The Private Palantir

AI Agents, Public Data, and the End of Practical Obscurity

Author: Benjamin Metzig

Affiliation: Wissenschaftswelle.de, independent science and knowledge magazine, Germany

Date: June 12, 2026

Published by: Wissenschaftswelle.de

Research report prepared for public discussion and further scholarly review.

The Private Palantir

A Research Report on AI Agents, Public Data, and the End of Practical Obscurity

Status: June 12, 2026
Purpose: To provide a robust research basis that can be used in public and policy debates on the risk of AI-assisted personal profiling by private, non-state actors.
Core finding: Comprehensive private surveillance should not be claimed as a present-day reality. What can be stated with confidence is this: the technical, organizational, and economic conditions for targeted automated personal profiling from public data are improving rapidly. This lowers the threshold for stalking, doxxing, intimidation, fraud, informal employer surveillance, and local forms of social power.

Author: Benjamin Metzig
Affiliation / project context: Wissenschaftswelle.de, independent science and knowledge magazine, Germany
Project site: https://www.wissenschaftswelle.de/
Recommended citation: Metzig, B. (2026). *The Private Palantir: A Research Report on AI Agents, Public Data, and the End of Practical Obscurity*. Wissenschaftswelle.de.
Transparency note: This report was developed by Benjamin Metzig for Wissenschaftswelle.de. Editorial responsibility, source assessment, final judgement, and any decision to publish remain with Benjamin Metzig.

Abstract

Public personal information is not equivalent to unrestrictedly usable information. Many privacy protections depend on practical obscurity: dispersion, effort, context, ambiguity, time, and limited linkability. This report argues that AI agents, multimodal models, long-context systems, memory architectures, tool use, and open-weight deployment are weakening that protection by lowering the cost of collecting, linking, storing, and interpreting public traces about identifiable people. The report defines **Private Intelligence Automation** as a distinct risk category: the non-state, automated, repeated, and often opaque aggregation of public or semi-public personal information for private purposes. The evidence does not establish widespread comprehensive private surveillance. It does establish relevant technical capabilities, adjacent harm domains, and governance gaps. Doxxing, technology-facilitated abuse, social engineering, informal employment screening, data brokerage, and chilling effects provide empirical anchors for the risk logic. A balanced pilot audit of 125 public municipal-publication PDFs from five German cities further shows that recurring city bulletins and official gazettes can contain durable combinations of contact patterns, roles, dates, school/youth/family-related terms, and metadata. The report proposes a non-operational research and policy agenda focused on prevalence measurement, synthetic agent evaluations, institutional exposure audits, comparative governance, and safeguards against repeated personal aggregation.

Keywords: AI agents; public data; practical obscurity; privacy; doxxing; technology-facilitated abuse; data brokers; OSINT; open-weight models; digital violence; personal profiling; AI governance.

Executive Summary

This report examines the thesis that AI agents, multimodal models, web- and computer-use systems, long context windows, memory architectures, and open-weight models are creating a new risk category: **Private Intelligence Automation**. The term refers to the non-state automated collection, linking, storage, and assessment of personal information that is publicly available.

The central danger does not lie in one "magical" AI capability. It lies in the coupling of already visible building blocks: research agents, document and image analysis, tool use, browser operation, network inference, profile building, memory, repetition, and falling costs. Scattered public data can become private files. Files can become social leverage.

The evidence for individual technical building blocks is strong: OpenAI, Anthropic, Google, Microsoft, Meta, Mistral, DeepSeek, and Qwen document or release systems with multi-step research, tool use, computer use, multimodal processing, long contexts, or open weights. At the same time, benchmarks such as OSWorld, WebArena, and Online-Mind2Web show that agents still make errors and remain slow and unreliable in real web and computer tasks. These limits dampen exaggerated claims, but they do not remove the risk: partial automation is enough for many abuse scenarios.

The evidence for social harms is moderate to strong: Europol, the German Federal Criminal Police Office (BKA), and the FBI already document that AI can professionalize fraud, social engineering, and cybercrime. HateAid and the Technical University of Munich show that digital violence already deters political engagement in Germany. From a data protection perspective, publicly available personal data is not free for unrestricted use; the GDPR, EDPB guidelines, and guidance from the German Data Protection Conference emphasize legal basis, purpose limitation, transparency, data minimization, and the special sensitivity of personal data processing.

The most important uncertainty is prevalence: there is still little robust empirical data on how often private actors already use AI systematically for personal research, stalking, doxxing, or informal profiling. The appropriate assessment is therefore sober: **not that comprehensive private surveillance is already everyday reality, but that the threshold for targeted profiling is falling, the forms of harm are plausible and partly connect to existing digital violence, and the governance gaps are large enough that research and policy should act now.**

Contribution and Scope

This report makes five contributions:

1. **Conceptual contribution:** It defines **Private Intelligence Automation** as a distinct risk category: non-state, automated, repeated, and often opaque personal aggregation from public or semi-public traces.

2. **Theoretical contribution:** It connects practical obscurity, contextual integrity, and online obscurity to the specific capabilities of agentic AI: search, synthesis, multimodal analysis, memory, tool use, and updating.
3. **Evidentiary contribution:** It separates evidence for technical capability from evidence for social harm and from evidence for prevalence. This avoids both overclaiming and premature reassurance.
4. **Governance contribution:** It proposes a non-operational research and policy agenda that can study and reduce harm without publishing abuse-enabling workflows or undermining legitimate OSINT, journalism, cybersecurity, and human-rights work.
5. **Empirical pilot contribution:** It includes a bounded audit of public local-institution PDFs to test whether institutional publication practices produce aggregable traces without targeting any individual person.

The scope is deliberately narrower than general AI safety and broader than any single harm category. The report does not claim that a new technology has created stalking, doxxing, fraud, or discrimination from scratch. It argues that agentic AI changes the cost, speed, repeatability, and apparent authority of practices that already exist.

Key Findings

1. **The building blocks exist.** Multi-step web research, document analysis, image understanding, tool calling, computer use, long contexts, and agent harnesses are documented current capabilities, even though their quality varies significantly.
2. **The logic of harm does not require perfect surveillance.** False, selective, or uncertain profiles can already cause harm in workplaces, local conflicts, political intimidation, stalking, and fraud before they are ever verified.
3. **Practical obscurity is a real protective layer.** Information can be public and still be functionally protected by dispersion, effort, context, and forgetting. Agentic systems reduce precisely this friction.
4. **Open-weight models are a distinct risk factor.** They matter for research, competition, and digital sovereignty, but they make centralized abuse control harder because local and adapted use is more difficult to observe.
5. **The main risk is not limited to public figures.** Especially relevant contexts include separation and stalking, politically engaged people, journalists, local officials, employees, job applicants, minors, queer people, people in small communities, and individuals with visible professional or social roles.
6. **Regulation has blind spots.** The GDPR and the EU AI Act provide important principles, but private, informal, local, and hard-to-prove profiling remains only partly addressed in practice.
7. **The research gap is itself a risk.** Without empirical data on use, harms, offender profiles, and effective countermeasures, the debate remains vulnerable to both panic and premature reassurance.

Evidence Classes

This report distinguishes four evidence classes:

Class	Meaning	Use in the report
Established	Supported by primary sources, official documentation, legal sources, scientific work, benchmarks, or robust situation reports.	Technical capabilities, data protection principles, known digital violence, cybercrime trends.
Plausible	Supported by established building blocks and traceable causal chains, but without direct prevalence data for the exact scenario.	The link between agent capabilities and private personal profiling.
Uncertain	Relevant indicators exist, but the data base, frequency, effect, or time horizon is unclear.	Spread of private agentic dossiers, effectiveness of protective measures.
Speculative	Possible, but currently not robust enough for strong claims.	Three- to five-year normalization of comprehensive local profiling agents.

Hypothesis and Research Logic

Main Thesis

H1: Over the next one to three years, the risk of non-state AI-assisted personal profiling from public data will increase significantly because technical capabilities, automation, and falling costs together weaken the protective effect of practical obscurity.

Sub-Hypotheses

Sub-hypothesis	Evidence class	Short rationale
H1a: Agents can already partially automate research, extraction, tool use, and synthesis.	Established	OpenAI Deep Research and ChatGPT Agent, Anthropic Computer Use, tool-use documentation, Google/Mistral Deep Research, Microsoft agent strategy.
H1b: Agents remain unreliable, but partial automation is enough for many abuse scenarios.	Established/Plausible	OSWorld, WebArena, and Online-Mind2Web show limits; stalking, fraud, and reputational harm do not require perfect full automation.
H1c: Open-weight models lower control and cost barriers.	Established/Plausible	Stanford AI Index, OECD, Meta Llama, DeepSeek-R1, Qwen3-Coder.
H1d: Public data loses part of its social protective effect through automated linking.	Established/Plausible	Practical obscurity, contextual integrity, data protection law, and experience with data brokers and facial recognition support the underlying logic.
H1e: The spread of private AI dossiers has not yet been measured sufficiently.	Uncertain	Robust case numbers, offender typologies, and longitudinal studies are missing.

Causal Chain

Stage	Description
1. Public and semi-public traces	Public records, institutional pages, old PDFs, photos, event listings, professional profiles, and social media traces.
2. AI research and document analysis	Search, extraction, OCR, summarization, and cross-source comparison reduce the effort of collection.
3. Normalization	Names, places, roles, dates, organizations, and repeated identifiers are made comparable.
4. Inference	Hypotheses are formed about a person, environment, interests, relationships, vulnerabilities, or routines.
5. Storage and updating	Episodic research becomes a persistent file that can be revisited or refreshed.
6. Use	Profiles, dossiers, or network maps can support stalking, doxxing, fraud, discrimination, or intimidation.
Cross-cutting driver	Falling costs and open-weight/local systems increase feasibility and reduce centralized oversight.

The logic is deliberately non-operational. It does not describe instructions, but the abstract risk chain.

