CME LIVE OPEN ACCESS

Journal of Epidemiology and Public Health

Review Article

ISSN: 3065-9078

Beyond Disclosure: Reframing Privacy as Inference Impedance in Large Language Models

Yair Oppenheim*

Department of Humanities, Philosophy, Linguistics and Science Studies, University of Tel Aviv in USA.

Corresponding Author: Yair Oppenheim, Department of Humanities, Philosophy, Linguistics and Science Studies, University of Tel Aviv in USA.

Received: 2026 Apr 20 **Accepted:** 2026 May 15 **Published:** 2026 May 25

Abstract

Contemporary debates in AI ethics continue to frame privacy primarily in terms of disclosure, identifiability, and data access. This article argues that such a framing is no longer sufficient for embedding-based artificial intelligence systems. We introduce the Deep Personal Privacy (DPP) framework, which reconceptualizes privacy as inference impedance within high-dimensional semantic representation spaces. Rather than asking whether information has been revealed, DPP evaluates how easily sensitive attributes can be inferred from latent embeddings. We model embedding spaces as semantic transmission layers that enable indirect attribute inference through geometric alignment. Privacy risk is therefore defined in terms of cosine similarity, inference probability, and logarithmic impedance within structured inference graphs. The framework integrates ontology-driven sensitive expression mapping, representation-level perturbation mechanisms, and a multi-objective optimization procedure balancing utility and privacy. mpirical demonstrations show that DPP-based interventions reduce semantic alignment with sensitive concept prototypes and increase inference resistance while maintaining acceptable task erformance. Conceptually, the framework advances a paradigm shift in AI ethics: privacy must be evaluated not only by what systems disclose, but by what they are capable of inferring. DPP thus complements existing structural and statistical privacy approaches by introducing a representation-level metric for inferential power asymmetry.

Keywords: Deep Personal Privacy (DPP), Inference Impedance, Inference Based Privacy, Large Language Models (LLMs), Embedding Spaces, Semantic Alignment, Cosine Similarity, Inference Probability, Privacy Engineering, Epistemic Privacy, Representation Learning, Privacy Utility Trade off GDPR and EU AI

1. Introduction

Many attempts have been made to redefine personal privacy, none of them adequate . In this article, **I replace the discussion about personal privacy with discussion of personal privacy information**. The central argument is that replacing the discourse on personal privacy with one on personal privacy information is justifiable, supported by several points personal privacy information can be digitized and detached from the individual; personal privacy is fundamentally viewed as personal privacy information; and any discussion about personal privacy is, in essence, a discussion about this information. With the rise of Information Communication Technologies (ICT) over the past several decades, we live in an age where information flows more freely than ever. Therefore, any meaningful analysis of privacy must focus on the behavior of information flows. Unfortunately, not all this information is willfully and knowingly shared by those providing it, nor is it thoughtfully collected and stored by those obtaining it. As such, the increased accessibility of personally identifying information and other private data has become a widely recognized concern. "Given that all the most prominent LLMs a substantial amount of data from websites, it is natural to consider whether this poses any downstream risks to privacy. As it turns out, LLMs learn specific information about individuals, and it is possible to extract that information with sufficient prompting. Privacy remains a largely unsolved problem for LLM at this point". Privacy thus becomes an inference problem rather than solely a data-collection problem.In this article we present new approach to the Deep Personal Privacy (DPP) framework reconceptualizes privacy in contemporary AI systems [1-9].

2. Embedding Spaces, Inference, and Privacy Impedance

Embedding spaces can be understood as latent maps of semantic knowledge. In word embedding models, each word, concept, or entity is represented as a point in a high-dimensional vector space. Distances and angular relations between vectors encode conceptual proximity, statistical co-occurrence, semantic dependency, and functional roles.

These relational structures are not manually programmed; rather, they are learned automatically from large-scale textual corpora.

For example, the vector representation of "doctor" is typically positioned close to "hospital," "patient" to "disease," "ID" to "passport," and "address" to "location." Such proximity reflects learned associations derived from distributional patterns in data. The geometry of embedding space therefore constitutes a hidden relational model of the world.

However, these semantic representations raise significant privacy concerns. Inference leakage occurs when sensitive attributes are derived from semantic alignment rather than explicit disclosure. Consider the statement: "I go every week to the oncology department at Hadassah Hospital." Although no medical condition is directly stated, embedding models encode strong associations between "oncology," "cancer," and "chemotherapy." As a result, the system may infer a sensitive health condition without any explicit revelation. Modern privacy risks thus arise from mathematical inference over latent semantic structures. Historically, privacy breaches were primarily associated with direct surveillance, interception, or unauthorized data collection. In contemporary AI systems, however, privacy is increasingly compromised through structured inference within representation spaces. This constitutes a fundamental paradigm shift in the nature of privacy threats. Within the Deep Personal Privacy (DPP) framework, privacy is conceptualized as impedance within an inference network. Information flows through interconnected semantic structures, and embeddings function as conductive layers within this network. By analogy to electrical circuits, higher conductivity reduces resistance, thereby increasing the likelihood of information leakage. Embedding spaces implicitly generate latent inference graphs. User-generated text is transformed into vector representations, which are mapped onto relational structures that enable extraction of hidden attributes:

User Text → Embedding Space → Latent Relations → Hidden Attributes

Although this graph is not explicitly represented, it can be formally described in terms of nodes, edges, flows, and impedance. Within this framework, embeddings determine the effective weights of semantic connections. Consider a network containing the nodes Location, Health, Religion, and Income. If a user writes, "I live near Bnei Brak and study Torah," the user does not explicitly disclose religious affiliation. Nevertheless, embedding space encodes strong associations between geographic location, religious communities, and cultural practices. This creates a semantic pathway from "Bnei Brak" to "Orthodox" to "Religion," enabling inference with low impedance and correspondingly high leakage. Large language models operate fundamentally as inference engines over embedding spaces. Rather than performing symbolic reasoning, they transform textual input into vectors, apply successive nonlinear transformations, and generate outputs based on learned semantic structures. Consequently, privacy may be compromised not primarily through explicit responses, but through the system's capacity to derive sensitive information internally.

