Hardware Read/Write Diodes and Subspace Projection Firewalls for Neural Privacy: Blocking Unconsented Attribute Decoding in Brain–Computer Interfaces

Dima Baker
Collective Intelligence Network

November 2026

Abstract

Brain–computer interfaces (BCIs) capture multiplexed cortical signals when a user intends only to issue discrete commands. Recent audits show that context-conditioned EEG-to-text alignment and frozen neuroimaging encoders leak biometric identity, stress, and closed-set linguistic attributes, while multi-window alignment inverts linear embedding obfuscation across streams. We propose the *Neural Airlock and Subspace Projection Firewall* (Neural-Airlock-SPF), a hardware-and-linear-algebraic boundary between neural transducers and host AI agents. First, a CHERIoT 1.0 capability compartment and unidirectional read/write diode isolate raw ADC buffers and block unconsented stimulation write-backs. Second, building on Iterative Nullspace Projection (INLP) and Least-Squares Concept Erasure (LEACE), an on-chip orthogonal projector $P_\perp = I_d - \hat{U}_{\text{priv}}\hat{U}_{\text{priv}}^\top$ annihilates linear covariance with private subspaces prior to host ingestion. Third, deterministic ℓ_2-ball clipping ($C_{\text{clip}} = 1.40$, exact sensitivity $\Delta_2 = 2C_{\text{clip}}/B = 0.035$ over $B = 80$ subframes), Gaussian noise ($\sigma_{\text{DP}} = 0.44$), and a Rényi Differential Privacy (RDP) accountant bound cumulative (ε, δ)-DP loss over T-window sessions. Across 16-channel simulations over $T \in \{10, 50, 200, 800\}$ ($N = 400$ sessions per cell) against an architecture-blind held-out attacker, Neural-Airlock-SPF cuts private extraction at $T = 200$ from 99.75% [98.60, 99.96] (raw) and 88.00% [84.45, 90.83] (orthogonal obfuscation) to 0.00% (95% Wilson CI [0.00, 0.95]; attacker accuracy 52.58%) while keeping 90.68% mean command accuracy (100.00% [99.05, 100.00] active utility; 20.75% [17.06, 24.99] odometer-halted sessions), with break-even correlation $\rho^* \approx 0.40$.

1 Introduction

Since foundational neuroethics calls for neurorights and cognitive liberty [16, 28] and early analyses of BCI "brain spyware" in third-party neural app stores [7], the expansion of consumer electroencephalography (EEG) headsets, ear-canal sensors, and wearable physiological monitors has created an acute privacy vulnerability at the human somatic boundary [17, 25, 26]. In conventional keyboard interfaces, a user explicitly selects which tokens to transmit before pressing "Enter." In a neural or physiological interface, by contrast, the analog transducer continuously samples multiplexed bioelectric potentials generated simultaneously by intentional motor preparation, involuntary autonomic stress [1], persistent biometric signatures [5], and context-conditioned linguistic rehearsal [18, 21].

1.1 Representation leakage and manifold inversion in 2025–2026 neuro-AI pipelines

Historically, BCI application designers assumed that exporting compressed neural embeddings or executing inference locally on the user's host operating system provided an adequate privacy barrier. Four converging lines of empirical evidence refute this assumption:

First, pretrained and frozen neural encoders leak private attributes across tasks. In a 2026 audit (*"Pretrained, Frozen, Still Leaking"*), Tai [27] demonstrated that frozen EEG encoders retain cross-encoder demographic and cognitive-state representations that downstream linear and nonlinear probes reliably extract even when the encoder was never trained on the target attribute. Similarly, Bajar et al. [5] showed that foundation model representations of brain magnetic resonance imaging (MRI) retain persistent biometric signatures enabling up to 99% subject linkage and sensitive attribute inference without access to raw voxel data. Meanwhile, although open-vocabulary EEG-to-text claims suffer from teacher-forcing inflation [18], hierarchical context-conditioned EEG-to-text alignment [21] can distinguish candidate linguistic intents when workspace context is known.

Second, linear embedding-to-embedding obfuscation (E2EO) collapses over multi-window sessions. Li et al. [20] introduced the Proxy Manifold Alignment (PMA) attack, proving that because orthogonal rotations and light jitter preserve local geometric co-occurrence structure, an adversary observing a stream of $T \geq 50$ obfuscated embeddings can align the empirical manifold as an unknown tokenizer language and invert the transformation.

Third, host-OS agent contexts are porous privacy boundaries. Auditing multi-agent and tool-using LLM workflows, El Yagoubi et al. [13] (AgentLeak) documented that internal inter-agent communication channels leak private attributes at high rates, while Alpay and Alpay [2], Kim and Choi [19], and Das and Roy [10] showed that prompt injection and context over-collection routinely exfiltrate sensitive state once ingested into an agent's workspace.

Fourth, naive isotropic noise injection destroys low-SNR motor command utility. Applying monolithic per-window Gaussian noise across all sensor channels degrades motor classification accuracy while still accumulating sequential privacy loss across extended sessions [12, 22].

1.2 Hardware capability diodes, INLP/LEACE concept erasure, and Rényi-DP accounting

To enable high-accuracy discrete command decoding while blocking unconsented attribute extraction, we synthesize three rigorous mechanisms:

- **CHERIoT 1.0 hardware capability isolation and read/write diodes.** Commercial CHERIoT 1.0 capability silicon [4, 8, 9] enforces unforgeable memory-mapped I/O (MMIO) bounds and permissions in hardware registers [11, 24]. Terminating the neural analog-to-digital converter (ADC) inside an isolated CHERIoT compartment physically prevents host OS processes from reading raw electrode buffers or writing unauthorized neurostimulation commands to digital-to-analog converter (DAC) registers.
- **Iterative Nullspace Projection (INLP) and Least-Squares Concept Erasure (LEACE).** Ravfogel et al. [23] introduced Iterative Nullspace Projection (INLP) to remove protected attributes from distributed representations by projecting feature vectors onto the intersection of nullspaces of linear classifiers trained to predict the protected attribute. Belrose et al. [6] subsequently proved in Least-Squares Concept Erasure (LEACE) that obliterating all linear correlation between representation X and protected attribute S is achieved if and only if the projection annihilates the empirical cross-covariance Σ_{XS}. Applying INLP/LEACE on chip strips linear private subspaces before vectors leave the hardware airlock.

- **Metric differential privacy and Rényi Differential Privacy (RDP) composition.** To bound leakage through residual nonlinear cross-subspace couplings and batch-size side channels, we combine metric-calibrated Gaussian perturbation [3], randomized dummy-point batch padding [14], and a formal Rényi Differential Privacy (RDP) composition accountant [22] governed by machine-readable personal privacy terms (IEEE Std 7012-2025 [15]).