Methodological Framework

This report is a synthetic risk analysis, not a prevalence study. It uses a structured narrative-synthesis approach: technical evidence, legal sources, policy reports, privacy theory, and documented harm domains are brought together to evaluate whether a new risk category is emerging.

Research Questions

The report is guided by five research questions:

1. Which AI capabilities are relevant to the automated aggregation of public personal information?
2. How reliable are these capabilities today, and which limits matter for risk assessment?
3. Which existing harm domains could be intensified by AI-assisted aggregation?
4. Which legal, institutional, and product-governance frameworks already address the risk, and where are their practical gaps?
5. Which empirical studies would be needed to measure prevalence, affected groups, and effective interventions?

Source Corpus

The analysis combines five source categories:

Source category	Examples	Role in the report	Main limitation
Provider documentation and model releases	OpenAI, Anthropic, Google, Microsoft, Meta, Mistral, DeepSeek, Qwen	Establish current product and model capabilities.	Providers have incentives to emphasize capability and underdescribe misuse.

Source category	Examples	Role in the report	Main limitation
Benchmarks and academic evaluations	OSWorld, WebArena, Online-Mind2Web, AI Agents That Matter	Establish limits, reliability problems, cost issues, and real-world transfer gaps.	Benchmarks may not match abusive workflows or low-stakes partial automation.
Legal and regulatory sources	GDPR, EDPB, German Data Protection Conference, EU AI Act, NIST AI RMF	Define principles for personal-data processing, risk management, and governance.	Law differs by jurisdiction and may lag informal private uses.
Situation reports and civil-society evidence	Europol, BKA, FBI IC3, HateAid/TUM, Privacy International	Connect the technical mechanism to existing harm domains.	Often not designed to isolate the marginal effect of AI agents.
Privacy and information theory	Practical obscurity, contextual integrity, online obscurity, OSINT ethics	Provide the conceptual basis for why publicness does not equal unrestricted use.	Conceptual frameworks need empirical operationalization.

Selection Criteria

Sources were prioritized when they met at least one of the following criteria:

- Primary documentation from a model provider, regulator, court, public authority, or recognized standards body.
- Peer-reviewed research, recognized preprints, or benchmark papers with clear methodology.
- Official law-enforcement, cybercrime, data-protection, or civil-society reports relevant to fraud, stalking, doxxing, digital violence, or profiling.
- Theoretical work that is widely cited in privacy, data protection, OSINT, or information-governance debates.
- Direct relevance to one of the report's core mechanisms: public-data aggregation, agentic automation, multimodal analysis, memory, open-weight deployment, or downstream harm.

Sources were not used as central evidence when they were purely anecdotal, promotional without technical detail, unverifiable, or operationally instructive for abuse. The report deliberately avoids concrete search operators, prompts, step-by-step workflows, database schemas, target-selection logic, or operational surveillance instructions.

Evidence Assessment

Each claim is assessed along four dimensions:

Dimension	Question	Possible values
Capability	Does the relevant technical function exist?	Established / plausible / unclear
Directness	Does the evidence address the exact scenario or only a related mechanism?	Direct / adjacent / indirect
Reliability	Is the evidence robust across sources, methods, and contexts?	Strong / moderate / weak
Prevalence	Is there evidence of widespread real-world use?	Measured / indicated / unknown

This creates an important distinction: a technical capability can be established while the prevalence of a specific abusive use remains unknown. Much of the report sits precisely in this space.

Terminology

Practical Obscurity

Practical obscurity describes the condition in which information is publicly accessible but functionally protected by dispersion, effort, context, age, ambiguity, or the lack of easy linking mechanisms. The term connects to the U.S. legal concept of "practical obscurity" in U.S. Department of Justice v. Reporters Committee for Freedom of the Press. That case concerned the difference between scattered public records and machine-centralized compilations.

Contextual Integrity

Helen Nissenbaum's theory of contextual integrity describes privacy not as mere secrecy, but as compliance with context-dependent informational norms: who may share or use which information, in which context, and for which purpose? This perspective is central because the private Palantir emerges above all through purpose shift: data intended for local publicity, association communication, or professional visibility becomes profiling material.

AI Agent

An AI agent is understood here as a system that can decompose a goal into subtasks, use tools, process intermediate results, possibly store them, and iteratively adapt its approach. OpenAI describes ChatGPT Agent as a system that can research and act, including through browsers, files, data sources, and forms. Anthropic documents computer-use capabilities in which Claude interprets screenshots and performs mouse and keyboard actions. Google and Mistral describe Deep Research features as multi-step research and synthesis systems.

Harness

A harness is the control and working environment around a model: memory, tool access, repetitions, evaluation, rules, logging, and, where appropriate, human approvals. For the private Palantir, the harness is often more important than the individual model because it turns episodic search into repeatable observation.

Private Intelligence Automation

Private Intelligence Automation refers to the automated collection, linking, storage, and assessment of public personal information by non-state actors. The term is analytical and is not intended as an established legal category.

Agentic Stalking

Agentic stalking refers to AI-assisted stalking that is persistent, memory-based, updating, and comparative. This term is also a proposal for describing a risk, not an already established criminal offense.

Conceptual Framework and Operationalization

The report treats the private Palantir not as a single technology, but as a socio-technical mechanism. The mechanism emerges when public traces, AI capabilities, storage, repeated updating, and a harmful or disproportionate purpose are combined.

Core Analytical Variables

Variable	Operational question	Observable indicators
Public-trace density	How many public or semi-public traces about a person exist across contexts and time?	Searchable profiles, old PDFs, local media items, social media traces, photos, event listings, institutional pages.
Linkability	How easily can traces be connected to the same person?	Unique names, locations, employers, photographs, repeated handles, contact details, co-occurring organizations.
Context collapse	Are traces interpreted outside the context in which they were created?	Old statements treated as current, professional visibility treated as private preference, event attendance treated as ideological commitment.
Automation depth	How much of collection, extraction, comparison, and summarization is automated?	One-off chatbot use, repeated agent workflow, persistent memory, local retrieval system, automated refresh.
Sensitivity of inference	Are protected or intimate attributes inferred?	Political views, religion, health, sexual orientation, union proximity, family situation, vulnerability.
Use context	How are outputs used?	Curiosity, employment screening, local conflict, harassment, fraud, doxxing, stalking, political intimidation.
Contestability	Can the affected person know, challenge, correct, or delete the profile?	Transparency, access rights, audit logs, documented decisions, reporting channels, deletion mechanisms.

What Would Support or Weaken the Main Thesis?

The thesis would be strengthened by empirical evidence that private actors are already using AI tools to aggregate public personal data in stalking, doxxing, employment screening, fraud preparation, political intimidation, or local reputational campaigns. It would also be strengthened by controlled studies showing that agentic workflows substantially reduce time, cost, or skill requirements compared with conventional search.

The thesis would be weakened if robust empirical studies showed that AI adds little practical advantage over ordinary search for abusive personal aggregation, that existing platform safeguards reliably block repeated personal profiling across models and providers, or that affected groups do not experience measurable additional harm or chilling effects.

The thesis would be falsified in its strongest form if agentic systems proved technically incapable of reliably collecting and linking public traces even at partial-automation levels, or if adoption costs and safeguards made such workflows impractical for ordinary private actors. Current evidence does not support that stronger dismissal.

Boundary Conditions

The report does not treat all public-interest research as suspect. Legitimate OSINT, investigative journalism, human-rights documentation, cybersecurity research, academic work, archival practice, and public-accountability reporting can rely on public information. The relevant distinction is not simply whether data is public, but whether aggregation is proportionate, lawful, transparent where required, verified, sensitive to bystanders, and connected to a legitimate purpose.

Relation to Existing Concepts

Private Intelligence Automation overlaps with several established concepts, but is not identical to any one of them.

Existing concept	Core focus	What Private Intelligence Automation adds
OSINT	Collection and analysis of publicly available information, often for legitimate investigative purposes.	Focus on private, opaque, non-public-interest profiling and the erosion of practical obscurity.
Doxxing	Disclosure of personal information to expose, intimidate, or harm.	Includes pre-disclosure dossier formation, monitoring, inference, and use without publication.
Cyberstalking / tech-facilitated abuse	Repeated harassment, monitoring, coercion, or control through technology.	Highlights agentic aggregation of scattered public traces as an enabling layer.
Data brokerage	Commercial collection, enrichment, and sale of personal data.	Includes non-commercial actors, local conflicts, interpersonal abuse, and bespoke dossiers.
Social-media screening	Use of online traces in hiring or institutional assessment.	Extends to cross-source AI synthesis, sensitive inference, memory, and informal decision influence.
State surveillance	Institutional monitoring by public authorities.	Focuses on private and non-state actors whose use may be invisible, informal, and difficult to contest.
Platform harassment	Abuse occurring on or through digital platforms.	Includes off-platform aggregation, local use, employment contexts, fraud preparation, and physical-world consequences.

The distinctive feature is the combination: **public traces + automated linkage + memory or repetition + private purpose + weak contestability**.