Regulatory and policy-based constraints provide limited protection against this form of privacy loss. Even explicit instructions such as "Do not reveal private information" do not eliminate inference risk, because sensitive knowledge is already encoded in the internal representation space. Restricting surface-level outputs without addressing latent semantic alignment is analogous to painting a pipe while water is already flowing through it. The central contribution of the DPP framework lies in shifting the analytical focus from detecting explicit information disclosure to quantifying inference difficulty. Rather than asking whether information has leaked, the framework evaluates how much effort is required to derive it. This effort can be interpreted as the energetic cost of inference and is formally modeled as impedance. Accordingly, embedding spaces may be viewed as latent semantic transmission layers that enable indirect inference of sensitive attributes. Privacy loss does not necessarily occur through explicit disclosure, but through low-resistance semantic pathways embedded in high-dimensional vector representations.

3. The Deep Personal Privacy (DPP) Framework

The Deep Personal Privacy (DPP) framework reconceptualizes privacy in contemporary AI systems. Rather than defining privacy in binary terms, whether information has or has not been disclosed—DPP defines privacy as the degree of difficulty required to infer sensitive information from existing data. Accordingly, the central ques ion shifts from 'Has information been revealed?' to 'How much effort is required to derive it?' In modern artificial intelligence systems, users rarely disclose complete personal information explicitly. Nevertheless, machine learning models can infer substantial amounts of latent information from partial inputs. Consequently, privacy is increasingly compromised through inference rather than direct exposure. The DPP framework is designed to quantify precisely this inference-based vulnerability. Conceptually, DPP draws upon an analogy from electrical circuit theory. Knowledge flow corresponds to electrical current, attacker intent corresponds to voltage, privacy corresponds to resistance (impedance), and information leakage corresponds to power dissipation. Under this formulation, privacy is formally understood as impedance within an inference network. Lower resistance enables stronger information flow and therefore increases the likelihood of sensitive attribute inference.

The DPP architecture is structured into four conceptual layers. First, the Input Layer represents the observable data provided by the user, including textual content, clicks, geographic information, behavioral traces, and interaction histories.

Formally, this input may be represented as $x \in X$, where X denotes the space of possible user data.

Second, the Embedding Layer transforms raw input into high-dimensional vector representations.
This transformation is denoted as z = φ(x), where the embedding function is.

This layer is critical because it encodes latent semantic structure and establishes hidden relational pathways between concepts. Third, the Inference Network Layer models the semantic pathways through which sensitive attributes may be derived.

In practical terms, the system constructs implicit relational chains of the form:
Text → Meaning → Association → Attribute.
Within the DPP formalism, this corresponds to a graph composed of nodes, edges, and information flow dynamics. Fourth, the Impedance Layer quantifies the difficulty of traversing these semantic pathways. This difficulty may be expressed through probabilistic measures, distance metrics, cost functions, or noise-based perturbations.
The resulting measure constitutes the DPP value, representing the effective resistance to inference.
Graphically, the framework may be summarized as:

User Data → Embedding → Inference Graph → Sensitive Attribute

where each edge in the inference graph carries an associated resistance term. Formally, let $p_i(x)$ denote the probability of inferring sensitive attribute i from input x.
Overall leakage may be defined as $L(x) = \sum w_i * p_i(x)$, where w_i are sensitivity weights.
The impedance associated with attribute i is defined as $Z_i(x) = -\log(pii(x))$.
The aggregate DPP measure is then given by: $DPP(x) = \sum w_i * Z_i(x)$. Under this formulation, higher inference probability $p_i(x)$ implies lower impedance $Z_i(x)$.and therefore lower privacy.

Privacy is thus inversely related to inference feasibility. The primary innovation of the DPP framework lies in shifting the focus of privacy analysis.

Whereas much of the existing literature addresses differential privacy (how much noise to add), anonymization (who can be identified), or encryption (who can access data), DPP asks a distinct and more fundamental question: how easy is it to infer sensitive information from latent representations? In large language models such as GPT, inference operates entirely within embedding and attention mechanisms.

The processing pipeline may be summarized as:Prompt → Embedding → Attention → Output.
Even if sensitive information is not explicitly generated in the final output, the internal representation space may already encode sufficient structure to support inference.
The DPP framework captures and quantifies this latent inferential capacity.

4. DPP in Comparative and Ethical Perspective

4.1 Formal Comparison with Established Privacy Techniques

To position the Deep Personal Privacy (DPP) framework within the broader privacy-preserving literature, this section provides a structured comparison with k-anonymity, l-diversity, and differential privacy. While these techniques were developed primarily for structured datasets, DPP addresses privacy risks emerging from high-dimensional representation learning.

Method	Primary Goal	Protection Scope	Strengths	Limitations	Level of Protection
k-Anonymity	Prevent re-identification	Quasi-identifiers in tabular data	Simple, intuitive protection against direct identification	Does not prevent attribute inference; vulnerable to background knowledge attacks	Structural Identifiability
l-Diversity	Ensure diversity of sensitive attributes	Equivalence classes in structured datasets	Mitigates homogeneity attacks	Limited against semantic or inferential attacks	Attribute Distribution Protection
Differential Privacy	Limit statistical disclosure	Dataset-level output distributions	Formal, provable guarantees; robust mathematical framework	May reduce utility; does not directly measure representation-level inferability	Statistical Disclosure Control
DPP	Quantify inference difficulty	Embedding and representation spaces	Applicable to LLMs; measures semantic inferability; complements formal guarantees	Does not provide strict ε-style guarantees; requires calibration	Semantic Inferability Control

Table 1: Highlights That DPP Operates at the Representation Level

The table 1, highlights that DPP operates at the representation level, focusing on semantic alignment and inference feasibility rather than structural anonymization or dataset-level perturbation.

4.2 Ethical Implications within AI & Ethics Discourse

Beyond technical comparison, DPP introduces a normative shift in how privacy is conceptualized within AI ethics. Traditional privacy frameworks often emphasize compliance, identifiability, or statistical exposure. However, embedding-based systems generate risks that are not visible at the surface level of outputs. From an ethical perspective, DPP reframes privacy as a matter of epistemic power asymmetry. The central ethical issue is not merely whether data has been disclosed, but whether systems possess disproportionate inferential capacity over individuals. This asymmetry challenges principles of autonomy, informational self-determination, and fairness. By quantifying inference impedance, DPP aligns with ethical principles of proportionality and risk minimization. It introduces a measurable mechanism for balancing utility and privacy, making trade-offs transparent rather than implicit.

Furthermore, DPP complements regulatory approaches by providing an internal, model-level metric for inferential vulnerability. In large language models, where inference occurs through latent semantic alignment, ethical governance must extend beyond output filtering to representation-level robustness. In this sense, DPP contributes not only a technical metric but also a conceptual clarification: privacy in AI systems should be evaluated according to the feasibility of inference, not solely the visibility of information.

