# Phase 1 — Technical Research Notebook

**Project:** *The Algorithmic Panopticon: How AI Amplifies Security Exploits in Online Communities* **Phase:** 1 — Technical & Systems Analysis **Owner / Team:** Mohammad Wael (and team) **Date Created:** 2025-10-18

**How to use this notebook:** - Treat this as your living lab notebook for Phase 1. Keep entries chronological and annotate every external source with URL + short note.  
- **DO NOT** include real users’ PII or run third-party plugins on real accounts. Work only with public code and controlled dummy accounts.

## Table of Contents

1. Scope & Research Questions
2. Annotated Bibliography (public sources)
3. Plugin Audit Table (static analysis)
4. Cache & Persistence Observations (sandbox)
5. AI Exploitation Model Templates
6. Findings Synthesis (2–3 page summary)
7. IRB Notes & Ethics Checklist
8. To‑Do / Action Items

## 1. Scope & Research Questions

**Scope (short):** Discord (primary), public client-side mods (e.g., Vencord), ephemeral messaging behaviors, local cache artifacts, and conceptual AI pipelines.

**Primary Research Questions:** - Which client-side plugin features explicitly or implicitly bypass platform security/privacy controls?  
- What artifacts of “deleted” or “ephemeral” messaging persist on local systems or in transit?  
- What data formats, identifiers, or metadata enable automated ingestion/correlation by an AI?  
- What constraints limit large-scale automated exploitation today?  
- What realistic AI pipelines (inputs → processing → outputs) could scale these vulnerabilities?

## 2. Annotated Bibliography (public sources)

Add at least 8–10 entries here. For each entry use this mini-template:

* **Title / Source:**  
  **URL:**  
  **Type:** (e.g., GitHub repo, DFIR blog, Discord API docs, Reddit thread)  
  **Short note (3 bullets):**

## 3. Plugin Audit Table (static analysis)

*Instructions:* Only perform **static** reviews of public repo code and docs. Record commit hashes or release tags when possible.

| Plugin | Version / Commit | Feature(s) | Hook(s) Observed | Stores Locally? (Y/N) | Data Types | Security Control Bypassed | Notes |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Example: MessageLogger | v1.2.0 / abc123 | Logs deleted messages | onMessageDelete | Y | Text, msgID, timestamp | Deletion expectation | public repo path: /src/plugins/MessageLogger |

**Audit checklist (for each plugin)** - Repo path(s) inspected:  
- Manifest / metadata (author, license, version):  
- Event hooks & API usage (list):  
- Local storage methods (filesystem, DB, IndexedDB):  
- Network calls (endpoints, hosts) — *only if visible in code*:  
- Any obfuscation / minified bundles:

## 4. Cache & Persistence Observations (sandbox)

*Instructions:* Use a private test server and dummy accounts. Do **not** enable or run third-party plugin binaries against other people’s systems. Capture OS paths, filenames, timestamps.

| Artifact Path | OS | Observed Content | Persisted After Deletion? (Y/N) | Recovery Method (public ref) | Notes |
| --- | --- | --- | --- | --- | --- |
| e.g., %AppData%/Discord/Cache/xxx | Windows | Partial message fragments | Y | DFIR blog post 2022 | timestamps retain message index |

**Sandbox checklist** - OS & client version used:  
- Dummy account IDs (anonymized labels):  
- Steps performed (send msg → delete → inspect cache):  
- Tools used (File Explorer, Wireshark passive capture, etc.):  
- Screenshot references (stored locally) — label each screenshot:

## 5. AI Exploitation Model Templates

For each conceptual model, complete the template below and draw a simple ASCII or Mermaid diagram if helpful.

### Model: Recon Agent (example)

**Inputs:** deleted message logs (text), message metadata (timestamps, msgIDs), user metadata (display names) — *all anonymized in research outputs.*

**Processing Steps (high level):** 1. Preprocessing: normalize text, remove stopwords, mask obvious PII. 2. Embedding & clustering: semantic embeddings → cluster by topic/personality indicators. 3. Behavioral inference: apply psycholinguistic models to infer traits. 4. Output: persona summaries, high‑value target lists.

**Outputs:** persona profiles (non-PII), risk scores, recommended social-engineering vectors (***do not*** implement or provide step‑by‑step attack instructions).

**Feasibility:** High/Medium/Low — justify with references.

**Ethical/Harms:** Describe potential harms and mitigation strategies.

**Mermaid diagram suggestion:**

flowchart LR  
 A[Deleted logs] --> B[Preprocess]  
 B --> C[Embeddings & Clustering]  
 C --> D[Behavioral Inference]  
 D --> E[Persona Profiles]

## 6. Findings Synthesis (2–3 page summary)

Use this section to draft the technical narrative you will hand off to Phase 2. Structure suggestions: - Executive summary (1 paragraph) - Key technical findings (bulleted) - What assumptions users make vs. reality - Potential AI amplification vectors (high-level, non-actionable) - Limitations & open questions

## 7. IRB Notes & Ethics Checklist

* Work restricted to public repos and controlled sandbox.
* No collection of real user data.
* All forum content paraphrased and anonymized.
* All outputs will be non-actionable and high-level.

**IRB draft bullet points:** - Study aims and non-malicious intent.  
- Data sources (public code, sandbox artifacts, paraphrased public forum posts).  
- Privacy & de-identification steps.  
- Risk mitigation and storage policies.

## 8. To‑Do / Action Items (editable checklist)

* Create sandbox server & 3 dummy accounts (anonymized labels).
* Populate annotated bibliography (≥8 sources).
* Fill plugin audit rows for MessageLogger, AlwaysTrust, and 3 other plugins.
* Run sandbox cache test and populate artifact table.
* Draft Recon Agent & Forensic Aggregator one-pagers.
* Draft 2–3 page technical summary for Phase 2.

**Appendix: Quick safe-research reminders** - Never store screenshots or raw logs that include real usernames or real server names in shared drives without anonymization.  
- When paraphrasing forum posts, change wording and remove unique phrases to prevent search-based re-identification.  
- If uncertain, include the question in IRB notes.

*End of notebook template.*