Intelligent Human Systems Integration (IHSI), Vol. 200, 2026, 84–94
https://doi.org/10.54941/ahfe1007067

Privacy at the Core: Toward Automated Detection of Privacy-Sensitive Content in an LLM-Based Care Documentation Support System

Reinhard Kletter and Sabine Theresia Koeszegi

Institute of Management Science, TU Wien, 1040 Vienna, Austria

ABSTRACT

Large language models (LLMs) introduce new opportunities in residential care, including the potential to assist with care documentation. However, if introduced unreflected, such technologies present challenges and potential harms to privacy and personal integrity. In this paper, we present a framework for automated filtering of privacy-sensitive content from LLM-supported care documentation. Our framework is based on Nissenbaum's theory of privacy as contextual integrity. As an initial step, we present the generation of a synthetic dataset derived from privacy-sensitive interactions between care workers and care recipients in the real world. We analyze the conversations by privacy categories and show that both care recipients and care workers are affected. Our contributions include a methodology for generating privacy-preserving synthetic datasets and insights into the content requirements of a dataset for fine-tuning an LLM to detect privacy-sensitive segments. In addition, we show that value-sensitive design can result in innovative approaches to creating technology that is safe, meaningful, and protective of important human values.

Keywords: Privacy, Synthetic data, Residential care, Value sensitive design

INTRODUCTION

Caring for older adults in institutional settings such as residential care homes is a stressful, physically and emotionally demanding job (Figueiredo et al., 2019). Technology is frequently introduced in this context, to reduce the workload of care workers while improving or at least maintaining the quality of care for residents (Bail et al., 2022). It is often overlooked that, through its active use in practice, technology assumes roles that extend beyond functioning merely as a tool (Orlikowski, 2007). Ultimately, it is essential to recognize that care constitutes a complex social interaction between care workers and care recipients (van Wynsberghe, 2013). Carelessly introducing technology might even have negative consequences for the quality of care or working conditions. This creates an inherent conflict: on the one hand, there are potential benefits from using new technologies; on the other hand, deterioration may manifest on multiple levels, as numerous values are embedded within care practices.

Received September 18, 2025; Revised November 10, 2025; Accepted November 25, 2025; Available online February 1, 2026

The technological push for artificial intelligence (AI) in recent years resulted in the release of technology that contains certain risks such as bias, fairness, harmful content, misinformation, privacy and security concerns (Li et al., 2025). Nevertheless, this technology has been made publicly accessible, thereby placing the responsibility for identifying malfunctions, limitations, and other issues on users.

In contrast, this work adopts a value sensitive design (VSD) approach. We present a novel approach to protecting the privacy of individuals that are using an LLM-based documentation support system in residential care homes. The proposed *privacy agent* is founded on the theory of privacy as contextual integrity and illustrates how VSD can enable innovative technological solutions without necessitating a trade-off between values and utility. Additionally, a methodology is introduced for generating a synthetic dataset that maintains privacy while being derived from real-world, privacy-sensitive conversations between care workers and care recipients. This dataset, which is available to the public (see Grabler et al. (2025)), is a first step toward enabling LLM-based automated detection of privacy-sensitive content. A detailed analysis of the dataset is conducted to identify the affected privacy categories to which the entries correspond. This analysis enhances our understanding of the nature of conversations occurring between care workers and care recipients and allows us to further refine our approach to protecting privacy in this context.

BACKGROUND

In this chapter, we explore the concept of privacy, with a particular focus on the theory of privacy as contextual integrity, and discuss the basic concept for synthetic data generation.