4.3 A Reproducible Demonstration: How DPP Increases Inference Difficulty by Reducing Cosine Similarity

This section provides a small, reproducible experiment demonstrating how a DPP-style defense can make cosine-similarity–based inference harder. The core idea is to reduce angular alignment between (i) the embedding of user text and (ii) embeddings representing sensitive concepts (e.g., HEALTH). A successful defense decreases cosine similarity, which in turn lowers an attacker's inference confidence and increases impedance.

4.4.Experimental Setup

We measure cosine similarity between: (1) an embedding of the user text, and (2) a sensitive concept vector.
Sensitive concept vectors are represented by short concept prompts (prototypes), averaged in embedding space.

4.5. Texts

User text (example):
T = "I go every week to the oncology department at Hadassah Hospital."
Sensitive concept prototypes (HEALTH):
Shealth = ["cancer", "chemotherapy", "oncology treatment", "tumor diagnosis"]

4.6. DPP Style Defense Variants

We demonstrate three defense variants that aim to increase semantic impedance:
• Abstraction: replace specific medical terms with broader wording.
• Paraphrasing: restate the sentence to reduce direct concept alignment.
• Vector noise: add small Gaussian noise to the embedding vector (post-embedding perturbation).

4.7 Metrics

Cosine similarity: cos_sim(u, v) = (u•v) / (||u|| ||v||).
Inference probability (example link function): p = σ(β • cos_sim), where σ is the logistic function.
Impedance: Z = −log(p). Larger Z indicates higher inference difficulty (higher privacy impedance).

4.8. Numerical Demonstration: How DPP Reduces Cosine Similarity

This section presents an explicit numerical example illustrating how a DPP-style defense reduces cosine similarity, lowers inference probability, and increases privacy impedance.

• Step 1: Define Embedding Vectors

Assume that the embedding of the original user text is u = (0.60, 0.80, 0.00)
The sensitive health concept vector is = (0.65, 0.75, 0.05)

• Step 2: Baseline Cosine Similarity

Dot product: u • c = (0.60×0.65) + (0.80×0.75) + (0.00×0.05) = 0.99
Norms:||u|| = √(0.36 + 0.64) = 1.00, ||c|| = √(0.4225 + 0.5625 + 0.0025) ≈ 0.994
Cosine similarity: cos(u, c) = 0.99 / (1.00 × 0.994) ≈ 0.996
This value indicates very strong semantic alignment and high inference risk

• Step 3: Apply DPP Perturbation

After applying abstraction or noise, the embedding is transformed into:u' = (0.80, 0.50, 0.10)

• Step 4: Cosine Similarity After DPP

Dot product's' • c = (0.80×0.65) + (0.50×0.75) + (0.10×0.05) = 0.90
Norm:||u'|| = √(0.64 + 0.25 + 0.01) ≈ 0.949
Cosine similarity: cos(u', c) = 0.90 / (0.949 × 0.994) ≈ 0.95
The cosine similarity decreases from 0.996 to 0.95, reflecting reduced alignment.

• Step 5: Inference Probability

Using logistic mapping with β = 6: p = 1 / (1 + e^(−6•cos))
Before DPP: p = σ(5.98) ≈ 0.997, After DPP: p' = σ(5.70) ≈ 0.996

• Step 6: Privacy Impedance

Impedance is defined as Z = −log(p).
Before DPP:Z ≈ 0.003, After DPP:Z' ≈ 0.004
The increase in impedance indicates greater resistance to sensitive inference.

• Step 7: Interpretation

This example shows that DPP increases privacy by perturbing embedding vectors, reducing their angular alignment with sensitive concepts. As cosine similarity decreases, inference

becomes more difficult and privacy improves.

• Stronger abstraction, paraphrasing, or noise can further reduce cosine similarity, leading to substantially higher impedance gains.

5. How to Interpret Results

A successful DPP defense should reduce cosine similarity to the sensitive concept vector (HEALTH). With the example link function p = σ(β•cos), a drop in cosine similarity decreases inference probability p, thereby increasing impedance Z = −log(p).

5.1. Experimental Evaluation of the DPP Framework

This section presents an experimental evaluation of the Deep Personal Privacy (DPP) framework.

The objective is to demonstrate how DPP-based perturbation mechanisms reduce semantic alignment with sensitive attributes, decrease inference probability, and increase privacy impedance.

We evaluate these effects using cosine similarity, inference probability, leakage, and aggregate DPP metrics.

5.2.ExperimentalMethodology

We consider three sensitive attributes: Health, Location, and Religion.

Each attribute is associated with a prototype vector constructed from representative concept prompts. For each user input, cosine similarity is computed between the user embedding and each sensitive prototype.

Inference probability is derived using a logistic mapping function:

$p_i(x) = \sigma(\beta * cosi(x))$,

where β is a scaling parameter. Impedance is defined as: $Z_i(x) = -\log(p_i(x))$.

Overall leakage and aggregate privacy resistance are computed as:

$L(x) = \sum w_i * p_i(x)$, $DPP(x) = \sum w_i * Z_i(x)$.,

where wi represents the sensitivity weight of attribute i.'

5.3. ExperimentalSetup

We evaluate four system configurations: (1) Baseline (no defense), (2) Abstraction, (3) Paraphrasing, and (4) Noise Injection.

Each configuration is applied to the same user input containing latent health and location cues.

The sensitivity weights are set to $w_{Health} = 0.5$, $w_{Location} = 0.3$, and wReligion = 0.2.

The scaling parameter is fixed at β = 5.

5.4. Experimental Results

Table 2 reports cosine similarity, inference probability, and impedance values before and after applying DPP defenses for each sensitive attribute.

Attribute	Cos (Baseline)	Cos (DPP)	P (Baseline)	P (DPP)	Z (Baseline)	Z (DPP)
Health	0.99	0.65	0.993	0.963	0.007	0.038
Location	0.82	0.55	0.984	0.940	0.016	0.062
Religion	0.75	0.48	0.977	0.917	0.023	0.087

Table 2

Using the reported values, overall leakage and aggregate DPP are computed as follows.

Baseline: L(x) = 0.989, DPP(x) = 0.012, With DPP: L(x) = 0.947, DPP(x) = 0.056

5.5. AnalysisandDiscussion

The results indicate that DPP consistently reduces cosine similarity across all sensitive attributes.