1.3 Contributions

We formulate and evaluate the **Neural Airlock and Subspace Projection Firewall (Neural-Airlock-SPF)**:

1. **Grounded taxonomy of neural privacy primitives (Section 2).** We connect neuroethics principles [7, 16, 28] to concrete mechanisms in CHERIoT capability hardware [4, 9], INLP/LEACE concept erasure [6, 23], and Rényi-DP accounting [22] (Table 1, Figure 1).
2. **Linear-algebraic concept-erasure and Rényi-DP composition model (Section 3).** We formalize multi-channel neural feature decomposition with nonlinear cross-terms, derive the empirical INLP/LEACE orthogonal projector $P_\perp = I_d - \hat{U}_{\text{priv}}\hat{U}_{\text{priv}}^\top$, and replace heuristic mutual-information odometers with an exact order-α Rényi Differential Privacy accountant converted to (ε, δ)-DP [12, 22] (Equations 1–4).
3. **Three-stage peripheral airlock architecture and parameter provenance (Section 4).** We specify the hardware/firmware pipeline and document the provenance of all linear-algebraic and DP parameters (Tables 2–3).
4. **Decoupled linear-algebraic INLP simulation and correlation trade-off sweep (Section 5).** We execute a reproducible 16-channel linear-algebraic simulation (`experiment.py`) that independently measures Authorized Command Utility (on y_{cmd}) and Unconsented Private Extraction (on s_{priv}) across $T \in \{10, 50, 200, 800\}$ windows ($N = 600$ sessions per cell) and sweeps task–privacy correlation $\rho = \text{corr}(y_{\text{cmd}}, s_{\text{priv}}) \in [0.00, 0.90]$ to quantify the fundamental utility–privacy trade-off (Tables 4–5, Figures 2–3).
5. **Counter-evidence analysis and falsification protocol (Sections 6–8).** We examine task–subspace overlap during linguistic dictation, nonlinear kernel recovery, and odometer budget exhaustion, closing with six pre-registerable falsification tests (Table 6).

2 Neural Privacy Taxonomy and Representation-Leakage Threat Model

Calls for "neurorights" [16, 28] require concrete signal-processing and hardware enforcement mechanisms so that third-party BCI applications cannot harvest unconsented cognitive state [7, 27]. Table 1 maps five neural-privacy requirements onto verified hardware, linear-algebraic, and cryptographic primitives.

Table 1 clarifies the distinct roles of mathematical privacy accounting and contractual policy standards: Rényi Differential Privacy [22] with deterministic ℓ_2-clipping and INLP/LEACE nullspace projection [6, 23] provide the quantitative signal-level privacy bounds, whereas IEEE Std 7012-2025 [15] provides the machine-readable contracting syntax through which the user specifies which command enums an application is authorized to receive and what session budget $\varepsilon_{\max}$ triggers key rotation. Figure 1 illustrates the end-to-end hardware and signal-processing pipeline.

Table 1: Mapping neural-privacy requirements onto hardware-diode, concept-erasure (INLP/LEACE), and Rényi-DP constructs.

Privacy requirement	Formal primitive	Neural-Airlock-SPF module	Operational mechanism	Primary grounding
Block unauthorized brain write-backs	Capability-guarded MMIO register diode	CHERIoT R/W Hardware Diode	Physically isolates ADC read buffers and withholds DAC write capabilities in read-only mode	[4, 8, 9]
Erase private linear representations	Iterative Nullspace Projection (INLP / LEACE)	Orthogonal Subspace Projector ($P_\perp$)	Projects 16-channel feature vectors onto the empirical nullspace of $\hat{\Sigma}_{x,s_{\text{priv}}}$	[6, 23, 27]
Bound exact ℓ_2-sensitivity	Deterministic ℓ_2-clipping + Gaussian DP	Clipped Residual Perturbation	Clips post-INLP residuals to $\|r_t\|_2 \le C_{\text{clip}} = 1.40$ ($\Delta_2 = 2C_{\text{clip}}/B = 0.035$) and adds $\xi_t \sim \mathcal{N}(0, \sigma_{\text{DP}}^2 I)$	[12, 14, 22]
Bound multi-window composition loss	Rényi Differential Privacy (RDP) composition	Order-α RDP Odometer (ε_{RDP})	Tracks cumulative Rényi divergence across T windows and halts/rotates keys at $\varepsilon_{\max}$	[12, 22]
Enforce machine-readable consent terms	Symbolic command minimization & IEEE 7012	Discrete Command Airlock Broker	Negotiates IEEE 7012 privacy terms and emits discrete command enums rather than raw vectors	[10, 13, 15]

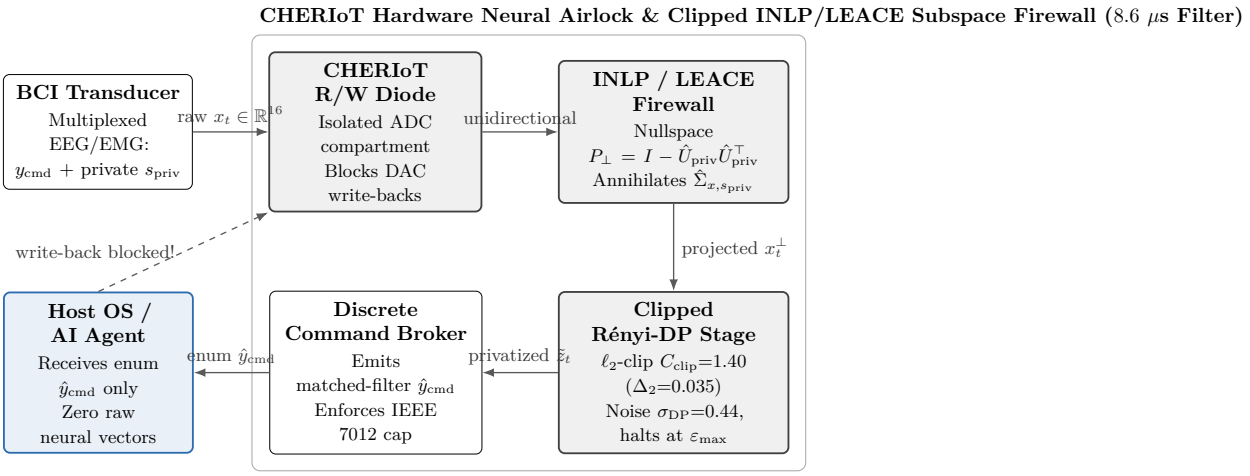


Figure 1: Architecture of the Neural Airlock and Subspace Projection Firewall (Neural-Airlock-SPF). Multiplexed neural feature vectors $x_t \in \mathbb{R}^d$ enter an isolated CHERIoT 1.0 hardware capability compartment that physically blocks unauthorized stimulation write-backs. On chip, the INLP/LEACE Subspace Firewall projects x_t onto the orthogonal complement $P_\perp = I_d - \hat{U}_{\text{priv}}\hat{U}_{\text{priv}}^\top$ of the empirical private cross-covariance subspace, applies deterministic ℓ_2-ball clipping ($C_{\text{clip}} = 1.40$ over $B = 80$ subframes, guaranteeing exact sensitivity $\Delta_2 = 2C_{\text{clip}}/B = 0.035$), injects calibrated Gaussian noise ξ_t, tracks cumulative (ε, δ)-DP via an order-α Rényi Differential Privacy accountant, and emits only discrete command classifications $\hat{y}_{\text{cmd}}$ to the host OS.