Privacy as Contextual Integrity

The conceptual understanding of privacy ranges from the fundamental *"right of the individual to be let alone"* (Brandeis & Warren, 1890) to dynamic processes of boundary management and interaction regulation (Altman, 1975). Other scholars lean towards a multi-dimensional model, along the dimensions of *informational*, *physical*, *social*, and *psychological* privacy (Burgoon, 1982). Nissenbaum (2004) introduces the concept of privacy as contextual integrity, emphasizing that privacy violations are context-dependent. According to this framework, different contextual norms apply based on the social roles individuals assume within a given interaction, and the type of information disclosed. A privacy-sensitive situation is conceptualized as a flow of personal information, characterized by five key parameters: the *sender*, the *recipient*, the *attribute*, the *subject*, and the *transmission principle*. If one of the parameters of an information flow does not align with the contextual privacy norms, a privacy violation occurs (Shvartzshnaider et al., 2019).

Synthetic Data Generation

Synthetic data generation (SDG) is an approach to address the challenge of limited dataset sizes (Douzas et al., 2022). Synthetic data is defined as data built by mathematical models or algorithms (Jordon et al., 2022). However, there are other approaches, such as virtual sample generation (VSG), which takes real samples and generates virtual ones (Wen et al., 2024), Diffusion-Neural-Network, or Mega-Trend-Diffusion (Douzas et al., 2022).

CONCEPTUAL FRAMEWORK

In the following, we outline a use case of LLM-based documentation support, including decisions for VSD. Moreover, a novel approach for detecting privacy-sensitive situations in conversations is proposed.

Use Case: Care Documentation Support

Across five participatory design workshops conducted in two residential care homes in Austria (refer to Frijns et al. (2024) for further details), participants were introduced to various technologies used in robots, including LLMs functioning as conversational agents. A use case of an LLM-based documentation support system for care workers emerged. Performed care tasks, such as administering medication, assisting with personal hygiene, and changing bandages, are extracted from audio recordings of interactions between care workers and residents. They are subsequently delivered to the care workers in the form of concise LLMs-generated summaries. However, to keep the human in control over the contents of the documentation, the care workers are asked to review the summary beforehand. We used local automatic speech recognition (ASR) of the interactions and created the summaries using a local model optimized for German language (for additional details see Hirschmanner et al. (2024)). In contrast to other LLM-supported documentation systems that depend on the verbal dictation of completed tasks, this design seeks to preserve the natural flow of interaction between care workers and residents by minimizing technological intrusiveness.

Value Sensitive Design Decisions

VSD is a human-centered design approach, which argues that morally important values should be preserve in technology design by principle, regardless of whether there are particular individuals or groups that do not share these values (Friedman et al., 2002). As we believe that designers are responsible for proactively considering VSD, several design decisions in the early stages were made with privacy in mind. In the LLM-based documentation support system, we use audio recordings only, instead of video-based approaches. We give care workers the ability to stop a recording at any time and store it locally on a microphone with internal storage, instead of uploading it directly. We also intentionally chose a clip-on microphone over an ambient recording installation to mitigate surveillance concerns. Furthermore, care workers can completely opt out of uploading audio to the server for processing. The LLM workstation uses only locally running models.

Proposal of a Privacy Agent

Despite the implementation of those privacy precautions, the technology in the use case inevitably functions as an intermediary in the interaction between care workers and residents. All verbal exchanges may be recorded and subsequently incorporated into LLM-generated documentation. This dynamic can adversely impact the quality of interaction, potentially discouraging care workers from sharing personal information with residents or shifting the focus of the interaction from social engagement to task-oriented communication.

Consequently, we propose adding a privacy agent (see Grabler et al. (2024) for further details) to the use case to proactively safeguard the privacy of all involved parties. Rather than automatically forwarding the raw transcript to the summarization LLM, this intermediary software agent should filter out privacy-sensitive information pertaining to care workers or care recipients (see Figure 1). This process is guided by the theory of privacy as contextual integrity, ensuring that only context-appropriate information is retained for summary generation. For the implementation we propose a two-step LLM-based solution. *Step 1* involves analyzing the information flows present in care interactions with respect to the key parameters defined by the theory of contextual integrity. *Step 2* entails evaluating these parameters against established contextual privacy norms to determine their appropriateness.