Central Mechanism

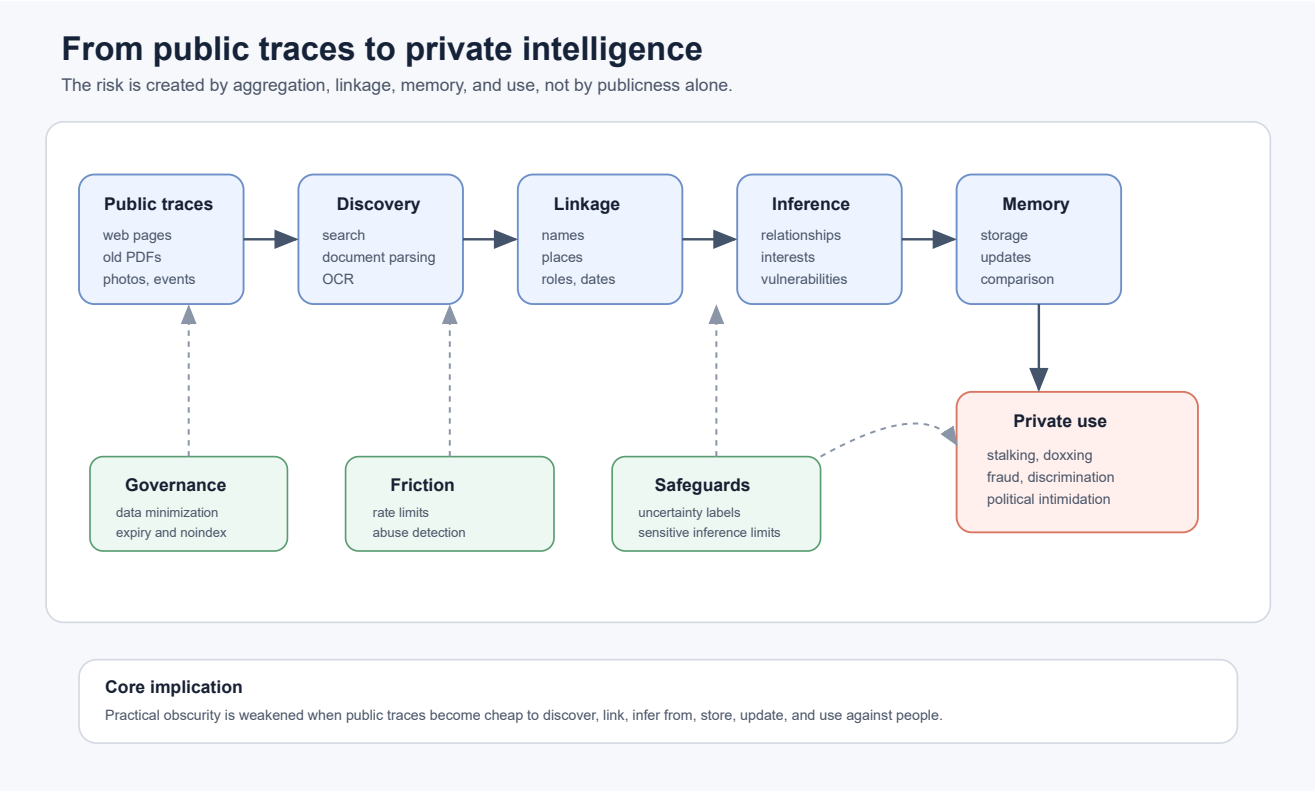

Figure 1. From public traces to private intelligence. Public traces become harmful private intelligence when discovery, extraction, linkage, inference, memory, and use are combined. Governance can intervene at multiple stages: by reducing institutional overexposure, preserving friction, limiting sensitive inference, marking uncertainty, and constraining repeated personal aggregation.

Capability Matrix

Capability	Status as of June 2026	Evidence base	1-3 years	Relevance for the private Palantir	Limit / uncertainty
Web search and Deep Research	Product category exists; systems can conduct multi-step research and synthesize sources.	Established: OpenAI Deep Research, ChatGPT Agent, Google/Mistral Deep Research.	Likely faster, better cited, and more integrated into workflows.	High relevance for dossiers from scattered sources.	Source errors, hallucinations, mistaken identity, platform rules.
Document analysis	Strong PDF, spreadsheet, text, and file analysis in frontier systems; OCR/parsing as standard components.	Established through provider features and practice.	Very likely cheaper and better, including locally.	Old association PDFs, official gazettes, event programs, and presentations become more analyzable.	Poor scans, missing context, outdated data.

Capability	Status as of June 2026	Evidence base	1-3 years	Relevance for the private Palantir	Limit / uncertainty
Image and screenshot understanding	Multimodal models can describe images, diagrams, and screenshots and extract clues.	Established: OpenAI, Anthropic, Meta Llama 4, multimodal benchmarks.	Further quality improvement is plausible.	Group photos, screenshots, logos, places, and visual clues can provide context.	Errors in identity, scene, time, and meaning.
Facial recognition	Technically available, legally highly sensitive, and restricted in many platforms and products.	Established: Clearview cases, data protection enforcement, biometric regulation.	Abuse remains plausible, especially outside central providers.	Very high harm potential when people can be identified across images.	Legal limits, provider policies, error rates, bias, access barriers.
Network mapping	Graphs from contacts, organizations, places, and events are technically straightforward, but not always reliable.	Plausible, supported by OSINT practice and graph analysis.	Strengthened by agent harnesses and memory.	Makes even cautious people visible through their environment.	High risk of misinterpretation: contact is not relationship; proximity is not consent.
Agents and tool use	Tool calling, browser agents, and computer use are documented, but error-prone.	Established: OpenAI, Anthropic, Microsoft, OSWorld, WebArena, Online-Mind2Web.	Progress likely, but not linear.	Automates repeated search, extraction, updating, and filing.	Benchmarks show major gaps compared with humans.
Memory and long contexts	Long contexts and project/workspace memory are becoming more important in products.	Established/Plausible: provider statements, agent architectures.	Higher relevance through more persistent agents.	Difference between one-off research and a file.	Memory can amplify errors and harden old hypotheses.
Local and open-weight models	Capable open weights are spreading; multimodal and agentic models are available.	Established: Stanford AI Index, OECD, Meta Llama, DeepSeek, Qwen.	Strong progress is plausible.	Circumvention of centralized platform control, lower costs, adaptability.	Local quality varies; hardware, know-how, and model size remain barriers.
Cost development	Inference costs and hardware efficiency have fallen sharply; small models are becoming more capable.	Established: Stanford AI Index 2025/2026.	Further cost reduction is likely, but market-dependent.	Scales from individual research to many people or regular monitoring.	Exact costs for abusive workflows are barely measured empirically.
Safety and policy boundaries	Central providers have rules against doxxing, stalking, biometric identification, and abuse.	Established: provider policies and product restrictions.	Stronger on central platforms, weaker in decentralized use.	Can slow abuse, but cannot prevent it completely.	Shadow use, local models, manual intermediate steps.

Prioritization of Technical Drivers

Not all AI capabilities contribute equally to the risk. A sharper causal model distinguishes primary drivers, amplifiers, special cases, and governance modifiers.

Priority	Capability	Why it matters
Primary driver	Search and synthesis	Converts scattered traces into readable narratives and reduces the effort of cross-source review.
Primary driver	Document parsing and OCR	Makes old PDFs, scanned documents, forms, newsletters, event programmes, and archives more accessible.
Primary driver	Memory and repeated workflows	Turns one-off curiosity into monitoring, comparison, and dossier maintenance.
Primary driver	Falling cost	Expands the number of people, sources, and repetitions that become feasible for private actors.
Secondary amplifier	Multimodal analysis	Adds photos, screenshots, logos, locations, and visual context to text-based traces.
Secondary amplifier	Entity resolution and network inference	Links people, organizations, places, and events, but carries high misinterpretation risk.
High-harm special case	Facial recognition	Can transform images into identity infrastructure; legally and ethically sensitive.
Governance modifier	Open-weight and local deployment	Does not create the harm by itself, but weakens centralized platform oversight and auditability.

This prioritization helps avoid vague claims about "AI" in general. The most immediate governance focus should be repeated personal aggregation, document extraction, memory, sensitive inference, and low-cost scaling.

Assessment Scale

The probability ratings in this report are qualitative risk assessments, not statistical forecasts.

Rating	Meaning
Low	The mechanism is technically possible, but currently substantially limited by cost, effort, access, quality, or enforcement.
Medium	Several building blocks are available; abuse is plausible without specialized infrastructure, but not shown to be widespread.
High	The mechanism connects to already observed digital violence, fraud, or profiling practices and is visibly facilitated by AI.
Very high	Occurrence or harm is especially likely in certain contexts, such as stalking, doxxing, sensitive data, or minors.

Threat Model

Protected Interests

Protected interest	Why it is affected
Privacy and practical obscurity	Public individual traces become personal profiles through aggregation.
Safety and freedom of movement	Stalking, doxxing, and intimidation can connect digital and physical danger.
Professional fairness	Informal AI research can influence hiring and career decisions.
Political participation	Profiling and digital violence can encourage withdrawal from public life and civic engagement.
Child protection	Traces of minors are often created by others and can later be aggregated.
Truth and reputation	Selective, false, or contextless profiles can cause reputational harm.

Actors, Motives, Capabilities

Actor	Motive	Capability today	Plausible development	Typical harms	Countermeasures
Ex-partner / stalker	Control, revenge, forced proximity	Manual research plus AI summarization; partial monitoring	More ready-made agents and local workflows	Fear, intimidation, physical escalation, contact with social environment	Stalking law, platform reports, evidence preservation, support services, visibility controls
Local conflict actor	Reputational harm, social power	Public profiles, association pages, local media, AI synthesis	Faster dossier building and context loss	Exclusion, rumors, pressure on family or employer	Local media standards, anti-doxxing rules, data protection in associations and schools
Employer / shadow recruiter	Risk selection, control	Informal web search, AI analysis of traces	Automated candidate profiles	Discrimination, chilling effect, lack of transparency	Employment data protection, audit duties, clear bans on informal profiling
Political extremist / intimidator	Deterrence, enemy marking	OSINT, social media analysis, campaign logic	Network maps, target prioritization, coordinated campaigns	Doxxing, threats, withdrawal from civic engagement	Law enforcement, platform coordination, victim protection, monitoring of coordinated attacks
Fraudster / social engineer	Money, access, identity abuse	Personalized messages using real details	Scaled and more credible attack preparation	Financial loss, identity theft, account takeover	Fraud prevention, MFA, warning systems, training, abuse detection
Data broker / gray-market service provider	Monetization	Aggregation of public and commercial data	AI-assisted profile products for niche markets	Discrimination, loss of control, opacity	Data protection enforcement, access rights, bans on sensitive inference
Curious individual	Curiosity, social control	Chatbots, search engines, screenshots	Lower barriers through ready-made assistants	Boundary violations, social surveillance	Norms, product limits, education about context and consent

Attack Path and Countermeasures

Phase	Risk	Friction / countermeasure
Collection	Public traces become discoverable at greater breadth.	Indexing options, robots.txt, platform limits, protection of old content, data minimization by institutions.
Normalization	Names, places, roles, and dates are unified.	Context labels, metadata reduction, source notes, expiry dates, institutional publication standards.
Inference	Clues become assumptions about relationships, political views, health, or vulnerabilities.	Ban sensitive inference, label uncertainty, audit duties, policy boundaries at AI providers.
Storage	Episodic research becomes a file.	Storage limitation, purpose limitation, deletion duties, logging, data subject rights.
Updating	Profiles are monitored over time.	Monitoring restrictions, abuse detection, rate limits, legal thresholds for repeated personal research.
Use	Dossiers are used for pressure, fraud, discrimination, or doxxing.	Criminal and civil law, platform enforcement, employer rules, victim support, evidence channels.

