TRAINED MODELS AS DERIVATIVE ARTEFACTS

Reframing the Object of Secondary Use in EHDS Scholarship

Subtitle: EHDS, secondary use, and **model-centred governance** (from datasets to exportable models)

Core question

If teams leave Secure Processing Environments with **trained models**, do "anonymous results" assumptions still hold under GDPR/EHDS?

Why it matters

Models can be interrogated to infer membership or reconstruct training data—challenging **opt-out** and **erasure** rights.

Author
Sumanta Narayan Podder
Presidency University, Kolkata, India/ Department of Economics

Date
24 March, 2026
sumantapodder@gmail.com

Motivation and Problem Statement

EHDS scholarship often assumes the **dataset** is the primary object of control—while modern ML workflows make the **trained model** the artefact that actually circulates.

ASSUMPTION

Dataset-centred control

Inherited lineage

Institutional vocabulary
Biobanks, clinical registries, and statistical offices shaped the idea that the **dataset** is the controllable entity.

Governance mechanics
Permits + Secure Processing Environments (SPEs) regulate **dataset access**; exports are framed as "anonymous results".

Subject-rights framing
Rights like access, erasure, restriction are operationalised as actions on **records in datasets**.

IMPLICIT RISK MODEL

Risk "stays" with data → Outputs presumed anonymous

PRACTICE

Model-centred workflows

Empirical shift

What exits the SPE
Research teams often leave with **trained weights** (exportable artefacts), not raw datasets.

Privacy-relevant properties
Models can encode training-set signals via memorisation and can be interrogated (membership inference, inversion, extraction).

Circulation pathways
Deployment, APIs, transfer learning, and federated learning create **downstream** exposure beyond permit controls.

OBSERVED RISK MODEL

Risk travels with model → Re-use at scale

Governance gap (problem)

EHDS discourse remains **dataset-centred** even when outputs are **model-centred**. Result: an **ontological mismatch**—treating trained models as "anonymous results" can leave privacy leakage undetected until after deployment.

KEY MOVE

Reframe the object of control → the model

EHDS Secondary-Use Governance Pipeline

The EHDS architecture is optimized for **dataset access control** inside Secure Processing Environments (SPEs), then exporting “**anonymous results**”. The blind spot: trained models may leave the SPE while retaining **privacy-relevant** signals.

1 **Data holders**
Providers & registries supply EHR/health datasets for secondary use.

→

2 **HDAB permit**
Health Data Access Body authorises access for specified purposes.

→

3 **SPE access**
Research in a controlled environment with **pseudonymised** data.

→

4 **Model training**
Analytics + ML produce parameters encoding population structure.

→

5 **Export** risk
“Anonymous results” released—may include **trained models**.

→

6 **Downstream use**
Deployment, APIs, transfer learning, federated consortia.

GOVERNANCE LEVERS (AS DESIGNED)

Permits, SPE controls, disclosure review at export

PERMIT SCOPE
Purpose limitation; approved users & projects.

SPE CONSTRAINTS
No raw export; controlled tooling; logging.

OUTPUT CHECK
Disclosure control assumes “results” are non-personal.

BLIND SPOT

Treating exported models as “anonymous results”

Trained models can be interrogated (membership inference, inversion, extraction) to recover training-set signals. Risk can therefore **circulate beyond** the SPE even when datasets do not.

IMPLICATION FOR EHDS

Output governance must become **model-aware** (classification, audit, provenance, rights).

Reading guide

Slides 6–9 summarise the technical evidence that makes “anonymous model export” a weak assumption in practice.

KEY QUESTION

What does “anonymous” mean when the exported artefact is a trained model?

CORE THESIS

Reframing the object of secondary use: from datasets to trained models

ONTOLOGICAL MISMATCH

Dataset-centred doctrine vs model-borne leakage

Core Thesis: The Trained Model Is the Circulating Artefact

In EHDS secondary-use practice, what leaves the Secure Processing Environment is often a **trained model**—a derivative artefact that can retain **privacy-relevant properties** even when datasets remain locked down.

BIG IDEA

Treat trained models as governable artefacts of secondary use—not merely "anonymous results".

Because models can be inverted, queried, or audited to infer membership or extract memorised content, they can function as a portable container of information about individuals in the training set.

WHAT CHANGES?

Object of control

Dataset → Model

CLAIM 1

Models are derivatives that circulate

Secondary-use workflows increasingly export **weights**, not data rows. Value and risk travel via model parameters across deployment, APIs, and re-use.