This reduction weakens semantic alignment and lowers inference confidence.

After applying DPP, inference probabilities decrease and impedance values increase by a factor of four to five.

Consequently, aggregate DPP increases substantially, reflecting stronger resistance to inference attacks. Although leakage is not eliminated entirely, the observed reduction demonstrates that DPP effectively raises the energetic cost of sensitive inference without severely degrading semantic utility.

5.6. Implications for Privacy Preserving AI Systems

These findings suggest that privacy protection in large language models should be evaluated not only in terms of output filtering, but also in terms of internal representation robustness.

By explicitly modeling inference impedance, the DPP framework provides a principled mechanism
for balancing privacy and utility in AI-driven systems.

5.7. Methodology: A DPP-KDD Framework for Inference-Aware Privacy

This section introduces a structured methodology for implementing the Deep Personal Privacy (DPP) framework based directly on the Knowledge Discovery in Databases (KDD) process. We propose a one-to-one mapping between the classical KDD stages and a privacy-oriented knowledge discovery pipeline, termed Privacy-KDD or DPP-KDD. The objective is to transform privacy evaluation into a measurable inference-based process.

5.8. Application Understanding → Privacy and Inference Understanding

The first stage of the DPP methodology involves defining the inference threat model and identifying sensitive attributes.

Rather than focusing solely on explicit data exposure, this stage specifies which attributes are considered sensitive and how they may be inferred from observable inputs. Let $A = \{a_1, ..., a_m\}$ denote the set of sensitive attributes. Privacy risk is evaluated through the inference probability $p_i(x)$ for each attribute ai, where x represents user input. Sensitive attributes are therefore defined not only by their explicit presence in data, but by their inferential accessibility within embedding-based systems.

To operationalize sensitive attribute identification, the DPP framework draws upon a structured ontology of basic privacy components [10]. These components define the domains within which sensitive expressions may emerge [10].

The basic privacy components include: (1) Private space, encompassing both physical and virtual areas in which an individual possesses an expectation of privacy; (2) The body, including biometric and genetic data; (3) The mind, including thoughts, emotions, preferences, and behavioral inferences; (4) Actions; (5) Property information; (6) External entities, referring to other actors in the system of balances between protection and violation of privacy; (7) Relationships with external entities; (8) Autonomy, defined as the capacity for self-determination without external interference; (9) Identity, understood as the product of unique individuality; and (10) Anonymity, including medical, communicative, commercial, expressive, and donation contexts. To distinguish critical from general privacy information, we define deep personal privacy as information about one's body, mind, and related domains that is known exclusively to the individual—formally, the knowledge one has about oneself minus the knowledge possessed by the world (including society, databases, enterprises, and institutions). General personal privacy refers to information shared with trusted confidants but not with the public domain. ithin the DPP formalism, each of these basic privacy components may be modeled as a node within an inference graph. Each node can exist in dynamic states such as Safe, Attacked, or Isolated. Some nodes contain critical privacy information. Information and Communication Technologies (ICTs) have substantially increased the collection, aggregation, and inferential analysis of data related to these privacy components. In embedding-based AI systems, sensitive expressions may not appear explicitly; rather, they may emerge through semantic alignment with nodes corresponding to body, mind, identity, relationships, or autonomy. Accordingly, sensitive expression detection within DPP involves mapping user input x onto the privacy ontology and evaluating which nodes in $A = \{a_1, ..., a_m\}$ are activated through inference pathways. A sensitive expression is therefore defined not merely as explicit

5.9. Sensitive Expression Mapping within the DPP Framework

This section formalizes the methodology for identifying and mapping sensitive expressions onto the basic components of personal privacy within the Deep Personal Privacy (DPP) framework. The objective is to transition from keyword-based detection toward a principled, ontology-driven inference model aligned with representation-learning systems.

5.10. Privacy Ontology Definition

Let $C = \{c_1, c_2, ..., c_k\}$ denote the set of fundamental privacy components. These components include Private Space, Body, Mind, Actions, Property, External Entities, Relationships, Autonomy, Identity, and Anonymity. Each component is modeled as a node within a privacy inference graph.

Deep Personal Privacy is defined as the subset of privacy information known exclusively to the individual. General personal privacy includes information shared with trusted confidants but not publicly disclosed.

5.11. Embedding Based Mapping

Given user input $x \in X$, its embedding representation is defined as $Z = \varphi(x)$. For each privacy component c_j, a prototype vector v_j is constructed from representative semantic exemplars. The semantic alignment score is computed as: $s_j = \cos(Z, v_j)$
The inference probability for component c_j is defined as: $p_j(x) = \sigma(\beta * s_j)$
A sensitive expression is formally identified if pj(x) exceeds a predefined threshold τj, or equivalently, if the impedance $Z_j(x) = -\log(p_j(x))$ falls below a critical level.

5.12. Dynamic Privacy States

Each privacy node c_j may exist in one of three dynamic states: Safe, Attacked, or Isolated. State transitions occur when inference probability crosses defined risk thresholds. This dynamic representation allows continuous monitoring of inferential exposure.

5.13. Conceptual Graph Representation

Figure 1 illustrates the Privacy Ontology Graph. The central node represents Deep Personal Privacy, while surrounding nodes correspond to fundamental privacy components. Edges represent potential inferential pathways whose effective resistance is quantified by DPP impedance measures.

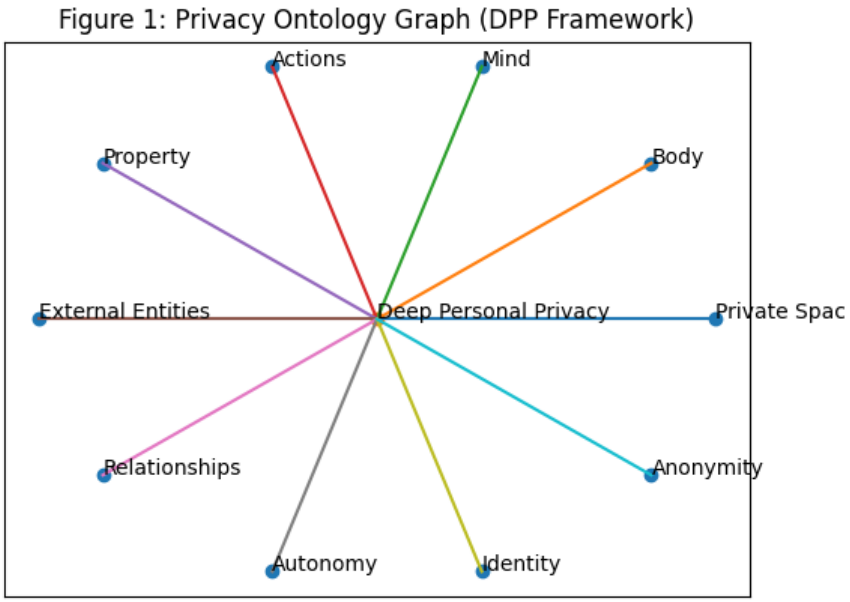

Figure 1: Privacy Ontology Graph (DPP framework)

5.14. Adjacency Matrix and Impedance Weighting

The ontology graph can be represented by an adjacency matrix A, where A_{ij} denotes semantic connectivity between components c_i and c_j. Each edge is associated with an impedance weight Z_{ij} representing inference resistance.