Risk Graphics

Risk Level by Combination of Data Situation and Automation

Automation / data situation	Few current traces	Many scattered traces	Historical traces + images + network
One-off manual search	low	low-medium	medium
AI-assisted synthesis	low-medium	medium	medium-high
Repeated agent workflow	medium	high	very high
Local/open-weight workflow with memory	medium	high	very high

Threat Level by Time Horizon

Period	Qualitative threat level	Interpretation
2024	1.5 / 5	Early agent and multimodal capabilities; significant friction remained.
2025	2.3 / 5	More public agent workflows, stronger document/image analysis, falling costs.
2026	3.0 / 5	Agentic research, long context, memory, and open-weight capabilities become more accessible.
2027-2029	4.0 / 5	Plausible normalization of lower-cost repeated personal aggregation if governance does not adapt.

The chart is a qualitative assessment, not a measurement series. It condenses the development derived in the report: rising technical capabilities, falling costs, better agents, and governance gaps.

Exposure by Scenario

Scenario	Plausibility of occurrence	Harm severity	Evidence base
Personalized fraud	high	high	established by cybercrime reports; AI link plausible to established
Political intimidation	medium-high	high	digital violence established, AI amplification plausible
Separation / stalking context	medium	very high	stalking/digital violence established, agentic automation plausible
Employer profiling	medium	medium-high	data protection risk established, informal use hard to measure
Local reputational harm	medium	medium-high	plausible harm logic, limited prevalence data
Fully automated comprehensive private surveillance	low to uncertain	very high	currently not robustly established

Technical Evidence

Research Agents and Deep Research

OpenAI describes Deep Research as an agentic capability that conducts multi-step internet research, analyzes large amounts of text, images, and PDFs, and synthesizes results. ChatGPT Agent integrates research and actions in an agent mode. Mistral describes Deep Research in Le Chat as multi-step public web research with sources. These sources do not prove that private surveillance is already happening at scale; they do show that the product category of "AI-assisted research and synthesis" is real.

Assessment: The capability for research support is established. Its abusive transfer to personal profiles is plausible. The actual frequency of use is uncertain.

Tool Use, Computer Use, and Long Action Chains

Anthropic documents tool use and computer use with screenshot, mouse, and keyboard interaction. OpenAI introduced Operator and later integrated it into ChatGPT Agent. In 2025, Microsoft described the "age of AI agents" with memory, tool use, and open agent standards. These developments show that agents increasingly do not merely generate text, but operate work environments.

Benchmarks urge caution: OSWorld showed in 2024 large gaps between humans and agents on real computer tasks; WebArena showed low success rates for GPT-4-based agents in realistic web environments; Online-Mind2Web warned in 2025 against overoptimistic conclusions from older web-agent benchmarks. "AI Agents That Matter" also criticizes the fact that agent evaluations often insufficiently account for cost, robustness, and real-world transferability.

Assessment: Progress and limits are both established. For this report, precisely this dual movement matters: agents are not reliable enough for comprehensive surveillance, but they are good enough to

reduce effort across many subtasks.

Multimodality and Documents

Multimodal models can analyze texts, images, screenshots, diagrams, and documents. Meta describes Llama 4 as open-weight and natively multimodal. Providers of frontier systems integrate vision, PDF processing, and file analysis into everyday products. For personal profiling, this matters because public traces do not exist only in text: association PDFs, flyers, group images, presentations, local websites, screenshots, logos, and event programs can contain clues.

Assessment: The growing analyzability of heterogeneous data is established. The reliability of concrete personal and contextual inference remains uncertain.

Open-Weight Models and Costs

The Stanford AI Index 2025 documented sharply falling inference costs and a narrowing gap between open and closed models on certain benchmarks. The AI Index 2026 describes open-weight models as more competitive and, at the same time, harder to evaluate. The OECD analyzes open-weight models as an engine of innovation while warning of risks such as misuse, privacy issues, and proliferation. Meta, DeepSeek, and Qwen show that capable open or open-weight models with multimodal, reasoning, or agentic capabilities are available.

Assessment: The spread and performance gains of open weights are established. It is plausible that local use weakens central platform control. Concrete cost and adoption curves for abusive personal profiling remain uncertain.

Academic Anchors in Adjacent Harm Domains

The report's risk claim is strengthened by the fact that adjacent harm domains are already studied empirically. These literatures do not prove the prevalence of private AI dossiers, but they establish the social mechanisms into which AI-assisted aggregation can plug.

Doxxing Research

Snyder, Doerfler, Kanich, and McCoy's quantitative study of doxing showed that malicious disclosure of personal information can be detected and characterized at scale, and that the practice is associated with identity theft, harassment, and online abuse. Later literature reviews and platform-safety research treat doxxing as a privacy and safety harm rather than merely offensive speech. This supports the report's claim that public or semi-public personal information becomes dangerous when assembled and deployed against a person.

Intimate Partner Surveillance and Technology-Facilitated Abuse

Research on intimate partner surveillance shows that abusers use spyware, account compromise, social engineering, and online resources to monitor and control partners or ex-partners. Technology-facilitated abuse research also emphasizes that digital tools can extend coercive control beyond physical proximity. Agentic AI is not required for this harm to exist; the concern is that it can lower the effort required to gather, compare, and preserve public traces.

Chilling Effects and Dataveillance

Empirical and theoretical work on chilling effects shows that perceived surveillance can inhibit lawful communication, information seeking, political participation, and self-expression. This literature matters because the private Palantir does not need to produce direct physical harm in every case. If people withdraw from public speech, local politics, journalism, science communication, or associational life because they expect automated profiling, democratic participation is already affected.

Employment Screening and Discrimination

Research on social-media screening in hiring shows that employers may search for applicants online and that protected or sensitive traits can become visible outside formal hiring channels. This supports the shadow-profiling scenario: even where formal AI hiring tools are regulated, informal AI-assisted background research can influence decisions without transparency, documentation, or contestability.

Data Brokers and Aggregation Markets

Data-broker research and policy analysis demonstrate that aggregation itself is a source of power. Data brokers combine public, commercial, inferred, and behavioral data, often with limited visibility for affected people. Private Intelligence Automation differs because it includes non-commercial, local, and interpersonal actors, but the data-broker literature establishes the broader point that aggregation changes the meaning and risk of individual data points.

Academic Source Coverage Audit

The source base is strongest where the report connects to established adjacent harms. It is weaker where the exact phenomenon -- private agentic dossiers from public traces -- has not yet become a mature research object.

Domain	Academic coverage in this report	Assessment	Further strengthening needed
Practical obscurity and contextual integrity	Strong: legal theory and privacy theory are included.	Sufficient for conceptual foundation.	Add more recent privacy scholarship on aggregation and inference if preparing journal submission.
AI agents and web/computer-use systems	Moderate to strong: benchmarks and provider documentation are included.	Sufficient for capability and limitation claims.	Add more peer-reviewed agent-evaluation work as the field matures.
Doxxing and harassment	Stronger after adding quantitative doxxing and literature-review anchors.	Sufficient for adjacent harm logic.	Add more cross-cultural and non-English doxxing research if available.
Technology-facilitated abuse and stalking	Moderate to strong: intimate partner surveillance and tech-abuse work are included.	Sufficient for separation/stalking scenario.	Add clinical, criminological, and victim-support studies for prevalence estimates.
Employment screening and discrimination	Moderate: social-media hiring discrimination and AI hiring literature are included.	Sufficient for risk plausibility.	Add jurisdiction-specific employment-law analysis and audit studies.
Data brokers and aggregation markets	Moderate: policy and scholarly analysis are included.	Sufficient for aggregation-market analogy.	Add empirical market mapping and data-broker enforcement studies.

Domain	Academic coverage in this report	Assessment	Further strengthening needed
Chilling effects and democratic participation	Moderate to strong: dataveillance and online-surveillance studies are included.	Sufficient for behavioural-risk framing.	Add studies on journalists, local politicians, scientists, and activists.
Prevalence of private AI dossiers	Weak by design: few direct studies exist.	Central research gap.	Requires new empirical work; cannot be solved by citation alone.

Overall, the report now has enough academic anchoring for a serious research report. For a peer-reviewed journal submission, the next improvement would not be simply adding more citations, but replacing some web-only provider and policy links with stable bibliographic entries, DOIs, and a reproducible literature corpus.

Pilot Study: Local Institutional Exposure

To move beyond conceptual synthesis, this report includes a balanced empirical pilot study on local institutional exposure. The study asks whether recurring public municipal publications create aggregable traces even when no sensitive database is breached and no individual person is targeted.

Pilot Research Question

How aggregable are public institutional PDFs from local civic contexts, and which document-level features increase the risk that public traces can later be converted into private intelligence?

Corpus and Ethics

The current pilot uses 125 publicly accessible municipal-publication PDFs from Heidelberg, Mannheim, Leipzig, Frankfurt am Main, and Nuremberg. The corpus is balanced across cities but not representative of all local institutions. Each city contributes 25 live-validated official serial PDFs, mainly city bulletins and official gazettes.

The study is designed as an institutional exposure audit, not as person research. No person was targeted; no person profiles were created; no names, raw PDF text, or raw PDFs were stored by default; no facial recognition or image analysis was conducted. School, youth, childcare, education, and family terms were counted only as document-level context indicators and were not mined for identities.