CLAIM 2

Models can carry subject-linked signals

Over-parameterisation enables **memorisation** without obvious overfitting; technical literature shows membership inference, inversion, and extraction attacks can recover training-set information.

CLAIM 3

Doctrine becomes reactive, not preventive

If governance only controls **dataset access**, it may detect harm only after models leak—undermining EHDS **opt-out** and GDPR **erasure** in practice.

WHY THE OLD FRAME FAILS

Dataset-centred "one-way" assumption

EHDS-style pipelines treat outputs as terminal: data go in, "anonymous results" come out. But models enable **information flow back** from model behaviour/parameters to training data.

OPERATIONAL IMPLICATION

"Anonymous results" requires a model test

Export review must assess whether a trained model poses more than an **insignificant risk** of membership or content leakage—before deployment.

Why Trained Models Can Be Privacy-Relevant

A trained model is not only a predictive tool. In modern ML, it can also be a **portable information container** whose parameters and outputs can reveal **facts about the training set**.

BIG IDEA

Treat a trained model as a privacy-bearing derivative artefact, not a neutral “result”.

Over-parameterisation enables **memorisation** without obvious overfitting; downstream access (weights or queries) can expose membership, sensitive features, or verbatim content.

TWO WAYS INFORMATION LEAKS

- **White-box**
  Inspect parameters / updates
- **Black-box**
  Query via API / labels

TRAINING DATA
Sensitive health records inside SPE

→ TRAINING
Optimisation updates weights to fit population patterns; capacity may also encode outliers.

→

TRAINED MODEL
Dual nature: function + information structure

RISK
Depends on access + attack

Over-parameterised
High capacity can store more detail than evaluation metrics reveal.

Can memorise
Specific records or rare strings may be retained (“remember too much”).

Probe-able
Behaviour under queries can expose training-set signatures.

CIRCULATION PATHWAYS
How model artefacts move after leaving the SPE

- **Exported weights** — to product / research repo
- **API access** — queries at scale
- **Transfer learning** — fine-tuning & re-release
- **Federated updates** — gradients/parameters shared

Once models circulate, **risk travels with parameters**—outside the EHDS permit/SPE boundary.

GOVERNANCE CONSEQUENCE
“Anonymous results” is not a safe default

If trained models can be probed or inspected to reveal training-set information, then the export object is **not automatically anonymous** and needs explicit assessment.

BRIDGE TO NEXT SLIDES
Membership inference • model inversion • training data extraction

Technical Evidence I — Membership Inference

An adversary can infer whether a specific individual's record was in the training set by exploiting subtle differences in model responses on **training** vs **non-training** inputs (Shokri et al., 2016).

1 PREPARE
Train an attack model
`offline`

Learn a "membership signature"
Use shadow models (or auxiliary data) to map output patterns to **in** vs **out** of training set.

INPUT	FEATURE
candidate record x	model output pattern

2 PROBE
Query the target model
`black-box`

Observe outputs under controlled queries
Even small differences in confidence/robustness can leak whether x was seen during training.

TYPICAL API OUTPUTS

Probabilities	Confidence	Label
softmax scores	top-k / logits	hard class only

3 INFER
Decide membership
`privacy risk`

Return: "in training set?"
Membership can reveal sensitive facts when the dataset itself encodes a condition (e.g., being in a rare disease cohort).

GOVERNANCE IMPLICATION
If model export is permitted as "anonymous results", the released artefact can still enable re-identification via membership signals.

LABEL-ONLY ADVANCE

Works even when probabilities are hidden

Choquette-Choo et al. (2021) show membership inference can succeed with **hard labels only** by probing decision-boundary robustness using strategically perturbed queries.

ACCESS	SIGNAL	RESULT
label-only API	robustness gap	membership guess

IMPLICATIONS FOR "ANONYMOUS RESULTS" REVIEW

Confidence masking is not sufficient

Masking / rounding
Removing probabilities may not prevent inference if an attacker can probe labels repeatedly.

Operational takeaway
EHDS export checks need a **model-specific anonymity assessment** (threat model + access pattern), not a default assumption that models are "results".