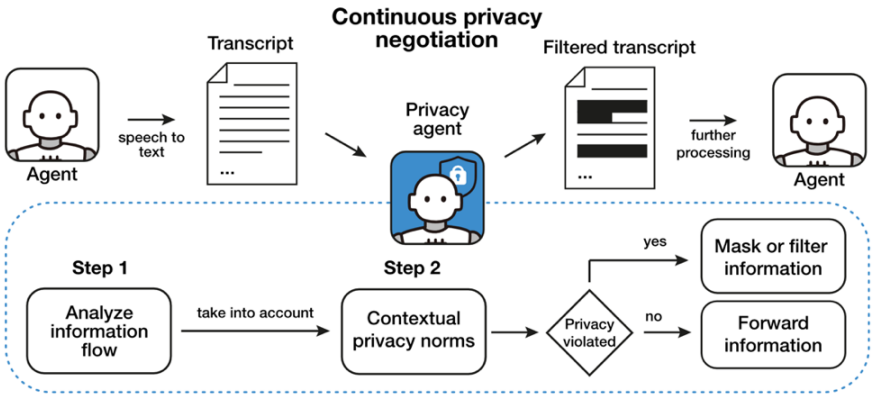

Figure 1: The privacy agent operates as an intermediary component within the LLM-based documentation support system. It facilitates a continuous negotiation of privacy values, resulting in the filtering of information deemed contextually inappropriate, in accordance with the theory of privacy as contextual integrity.

Example Privacy Negotiation. To enable the functioning of a privacy agent, information flows must be systematically extracted and classified from the conversation transcripts. Figure 2 presents a sample dialogue representative of a typical care interaction. The conversation has been manually annotated to demonstrate the identification of key information flow parameters.

Step 1: Analyze Information Flows. In the example excerpt, the care worker simultaneously serves as both the sender and the subject by disclosing personal details about their own life. This highlights that individual text

segments can be annotated with multiple contextual integrity key parameters. Several distinct attributes are revealed – such as the care worker's sleep and wake times, and their choice of breakfast beverage – demonstrating that parameter assignment can occur multiple times within a single information flow. While the resident is the immediate recipient of the conversation at the time of recording, the excerpt may later be transferred to documentation accessible by other care workers or managerial staff. Consequently, these additional individuals must also be identified as recipients. The transmission principle governing this interaction – confidentiality between care worker and resident – applies to the entire excerpt and must be annotated accordingly.

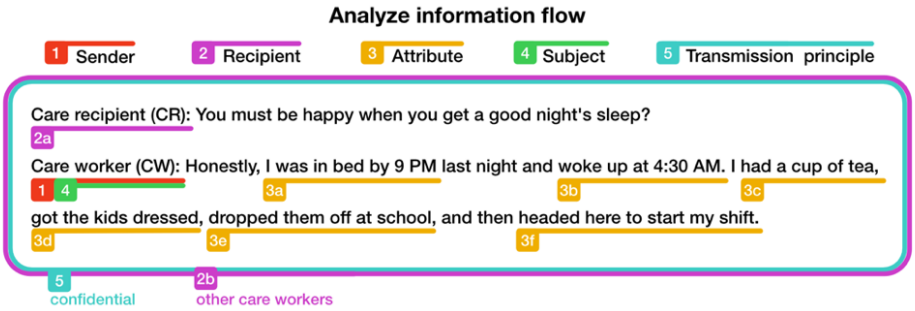

Figure 2: An example information flow originating from an interaction of a care worker and a resident is annotated with the key parameters of contextual integrity information flows.

Step 2: Consider Contextual Privacy Norms. The identified key parameters must be assessed in relation to applicable privacy norms. In the example presented in Figure 2, a privacy violation is evident: although the information exchanged concerns the care worker who is sharing the information and is directed toward the resident, it becomes accessible to other care workers or managerial staff through its inclusion in the documentation system. There is no justifiable interest for these additional parties to know personal details such as the care worker's sleep schedule or breakfast choices. Given that the conversation was intended to remain confidential between the care worker and the resident, and that there is no legitimate purpose for processing this information within the documentation, such inclusion constitutes a breach of established privacy norms.