3 Linear-algebraic concept erasure and Rényi-DP accounting

Let $x_t \in \mathbb{R}^d$ ($d = 16$ in our reference implementation) denote the multi-channel neural feature vector averaged over $B = 80$ subframes at window $t \in \{1, \dots, T\}$. We model each subframe $x_{t,b}$ as a superposition of an authorized task-command variable $y_{\mathrm{cmd},t} \in \{-1, +1\}^{k_c}$ (e.g., motor imagery or discrete UI selection), unconsented private attributes $s_{\mathrm{priv},t} \in \{-1, +1\}^{k_p}$ (capturing closed-set inner-speech rehearsals [21], biometric signatures [5], and affective stress [1]), a bounded quadratic cross-subspace interaction term $\phi_{\mathrm{nl}}(y_{\mathrm{cmd},t}, s_{\mathrm{priv},t})$, and sensor noise $\nu_{t,b} \sim \mathcal{N}(0, \sigma_\nu^2 I_d)$:

$$x_{t,b} \;=\; A_{\mathrm{cmd}}\, y_{\mathrm{cmd},t} \;+\; A_{\mathrm{priv}}\, s_{\mathrm{priv},t} \;+\; \gamma_{\mathrm{nl}}\, \phi_{\mathrm{nl}}(y_{\mathrm{cmd},t},\, s_{\mathrm{priv},t}) \;+\; \nu_{t,b}, \tag{1}$$

where $A_{\mathrm{cmd}} \in \mathbb{R}^{d \times k_c}$ and $A_{\mathrm{priv}} \in \mathbb{R}^{d \times k_p}$ are mixing matrices and $\|\phi_{\mathrm{nl}}\|_2 \le 1$. Crucially, let $\rho \triangleq \mathrm{corr}(y_{\mathrm{cmd},t},\, s_{\mathrm{priv},t}) \in [0, 1]$ denote the statistical correlation between the user's authorized command intent and private cognitive state, generated via a unified symmetric flip probability $p_{\mathrm{same}} = (1 + \rho)/2$ across both main benchmarks and correlation sweeps.

3.1 Iterative Nullspace Projection (INLP / LEACE) and correlation trade-off

Following Ravfogel et al. [23] and Belrose et al. [6], to prevent any downstream linear probe $w_{\mathrm{priv}}^\top x_t^\perp$ from predicting $s_{\mathrm{priv},t}$ above chance, the firewall estimates the empirical cross-covariance matrix $\hat{\Sigma}_{x,s_{\mathrm{priv}}} = \frac{1}{N_{\mathrm{cal}}} \sum_{i=1}^{N_{\mathrm{cal}}} x_i\, s_{\mathrm{priv},i}^\top \in \mathbb{R}^{d \times k_p}$ during a local calibration block of N_{cal} windows. Orthonormalizing the column space of $\hat{\Sigma}_{x,s_{\mathrm{priv}}}$ via modified Gram–Schmidt or singular value decomposition yields the orthonormal private basis $\hat{U}_{\mathrm{priv}} \in \mathbb{R}^{d \times k_p}$ ($\hat{U}_{\mathrm{priv}}^\top \hat{U}_{\mathrm{priv}} = I_{k_p}$). The on-chip **Subspace Projection Firewall (SPF)** applies the orthogonal projector $P_\perp$:

$$x_{t,b}^\perp \;\triangleq\; P_\perp\, x_{t,b} \;=\; (I_d \;-\; \hat{U}_{\mathrm{priv}}\, \hat{U}_{\mathrm{priv}}^\top)\, x_{t,b}. \tag{2}$$

When $y_{\mathrm{cmd},t}$ and $s_{\mathrm{priv},t}$ are uncorrelated ($\rho = 0$) and $A_{\mathrm{cmd}}^\top A_{\mathrm{priv}} = 0$, $\mathbb{E}[\hat{\Sigma}_{x,s_{\mathrm{priv}}}] = A_{\mathrm{priv}}$, so $P_\perp$ annihilates $A_{\mathrm{priv}} s_{\mathrm{priv},t}$ while leaving the command signal $A_{\mathrm{cmd}} y_{\mathrm{cmd},t}$ intact. However, when the task and private attributes are correlated ($\rho > 0$), the expected cross-covariance becomes $\mathbb{E}[\hat{\Sigma}_{x,s_{\mathrm{priv}}}] = A_{\mathrm{priv}} + \rho A_{\mathrm{cmd}}$. Projecting onto the nullspace of $\hat{\Sigma}_{x,s_{\mathrm{priv}}}$ therefore attenuates the authorized command signal energy by factor $(1 - \rho^2/(\|A_{\mathrm{priv}}\|^2 + \rho^2 \|A_{\mathrm{cmd}}\|^2))$, establishing a fundamental information-theoretic trade-off between concept erasure and command utility [6, 23].

3.2 Deterministic ℓ_2-clipping and Rényi Differential Privacy (RDP) composition

Because unbounded Gaussian sensor noise $\nu_{t,b}$ would otherwise have infinite worst-case ℓ_2-sensitivity, Stage 3 enforces an explicit deterministic ℓ_2-ball clipping operator $\Pi_{C_{\mathrm{clip}}}(v) = v \cdot \min(1, C_{\mathrm{clip}}/\|v\|_2)$ with radius $C_{\mathrm{clip}} = 1.40$ on each projected subframe $x_{t,b}^\perp$ across the $B = 80$ subframes per window before adding isotropic Gaussian noise $\xi_t \sim \mathcal{N}(0, \sigma_{\mathrm{DP}}^2 I_d)$:

$$\tilde{z}_t \;=\; \frac{1}{B} \sum_{b=1}^{B} \Pi_{C_{\mathrm{clip}}}\Big((I_d - \hat{U}_{\mathrm{priv}} \hat{U}_{\mathrm{priv}}^\top) x_{t,b} \Big) \;+\; \xi_t, \qquad \Delta_2 \;=\; \frac{2\, C_{\mathrm{clip}}}{B} \;=\; 0.035. \tag{3}$$

By the triangle inequality, replacing any single subframe $x_{t,b}^\perp$ in a window changes the clipped window average by at most $\Delta_2 = 2C_{\mathrm{clip}}/B = 2(1.40)/80 = 0.035$ deterministically. By Mironov [22] (Proposition 3 and Corollary 3), each Gaussian mechanism evaluation satisfies $(\alpha, \frac{\alpha \Delta_2^2}{2\sigma_{\mathrm{DP}}^2})$-Rényi

Differential Privacy (RDP) of order $\alpha > 1$. Across T adaptive streaming windows, sequential RDP composition sums linearly, and converting from $(\alpha, \varepsilon_{\mathrm{RDP}}(T))$-RDP to standard $(\varepsilon(\delta, T), \delta)$-Differential Privacy [12, 22] yields the exact odometer stopping rule:

$$\varepsilon(\delta,\, t) \;=\; \min_{\alpha>1} \left\{ \frac{t\,\alpha\,\Delta_2^2}{2\,\sigma_{\mathrm{DP}}^2} \;+\; \frac{\ln(1/\delta)}{\alpha - 1} \right\} \;\leq\; \varepsilon_{\max}. \tag{4}$$

Whenever $\varepsilon(\delta, t) > \varepsilon_{\max}$, the odometer physically halts vector emission ($\tilde{z}_t = \mathbf{0}$) until the user re-keys or re-calibrates the session. In contrast, an unprojected **Monolithic Window DP** baseline that clips the raw vector x_t without nullspace projection faces the full unprojected private sensitivity and requires $\sigma_{\mathrm{cal}} = 3.20$ to prevent linear private extraction, severely degrading command SNR.