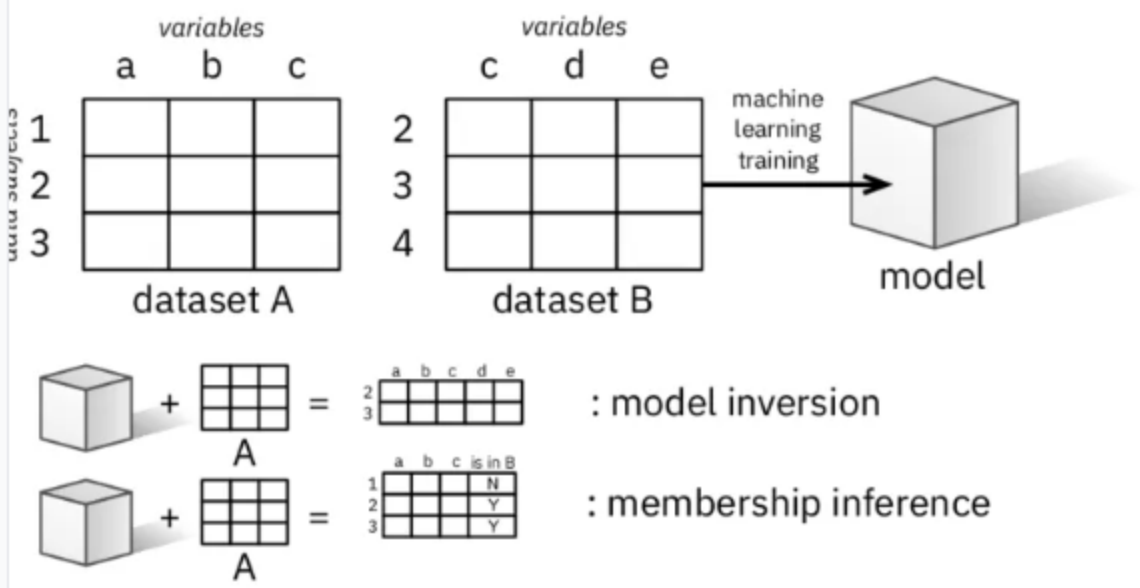
CITED
Shokri et al. (2016) • Choquette-Choo et al. (2021)

Technical Evidence II — Model Inversion

Model inversion attacks show that query access to a trained model can enable reconstruction of **sensitive features**—and in some settings, **recognisable inputs** (Fredrikson et al., 2014; 2015).

ILLUSTRATION
Inversion + membership inference (conceptual)

black-box



Attack intuition: use model responses to iteratively refine an input guess until it maximises the model's score for a target class/attribute.

SOURCE IMAGE
ResearchGate (conceptual diagram)

WHAT THE ATTACKER NEEDS
Repeated queries

Access can be via a prediction API; no dataset access is required.

GOVERNANCE IMPLICATION
"Anonymous export" is contestable

The ability to reconstruct sensitive approximations undermines treating a deployed model as a neutral, anonymous "result".

WHAT CAN BE RECONSTRUCTED
Sensitive features and recognisable inputs

Attribute inference
Fredrikson et al. (2014) demonstrate recovering sensitive input features (e.g., medical attributes) from model access.

Recognisable reconstructions
Fredrikson et al. (2015) show that confidence-bearing outputs can enable reconstructions that human evaluators recognise (notably in face-recognition settings).

WHY CONFIDENCE SCORES MATTER
They turn inversion into optimisation

MECHANISM (HIGH LEVEL)

Guess input x → Maximise score f(x) → Iterate

Confidence vectors provide a smooth signal to guide updates; with only hard labels, optimisation becomes harder (but not always impossible).

HEALTH RELEVANCE
Imaging + clinical prediction models

Medical imaging models
If a model is trained on scans, inversion-style probing may reveal approximations of anatomy or pathology-relevant patterns (depending on access and model design).

EHDS implication
Export control should evaluate whether deployed models (or APIs) enable inversion risks—rather than assuming outputs are anonymous because raw data never leave the SPE.

CITED
Fredrikson et al. (2014) • Fredrikson et al. (2015)

Technical Evidence III — Training Data Extraction

Large language models can emit **verbatim fragments** from their training corpus, including **personally identifiable information (PII)** (Carlini et al., 2021).

ILLUSTRATION
How extraction works (conceptual)

black-box

PROMPT → SAMPLE → RANK → SELECT

Generate
Sample many completions with diverse prompts.

2 **Rank**
Identify high-likelihood outliers (target vs reference model).

3 **Extract**
Verbatim strings can surface—sometimes unique in the corpus.

Illustrative output pattern

example

```
"... {verbatim span resembling an address/phone/email} ..."
```

Note: exact extracted PII varies by model and corpus; this box illustrates the **risk form**, not a specific leaked record.

