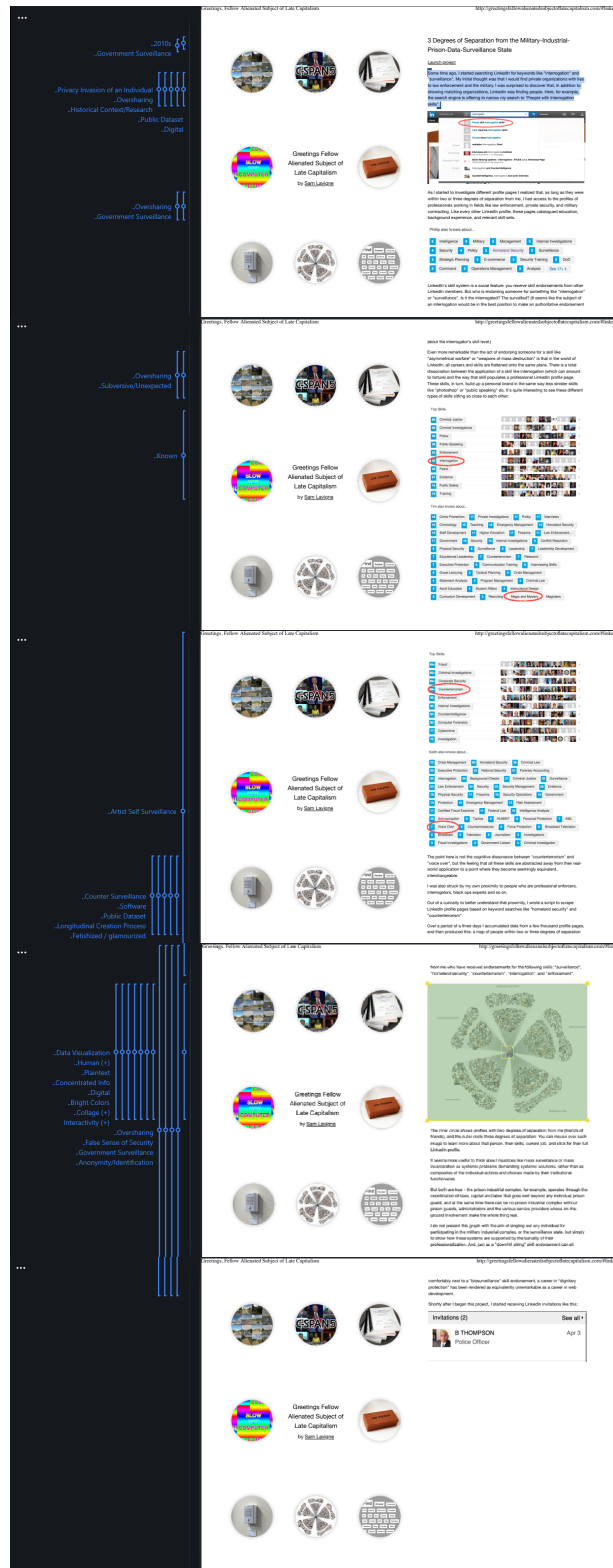


Figure I.1: An example of a coded artwork, *3 Degrees of Separation* [160] ((©Sam Lavigne), in the coder window of MAXQDA.