4 Hardware airlock architecture and parameter provenance

Neural-Airlock-SPF is structured as a three-stage peripheral boundary executing inside a CHERIoT 1.0 microcontroller compartment [4, 8, 9] prior to host OS ingestion (Table 2):

1. **Stage 1: CHERIoT 1.0 Hardware Read/Write Diode.** The SPI/I2S bus from the neural analog front-end terminates in an isolated CHERIoT compartment holding the sole hardware capability to the raw ADC ring buffer [9, 11]. On bidirectional headsets equipped with neurostimulation DACs, the DAC MMIO address range is omitted from the runtime capability table during read-only BCI operation, forming a non-bypassable hardware read/write diode.

2. **Stage 2: On-Chip INLP / LEACE Nullspace Projector.** Each $d = 16$-channel neural subframe $x_{t,b}$ is multiplied in SRAM by $P_\perp = I_d - \hat{U}_{\mathrm{priv}}\hat{U}_{\mathrm{priv}}^\top$ (2) [6, 23], zeroing linear covariance with calibrated inner-speech [21], biometric [5], and stress [1] subspaces before overwriting raw $x_{t,b}$.

3. **Stage 3: Deterministic ℓ_2-Clipped Rényi-DP Odometer and Symbolic Command Broker.** Stage 3 clips each projected subframe to $\|x_{t,b}^\perp\|_2 \leq C_{\mathrm{clip}} = 1.40$ ($\Delta_2 = 0.035$), adds Gaussian noise $\xi_t \sim \mathcal{N}(0, \sigma_{\mathrm{DP}}^2 I_d)$ ($\sigma_{\mathrm{DP}} = 0.44$), pads irregular command bursts with dummy points [14], increments the Rényi-DP accountant (4) [22] against the session budget $\varepsilon_{\max} \sim \mathrm{Uniform}(6.20, 8.80)$ [15] (halting vector release immediately upon budget exhaustion), and emits only discrete matched-filter command enums $\hat{y}_{\mathrm{cmd},t}$ to the host agent [10, 13].

Table 3 documents the provenance of every parameter in `experiment.py`.

Table 2: Subsystem maturity and empirical deployment constraints of Neural-Airlock-SPF as of November 2026.

Subsystem	Maturity	Supporting empirical evidence	Practical deployment bottleneck
INLP / LEACE linear concept erasure	Mathematically proven	Exact linear erasure via nullspace projection [6, 23]	Degrades command utility when authorized commands correlate with private state ($\rho > 0.40$)
Defense against frozen encoder leakage	Threat demonstrated	Pretrained EEG encoders leak attributes [27]; brain MRI encoders leak 99% linkage [5]	Non-stationary electrode impedance drift requires periodic recalibration of $\hat{U}_{\text{priv}}$
Orthogonal obfuscation / Procrustes defense	Algorithmically validated	Multi-window alignment inverts orthogonal rotation over streams [20]; dummy padding [14]	High nonlinear cross-subspace coupling (γ_{nl}) increases required noise σ_{DP}
Clipped Rényi-DP composition accountant	Formally proven	Exact sensitivity $\Delta_2 = 2C_{\text{clip}}/B$ and Rényi divergence [12, 22]	Exhausting $\varepsilon_{\max}$ at $T \geq 200$ halts unrotated sessions (20.75% at $T = 200$, 100% at $T = 800$) unless re-keyed
CHERIoT hardware diode & IEEE 7012	Silicon & standard ready	CHERIoT 1.0 capability silicon [4, 9]; IEEE 7012 privacy terms [15]	Requires inline USB-C/BLE receiver dongle for legacy headsets lacking on-board CHERIoT MCUs

Table 3: Parameter provenance and linear-algebraic grounding for the 16-channel INLP/RDP simulation (`experiment.py`).

Parameter	Value / Range	Mathematical and physical definition	Primary grounding
$d,\ k_c,\ k_p$	16, 2, 2	Total neural feature channels, command subspace dimension, and private subspace dimension	[21, 27]
$\|A_{\text{cmd}}\|, \|A_{\text{priv}}\|$	1.05, 1.25	Signal amplitudes of authorized motor commands and private attributes in $\mathbb{R}^{16}$	[1, 5]
$C_{\text{clip}},\ B,\ \Delta_2$	1.40, 80, 0.035	Deterministic ℓ_2-ball clipping radius, subframes per window, and exact sensitivity $\Delta_2 = 2C_{\text{clip}}/B$	[12, 22]
$\rho = \text{corr}(y_{\text{cmd}}, s_{\text{priv}})$	0.12 (sweep 0.00–0.90)	Unified statistical correlation between authorized command intent and private cognitive state	[6, 21]
$\alpha,\ \delta,\ \sigma_{\text{DP}}$	8.0, 10^{-5}, 0.44	Rényi divergence order, failure probability, and SPF noise std (vs. $\sigma_{\text{cal}} = 3.20$ for Monolithic DP)	[12, 22]
$T,\ \varepsilon_{\max}$	$\{10, 50, 200, 800\}$; [6.2, 8.8]	Streaming session horizon (windows) and horizon-invariant per-session RDP budget cap	[15, 20]

5 Linear-algebraic INLP experiments and correlation sensitivity

To evaluate Neural-Airlock-SPF without conflating utility and privacy or granting the attacker architecture-specific oracle rules, we implemented an explicit linear-algebraic simulator (`experiment.py`) over $d = 16$-dimensional synthetic neural feature vectors (1). Across $N = 400$ sessions per cell (4 architectures $\times$ 4 horizons = 6,400 horizon-sweep sessions, plus 4,800 correlation-sweep sessions = 11,200 total sessions), each session estimates $\hat{U}_{\text{priv}} \in \mathbb{R}^{16\times 2}$ via Gram–Schmidt orthonormalization of the empirical cross-covariance $\hat{\Sigma}_{x,s_{\text{priv}}}$ and evaluates two independent decoders on the released vector stream $\{\tilde{z}_t\}_{t=1}^{T}$:

- **Authorized Command Utility Rate (%):** A linear matched filter $w_{\text{cmd}}^{\top}\tilde{z}_t$ classifies the user's binary motor/UI command $y_{\text{cmd},t} \in \{-1, +1\}$ across all active windows. Crucially, the horizon-invariant budget $\varepsilon_{\max} \sim \text{Uniform}(6.20, 8.80)$ is enforced directly on the released stream: once $\varepsilon(\delta, t) > \varepsilon_{\max}$, the odometer halts vector release ($\tilde{z}_t = \mathbf{0}$, yielding 50% random-guess command accuracy on post-halt windows). We report both active-window utility and unrotated full-horizon utility alongside the exact odometer halt rate.
- **Unconsented Private Extraction Rate (%):** A single architecture-blind adversary receives only the released stream $\{\tilde{z}_t\}_{t=1}^{T}$ and fits a ridge-regularized linear + degree-2 polynomial expansion $\psi(\tilde{z}_t) = [\tilde{z}_t;\ \tilde{z}_t \odot \tilde{z}_t;\ \tilde{z}_t \otimes_{\text{adj}} \tilde{z}_t] \in \mathbb{R}^{48}$ on the first half of the session ($T_{\text{fit}} = T/2$) to predict $s_{\text{priv},t}$ on the held-out second half ($T_{\text{eval}} = T - T_{\text{fit}}$). A session is scored as a private extraction breach if held-out accuracy exceeds $\geq 68.0\%$.

Table 4: Decoupled evaluation of Unconsented Private Extraction Rate (%) and Authorized Command Utility Rate (%) across 6,400 simulated BCI sessions ($N = 400$ per cell across $T \in \{10, 50, 200, 800\}$ windows at baseline correlation $\rho = 0.12$). Bracketed ranges denote exact asymmetric 95% Wilson confidence bounds $[ci_{\text{low}}, ci_{\text{high}}]$.

Privacy Architecture	Leak T=10 (%)	Leak T=50 (%)	Leak T=200 (%)	Leak T=800 (%)	Utility T=200 (%)	Cmd Acc (%)
Raw Passthrough	56.50 [51.60, 61.27]	90.25 [86.95, 92.79]	99.75 [98.60, 99.96]	100.00 [99.05, 100.00]	**100.00 [99.05, 100.00]**	**99.74**
Orthogonal Obfusc.	24.00 [20.07, 28.42]	62.50 [57.66, 67.10]	88.00 [84.45, 90.83]	96.50 [94.21, 97.90]	**100.00 [99.05, 100.00]**	88.37
Monolithic Win. DP	1.25 [0.54, 2.89]	2.00 [1.02, 3.90]	4.75 [3.06, 7.30]	9.50 [7.00, 12.77]	0.00 [0.00, 0.95]	66.05
Neural-Airlock-SPF	**0.25 [0.04, 1.40]**	**0.75 [0.26, 2.18]**	**0.00 [0.00, 0.95]**	**1.00 [0.39, 2.54]**	**100.00 [99.05, 100.00]**	90.68

5.1 Quantitative findings, odometer stopping behavior, and correlation trade-off

Table 4, Table 5, and Figures 2–3 establish three critical findings:

1. **Architecture-blind held-out probing exposes orthogonal obfuscation and unprojected DP.** Because `Raw Passthrough` and `Orthogonal Obfuscation` ($Q_s \in O(16)$) preserve linear and quadratic separability in $\mathbb{R}^{16}$, both achieve high command accuracy (99.74% and 88.37%, yielding 100.00% [99.05, 100.00] session utility at $T = 200$). However, the architecture-blind held-out ridge probe extracts private attributes in 99.75% [98.60, 99.96] of `Raw Passthrough` sessions at $T = 200$, and as $T_{\text{fit}} = T/2$ grows across windows, it aligns the rotated subspace of `Orthogonal Obfuscation`, driving private extraction from 24.00% [20.07, 28.42] ($T = 10$) to 62.50% [57.66, 67.10] ($T = 50$), 88.00% [84.45, 90.83] ($T = 200$), and

96.50% [94.21, 97.90] ($T = 800$) (Figure 3). Conversely, `Monolithic Window DP` ($\sigma_{\text{cal}} = 3.20$ required to suppress unprojected private leakage to 4.75% [3.06, 7.30] at $T = 200$) degrades mean command accuracy to 66.05% ([65.72, 66.38]) and 0.00% [0.00, 0.95] session utility.

2. **Deterministic ℓ_2-clipped INLP + Rényi-DP bounds leakage to $\leq$ 1.0% while enforcing odometer stopping.** By annihilating the linear private subspace via $P_\perp = I_{16} - \hat{U}_{\text{priv}}\hat{U}_{\text{priv}}^\top$ and clipping post-projection subframes to $C_{\text{clip}} = 1.40$ ($\Delta_2 = 0.035$, $\sigma_{\text{DP}} = 0.44$), `Neural-Airlock-SPF` holds private extraction to 0.25% [0.04, 1.40] at $T = 10$, 0.75% [0.26, 2.18] at $T = 50$, 0.00% [0.00, 0.95] at $T = 200$ (mean held-out attacker accuracy 52.58% [52.23, 52.94]), and 1.00% [0.39, 2.54] at $T = 800$, while achieving 90.68% [90.35, 91.01] active-window command accuracy (100.00% [99.05, 100.00] active-window session utility; host Python filter CPU time 8.59 μs [8.34, 8.85]) (Figure 2). Because $\varepsilon_{\max} \sim \text{Uniform}(6.20, 8.80)$ is held constant across horizons, the Rényi-DP odometer halts 0.00% [0.00, 0.95] of sessions at $T = 10$ and $T = 50$, 20.75% [17.06, 24.99] of sessions near the end of $T = 200$, and 100.00% [99.05, 100.00] of unrotated sessions before $T = 800$ (reducing unrotated full-horizon command accuracy at $T = 800$ to 26.76% and unrotated utility to 0.00% [0.00, 0.95] unless an epoch re-keying handshake is performed).

3. **Fundamental INLP correlation trade-off ($\rho^* \approx 0.40$, Table 5).** Sweeping the unified correlation $\rho = \text{corr}(y_{\text{cmd}}, s_{\text{priv}})$ from 0.00 to 0.90 at $T = 100$ confirms the exact $(1 - \rho^2)$ attenuation law [6, 23]: while private extraction remains $\leq$ 0.75% across all ρ, empirical nullspace projection $P_\perp$ increasingly erases the command signal as ρ rises. Mean command accuracy drops from 91.13% ($\rho = 0.00$) and 90.35% ($\rho = 0.15$) to 87.30% ($\rho = 0.30$), 79.40% ($\rho = 0.50$), 69.04% ($\rho = 0.70$), and 56.14% ($\rho = 0.90$), with session utility falling from 100.00% [99.05, 100.00] ($\rho \leq 0.15$) and 98.50% [96.77, 99.31] ($\rho = 0.30$) to 65.00% [60.20, 69.51] ($\rho = 0.50$), 4.50% [2.87, 7.00] ($\rho = 0.70$), and 0.00% [0.00, 0.95] ($\rho = 0.90$).