Within the LLM-based documentation system, we acknowledge that care workers could be tasked with manually removing privacy-sensitive content. However, guided by the principle of VSD, the goal is to ensure that such conversational snippets are automatically filtered. This approach fosters trust in the system and helps preserve the human quality of interaction between care workers and residents. Ultimately, both parties should feel at ease sharing any information they deem appropriate within the context of their interaction.

GENERATION OF THE SYNTHETIC DATASET

The first step in the automated identification of parameters of an information flow is generating a dataset. Various datasets enable LLMs to recognize personally identifiable information (PII), for example, the

dataset pii-masking-200k (ai4Privacy, 2023). However, there is no database of privacy-sensitive conversation fragments which – using the theory of privacy as contextual integrity – we aim to use in the fine-tuning of an LLM to detect privacy-sensitive contents more generally. We are addressing this issue by generating a synthetic dataset from real-world recordings. We adopted an approach loosely based on the virtual sample generation (VSG) method, using real samples to generate synthesized data (Wen et al., 2024).

Initial Data Collection

The raw data was collected in the care context, continuing the work of Hirschmanner et al. (2024) to test the prototype of LLM-based documentation support with real-world interactions. 23 audio recordings of care interactions were made over an 11-week period in a residential care home. 8 care workers and 6 care recipients participated voluntarily. Durations of the interactions range from 1 min 28 s to 27 min 53 s, the average duration was M=13 min 16 s (SD=9 min 9 s) and the performed care actions included partial and full body washing, giving of medication, the morning routine, as well as wound care. Audio recordings were transcribed with WhisperX (Bain et al., 2023) locally and without speaker diarization due to poor results in initial tests at the time of conducting the study.

Data Preparation

The audio recordings were listened to, and speaker diarization was added manually. Minor corrections were made to the transcripts, such as if a name or a location was incorrectly transcribed. Nevertheless, the objective was not to generate a flawless transcript, as real-world transcription systems are also expected to produce similar errors. Due to the use of clip-on microphones, some of what the residents have said was unintelligible. Missing responses from residents have been added where possible. Moreover, the transcripts were pseudonymized, by replacing personally identifiable data with placeholder labels. This included among other names, locations, birth dates, ages, and vital signs. The transcripts were reviewed for privacy-sensitive conversation parts and coded in the qualitative data analysis software MAXQDA. Subsequently, n=86 conversations were exported for further processing.

Synthesizing Data

The placeholders to ensure pseudonymity were replaced with randomly assigned values with a search and replace action. Consistency in the data was considered, e.g., if a city name was replaced to Rome, the corresponding placeholder for the country was also changed to Italy. The sections were translated locally from German to U.S. English using icky/translate (icky/translate 2024), a small LLM for translations, for broad appeal of the published dataset. Each speaker change was handled as a separate request to the LLM, to avoid losing the correct speaker diarization in the process.

To generate data that is similar, yet distinct from the original data, the model llama3.1:70b (Meta AI 2024) was queried with the following system prompt:

"You are an AI assistant specializing in synthetic data generation (SDG). Your task is to generate a single short, realistic conversation between a care worker (CW) and a care recipient (CR), an older adult living in a care home, based on the conversation you receive. Your output must: 1. Be concise, realistic, and reflective of the tone, context, and nature of the original dialogue. 2. Not repeat the original conversation exactly but instead provide a unique yet similar variation. 3. Include only the generated conversation, with no additional text or commentary. Focus on maintaining realism and brevity in your responses."

Several versions of the prompt were tested, until this version produced satisfying results. We used zero-shot prompting, ran each conversation as a separate request to the LLM without conversation history and used a temperature of 0.7 to ensure some variability in the output. The resulting synthetic data was manually reviewed, the categorization of privacy violations were revised, and the affected individuals were added. Moreover, n=1 entry was deleted, as it did not yield any privacy-sensitive content after synthesizing. Some conversations were split into multiple entries as some segments captured content fitting for a certain privacy category, but as a whole the conversation was a different category. Overall n=95 entries were obtained.