Indicators

The audit helper fetched each public PDF in memory, extracted text and PDF metadata, and wrote only document-level aggregate indicators:

- name-like candidate density per 10,000 characters;
- email and phone-pattern density;
- date-pattern density;
- role-term density;
- minor-context-term density;

- archive age;
- PDF author/creator metadata;
- document type;
- extraction quality.

The resulting aggregability score is a heuristic index for documents, not a legal judgement and not a score for people. It estimates how much friction a document removes for later aggregation. Compared with the earlier exploratory audits, the current corpus reaches 125 documents, balances the city counts, removes live-fetch failures, reports extraction quality, and exports score components for transparency.

Pilot Results

All 125 documents were successfully processed. No document was manually flagged as a dedicated minor-related context, but school, youth, childcare, education, and family terms appeared often enough to contribute to the minor-context term component. All 125 documents contained email patterns and phone-number patterns, and 79 contained PDF author metadata. The mean aggregability score was 8.92, with a median of 9 and a maximum of 13. Extraction quality was high for all 125 documents. The risk-level distribution was:

Risk level	Documents
High	48
Very high	77

City-level descriptive results show why the balanced design is useful:

City	Documents	Mean score	Median score	High	Very high	Author metadata
Frankfurt am Main	25	10.16	11	6	19	25
Heidelberg	25	7.68	8	23	2	0
Leipzig	25	9.28	9	0	25	25
Mannheim	25	8.80	9	8	17	25
Nuremberg	25	8.68	9	11	14	4

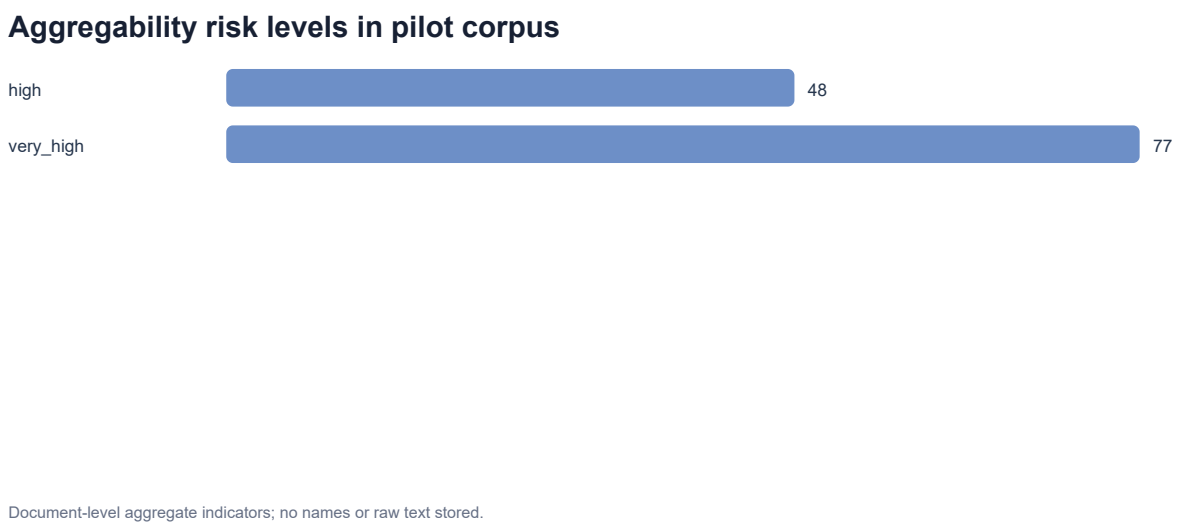

Figure 2. Aggregability risk levels in the local institutional exposure pilot. The corpus is balanced across five city-publication series but is not a population estimate for all local institutions.

Interpretation

The pilot supports the report's core mechanism. Practical obscurity is not only weakened downstream by a malicious actor using AI. It can be weakened upstream by institutional publication practices that make people, roles, places, dates, organizations, and contact channels persistently linkable.

The finding is not that the audited documents are unlawful or improper. Most have legitimate public functions. The finding is that public-interest publication and long-term machine aggregability can diverge over time. A document can be appropriate at publication and still become a component in future profiling when combined with search, OCR, long-context analysis, memory, and repeated updating.

Pilot Limitations

The sample is balanced across five cities but limited to serial municipal publications. This improves comparability and reproducibility while narrowing institutional diversity. The indicators are heuristic and include false positives. Name-like candidates are not verified names. The study did not analyze images, scans, photo captions, image-text coupling, search-engine indexing, robots.txt behavior, platform caching, web-archive persistence, or actual misuse. It also does not measure whether AI agents outperform ordinary search in this specific corpus. It should therefore be read as a proof-of-plausibility audit and a methodological prototype, not as a prevalence estimate.

Pilot Publication Package

For PDF-only publication, the pilot is provided as a standalone manuscript and a separate PDF data appendix. The data appendix contains the public source URL list, corpus summary, city-level results, and document-level aggregate indicators needed to assess the study. Raw PDFs, raw extracted text, names, and person-level records are not included as publication materials.

Open-Weight Models as a Distinct Risk Factor

Open-weight models are not "the problem." They have strong legitimate reasons: research, transparency, competition, digital sovereignty, local processing, privacy through on-device use, and reduced dependence on a few platforms. The risk factor arises from the same properties:

1. **Local executability:** Abuse does not have to pass through central providers that can observe, slow, or block queries.
2. **Adaptability:** Models can be optimized for languages, local sources, specific formats, names, or document types.
3. **Combinability:** Even weaker models can be combined with OCR, search indexes, databases, browser automation, and simple scripts into effective workflows.
4. **Cost and scale effects:** When analysis becomes cheaper, not only better analysis becomes possible, but more analysis: more people, more frequent repetition, longer storage.
5. **Governance gap:** Platform rules address centrally operated models. Local models, fine-tunes, private harnesses, and manual intermediate steps are harder to regulate and audit.

The fair conclusion is this: open-weight models should not be stigmatized wholesale. But a policy approach that looks only at large platform providers misses the private Palantir.

Scenarios

1. Separation and Stalking Context

An ex-partner wants to know who the former partner is in contact with, where she appears, and which new social circles become visible. Previously, this required repeated manual searching. AI can summarize public clues, compare old findings with new ones, and flag apparent changes.

Evidence class: Plausible, supported by established agent capabilities and known stalking/digital-violence dynamics.
Harm: Fear, feeling of being controlled, contact with the target's environment, escalation into physical proximity.
Uncertainty: Little empirical data on AI-assisted stalking as a distinct phenomenon.

2. Local Reputational Harm

In a neighborhood, association, school, or small town, old statements, photos, professional roles, and contacts are collected to portray someone as contradictory, extremist, unserious, or morally questionable.

Evidence class: Plausible.
Harm: Social isolation, rumors, conflict escalation, pressure on family.
Critical mechanism: Context loss. An old photo or one-time contact becomes an alleged character statement.

3. Shadow Employer Profiling

An employer, recruiter, or supervisor uses AI to gain informal assessments of applicants or employees: political activity, family burdens, health clues, side jobs, or proximity to unions.

Evidence class: Plausible to uncertain. The data protection risk is clear; empirically, informal use is hard to prove.
Harm: Discrimination without notice, chilling effect, hidden purpose shift of public information.
Countermeasure: Clear employment-law rules, documentation duties, prohibition of sensitive inference, auditability of recruiting tools.

4. Political Intimidation

Journalists, activists, science communicators, local politicians, or local initiatives are analyzed in a targeted way: public appearances, contacts, old statements, employers, events, private vulnerabilities. HateAid/TUM show that digital violence already changes political engagement today and produces withdrawal effects.

Evidence class: Digital violence established, AI amplification plausible.
Harm: Self-censorship, withdrawal from public life, democratic distortion.
Uncertainty: The exact additional contribution of agentic AI is still barely measured.

5. Personalized Fraud

Fraudsters use public information to appear more credible: real names, places, employers, relationships, interests, events, and language style. Europol, the BKA, and the FBI already document that AI can make fraud and social engineering more scalable and credible.

Evidence class: Established/Plausible.
Harm: Financial loss, identity theft, account takeover, emotional manipulation.
Critical mechanism: Even a few real details increase trust.

6. Network Mapping of Third Parties

A target person is analyzed, but their environment is captured as well: partners, children, colleagues, association contacts, neighbors. People who post little themselves can become visible through others.

Evidence class: Plausible.
Harm: Pressure on social environment, indirect attacks, identification of vulnerabilities.
Limit: Network maps are interpretively risky: visible contact does not prove closeness, attitude, or responsibility.

7. False Suspicion

An agent confuses people with the same name, interprets irony as conviction, treats old data as current, or constructs closeness from a one-time contact. The result sounds factual and is passed on.

Evidence class: Established/Plausible, because LLM errors and mistaken identity are known limits.
Harm: Reputational harm, discrimination, conflict escalation.
Critical point: Errors are not only reassuring limits; they are part of the risk.

International Comparative Lens

The private Palantir is not a uniquely German or European problem. Its legal interpretation differs by jurisdiction, but the underlying mechanism is transnational: public traces circulate globally, search and AI systems cross borders, open-weight models can be run locally, and harms can occur in private social contexts that law often sees only after escalation.

European Union

The EU has the strongest general-purpose data protection framework relevant to this report. The GDPR's principles of legal basis, purpose limitation, data minimization, transparency, accuracy, storage limitation, and data subject rights directly challenge the assumption that public personal data can be freely repurposed. The EU AI Act adds risk-based obligations, but many private intelligence-automation harms remain below the threshold of formal high-risk systems or automated decisions.

United States

The United States combines strong speech protections, sectoral privacy rules, state-level privacy laws, tort concepts, anti-stalking law, and a large data-broker economy. This creates a different balance: public-record access and speech interests are often weighted heavily, while personal-data aggregation may be regulated unevenly. The U.S. context is especially important for data brokers, people-search services, platform practices, and First Amendment-sensitive debates about OSINT and public information.