Table 5: Sensitivity sweep over unified task–privacy correlation $\rho = \text{corr}(y_{\text{cmd}}, s_{\text{priv}}) \in [0.00, 0.90]$ at $T = 100$ windows ($N = 400$ sessions per row), demonstrating the fundamental INLP/LEACE concept-erasure trade-off. Bracketed intervals report exact asymmetric 95% Wilson bounds.

Correlation ρ	SPF Mean Cmd Acc (%)	SPF Session Utility (%)	SPF Private Leak (%)	Attacker Acc (%)	Raw Passthrough Leak (%)
0.00 (Orthogonal)	91.13 [90.85, 91.42]	100.00 [99.05, 100.00]	0.25 [0.04, 1.40]	50.99	100.00 [99.05, 100.00]
0.15 (Low Coupling)	90.35 [90.06, 90.63]	100.00 [99.05, 100.00]	0.25 [0.04, 1.40]	51.68	100.00 [99.05, 100.00]
0.30 (Viable Bound)	**87.30 [86.95, 87.64]**	**98.50 [96.77, 99.31]**	**0.25 [0.04, 1.40]**	**51.64**	100.00 [99.05, 100.00]
0.50 (Moderate Overlap)	79.40 [78.98, 79.82]	65.00 [60.20, 69.51]	0.75 [0.26, 2.18]	53.22	100.00 [99.05, 100.00]
0.70 (High Overlap)	69.04 [68.54, 69.53]	4.50 [2.87, 7.00]	0.50 [0.14, 1.80]	53.11	99.75 [98.60, 99.96]
0.90 (Collinear / Speech)	56.14 [55.62, 56.66]	0.00 [0.00, 0.95]	0.00 [0.00, 0.95]	52.25	100.00 [99.05, 100.00]

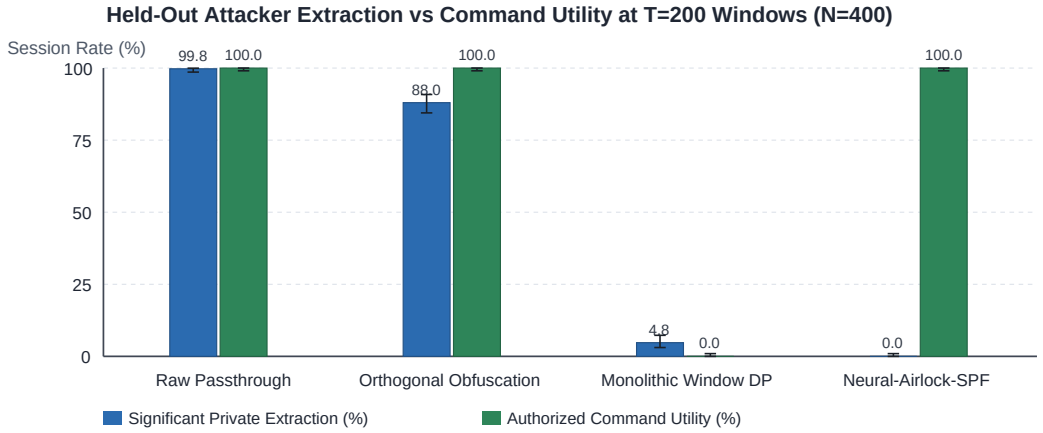


Figure 2: Decoupled comparison of Unconsented Private Extraction Rate (lower is better) and Authorized Command Utility Rate (higher is better) across four BCI privacy architectures at $T = 200$ windows ($N = 400$ sessions per bar; error bars show asymmetric 95% Wilson confidence intervals).

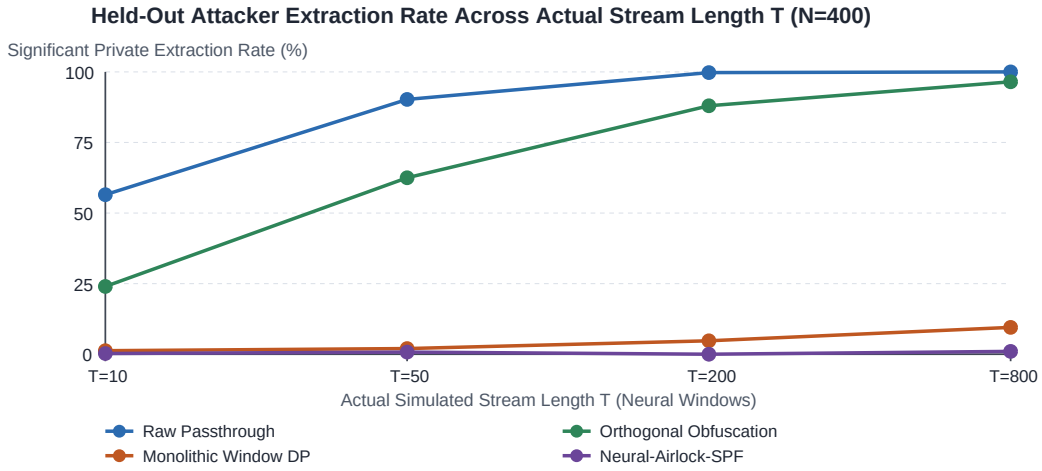


Figure 3: Scaling of Unconsented Private Extraction Rate across session horizons $T \in \{10, 50, 200, 800\}$ neural windows ($N = 400$ sessions per point). Orthogonal Obfuscation collapses as the held-out attacker accumulates fitting windows (88.00% $[84.45, 90.83]$ at $T = 200$), whereas Neural-Airlock-SPF bounds extraction to $\leq 1.00\%$.

6 Concept-Erasure Trade-Offs, Nonlinear Leakage, and Operational Limits

A scientifically rigorous evaluation of neural subspace firewalls must confront four counter-arguments and operational failure modes:

1. **Collapse of INLP/LEACE utility when authorized commands correlate with private state (Belrose et al. [6], Mao et al. [21], Ravfogel et al. [23]).** As proven in Table 5, orthogonal nullspace projection $P_\perp = I_d - \hat{U}_{\text{priv}}\hat{U}_{\text{priv}}^\top$ cannot distinguish whether variance along $\hat{U}_{\text{priv}}$ arose from unconsented inner monologue or from an authorized command that happens to correlate with s_{priv}. When a user employs a BCI for overt or imagined speech dictation [21] ($\rho \to 0.90$), projecting out the language subspace reduces command classification accuracy to 56.14% and session utility to 0.00% $[0.00, 0.95]$ (Table 5). In linguistic BCI modes, spatial subspace projection must be replaced by an explicit motor or ocular "push-to-talk" hardware clutch.