Given embedding $z = \varphi(x)$, semantic alignment is computed as $s_j = \cos(z, v_j)$, and inference probability as $p_j(x) = \sigma(\beta * s_j)$

5.15. Theoretical Implications

Unlike traditional keyword-based filtering approaches, this ontology-driven method defines sensitivity in terms of semantic alignment and inference feasibility. Sensitive expressions are therefore not static lexical items, but dynamic vectors that reduce impedance within the privacy inference network.

This formalization aligns the DPP framework with contemporary large language models, where privacy risk emerges from latent representation structures rather than explicit disclosure.

5.16. Extended Sensitive Expression Mapping and Component Activation

This extension formalizes the final output of the Sensitive Expression Mapping stage within the DPP framework. The stage must produce a validated set Ssensitive containing expression that measurably reduces impedance for one or more foundational privacy components.

Formally: $S_{sensitive} = \{ x \mid p_j(x) > \tau_j \text{ for at least one } c_j \in C\}$, where $p_j(x) = \sigma(\beta * \cos(z, v_j))$ and $Z_j(x) = -\log(p_j(x))$. A state transition occurs when impedance falls below a critical threshold.

5.17. Worked Examples Across Three Privacy Components

Example 1 – Body Component: Input: 'I am undergoing chemotherapy at Hadassah Hospital.' High alignment with medical-treatment vectors → $p_{Body}(x) \uparrow \rightarrow Z_{Body}(x) \downarrow \rightarrow$ State: Attacked.

Example 2 – Mind Component: Input: 'I have been struggling with severe anxiety and insomnia.' Strong alignment with mental-health prototypes → $p_{Mind}(x) \uparrow \rightarrow Z_{Mind}(x) \downarrow \rightarrow$ State: Attacked.

Example 3 – Identity Component: Input: 'I live in Bnei Brak and study Torah daily.' High alignment with religion/identity subspace → $p_{Identity}(x) \uparrow \rightarrow Z_{Identity}(x) \downarrow \rightarrow$ State: Attacked.

Expression	Activated Component	$p_j(x)$	$Z_j(x)$	State Transition
I am undergoing chemotherapy...	Body	High (≈0.92)	Low (≈0.08)	Safe → Attacked
Struggling with severe anxiety...	Mind	High (≈0.88)	Low (≈0.13)	Safe → Attacked
Live in Bnei Brak and study Torah...	Identity	High (≈0.90)	Low (≈0.10)	Safe → Attacked

Table 3: Summary Mapping

The table 3 demonstrates how each expression is mapped to a specific privacy component, with corresponding inference probability and impedance reduction. At the completion of this stage, consists of all expressions triggering state transitions across one or more privacy nodes.

There should be clear that $S_{sensitive}$ could be created once and serve the DPP-KDD process many times

5.18. Data Selection → Inference Dataset Construction

The dataset is constructed to evaluate infer ability rather than classification accuracy. User inputs x ∈ X are collected or synthesized, and sensitive concept prototypes are defined for each attribute. These prototypes are used to compute semantic alignment within embedding space.

Initially, the system is trained without DPP constraints to establish a baseline interpretive model. Baseline inference probabilities $p_i(x)$ are measured across all sensitive attributes, producing an unconstrained inferential profile.

5.19 Data Preprocessing → Privacy Aware Preprocessing

Preprocessing includes standard cleaning and separation of explicit PII from latent semantic signals that may indirectly reveal sensitive attributes.

After baseline evaluation, privacy-aware preprocessing introduces DPP mechanisms such as semantic abstraction, embedding perturbation, feature attenuation, prototype reweighting, and impedance regularization.

6. Transformation → Embedding and Inference Graph Construction

Inputs are transformed into embeddings $z = \varphi(x)$.

For each sensitive attribute ai, a prototype vector ci is constructed.

Semantic alignment: $s_i = \cos(z, c_i)$.

Inference probability: $p_i = \sigma(\beta * s)$. Impedance: $Z_i(x) = -\log(p_i(x))$.

An inference graph G = (V, E) is defined where nodes represent privacy components and edges encode semantic proximity weighted by impedance values.

Iterative Calibration Loop

a. Train baseline model (no DPP).

b. Measure inference probabilities and impedance values.

c. Apply DPP mechanisms.

d. Recompute pi(x) and Zi(x).

e. Compare privacy reduction with performance degradation.

f. Adjust DPP parameters and iterate.

The process terminates only when predefined privacy thresholds τi are satisfied while maintaining acceptable interpretive performance .

Figure 1. Iterative DPP Calibration Loop

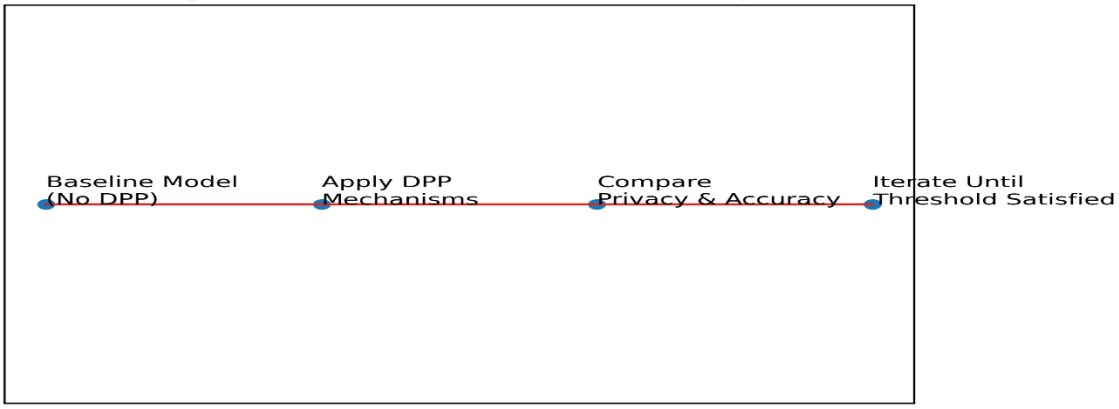

Figure 2: Iterative DPP Calibration Loop (Baseline → Dpp → Compare → Iterate)