Key finding (Carlini et al., 2021): extraction succeeds even when training loss does not indicate obvious overfitting—specific outliers may still be memorised.

CITATION
USENIX Sec 2021

ATTACKER ACCESS
Model interface is enough
Extraction can be done via local weights or a text-generation API (depending on rate limits and logging).

GOVERNANCE IMPLICATION
"Anonymous result" is fragile
If a trained model can reproduce identifiable text, exporting the model (or deploying it) resembles further processing—not a neutral, anonymous output.

WHAT CAN BE EXTRACTED
Verbatim sequences + PII

Personally identifiable information
Carlini et al. report recovered **names, addresses, emails, and phone numbers** from LLM outputs.

Uniqueness is not a safeguard
Extracted spans can originate from **single-occurrence** documents—memorisation is example-specific.

WHY HEALTH CONTEXTS ARE EXPOSED
Clinical notes contain rare, distinctive strings

EHR free text
Even "de-identified" corpora can retain quasi-identifiers, rare diagnoses, dates, or idiosyncratic phrasing that increases memorisation risk.

Rare presentations
Outliers are more likely to be memorised, aligning with extraction methods that search for high-likelihood outliers.

DOCTRINAL IMPLICATION
Export ≠ end of risk

EHDS STRESS POINT

EHDS governance relies on export of "anonymous results" from Secure Processing Environments. If models can leak training data, then model export demands **risk-based assessment** rather than categorical anonymisation.

CITED
Carlini et al. (2021)

Why Simple Defences Fail (and the Differential Privacy Trade-off)

Technical mitigations can reduce leakage, but the evidence base suggests there is **no default “anonymous model” state**. Differential privacy offers a formal guarantee, but introduces a **utility–privacy budget** trade-off (Dwork, 2006; Abadi et al., 2016).

OFTEN INSUFFICIENT

Mitigations that do not eliminate inference/extraction

pragmatic but brittle

Confidence masking / rounding

Hiding probabilities is frequently treated as a quick fix for membership inference. But **label-only attacks** show that hard labels can still leak training-set presence via robustness probing.

IMPLICATION

“API returns labels only” ≠ anonymity guarantee.

Adversarial regularisation

Can reduce leakage signals, but is often **partial**, threat-model-dependent, and may degrade performance—especially when models need high accuracy for clinical tasks.

IMPLICATION

Not a stable basis for “anonymous results” export.

Interface restrictions

Rate limits, output throttling, and logging can help operationally, but do not remove the underlying possibility of **model inversion** or **extraction** given sufficient access or model release.

IMPLICATION

Security controls complement—but cannot replace—privacy assessment.

“No overfitting” heuristics

Training vs test loss parity does not imply non-memorisation; outliers can still be retained. Therefore, standard ML evaluation is a weak proxy for privacy.

IMPLICATION

“Looks well-generalised” ≠ safe to export.

BRIDGE TO EHDS

If export review treats trained models as “anonymous results” by default, these mitigations can create a false sense of closure. A model’s anonymity should be **demonstrated** under a stated threat model.

DIFFERENTIAL PRIVACY (DP)

A formal guarantee—at a cost

DP bounds how much a model can change if any single person’s record is removed or replaced. It therefore directly targets membership inference risk, but requires choosing a **privacy budget (ε)**.

STRONGER PRIVACY

Lower ε

Less influence per individual record; generally lower utility.

HIGHER UTILITY

Higher ε

Better performance; weaker protection against single-record leakage.

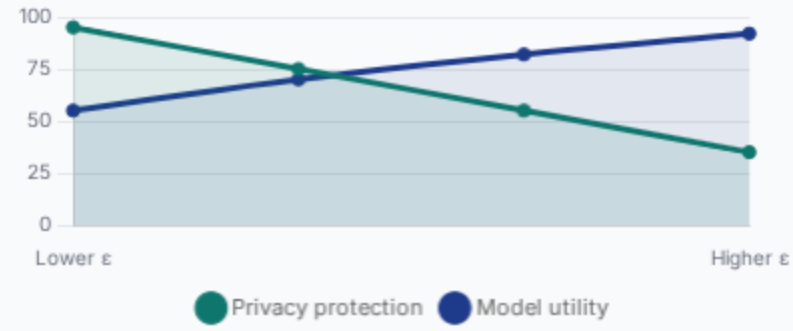
KEY POINT

DP supports **ex ante governance** (set guarantees before export), but it is not “free”—legal frameworks must account for performance impacts in clinical contexts.