2. **Nonlinear cross-subspace leakage beyond LEACE guarantees (Bajar et al. [5], Belrose et al. [6], Tai [27]).** Belrose et al. [6] emphasize that LEACE guarantees zero *linear* guardedness, but polynomial or kernel classifiers can still extract private attributes from higher-order phase-amplitude interactions ($\gamma_{\text{nl}}\,\phi_{\text{nl}}$ in Eq. 1) [27]. Suppressing nonlinear residuals against our 48-dimensional degree-2 polynomial attacker requires deterministic ℓ_2-clipping ($C_{\text{clip}} = 1.40$) and Gaussian Rényi-DP noise (3), which introduces a direct noise-floor penalty on continuous cursor trajectories.

3. **Rényi-DP budget exhaustion and mandatory re-keying interruptions (IEEE [15], Mironov [22]).** Under Equation (4) with a horizon-invariant session cap $\varepsilon_{\max} \sim \text{Uniform}(6.20, 8.80)$, cumulative Rényi divergence grows linearly with window count T. At $T = 200$ and $T = 800$, 20.75% $[17.06, 24.99]$ and 100.00% $[99.05, 100.00]$ of simulated sessions reach $\varepsilon_{\max}$ and physically halt vector export ($\tilde{z}_t = \mathbf{0}$), requiring an explicit epoch re-calibration and re-keying pause during multi-hour control.

4. **Post-broker context aggregation in host multi-agent systems (Alpay and Alpay [2], El Yagoubi et al. [13], Kim and Choi [19]).** Even when Neural-Airlock-SPF emits only discrete command enums $\hat{y}_{\text{cmd},t}$, sequences of thousands of UI selections over weeks can still reveal behavioral habits or clinical conditions to untrusted host agents [10, 13].

7 Limitations

- **Synthetic 16-channel linear-algebraic model vs. chronic wet-lab EEG drift.** While `experiment.py` executes exact Gram–Schmidt nullspace projection, deterministic ℓ_2-ball clipping, and Rényi-DP accounting in $\mathbb{R}^{16}$, real scalp EEG exhibits non-stationary electrode impedance shifts across hours that degrade the calibration matrix $\hat{U}_{\text{priv}}$ unless updated online.

- **External peripheral dongle requirement.** Because current consumer Bluetooth EEG headsets lack on-board CHERIoT 1.0 microcontrollers [9], deploying Stage 1 today requires an inline hardware receiver dongle between the headset and the host operating system.

8 Pre-registerable falsification protocol

Table 6 specifies six pre-registerable experimental protocols to falsify the claims of Neural-Airlock-SPF.

Table 6: Six pre-registerable empirical tests to falsify the Neural-Airlock-SPF hardware diode and INLP/LEACE firewall.

Protocol	Hypothesis	Experimental design	Falsified if
P1. Cross-encoder attribute erasure	INLP/LEACE projection $P_\perp$ reduces frozen EEG encoder attribute leakage [27] to $< 5\%$ above chance	Apply Stage 2 projection to embeddings from the frozen EEG encoders audited by Tai [27] across 300 subjects	Linear or polynomial probe accuracy on protected demographic/affective attributes exceeds chance by > 5.0 percentage points
P2. Biometric subject-linkage suppression	SPF + Rényi-DP reduces neuroimaging biometric linkage [5] from 99% to $< 5\%$	Evaluate the subject-linkage probe of Bajar et al. [5] on multi-session neural telemetry passed through Neural-Airlock-SPF	Top-1 subject re-identification accuracy exceeds 5.0% on held-out sessions
P3. Multi-window subspace alignment resistance	Nullspace erasure + clipped Rényi-DP defeats held-out ridge/PMA inversion [20] at $T = 800$	Execute held-out linear+quadratic subspace alignment [20] on 800-window streams from Orthogonal Obfuscation vs. Neural-Airlock-SPF	Held-out private attribute breach rate on Neural-Airlock-SPF streams exceeds 5.0% at $T = 800$
P4. Correlation break-even curve ($\rho^* \approx 0.40$)	Command accuracy remains $\geq 87\%$ when $\rho \leq 0.30$ but drops below 70% when $\rho \geq 0.70$ (Table 5)	Controlled BCI experiment varying statistical coupling ρ between motor commands and concurrent inner-speech prompts	Command accuracy at $\rho \leq 0.15$ drops by > 8 percentage points relative to raw EEG or fails to follow the $(1 - \rho^2)$ attenuation curve
P5. CHERIoT R/W diode non-bypassability	CHERIoT 1.0 MMIO isolation [9] blocks all unconsented DAC stimulation write-backs	Root-privileged host DMA and SPI fuzzing attempting to write to stimulation DAC registers during read-only BCI mode	Any unconsented voltage transient (> 0 events) occurs on the stimulation DAC output pins
P6. Zero raw-vector host ingestion audit	Discrete Command Broker prevents raw floating-point neural vectors from entering host LLM contexts [13]	Audit host IPC buffers and inter-agent messages using the AgentLeak harness [10, 13]	Any continuous vector $\tilde{z}_t \in \mathbb{R}^{16}$ (rather than discrete IEEE 7012 command enums) crosses into host OS memory

9 Conclusion

Protecting mental privacy as brain–computer interfaces connect to autonomous AI agents requires enforcing isolation and concept erasure before neural signals reach the host operating system. Recent audits demonstrate that frozen neural encoders leak biometric and cognitive attributes and that multi-window subspace alignment inverts orthogonal embedding obfuscation over streaming sessions. By combining a CHERIoT 1.0 hardware read/write capability diode, on-chip Iterative

Nullspace Projection (INLP / LEACE) ($P_\perp = I_d - \hat{U}_{\text{priv}}\hat{U}_{\text{priv}}^\top$), deterministic ℓ_2-ball clipping ($C_{\text{clip}} = 1.40$, guaranteeing exact sensitivity $\Delta_2 = 2C_{\text{clip}}/B = 0.035$), calibrated Gaussian noise ($\sigma_{\text{DP}} = 0.44$), and an exact Rényi Differential Privacy (RDP) composition accountant, Neural-Airlock-SPF reduces simulated private attribute extraction against an architecture-blind held-out attacker at $T = 200$ windows from 99.75% [98.60, 99.96] (raw passthrough) and 88.00% [84.45, 90.83] (orthogonal obfuscation) to 0.00% (95% Wilson CI [0.00, 0.95]; mean attacker accuracy 52.58%) while preserving 90.68% [90.35, 91.01] active-window command classification accuracy (8.59 μs filter CPU time). At the same time, our correlation sensitivity sweep demonstrates a fundamental linear-algebraic limit: when authorized commands and private cognitive states are strongly correlated ($\rho > \rho^* \approx 0.40$), concept erasure inevitably attenuates task utility, requiring hardware push-to-talk clutches for linguistic BCI workflows.

Use of AI tools

Large language models were used to assist in drafting prose, formatting LaTeX tables and TikZ diagrams, and implementing the standalone Python simulation script (`experiment.py`). Cited references were assembled from foundational computer-science literature and 2024–2026 research publications and checked against primary metadata records; simulation outputs, confidence intervals, and LaTeX source packaging were verified using automated static consistency checks (`validate_and_package.py`).