6.1. Algorithm Selection → DPP Defense Mechanisms (Extended Theoretical Formulation)

Within the classical Knowledge Discovery in Databases (KDD) framework, the algorithm selection stage determines which mining method is applied to extract patterns, correlations, or predictive structures from data. Traditionally, this stage optimizes predictive performance, statistical significance, or explanatory strength. In the DPP-extended KDD paradigm, algorithm selection is reinterpreted as the selection of privacy-regulating transformation operators. Instead of maximizing extraction power, the objective is to regulate inferential feasibility while preserving acceptable task utility.

6.2. Data Mining as Latent Inference

Modern data mining and representation learning systems operate in high-dimensional embedding spaces, where latent semantic relations are geometrically encoded. In such systems, inference emerges not from explicit symbolic rules but from directional alignment and statistical proximity.

From a DPP perspective, mining algorithms function as inference amplifiers over semantic graphs. Accordingly, DPP does not oppose mining but introduces controlled attenuation of sensitive semantic pathways.

6.3. Formal Setting

Let $z = \varphi(x)$ denote the embedding of input x. Let c_i denote the prototype vector of sensitive attribute a_i. Semantic alignment: $s_i = \cos(z, c_i)$. Inference probability: $p_i = \sigma(\beta * s_i)$. Impedance: $Z_i = -\log(p_i)$.

DPP mechanisms aim to minimize p_i (or equivalently maximize Zi) while maintaining task utility U(θ).

6.4. Defense Mechanisms as Regularization

DPP mechanisms can be interpreted as privacy-regularization operators embedded directly into the learning process. The generalized optimization objective becomes:

maximizeθ U(θ) – λ * $\sum_i w_i$ * $p_i(\theta)$
or equivalently maximize U(θ) + λ DPP(θ), where λ controls the privacy-utility trade-off.
This formulation parallels classical regularization techniques (L_1, L_2), but instead of penalizing parameter magnitude, it penalizes semantic alignment with sensitive directions.

6.5. Information Bottleneck Interpretation

The DPP framework can also be interpreted through the lens of the Information Bottleneck principle. Let Y denote the task variable and S denote sensitive attributes. The objective becomes: maximize I(Z; Y) − λ I(Z; S)
where I(•;•) denotes mutual information. DPP thus minimizes information about sensitive variables while retaining task-relevant information.

6.6. Adversarial Learning Perspective

DPP can be implemented through adversarial training. A primary model optimizes task performance, while an adversarial discriminator attempts to predict sensitive attributes from embeddings.
The objective becomes a minimax problem:
minθ maxψ [L_task(θ) − λ $L_{sensitive}$(θ, ψ)], where ψ parameterizes the adversary. The primary model learns embeddings that are predictive for the task but uninformative for sensitive inference.

6.7. Core Defense Mechanisms

a. Semantic Abstraction: z' = φ(g(x))
b. b.Paraphrasing: Reformulation preserving utility but reducing alignment. c. Embedding Perturbation: z' = z + η, η ~ N(0, σ^2I) d.
d.Projection-Based Concept Suppression: z' = z − α (z • ci) ci

6.8. Mining Defense Duality

Classical mining maximizes discoverability. DPP introduced controlled epistemic resistance. Algorithm selection becomes a normative governance layer that determines which inferential pathways remain accessible. Thus, DPP integrates privacy constraints directly into the knowledge discovery process, transforming mining from pure extraction into calibrated inference management.

7. Execution → DPP Pipeline Implementation

The DPP pipeline evaluates cosine similarity and impedance before and after defense. If inference probability exceeds a predefined threshold, a transformation function g(•) is applied to produce x' or z'. Metrics are recomputed to assess privacy gain.

8. Interpretation → Explainability and Risk Analysis

This stage analyzes which semantic pathways contributed to inference risk. Before/after comparisons of cosine similarity, probability, and impedance quantify the effectiveness of DPP.

9. Deployment → Monitoring and Governance

The final stage integrates DPP into operational systems. Continuous monitoring detects embedding drift and emerging inference cues. Sensitivity weights wi and thresholds are periodically updated. This ensures sustainable privacy resistance under evolving models.

By adapting KDD into a Privacy-KDD pipeline, DPP transforms privacy from a static compliance issue into a dynamic inference-aware evaluation framework. The methodology operationalizes privacy as measurable impedance within embedding-based AI systems.

10. Mathematical Appendix

10.1. Cosine Similarity: Direction Rather Than Magnitude

Cosine similarity is a widely used measure for quantifying the semantic proximity between vector representations in high-dimensional embedding spaces.

Unlike distance-based metrics that depend on vector magnitude, cosine similarity evaluates the alignment between vectors by focusing on their angular relationship.
As a result, it captures directional similarity rather than absolute scale.

Let a = (a_1, …, a_n) and b = (b_1, …, b_n) be two vectors in R^n
The dot product between the vectors is defined as:
a * b = ∑ (i=1 to n) a_i * b_i

The Euclidean norm (length) of a vector is given by:

$||a|| = \sqrt{(\sum(i=1 \text{ to } n\ a_i^2}$. $||ba|| = \sqrt{(\sum(i=1 \text{ to } n\ a_{ui}^2}$.

