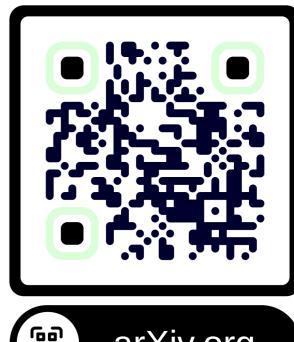
## PETLP: A Privacy-by-Design Pipeline for Social Media Data in AI Research

**Co-authors** Nick Oh socius labs Giorgos D Vrakas\* University of Cyprus Siân JM Brooke University of Amsterdam, LSE Sasha Morinière Conspiracy Watch (wcda ODI) Toju Duke Bedrock Al



from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No 101034403



















9 Decision Trees for Legal Compliance in Social Media Al Research (to-be-printed in A2 booklet; 42 x 59.4)

## TL;DR

We've mapped the maze of GDPR + IP rights (copyright, database) + Platform ToS across the entire AI research lifecycle.

From legally acquiring social media data (Extract) to confidently publishing your models, datasets, and papers (Present).

## MOTIVATION: AI Research Should Not Require a Lawyer

Social media data is essential for AI research, yet intersecting obligations create insurmountable barriers

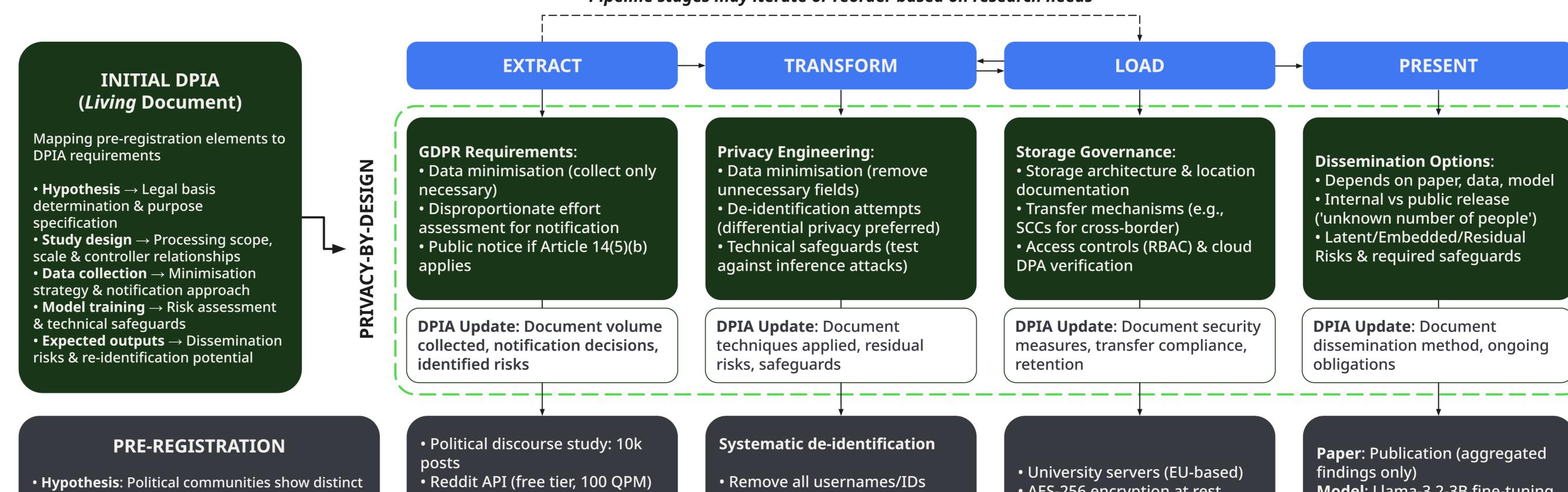
- GDPR: Core concepts undefined 'research purposes', 'personal data' in social contexts, 'scientific research' qualification
- Copyright: EU TDM permits data extraction but leaves AI model training in legal limbo
- Platform Governance: Privacy weaponised to restrict access despite public nature of posts
- Platform ToS: Prohibit automated access, redistribution, and AI training conflicting with mandatory research exceptions
- Ethics Gap: <10% of studies address these complexities beyond IRB checkboxes

Without unified framework, researchers operate in legal grey zones or abandon critical research.

## KEY QUESTIONS

- **Q1.** How can researchers navigate multiple, overlapping legal frameworks?
- Q2. When can research organisations legally override platform restrictions?
- **Q3.** How to ensure compliance throughout the entire research pipeline?





- linguistic patterns
- Study Design: Observational, 3 subreddits, Jan 2023-Dec 2024
- Collection: 10k posts via API, no usernames • Model: Llama-3.2-3B
- Outputs: Paper, fine-tuned model, methods • Status: University researcher (DSM Art. 3 qualified, GDPR legal basis as Art. 6(1)(e) public

interest)

- 3 subreddits: r/politics,
- r/conservative, r/democrats Only necessary data collected (fields: selftext, created\_utc,
- score) • Art. 9(2)(j) basis for political
- opinions Public notice: [web link]
- Hash submission IDs (prevent
- search) Generalise timestamps to
- week-level Remove posts <10 words</li>
- (uniqueness) Strip quoted usernames from content
- AES-256 encryption at rest
- Access: Researcher +
- supervisor only • Full audit logging enabled
- 5-year retention (DSM Art. 3(2)) Raw JSON preserved for 2 years for verification

Model: Llama-3.2-3B fine-tuning (Reddit permission required for model release; could be potentially fine-tuned with Opacus) **Dataset**: Not released (re-identification risk) Methods: Shared openly