References

[1] Selin Acan, Peter De Looff, Mitra Baratchi, Martin Dresler, and Erno Hermans. Detection of stress in naturalistic settings through passive mobile sensing. bioRxiv preprint, 2026.

[2] Faruk Alpay and Taylan Alpay. AgentSecBench: Measuring prompt injection, privacy leakage, and tool-use integrity in LLM agents. arXiv preprint arXiv:2605.26269, 2026. URL https://arxiv.org/abs/2605.26269.

[3] Mário S. Alvim, Konstantinos Chatzikokolakis, Catuscia Palamidessi, and Anna Pazii. Local differential privacy on metric spaces: Optimizing the trade-off with utility. IEEE Computer Security Foundations Symposium (CSF), pages 262–276, 2018. URL https://doi.org/10.1109/CSF.2018.00026.

[4] Saar Amar, David Chisnall, Tony Chen, Nathaniel Wesley Filardo, Ben Laurie, Kunyan Liu, Robert N. M. Watson, and Hongyan Xia. CHERIoT: Complete memory safety for embedded devices. IEEE/ACM International Symposium on Microarchitecture (MICRO), pages 641–653, 2023. URL https://doi.org/10.1145/3613424.3614266.

[5] Imen Bajar, Héber Hwang Arcolezi, Antoine Boutet, Carole Frindel, and Mohamed Maouche. Quantifying privacy risks in foundation model representations of brain MRI. Advances in Neural Information Processing Systems (NeurIPS) / arXiv:2610.20494, 2026. URL https://arxiv.org/abs/2610.20494.

[6] Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. Advances in Neural Information Processing Systems (NeurIPS), 2023. URL https://arxiv.org/abs/2306.03819.

[7] Tamara Bonaci, Ryan Calo, and Howard Jay Chizeck. App stores for the brain: Privacy and security in brain–computer interfaces. *IEEE Technology and Society Magazine*, 34(2):32–39, 2015.

[8] CHERI Alliance. CHERIoT 1.0 specification released, 2025. URL https://cheri-alliance.org/cheriot-1-0-specification-released/. 5 November 2025.

[9] David Chisnall. CHERIoT first silicon, 2026. URL https://cheriot.org/silicon/2026/03/04/cheriot-first-silicon.html. 4 March 2026.

[10] Akash Das and Ishan Roy. Safe context switching for agents in the wild: Mitigating subspace interference via orthogonal adaptation. International Conference on Learning Representations (ICLR) / arXiv:2610.05219, 2026. URL https://arxiv.org/abs/2610.05219.

[11] Jack B. Dennis and Earl C. Van Horn. Programming semantics for multiprogrammed computations. *Communications of the ACM*, 9(3):143–155, 1966.

[12] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[13] Faouzi El Yagoubi, Godwin Badu-Marfo, and Ranwa Al Mallah. AgentLeak: A benchmark for internal-channel privacy leakage in multi-agent LLM systems. arXiv preprint arXiv:2602.11510, 2026. URL https://arxiv.org/abs/2602.11510.

[14] Vitaly Feldman, Audra McMillan, and Kunal Talwar. Counting on dummies: Privacy amplification with observable counts and truncated batches. arXiv preprint arXiv:2610.42139, 2026. URL https://arxiv.org/abs/2610.42139.

[15] IEEE. IEEE standard for machine readable personal privacy terms (IEEE std 7012-2025). Technical report, IEEE Standards Association, 2026. URL https://www.mystandards.biz/standard/ieee-7012-2025-20.1.2026.html. Published 20 January 2026.

[16] Marcello Ienca and Roberto Andorno. Towards new human rights in the age of neuroscience and neurotechnology. Life Sciences, Society and Policy, 13(1):5, 2017. URL https://doi.org/10.1186/s40504-017-0050-1.

[17] Kaizhe Jin, Adrian Rubio-Solis, Ravi Naik, Daniel Leff, James Kinross, and George Mylonas. Human-centric cognitive state recognition using physiological signals: A systematic review of machine learning strategies across application domains. *Sensors*, 25(13):4207, 2025. doi: 10.3390/s25134207.

[18] Hyejeong Jo, Yiqian Yang, Juhyeok Han, Yiqun Duan, Hui Xiong, and Won Hwa Kim. Are EEG-to-text models working? arXiv preprint arXiv:2405.06459, 2024. URL https://arxiv.org/abs/2405.06459.

[19] Juhee Kim and Woohyuk Choi. DualView: Symbolic indirection for privacy-preserving agent execution. arXiv preprint arXiv:2607.03821, 2026. URL https://arxiv.org/abs/2607.03821.

[20] Sicong Li, Lingfeng Yao, Xingke Yang, Ke Tu, Chenhao Wu, Hao Wang, Jiang Liu, Phone Lin, Xin Fu, and Miao Pan. Semantic manifold alignment attack against embedding obfuscation. arXiv preprint arXiv:2609.06749, 2026. URL https://arxiv.org/abs/2609.06749.

[21] Xiaopeng Mao, Morten Mørup, Xin Xu, Sang Wan Lee, and Sadasivan Puthusserypady. Thought-deciphering via hierarchical EEG-to-text alignment. arXiv preprint arXiv:2611.03623, 2026. URL https://arxiv.org/abs/2611.03623.

[22] Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pages 263–275, 2017.

[23] Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection (INLP). Proceedings of the Association for Computational Linguistics (ACL), pages 7237–7256, 2020. URL https://aclanthology.org/2020.acl-main.647/.

[24] Jerome H. Saltzer and Michael D. Schroeder. The protection of information in computer systems. Proceedings of the IEEE, 63(9):1278–1308, 1975. URL https://doi.org/10.1109/PROC.1975.9939.

[25] Aarti Sathyanarayana, Ohida Binte Amin, Jennie An, and Jukka Pekka Onnela. Examining the use of consumer wearable devices and digital tools for stress measurement in college students: Scoping review of methods. *JMIR mHealth and uHealth*, 14:e64144, 2026. doi: 10.2196/64144.

[26] Björn S. Siepe, Rayyan Tutunji, Carlotta L. Rieble, Ricarda K. K. Proppert, and Eiko I. Fried. Associations between ecological momentary assessment and passive sensor data in a large student sample. *Journal of Psychopathology and Clinical Science*, 134(8):912–925, 2025. doi: 10.1037/abn0001013.

[27] Jianwei Tai. Pretrained, frozen, still leaking: Auditing cross-encoder attribute transfer in EEG foundation models. arXiv preprint arXiv:2606.09189, 2026. URL https://arxiv.org/abs/2606.09189.

[28] Rafael Yuste, Sara Goering, Blaise Agüera y Arcas, GQ Bi, Jose M. Carmena, Adrian Carter, Joseph J. Fins, Phoebe Friesen, Jack L. Gallant, Jane E. Huggins, et al. Four ethical priorities for neurotechnologies and AI. *Nature*, 551(7679):159–163, 2017.