ILLUSTRATIVE TRADE-OFF

Conceptual relationship (not empirical)

schematic



CITED

Dwork (2006) • Abadi et al. (2016)

GOVERNANCE TAKEAWAY

Treat model privacy risk as a policy variable

No default anonymity

“Anonymous results” cannot be assumed for exported models; it must be assessed and documented.

Residual risk persists

Where DP is not used (or is weak), rules must address monitoring, incident response, and downstream circulation.

Export review needs a threat model

Evaluate access pattern (weights vs API), query scale, and known attack classes before approving release.

Subject Rights Stress Test

Removing a record from a dataset is straightforward; removing a person's influence from a trained model is not. The result: EHDS opt-out and GDPR erasure risk becoming **procedurally satisfied** but **substantively ineffective**.

COMPARISON MATRIX

Dataset-centred rights vs model-centred reality

rights → operations

	OBJECT: DATASET	OBJECT: TRAINED MODEL
	Record deletion is legible	**"Forgetting" is contested**
RIGHT: **GDPR Article 17 — Right to erasure**	**Operationally clear** Delete or suppress the person's record; future processing excludes it. Auditable via logs and dataset versions. ACTION: remove row(s) · VERIFY: dataset diff	**Operationally costly / hard to verify** If the model has learned from the record, deletion from the dataset does not undo information encoded in parameters. Reliable erasure may require **retraining** or **machine unlearning**. ACTION: retrain / unlearn · COST: compute + time · VERIFY: non-trivial
RIGHT: **EHDS opt-out of secondary use**	**Prospective control fits datasets** Mark record as excluded; HDAB permits and SPE access can prevent future reuse of that record. GOVERNANCE LEVER: permit + access controls	**Models trained pre-opt-out can persist** If trained models are exported/deployed before opt-out, the person's contribution may continue to circulate even when dataset access is blocked. GOVERNANCE GAP: no default recall/unlearning path

stress point

risk: "empty right"

OPERATIONAL OPTIONS

If rights are to bite in model workflows

These are governance levers (not "magic fixes") that align subject rights with model-centred practice.

DP-by-default for training
Reduces single-record influence (utility trade-off).

Unlearning obligations
Define when retraining/unlearning is required.

Model provenance & recall
Track which models used which data versions.

Auditability of "forgetting"
Evidence standards for compliance claims.

KEY TAKEAWAY

Rights designed for datasets do not automatically transfer to models

Treating trained models as "anonymous results" can make erasure/opt-out rights **hard to operationalise** and **hard to verify**.

BRIDGE TO NEXT

Classification across legal regimes

Whether a model is "personal data" (and under which regime) determines which rights, duties, and remedies attach.

CLASSIFICATION ACROSS LEGAL REGIMES
How a "trained model" is classified determines which duties & rights attach

KEY TENSION
Performance, safety, and privacy obligations can pull in different directions

Cross-Regime Classification and Tensions

Once models become the circulating artefact of secondary use, governance depends on whether the exported model is treated as **personal data**, a **high-risk AI system**, and/or a **medical device**.

GDPR

Are models "personal data"?

classification gate

REGULATORY SIGNAL

EDPB Opinion 28/2024: AI models trained on personal data are **not automatically anonymous**; anonymity is **case-by-case** based on "insignificant risk" of (re)identification and membership inference / memorisation.

PRACTICAL IMPLICATION

If "personal data"
GDPR obligations attach (DPIA, purpose limits, breach notification, subject rights).

If "anonymous"
Many GDPR duties drop away—so the threshold choice is legally consequential.

EHDS STRESS POINT

Export review cannot presume "anonymous results" when the artefact is a model; it must specify a threat model and evaluate model-borne leakage risk.

CITED
EDPB Opinion 28/2024

AI ACT

Is the system high-risk?

risk tier

TYPICAL EHDS PATHWAY

Clinical decision support, triage, diagnosis/prognosis tools trained on EHDS data are likely candidates for **high-risk** classification (health context; decisions affecting persons).

OPERATIONAL REQUIREMENTS (HIGH LEVEL)

Data governance
Quality, relevance, bias control, documentation.

Transparency & oversight
Human oversight; logs; instructions for use.

TENSION WITH PRIVACY

Strong privacy training (e.g., differential privacy) can reduce accuracy; yet clinical systems are evaluated on performance and safety. Where **accuracy ↔ privacy** trade off, compliance strategy is unclear.