RESULTS: SYNTHETIC DATASET

Figure 3 shows the frequency of privacy-related categories in the conversations, with bar segments indicating the share of affected individuals. The most frequent category is personal information, encompassing details about vacations, events, residences, and care workers' schedules. Although residents' personal information is disclosed most often, a substantial share also concerns care workers and other third parties. The health category primarily concerns residents and includes topics such as pain, hygiene, illnesses, mobility, medication, exhaustion, and sleep. Family covers the well-being of family members, their visits, and personal details, with about two-thirds relating to residents' families. Values refer to conversations reflecting underlying beliefs, such as views on vaccination or freedom, while thoughts involve openly expressed opinions on diverse topics. Other categories include acquaintances (not family) and appointments, such as with therapists.

Overall, the privacy of care recipients is most frequently affected in the conversations (64.21%, n=61), followed by care workers (22.11%, n=21). In some cases, both groups are affected simultaneously (6.32%, n=6). Additionally, the privacy of other third parties is also impacted (7.37%, n=7) in the conversations.

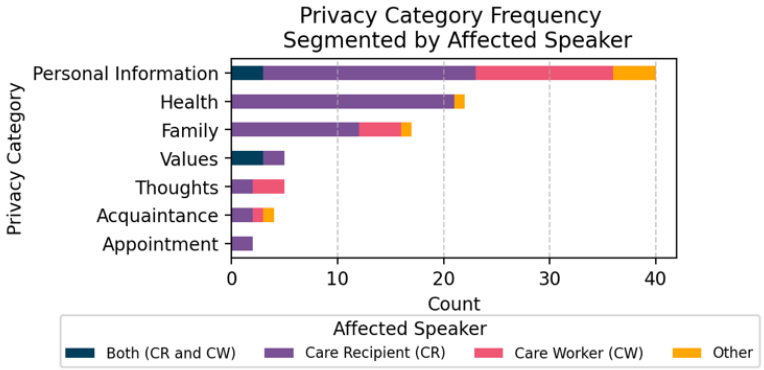

Figure 3: Frequencies of dataset entries labeled with the given privacy categories, segmented by the affected speakers.

DISCUSSION

We show an implementation of VSD with the first step toward creating an LLM-based privacy agent to detect privacy-sensitive content. Through pseudonymization, translation and synthesization, a privacy-preserving labeled dataset of care interactions was created.

Value Sensitive Design for Technology Used in Care Interactions

Our commitment to the VSD approach shows that several steps can be taken early in the design. Not only did we take several design decisions (clip-on microphones instead of ambient microphones, local LLM processing, etc.), but we demonstrate that novel approaches can be implemented with the rise of new technologies such as LLMs. By recording real care worker-resident interactions and synthesizing the data we took a first step toward creating a training dataset for a privacy agent in this use case. Other than dialogues between doctors and patients for diagnostic, which are goal oriented, care worker–resident dialogues are longer and contain more small talk (see also e.g., Macdonald (2016)). Conversations help building a relationship to the residents, helping with trust and confidence, and overall creating a team spirit in the residential care home (Sundler et al., 2020; Höglander et al., 2023). After all, care workers are also the daily contact persons as the virtually live with the residents (Wilson et al., 2009). As demonstrated by our synthetic dataset, the privacy of both care workers and care recipients is impacted. Thus, in this context, embedding privacy at the core is crucial, as an unsafe system may alter the content and quality of these social interactions.