United Kingdom

The UK combines data protection law derived from the GDPR tradition with online-safety and communications frameworks. The practical challenge is similar to the EU: harmful profiling can occur informally, privately, and without a formal automated decision. The UK is also relevant because many OSINT, journalism, and online-harms debates explicitly consider the tension between public-interest investigation and personal safety.

Other Democratic Contexts

Countries such as Canada, Australia, Brazil, Japan, South Korea, and India have different privacy, platform, and cybercrime regimes, but face comparable pressures: public visibility is increasingly necessary for work and civic life, while AI tools make aggregation cheaper. Comparative research should examine how different combinations of data protection, anti-stalking law, platform regulation, employment law, and data-broker rules affect private profiling risks.

Authoritarian and Hybrid Contexts

In authoritarian or hybrid regimes, the line between private and state surveillance may be less stable. Private actors may be politically aligned with state interests, data brokers may support repression, and public participation may already carry high surveillance risk. The private Palantir mechanism can therefore amplify existing repression, even when the immediate actor is nominally non-state.

Cross-Border Problems

Several governance problems are inherently cross-border:

- Public traces can be collected in one jurisdiction, processed in another, and used in a third.
- Open-weight models and local tools can bypass platform-level restrictions.
- Doxxing, harassment, and fraud campaigns can target people across borders.
- Data subject rights may be difficult to exercise against anonymous, informal, or local actors.
- Legal definitions of public interest, journalism, research, and personal data differ substantially.

For international policy, the core question is therefore not whether one legal system already has the answer. It is how to preserve legitimate public-interest uses of public data while constraining disproportionate, sensitive, repeated, opaque, and harmful personal aggregation.

Comparative Governance Table

Context	Relevant framework	Strength	Gap for private profiling
European Union	GDPR, EU AI Act, platform regulation, data-protection authorities.	Strong general principles for personal-data processing and purpose limitation.	Informal private profiling, hidden use, and non-decision harms remain difficult to detect and enforce.
United States	Sectoral privacy law, state privacy laws, tort law, anti-stalking law, First Amendment protections, data-broker rules in some states.	Strong speech protections and some targeted remedies; active debate on data brokers.	Fragmented privacy regime, broad public-record access, uneven data-broker accountability, limited transparency for affected people.
United Kingdom	UK GDPR, Data Protection Act, Online Safety Act, communications and harassment law.	Combines data protection with online-harms frameworks.	Informal aggregation and private social use may fall below formal enforcement thresholds.
Canada / Australia	Privacy law, cybercrime law, technology-facilitated abuse research, platform and safety regulators.	Stronger policy attention to technology-facilitated abuse and online safety in some areas.	Cross-source AI profiling and public-trace aggregation are not always treated as distinct risks.
Brazil / India	Emerging or evolving data-protection and platform-governance frameworks.	Large democratic contexts where public visibility, platform use, and AI adoption are highly consequential.	Enforcement capacity, informal profiling, cross-border tools, and uneven public awareness may limit protection.
Authoritarian or hybrid contexts	Surveillance law, security institutions, platform controls, weak rule-of-law constraints.	Some formal control over platforms and data flows may exist.	Private and state interests can blur; personal aggregation can support repression or intimidation.

Legal and Ethical Assessment

GDPR and Public Data

Publicly available personal data is not outside the law. The GDPR requires a legal basis, purpose limitation, data minimization, transparency, accuracy, storage limitation, and security. The European

Data Protection Board emphasizes in its Guidelines 1/2024 on legitimate interest that a concrete necessity and balancing assessment is required. The German Data Protection Conference underlines data protection requirements for AI applications.

The private Palantir typically conflicts with several principles: unclear purposes, lack of transparency, excessive collection, sensitive inference, insecure storage, and missing data subject rights.

Sensitive and Derived Data

Political opinions, religion, health, sex life, sexual orientation, and biometric data are especially protected. A problem arises when such data is not collected directly, but inferred from event attendance, group photos, association roles, donations, comments, or contacts.

Facial Recognition

Facial recognition is especially risky because the face is an identifier that is difficult to avoid. Clearview AI illustrates the governance problems: Privacy International reports that several European authorities considered Clearview's practices unlawful, imposed fines, or ordered deletion and processing bans. The example does not prove that private actors use facial recognition at scale, but it shows how public images can be transformed into biometric infrastructure.

EU AI Act

The EU AI Act has been in force since August 1, 2024 and applies in stages. It addresses high-risk systems, prohibited practices, transparency duties, and general-purpose AI. For the private Palantir, gaps remain: many harms arise informally, below the level of formal automated decisions, in private contexts, through local models, or through the social use of outputs.

Digital Violence

The German draft law against digital violence from April 2026 addresses, among other things, cyberstalking, deepfakes, and violations of personality rights. It is relevant, but it often intervenes only once visible legal violations have occurred. Private Intelligence Automation begins earlier: with collection, condensation, observation, and preparation.

Counterarguments and Technical Limits

Counterargument 1: "Search engines can already do this."

Partly. Search engines find hits. Agents can compare hits, summarize them, store them, update them, and embed them in workflows. The difference is not the individual finding, but repeatable condensation.

Assessment: This counterargument is important because not every AI function is new. But it underestimates the automation and memory effect.

Counterargument 2: "Everything is public."

Publicness is not a permission slip for purpose shift. Data protection law, practical obscurity, and contextual integrity show that dispersion, context, and purpose remain socially relevant.

Assessment: A strong everyday argument, but a weak legal and ethical argument.

Counterargument 3: "Agents are too bad."

Agents are indeed error-prone. OSWorld, WebArena, Online-Mind2Web, and AI Agents That Matter show clear limits. But many harms do not require perfect agent performance: fraud needs credible details, stalking needs intimidation, and employer discrimination often needs no documented justification.

Assessment: The limit is real, but not a reason for reassurance.

Counterargument 4: "Platforms will ban this."

Central providers can limit personal profiling, doxxing, stalking, and biometric identification. That matters. But open-weight models, local workflows, manual intermediate steps, and decentralized tools move abuse into less observable spaces.

Assessment: Platform governance is necessary, but not sufficient.

Counterargument 5: "Regulation will solve this."

Regulation is necessary. But private, informal, and hard-to-prove use remains difficult. Affected people often do not learn that they were profiled. Harms can occur without a dossier ever being published.

Assessment: Law must be connected to technical, organizational, and social protective mechanisms.

Counterargument 6: "OSINT is legitimate and important."

Correct. OSINT is important for journalism, human rights work, cybersecurity, war-crimes documentation, and research. This report is not directed against legitimate OSINT, but against opaque private profiling without public interest, proportionality, verification, protection of third parties, and data subject rights.

Assessment: Good regulation must protect legitimate research while limiting abusive profiling.

Strongest Objections and Responses

This section formulates the strongest objections in the form a skeptical reviewer might use.

Objection	Why it is serious	Response / strengthening move
"This is just OSINT with a new name."	Public-source investigation already exists and has legitimate uses.	The report should emphasize the distinct combination of private purpose, repeated automation, memory, sensitive inference, and weak contestability.
"AI adds little beyond search engines."	Search engines already collapsed much practical obscurity.	The relevant delta is not discovery alone, but synthesis, updating, document parsing, multimodal analysis, and workflow integration.
"There is no prevalence evidence."	Without prevalence, the claim can look speculative.	The report frames the thesis as risk emergence, not widespread present use, and proposes specific prevalence studies.

Objection	Why it is serious	Response / strengthening move
"Existing stalking, doxxing, and data-protection law already cover this."	Some harmful uses are already unlawful.	The gap is practical detectability, pre-disclosure dossier formation, informal use, and harms that occur without a formal automated decision.
"Regulating this could chill journalism and OSINT."	Overbroad rules could damage public-interest work.	The report focuses on purpose, proportionality, sensitivity, repetition, storage, verification, and harm rather than public data as such.
"Open-weight models are being unfairly blamed."	Open models have legitimate social, scientific, and sovereignty benefits.	The report treats open weights as a governance-perimeter issue, not as the cause of harm.
"The concept is too broad to operationalize."	Broad concepts can become rhetorically useful but empirically weak.	The operational variables, evidence matrix, and research programme translate the concept into measurable components.
"Agent errors reduce the risk."	Unreliable agents may be bad at sustained profiling.	Errors reduce some capabilities but create others: mistaken identity, false suspicion, context collapse, and plausible but wrong accusations.

Limitations

This report deliberately makes a threshold argument, not a prevalence claim. It argues that the conditions for private AI-assisted personal profiling are becoming more accessible. It does not prove that private AI dossiers are already widespread.

No Direct Prevalence Measurement

The most important limitation is the absence of robust prevalence data. Existing reports on cybercrime, digital violence, stalking, data brokers, and AI misuse are relevant, but they generally do not isolate the specific practice of agentic personal aggregation from public data. Hidden private use is especially difficult to measure because affected people may never learn that a profile exists.

Fast-Moving Technical Baseline

Agent capabilities, model prices, open-weight releases, and provider policies change quickly. A report dated June 12, 2026 can identify mechanisms and trajectories, but individual product details may become outdated. The more durable contribution is therefore conceptual and methodological: the risk chain, evidence classes, and research agenda.

Provider-Documentation Bias

Some capability evidence comes from provider documentation. Such sources are useful because they describe real product categories, but they are not neutral. They may emphasize capability, omit failure modes, or describe safety measures at a high level. This is why benchmark evidence and independent evaluations are treated as necessary counterweights.

European Legal Emphasis

The legal analysis is strongest for the European Union and Germany because the GDPR, EDPB guidance, German data-protection materials, and the EU AI Act provide explicit frameworks for personal-data processing. The report's mechanism is international, but legal remedies and constitutional balances differ across jurisdictions.

Ambiguous Boundary with Legitimate OSINT

The same technical acts -- searching, linking, archiving, verifying -- can be legitimate or harmful depending on purpose, proportionality, safeguards, and use. Overbroad regulation could chill journalism, research, cybersecurity, and human-rights documentation. The report therefore emphasizes harmful personal aggregation rather than public-data work as such.