OPEN GOVERNANCE QUESTION

If a model is both **high-risk** (AI Act) and **personal data** (GDPR), how should requirements be reconciled in practice?

NOTE

High-level mapping for conference discussion (not legal advice).

MDR (SAMD)

Is it a medical device?

safety lens

TRIGGER

A model deployed for diagnosis, treatment decisions, or clinical guidance may qualify as **software as a medical device** (SaMD), invoking conformity assessment and safety obligations.

ADAPTIVE MODELS PROBLEM

Continuous learning
If models update on new data, when does an update become a "significant change" requiring re-certification?

Rights vs recertification
Retraining/unlearning to satisfy erasure or opt-out may collide with medical-device change control.

SYSTEM DESIGN PRESSURE

Safety governance can favour "locked" models; privacy governance may require retraining/unlearning. This creates design choices with regulatory consequences.

OPEN QUESTION

How should post-market monitoring handle discovered model leakage vulnerabilities—are they treated like a data breach, a safety issue, or both?

SCOPE NOTE

Illustrative MDR issues for adaptive ML; exact classification depends on intended use.

OPEN QUESTIONS FOR EHDS SCHOLARSHIP

Where responsibilities split (and who answers to whom)

Priority rules
When privacy protection reduces clinical performance, how should conflicts across regimes be resolved?

Joint control in model workflows
In federated learning and transfer learning, who is the controller/provider of the resulting model artefact?

IMPLICATION

Model provenance (what data shaped which model; who exported it; where it is deployed) becomes central to enforcing rights and allocating liability.

BOTTOM LINE

Classification drives governance

EHDS export controls and "anonymous results" assumptions must be re-evaluated when the artefact is a trained model that can leak training data through known attack classes.

BRIDGE TO NEXT

What research agenda closes the doctrinal gap?

Conclusions and Research Agenda

The EHDS secondary-use framework is largely **dataset-centred**, but contemporary practice is **model-centred**. Closing this doctrinal gap requires treating trained models as **derivative artefacts of secondary use** with privacy-relevant properties.

CORE TAKEAWAY

Shift the object of governance from "data access" to "model circulation".

If trained models can be interrogated to reveal training-set membership or memorised content, then "export of anonymous results" cannot be assumed when the export object is a model.

WHY THIS MATTERS

Rights, duties, and breach logic follow the artefact that circulates.

RESEARCH AGENDA

Five concrete directions for EHDS scholarship

doctrinal + operational

1 Classification criteria for models as personal data

Develop workable thresholds for when a trained model should be treated as **personal data** (risk-based, threat-model explicit).

OUTPUT

A doctrinal test grounded in technical evidence (memorisation/inference).

2 Model provenance, registries, and recall mechanisms

Track which datasets (and versions) contributed to which exported models; enable **notification**, **patching**, and **withdrawal** when risks surface.

OUTPUT

A governance model for downstream circulation (beyond SPE walls).

3 Operationalising opt-out and erasure for models

Specify when **retraining** or **machine unlearning** is required, and what counts as evidence of "forgetting" (verification & auditability).

OUTPUT

A rights-to-operations pathway that avoids "procedural but ineffective" compliance.

4 DPIAs for model-centric & federated workflows

Adapt DPIAs to assess privacy risks in exported weights, APIs, federated learning parameter exchange, and transfer-learning pipelines.

OUTPUT

Threat-model templates for HDAB export review and downstream monitoring.

5 Liability & allocation for joint training (EHDS + proprietary data)

Address accountability when models are fine-tuned, combined, or co-trained: who is responsible for privacy leakage, subject-rights execution, and incident response when provenance is mixed?

WHERE IT BITES

Federated + transfer learning

OUTPUT

A doctrine of shared responsibility for model artefacts and their deployments.

FRAMING

"Anonymous results" is an export claim that must be justified.

Once models circulate, risk assessment must follow the model—across APIs, deployments, and re-use.

CONFERENCE QUESTION

What institutional capacity (HDABs, DPAs, notified bodies) is needed to audit models as artefacts of secondary use?

FINAL NOTE

Until governance shifts, EHDS may under-protect fundamental rights.

Dataset controls can prevent raw data exfiltration, yet models trained before opt-out/erasure may continue to circulate and leak information about the individuals whose data shaped them.

PRACTICAL IMPLICATION

Model export and deployment should be treated as a continuing governance domain—not the endpoint of EHDS safeguards.

DISCUSSION

Questions / collaboration

sumantapodder@gmail.com