Limitations of Synthesized Data

Translation and synthetic data generation using the LLM distorted data quality. The native languages spoken by the care workers in the study included Bulgarian, Croatian, Polish, Serbian, Romanian, Spanish, and

Hungarian – all while the conversations were conducted in German. Due to this linguistic discrepancy, the original transcripts are somewhat more difficult to understand than the sections synthesized by the LLM. Moreover, we found that creating a synthetic dataset from real-world data while preserving the privacy of individuals requires substantial effort. Due to poor speaker diarization, manual identification was necessary, along with error correction and transcript refinement. This manual approach is not scalable for achieving the diversity required in a training dataset, making the generation of synthetic data from a small initial sample a necessary next step. Nevertheless, the yielded insights into the conversational content serve as an essential prerequisite in the roadmap toward implementing the *privacy agent*.

CONCLUSION

We demonstrate how incorporating VSD approaches into early design decisions can improve and enhance the safety of technology. We present a novel approach based on LLMs to embed privacy considerations at the core. The proposed privacy agent is based on the concept of privacy as contextual integrity and can be implemented in two steps. *Step 1* involves identifying the key parameters of an information flow, while *step 2* requires evaluating these parameters against established privacy norms within the specific context. In our publication, we took the first step toward identifying privacy-sensitive conversations. We contribute with an initial synthetic dataset from real interactions between care workers and care recipients in a residential care home and classified it. While the privacy of care recipients is primarily affected, the privacy of care workers is also significantly impacted. This knowledge is valuable for refining the privacy agent's design requirements.

However, the presented dataset is not sufficient to fine-tune an LLM for classifying privacy-sensitive content. The variability of the content is very high, making it difficult to determine why a particular piece of information is privacy-relevant or not, calling for the development of an Inter-Annotator Agreement (IAA) metric for labeling consistency. We found that collecting real data is time-consuming and challenging. After several weeks of data collection, we only have small amounts of data to work with. Additionally, synthesizing the data to protect the privacy of the individuals involved required considerable time and effort.

Future Work

Future research should investigate approaches for generating large volumes of synthetic data from a limited set of real data, with an emphasis on minimizing the required effort. Moreover, as pointed out in Grabler et al. (2024), the key parameters of information flows must be annotated manually through crowdsourcing. This should facilitate an IAA metric that yet has to be developed. Additionally, semi-automated annotation methods need to be explored. To apply the theory of contextual integrity and assess whether a conversation constitutes a privacy violation, it is moreover necessary to examine the privacy norms specific to the given context. Following this roadmap, a proof-of-concept version of the *privacy agent* is to be created and critically examined with qualitative evaluations.

ACKNOWLEDGMENT

This work was funded by a #ConnectingMinds grant to the project Caring Robots // Robotic Care (CM 100-N). We would like to thank Michael Starzinger for his assistance in annotating the data, and Muhammad Saleemi, who assisted with the editing of the transcripts. Furthermore, we would like to express our deep appreciation for the work of Matthias Hirschmanner, Helena Anna Frijns, and Evelyn Mayer-Haas, who were involved in the design of the LLM-based documentation support system and the data collection. The vector icons in Figure 1 are by Soco St in CC Attribution License via SVG Repo.

REFERENCES

ai4Privacy (2023), 'Pii-masking-200k (Revision 1d4c0a1)'.

Altman, I. (1975), 'The environment and social behavior: Privacy, personal space, territory, and crowding.'.

Bail, K., Gibson, D., Acharya, P., Blackburn, J., Kaak, V., Kozlovskaia, M., Turner, M. & Redley, B. (2022), 'Using health information technology in residential aged care homes: An integrative review to identify service and quality outcomes', International journal of medical informatics 165, 104824.

Bain, M., Huh, J., Han, T. & Zisserman, A. (2023), 'WhisperX: Time-accurate speech transcription of long-form audio', INTERSPEECH 2023 .

Brandeis, L. & Warren, S. (1890), 'The right to privacy', Harvard law review 4(5), 193–220.

Burgoon, J. K. (1982), 'Privacy and communication', Annals of the International Communication Association 6(1), 206–249.

Douzas, G., Lechleitner, M. & Bacao, F. (2022), 'Improving the quality of predictive models in small data GSDOT: A new algorithm for generating synthetic data', Plos one 17(4), e0265626.