Measurement Risk

Studying the private Palantir can itself create risks if research designs become operationally instructive. Empirical work should avoid publishing replicable abuse workflows, target-selection methods, or detailed prompts. Ethics review, red-team containment, and harm minimization are essential.

Open Empirical Questions and Research Gaps

1. How often do private actors already use AI for personal research, stalking, doxxing, reputational harm, or informal employer assessment?
2. Which model capabilities are actually decisive for abuse: web search, document analysis, vision, facial recognition, memory, network modeling, tool use, or cost?
3. How large is the added value of AI compared with classic search-engine and social-media research?
4. Which groups are empirically most affected: women, queer people, politically engaged people, journalists, minors, employees, local officials, science communicators?
5. Which protective measures work in practice: rate limits, abuse detection, contextual warnings, exclusion of personal aggregation, expiry dates, noindex, platform reporting, legal access rights?
6. How can practical obscurity be operationalized without disproportionately restricting legitimate archiving, press freedom, and research?
7. How can affected people detect that they have been automatically profiled when the use remains private and invisible?
8. What role is played by local media, association websites, school websites, event PDFs, official gazettes, and old archives?
9. Which audit methods are suitable for agents that process public personal data?
10. How can policy protect open-weight innovation while addressing the consequences of abuse?

Proposed Research Programme

The research gap should be addressed through a combination of qualitative, quantitative, legal, and technical studies. No single method will be sufficient because private profiling is often hidden, informal, and difficult to attribute.

1. Prevalence and Case Documentation

Objective: Estimate how often AI-assisted personal aggregation appears in stalking, doxxing, reputational harm, fraud preparation, or employment screening.

Possible methods: victim-support intake coding, law-enforcement case review, platform abuse-report analysis, surveys of affected groups, interviews with digital-violence counselors, and structured case repositories.

Ethical constraint: Case documentation must protect affected people from renewed exposure and should avoid reproducing dossiers or sensitive personal details.

2. Capability-Decomposition Studies

Objective: Identify which technical components most increase risk.

Possible variables: web search, document parsing, image analysis, entity resolution, long context, memory, tool use, local deployment, cost, and repeated updating.

Scientific value: This would prevent vague claims that "AI" as a whole causes the risk and would help target interventions to the highest-impact components.

3. Non-Operational Agent Evaluations

Objective: Evaluate whether agentic systems enable personal aggregation without publishing abuse-enabling workflows.

Possible methods: synthetic personas, fictional public traces, controlled web environments, redacted benchmark tasks, and scoring for context collapse, mistaken identity, sensitive inference, and collateral exposure.

Safety requirement: Benchmarks should use synthetic data and avoid real private individuals.

4. Institutional Exposure Audits

Objective: Measure how schools, associations, municipalities, universities, employers, and local media contribute to long-term aggregability.

Possible methods: publication-policy review, metadata audits, old-PDF sampling, retention-policy analysis, and consent-form review.

Outcome: Practical standards for names, photographs, minors, event lists, archives, and expiration practices.

5. Comparative Legal and Policy Analysis

Objective: Compare how jurisdictions handle public personal data, data brokers, stalking, doxxing, employment screening, and automated inference.

Possible comparisons: EU, United States, United Kingdom, Canada, Australia, Brazil, India, Japan, and selected authoritarian or hybrid contexts.

Outcome: Identify which legal combinations best protect practical obscurity while preserving legitimate public-interest research.

6. Intervention Testing

Objective: Test which safeguards reduce harm without blocking legitimate uses.

Candidate interventions: rate limits for repeated personal queries, uncertainty labels, source-age warnings, restrictions on sensitive inference, audit logs, visibility controls, noindex tools, deletion workflows, institutional retention limits, and anti-doxxing escalation channels.

Evaluation criteria: harm reduction, false positives, impact on legitimate OSINT, user comprehension, enforceability, and resilience against local/open-weight workarounds.

Recommendations

Priority and Feasibility

The recommendations below should not be treated as equal in urgency or difficulty. A defensible policy agenda should begin with low-regret measures that protect practical obscurity without restricting legitimate public-interest work.

Priority	Measure type	Examples	Rationale
Immediate low-regret	Institutional data minimization	Review old PDFs, event lists, photos, minors' data, metadata, and long-term discoverability.	Reduces exposure without waiting for new AI law.
Immediate low-regret	Transparency and uncertainty	Source-age warnings, uncertainty labels, identity-ambiguity notices, context labels.	Reduces false authority and context collapse.
Immediate low-regret	Victim-support readiness	Digital-violence counseling, evidence preservation, reporting templates for AI-assisted observation.	Helps affected people before prevalence is fully measured.
Medium-term governance	Employment and institutional profiling rules	Ban or document informal AI-assisted screening and sensitive inference.	Targets high-impact decisions where affected people may never learn the cause.
Medium-term governance	Platform and AI-provider safeguards	Detect repeated personal aggregation, dossier building, network mapping, and sensitive inference.	Focuses product governance on patterns rather than isolated prompts.
Research infrastructure	Synthetic benchmarks and prevalence studies	Fictional-persona evaluations, case repositories, surveys, legal comparison.	Converts plausible risk into measurable evidence.

Priority	Measure type	Examples	Rationale
Contested high-impact	Restrictions on repeated personal aggregation	Legal thresholds, audit duties, data-broker limits, stronger remedies.	Potentially powerful, but must be designed to protect journalism and legitimate OSINT.

For Policymakers and Supervisory Authorities

- Treat automated condensation of public personal data as a distinct data protection risk category.
- Strengthen employment data protection against informal AI-assisted applicant and employee research.
- Strengthen protection of minors against later aggregation of public traces.
- Extend doxxing, cyberstalking, and digital violence frameworks to preparation, storage, and repetition logics.
- Improve access and evidentiary rights without introducing blanket real-name requirements or disproportionate data retention.
- Equip data protection authorities with resources for AI-assisted profiling and scraping cases.

For Platforms and Search Systems

- Recognize repeated mass personal queries, dossier building, and network mapping as abuse patterns.
- Better protect context-sensitive old content, local documents, and information about minors.
- Make noindex, visibility, and deletion options easier to understand.
- Detect aggregated harm patterns, not only individual abusive posts.
- Strengthen protections against image-based identification and doxxing chains.

For AI Providers

- Explicitly address stalking, doxxing, private personal profiles, sensitive inference, and political intimidation in safety rules.
- Design agents so that personal research is more strongly limited, logged, and marked with uncertainty.
- Require Deep Research and agent systems to handle sources, uncertainty, mistaken identity, and data protection risks.
- Orient abuse detection not only toward illegal content, but toward repeated personal aggregation.
- Accompany open-weight releases with risk assessment, evaluation data, and misuse impact assessment.

For Public Institutions, Associations, Schools, and Local Media

- Review publications containing personal names, images, and data about minors for later aggregability.
- Regularly review old PDFs, event lists, and photo archives.
- Establish expiry dates, context notes, and data-minimizing publication standards.

- Make consent understandable not only for publication, but also for long-term discoverability.

For Journalism and Research

- Test agents not only for productivity, but also for abuse potential.
- Document cases in which public data has been aggregated for intimidation, reputational harm, or fraud.
- Conduct prevalence studies on AI-assisted personal research and agentic stalking.
- Clearly distinguish legitimate OSINT from private surveillance automation.

For Civil Society and Affected People

- Expand support services for digital violence to include AI-assisted observation.
- Develop protective guidance that does not shift responsibility onto victims.
- Develop collective protection strategies for parties, newsrooms, associations, schools, and activist groups.
- Make evidence preservation, platform reporting, and legal help more accessible.

FAQ

Is the private Palantir already real?

Not in the sense of comprehensive private surveillance. What is established is this: the technical building blocks exist, digital violence and AI-assisted fraud are real problem areas, and costs are falling. What remains uncertain is how often private AI dossiers are already used systematically.

Is this about secret data?

No. The core issue is public and semi-public traces: websites, old PDFs, local media, social media, images, event programs, legal notices, professional profiles, and publications caused by third parties.

Is everyone not responsible for themselves if they post publicly?

No. Many traces are created by others: employers, associations, schools, media, friends, public authorities, event organizers. Public participation is also necessary for work, politics, science, culture, and community. A democratic society cannot demand that people disappear in order to be protected.

Why are AI errors not reassuring?

Because many harms arise from plausible claims, not from court-proof truth. A false profile can damage applications, relationships, local reputation, or safety.

Is open source to blame?

No. Open-weight models have legitimate benefits. But they are a distinct governance factor because local and adapted use is harder to control than central platform use.

What is the difference from OSINT?

Legitimate OSINT pursues a public interest, works proportionately, verifies sources, protects third parties, and reflects on harms. Private Intelligence Automation pursues private control, intimidation, profit, or curiosity without transparency and data subject rights.

What should happen first?

Empirical research, clear rules against personal aggregation, protection of minors and old content, better platform and AI-provider boundaries, employment-law clarification, and specialized help for people affected by digital violence.

Evidence Matrix

Claim	Capability evidence	Harm evidence	Directness	Prevalence evidence	Scientific assessment
Agents can conduct multi-step research and synthesis.	Strong: provider documentation and product releases.	Indirect: relevant to dossier formation.	Adjacent to private profiling, direct for research capability.	Product availability established; abusive prevalence unknown.	Current capability, not future speculation, but misuse frequency not measured.
Agents can use tools and computer interfaces.	Strong: Anthropic Computer Use, Tool Use, OpenAI Operator/Agent, Microsoft agent strategy.	Indirect: supports repeated workflows.	Adjacent.	Unknown for private profiling.	Real capability, but error-prone and uneven across tasks.
Web and computer agents remain limited.	Strong: OSWorld, WebArena, Online-Mind2Web, AI Agents That Matter.	Directly relevant as a constraint.	Direct for reliability limits.	Not a prevalence claim.	Dampens claims of comprehensive surveillance; does not eliminate partial-automation risk.
Multimodal models make images, screenshots, and documents easier to analyze.	Strong: provider features, multimodal model releases, document-analysis tools.	Plausible: expands analyzable public traces.	Adjacent to private profiling.	Unknown.	Increases the amount of usable public material; identity and context errors remain critical.
Open-weight models are becoming more capable and more widespread.	Strong: Stanford AI Index, OECD, Meta, DeepSeek, Qwen.	Plausible: weakens centralized control.	Adjacent.	Unknown for abusive workflows.	Central governance factor, but not inherently harmful.