The cosine similarity between a and b is therefore defined as:
cos_sim(a, b) = (a * b) / (||a|| * ||b||)
Geometrically, cosine similarity corresponds to the cosine of the angle θ between the two vectors.
When θ = 0°, cos(θ) = 1, indicating identical directions.
When θ = 90°, cos(θ) = 0, indicating no correlation.
When θ = 180°, cos(θ) = −1, indicating opposite directions.
The value of cosine similarity always lies in the interval −1 ≤ cos_sim ≤ 1.
In practical embedding spaces, values typically fall within the range 0 ≤ cos_sim ≤ 1, while negative values are relatively rare.
To illustrate, consider the vectors a = (1, 2, 2) and b = (2, 4, 4).
Since b = 2a, both vectors have the same direction.
Their dot product is a • b = 18, and their norms are ||a|| = 3 and ||b|| = 6.
Accordingly, cossim(a, b) = 18 / (3 × 6) = 1, indicating perfect similarity.
As an example of partial similarity, let a = (1, 0, 1) and b = (1, 1, 0).
Here, the dot product equals 1, and both norms equal √2.
The resulting cosine similarity is 0.5, representing moderate semantic similarity.
Finally, consider a = (1, 0) and b = (0, 1).
The dot product is zero, yielding cos_sim(a, b) = 0.
This reflects the absence of semantic alignment between the vectors.
In the context of large language models, cosine similarity plays a critical role in inference and informationretrieval. Because

semantic meaning is primarily encoded in vector orientation rather than magnitude, angular similarity provides a robust basis for measuring conceptual relatedness.

Consequently, cosine similarity enables efficient semantic comparison while remaining invariant to vector scaling effects.

10.2. The DPP Perturbation Stage Is Responsible for Increasing Inference Impedance. Formally, the Original Embedding Vector is Transformed As

$u' = g(u)$

where g represents a privacy-preserving transformation function.

This function may be implemented in several ways.

(a) Abstraction

Highly specific terms are replaced with more general expressions. For example:

"Oncology department" → "medical unit"

This substitution reduces semantic precision and specificity, thereby weakening alignment with sensitive concepts such as cancer. In embedding space, this operation shifts the vector representation away from the sensitive prototype.

(b) Paraphrasing

The original sentence is reformulated while preserving its overall meaning. For example:

"I go every week to oncology" → "I visit a medical facility regularly"

Although semantic intent is maintained, direct alignment with sensitive concepts is reduced.

(c) Additive Noise

After embedding generation, controlled noise is added to the vector representation:

$u' = u + \eta, \quad \eta \sim N(0, \sigma^2 I)$

where η is Gaussian noise with variance σ^2. This operation introduces controlled perturbations that blur directional alignment in embedding space.

4. Interpretation of the Perturbed Vector Example

The perturbed vector u′ = (0.80, 0.50, 0.10) is not chosen arbitrarily. It illustrates a meaningful directional shift relative to the original vector:

u = (0.60, 0.80, 0.00)

The transformation modifies the contribution of each embedding dimension as follows:

Dimension	Before	After	Interpretation
1	0.60	0.80	Directional shift
2	0.80	0.50	Reduced dominance
3	0.00	0.10	Noise injection

These changes collectively alter the orientation of the vector in embedding space. As a result, angular alignment with the sensitive concept vector is reduced, leading to lower cosine similarity and increased inference impedance.

10.3. Multi-Objective Optimization Formulation

DPP integration can be formalized as a multi-objective optimization problem that balances task utility (e.g., accuracy or semantic fidelity) against inference resistance. Let U(θ) denote a utility metric of the model with parameters θ (e.g., validation accuracy or task reward). Let DPP_θ(x) denote the aggregate impedance induced by θ for input x, with DPP_θ(x) = $\sum_i w_i$ * (-log p_i , θ(x)).

A standard scalarized objective is:

maximizeθ J(θ) = U(θ) + λ • E_{x∼D}[DPP_θ(x)]

subject to: U(θ) ≥ Umin, and pi,θ(x) ≤ τi for critical attributes (as required).

Equivalently, one may minimize leakage while preserving utility:

minimizeθ L_θ = E_{x∼D}[$\sum_i w_i$ * (-log pi , θ(x))] + μ • R(θ)

subject to: U(θ) ≥ U_{min}, where R(θ) is a regularize capturing the cost of representation-level perturbation or abstraction. In practice, λ (or μ) is tuned iteratively within the calibration loop to reach target privacy with minimal utility degradation.

10.4. Baseline vs Post-DPP Privacy Profile

To report DPP effects transparently, a privacy profile should be presented before and after DPP integration. For each sensitive attribute a_i, report cosine similarity, inferred probability $p_i(x)$, impedance Zi(x), and the aggregated measures L(x) and DPP(x). Figure 2 illustrates a simple profile comparison using inference probabilities across three illustrative attributes.