Figueiredo, M., Teixeira, L. & Paúl, C. (2019), 'StressadaMente: Mental health promotion program for direct care workers of older people', SAGE Open Medicine 7, 2050312119834116.

Friedman, B., Kahn, P. & Borning, A. (2002), 'Value sensitive design: Theory and methods', University of Washington technical report 2(8), 1–8.

Frijns, H. A., Vetter, R., Hirschmanner, M., Grabler, R., Vogel, L. & Koeszegi, S. T. (2024), Co-design of robotic technology with care home residents and care workers, in 'Proceedings of the 17th International Conference on Pervasive Technologies Related to Assistive Environments', Petra '24.

Grabler, R., Hirschmanner, M., Frijns, H. A. & Koeszegi, S. T. (2024), 'Privacy Agents: Utilizing Large Language Models to Safeguard Contextual Integrity in Elderly Care', Privacy-Aware Robotics Workshop, HRI 2024 .

Grabler, R., Starzinger, M., Hirschmanner, M. & Frijns, H. A. (2025), 'Privacy-sensitive conversations between care workers and care home residents in a residential care home (Dataset, Version 1.0.0)'.

Hirschmanner, M., Grabler, R., Frijns, H. A., Mayer-Haas, E. & Vincze, M. (2024), 'Prototype of a Care Documentation Support System Using Audio Recordings of Care Actions and Large Language Models'.

Höglander, J., Holmström, I. K., Lövenmark, A., Van Dulmen, S., Eide, H. & Sundler, A. J. (2023), 'Registered nurse–patient communication research: An integrative review for future directions in nursing research', Journal of advanced nursing 79(2), 539–562.

icky/translate (2024), 'Icky/translate'. URL: https://ollama.com/icky/translate

Jordon, J., Szpruch, L., Houssiau, F., Bottarelli, M., Cherubin, G., Maple, C., Cohen, S. N. & Weller, A. (2022), 'Synthetic Data–what, why and how?', arXiv preprint arXiv:2205.03257 .

Li, M., Bickersteth, W., Tang, N., Hong, J., Cranor, L., Shen, H. & Heidari, H. (2025), 'A Closer Look at the Existing Risks of Generative AI: Mapping the Who, What, and How of Real-World Incidents', arXiv preprint arXiv:2505.22073 .

Macdonald, L. M. (2016), 'Expertise in everyday nurse–patient conversations: The importance of small talk', Global qualitative nursing research 3, 2333393616643201.

Meta AI (2024), 'Llama 3.1-70B'. URL: https://huggingface.co/meta-llama/Llama-3.1-70B

Nissenbaum, H. (2004), 'Privacy as contextual integrity', Wash. L. Rev. 79, 119.

Orlikowski, W. J. (2007), 'Sociomaterial practices: Exploring technology at work', Organization studies 28(9), 1435–1448.

Shvartzshnaider, Y., Apthorpe, N., Feamster, N. & Nissenbaum, H. (2019), Going against the (appropriate) flow: A contextual integrity approach to privacy policy analysis, in 'Proceedings of the AAAI Conference on Human Computation and Crowdsourcing', Vol. 7, pp. 162–170.

Sundler, A. J., Hjertberg, F., Keri, H. & Holmström, I. K. (2020), 'Attributes of person-centred communication: A qualitative exploration of communication with older persons in home health care', International Journal of Older People Nursing 15(1), e12284.

van Wynsberghe, A. (2013), 'Designing Robots for Care: Care Centered Value-Sensitive Design', Science and Engineering Ethics 19(2), 407–433.

Wen, J., Su, A., Wang, X., Xu, H., Ma, J., Chen, K., Ge, X., Xu, Z. & Lv, Z. (2024), 'Virtual sample generation for small sample learning: A survey, recent developments and future prospects', Neurocomputing p. 128934.

Wilson, C. B., Davies, SUE. & Nolan, M. (2009), 'Developing personal relationships in care homes: Realising the contributions of staff, residents and family members', Ageing & Society 29(7), 1041–1063.