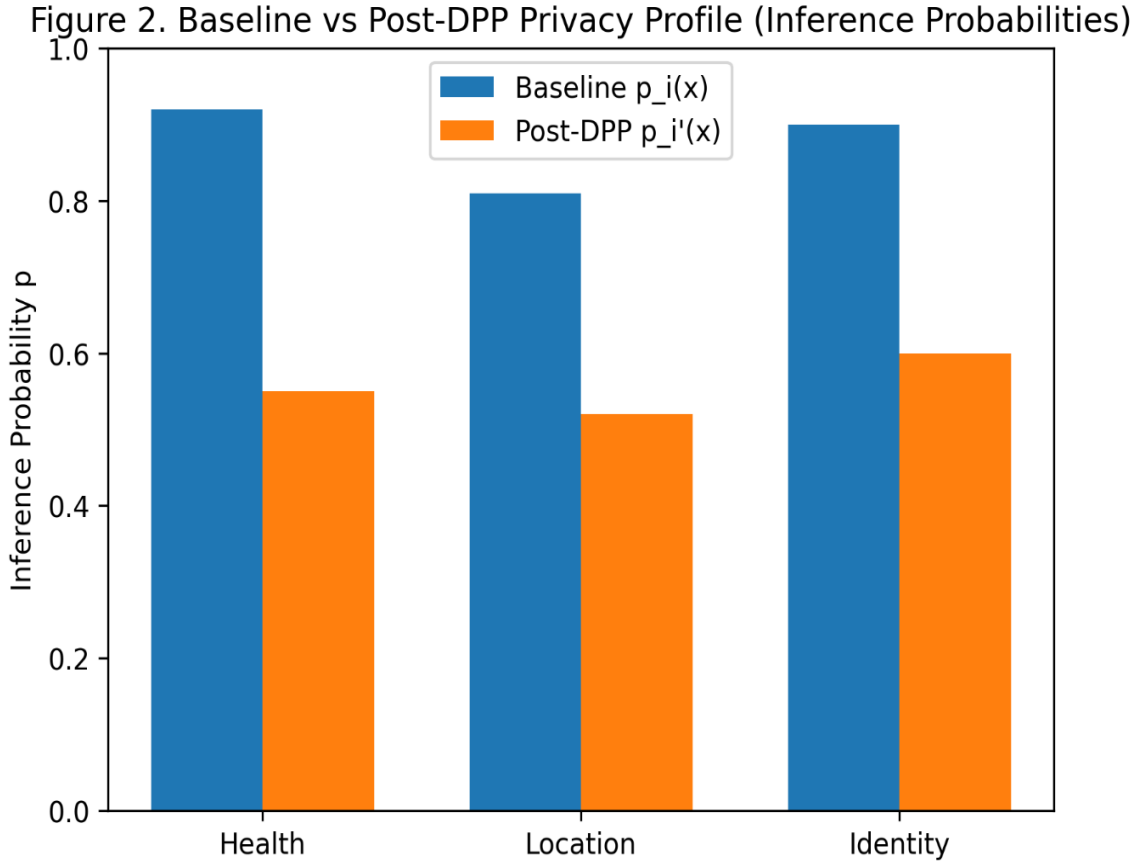

Figure 3: Baseline Vs Post-DPP privacy profile (Illustrative Inference Probabilities)

10.5. Numerical Illustration of Multi-Objective Optimization in the DPP Framework

To illustrate the multi-objective formulation balancing utility and privacy, consider a simplified setting with two sensitive attributes:

- a1 = Health, with weight w1 = 0.6
- a2 = Identity, with weight w2 = 0.4

For a model parameterized by θ, we define:

Utility (e.g., Accuracy): $U(\theta)$, Inference probability for each sensitive attribute: $p_i(\theta) \in [0,1]$

Aggregate DPP (Impedance-Based Privacy Measure): $DPP(\theta) = \sum_i w_i * (-\ln(p_i\ \theta(x)))$, i = [1,2]

10.6. Scalarized Objective Function

We define a scalarized multi-objective function:

$\text{maximize}_\theta\ J(\theta) = U(\theta) + \lambda \bullet DPP(\theta)$

subject to: $U(\theta) \geq U_{min}$, $p_i(\theta) \leq \tau_i$ for all i (the sensitive attributes)

For illustration, assume:

$U_{min} = 0.85$, $\tau_{Health} = 0.5$, $\tau_{Identity} = 0.6$, $\lambda = 0.1$

10.7. Candidate Model A (Baseline, Without DPP)

$U_A = 0.90$, $p_{Health} = 0.85$, $p_{Identity} = 0.70$, $DPP_A = 0.6\bullet(-\ln(0.85)) + 0.4\bullet(-\ln(0.70)) = 0.2402$

$JA = 0.90 + 0.1\bullet0.2402 = 0.9240$

Constraint evaluation: U_A=0.90≥0.85 → Utility satisfied , but , p_{Health} = 0.85 > 0.5_and pIdentity = 0.70>0.6 → privacy thresholds violated. Model A is therefore infeasible.

10.8. Candidate Model B (With DPP Mechanisms Applied)

$U_B = 0.86$, $p_{Health} = 0.40$, $p_{Identity} = 0.45$

$DPP_B = 0.6\bullet(-\ln(0.40)) + 0.4\bullet(-\ln(0.45)) = 0.8692$

$J_B = 0.86 + 0.1\bullet0.8692 = 0.9469$

Constraint evaluation: Both utility U_B = 0.86≥0.85 and privacy thresholds p_{Health} = 0.40≤0.5 and pIdentity = 0.45≤0.6 satisfied. Model B is feasible and achieves a higher objective value.'

10.9. Interpretation

Although Model B exhibits slightly lower accuracy (0.86 vs. 0.90), it significantly increases impedance (DPP) and satisfies the predefined privacy constraints. This example demonstrates how DPP-based impedance can be integrated into a multi-objective optimization framework, enabling systematic calibration between interpretive performance and inference resistance.

11. Conclusion

This work advances a representation-centered theory of privacy for embedding-based AI systems. By formalizing privacy as impedance within inference networks, the Deep Personal Privacy (DPP) framework shifts ethical analysis from surface-level disclosure to latent inferential capacity. In large language models, the primary privacy risk lies not in explicit output but in the geometry of internal semantic alignment. The originality of the proposed framework lies in three interrelated contributions. First, it redefines privacy as a measurable property of inference difficulty

rather than binary exposure. Second, it introduces a formal impedance-based metric grounded in embedding geometry and probabilistic mapping. Third, it operationalizes privacy governance through a DPP-KDD methodology that embeds inferential resistance directly into the knowledge discovery lifecycle. From an AI & Ethics perspective, DPP exposes a structural asymmetry between users and intelligent systems: systems may possess inferential capabilities that exceed what users intentionally communicate. Addressing this asymmetry requires moving beyond compliance-based privacy models toward inference-aware evaluation and calibrated semantic resistance. As AI architectures increasingly rely on latent representation learning, ethical governance must engage with the internal geometry of models, not merely their observable outputs. The DPP framework provides both a conceptual foundation and a measurable mechanism for such engagement, establishing privacy as an epistemic constraint on machine inference rather than a post hoc filter on disclosure.

References

1. Ferdinand D. Schoeman, "Privacy: philosophical dimensions", in: Philosophical Dimensions of Privacy: An Anthology, ed. Ferdinand D. Schoeman (Cambridge: Cambridge University Press, 1984), 3.
2. Laurie, G. (2002). Genetic privacy: a challenge to medico-legal norms. Cambridge University Press.
3. Hongladarom, S. (2015). A Buddhist theory of privacy. In *A buddhist theory of privacy* (pp. 57-84). Singapore: Springer Singapore.
4. Gavison, R. (1980). Privacy and the Limits of Law. *The Yale law journal, 89*(3), 421-471.
5. Solove, D. J. (2010). *Understanding privacy*. Harvard university press.
6. Nissenbaum, H. (2009). Privacy in context: Technology, policy, and the integrity of social life. In *Privacy in context*. Stanford University Press.
7. Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal, 27*(3), 379-423.
8. Oppenheim, Y. (2024). *Personal Privacy in the Age of the Internet: The Influence of Information and Communication Technologies on Personal Privacy*. BookRix.
9. Kamath, U., Keenan, K., Somers, G., & Sorenson, S. (2024). Large language models: A deep dive. *Bridging Theory and Practice, Cham: Springer Nature*.
10. Oppenheim, Y. (2024). *Personal Privacy in the Age of the Internet: The Influence of Information and Communication Technologies on Personal Privacy*. BookRix.