Claim	Capability evidence	Harm evidence	Directness	Prevalence evidence	Scientific assessment
AI can amplify social engineering and fraud.	Moderate to strong: law-enforcement and cybercrime reports.	Strong for fraud; AI-specific contribution varies by report.	Direct for fraud, adjacent for public-data aggregation.	Indicated, but not fully quantified for agentic personal dossiers.	High relevance because a few accurate personal details can increase credibility.
Doxxing is a documented privacy and safety harm.	Not primarily technical.	Strong: quantitative doxxing research and literature reviews.	Direct for disclosure harm, adjacent for pre-disclosure dossier formation.	Measured in specific platforms and datasets; broader prevalence varies.	Supports the claim that aggregation and exposure of personal information can lead to harassment and safety risks.
Technology-facilitated abuse and intimate partner surveillance are established harm domains.	Moderate: spyware, account compromise, social engineering, monitoring tools.	Strong for coercive control and partner surveillance.	Direct for stalking context, adjacent for AI-assisted public-trace aggregation.	Measured in specific studies; AI-specific contribution unclear.	Supports the separation/stalking scenario and shows why partial automation can matter.
Social-media screening can affect hiring and expose protected traits.	Moderate: online search and screening practices.	Strong for discrimination risk and informal assessment.	Direct for employment context, adjacent for AI-assisted synthesis.	Some empirical evidence; hidden informal use remains hard to measure.	Supports the shadow employer-profiling scenario.
Surveillance and dataveillance can produce chilling effects.	Not primarily technical.	Moderate to strong: empirical and theoretical literature.	Direct for behavioural withdrawal, adjacent for AI-specific profiling.	Measured in selected contexts; AI-specific effect unknown.	Supports democratic-harm claims even when no dossier is publicly released.
Digital violence can chill political participation.	Not primarily technical.	Strong: civil-society and academic evidence.	Direct for harm domain, adjacent for AI amplification.	Measured for digital violence; AI-specific prevalence unclear.	Democratic harm is already observable; AI contribution requires further study.
Public personal data is not freely usable under data protection law.	Not technical.	Legal harm logic established.	Direct in EU/GDPR context.	Not a prevalence claim.	Legal core of the argument; jurisdiction-specific outside Europe.
Private agentic personal profiles are already widely used.	Technical capability exists.	Harm logic plausible.	Indirect.	Weak / unknown.	Central research gap; should not be asserted as established.
Fully automated comprehensive private surveillance is present reality.	Insufficient.	Potential harm high.	Indirect.	No robust evidence.	Should not be claimed; stronger than the evidence permits.

Interpretation of the Matrix

The matrix shows a layered evidentiary structure. The strongest evidence concerns technical building blocks and legal principles. The moderate evidence concerns adjacent harm domains such as fraud, digital violence, doxxing, stalking, and reputational harm. The weakest evidence concerns prevalence of the exact phenomenon: private, agentic, repeated, non-state personal dossier formation from public data. This does not invalidate the risk claim, but it determines its correct strength.

Conclusion

The private Palantir is not proof that ubiquitous private surveillance already exists. It is a warning about a plausible and rapidly approaching risk category: the protective effect of dispersion, effort, context, and forgetting is weakened by AI agents, multimodal analysis, memory, and falling costs.

The core policy question is therefore not only: "Which data is public?" It is: **Who may automatically connect, evaluate, store, and use public traces against people?**

A democratic response does not have to romanticize practical obscurity, but it must defend it: through data protection, technical friction, clear limits on personal aggregation, protection for people affected by digital violence, fair rules for OSINT, and research that neither dramatizes nor downplays the problem.

Methods Appendix: Search Strategy and Review Protocol

This appendix documents the search logic used for this report and provides a reproducible protocol for further development. The report is a structured narrative synthesis rather than a systematic review. The protocol below is intended to make source selection more transparent and to support future replication or expansion.

Review Aim

The review aimed to identify sources relevant to five linked questions: current agent capabilities, limits of those capabilities, privacy theory around public information, adjacent harm domains, and governance frameworks for personal-data aggregation.

Time Frame

The report reflects sources available and reviewed up to June 12, 2026. Older foundational sources were included where they remain theoretically or legally central, such as practical obscurity, contextual integrity, and early doxxing measurement work.

Source Types

The search prioritized:

- peer-reviewed articles and conference papers;
- recognized preprints from arXiv or equivalent repositories;
- official documentation from AI providers where it establishes product capabilities;

- legal sources, regulatory guidance, and standards documents;
- law-enforcement, civil-society, and public-interest reports on cybercrime, digital violence, data brokers, and privacy harms;
- high-quality policy analysis where peer-reviewed literature is not yet available.

Search Locations

The search strategy used a combination of:

- academic search engines and publisher pages;
- arXiv and conference-paper repositories;
- official provider documentation;
- EU, EDPB, national data-protection, and standards-body websites;
- law-enforcement and civil-society report repositories;
- targeted web search for known titles, authors, and institutions.

Indicative Search Terms

Search terms were combined across conceptual clusters. The following are examples of non-operational literature-search terms, not instructions for profiling individuals:

Cluster	Indicative terms
AI agents	"AI agents web benchmark", "computer use agent evaluation", "web agent benchmark", "tool use large language models", "multimodal agent benchmark"
Privacy theory	"practical obscurity", "contextual integrity", "online obscurity", "public records aggregation privacy", "privacy aggregation inference"
Doxxing and harassment	"doxing detection characterization", "doxxing literature review", "online harassment personal information disclosure", "networked harassment privacy harm"
Technology-facilitated abuse	"intimate partner surveillance", "technology facilitated abuse", "cyberstalking digital abuse", "coercive control technology"
Employment and screening	"social media screening hiring discrimination", "algorithmic hiring bias", "employment screening online profiles", "workplace surveillance privacy"
Data brokers	"data brokers privacy harms", "data broker surveillance capitalism", "people search services privacy", "data broker regulation"
Chilling effects	"online surveillance chilling effects", "digital dataveillance chilling effects", "surveillance self-censorship online", "political participation digital surveillance"
Governance	"GDPR public personal data legitimate interest", "AI Act personal data profiling", "OSINT ethics privacy", "data protection public data aggregation"

Inclusion Criteria

Sources were included when they met at least one of the following criteria:

1. They directly established or evaluated an AI capability relevant to public-data aggregation.
2. They documented a harm domain into which AI-assisted aggregation could plausibly plug.

3. They provided legal, regulatory, or theoretical grounding for the claim that publicness does not equal unrestricted use.
4. They helped distinguish established evidence from plausible risk and unknown prevalence.
5. They supported a concrete governance or research recommendation.

Exclusion Criteria

Sources were excluded or avoided as central evidence when they were:

- purely anecdotal without verifiable context;
- promotional without technical detail;
- operationally instructive for abuse;
- redundant with stronger primary or academic sources;
- focused on unrelated AI risks without a link to personal-data aggregation;
- too jurisdiction-specific to support a general claim without careful qualification.

Evidence Coding

For each major claim, sources were coded informally along four dimensions:

Dimension	Coding question
Capability	Does the source establish that the relevant technical function exists?
Directness	Does it address the exact scenario or an adjacent mechanism?
Reliability	Is the source robust, independent, and methodologically clear?
Prevalence	Does it measure real-world frequency, merely indicate cases, or leave frequency unknown?

Pilot Audit Protocol

The local institutional exposure pilot used a bounded corpus of public municipal-publication PDFs documented in the accompanying PDF data appendix. The audit protocol:

1. uses a human-curated list of public institutional PDF URLs;
2. fetches each document in memory;
3. extracts text and PDF metadata;
4. computes document-level aggregate indicators;
5. writes only processed metrics, summary statistics, and figures;
6. avoids storing raw PDFs, raw text, names, or person-level records.

The pilot indicators are deliberately conservative and non-identifying. Name-like candidate counts are treated as noisy exposure proxies, not as verified personal names.

The current corpus contains 125 public municipal-publication PDFs from five German cities: Heidelberg, Mannheim, Leipzig, Frankfurt am Main, and Nuremberg. Each city contributes 25 official serial PDFs. The corpus was assembled from human-defined official source patterns and validated as live PDF

responses before inclusion. The final audit processed 125 of 125 documents successfully and records extraction quality for each document.

The scoring model exports component-level contributions for name-like density, contact density, date density, role density, minor context, archive age, PDF metadata, document type, and low extraction quality. These components are intended to make the heuristic inspectable, not to imply statistical calibration or legal classification.

Reproducibility Limitations

This is not yet a full systematic review. The search process did not produce a PRISMA flow diagram, a complete screened corpus, inter-coder reliability, or a public bibliography database. The pilot corpus is balanced across five city-publication series but not representative of all local institutions. Its size and city balance are stronger for descriptive comparison, while its focus on serial municipal publications narrows the document-type range. The PDF audit does not analyze images, scans, photo captions, web-indexing behavior, platform caching, web archives, or actual malicious use. A future journal submission should add a structured bibliography file, explicit database queries, date-stamped search logs, a source-screening table, external scoring validation, and a larger pre-registered institutional audit sample.

References

Academic and Theoretical Literature

Acquisti, A. & Fong, C. M. (2020). An experiment in hiring discrimination via online social networks. *Management Science*, 66(3), 1005-1024. https://doi.org/10.1287/mnsc.2018.3269

Brown, C. & Hegarty, K. (2024). Fear and distress: How can we measure the impact of technology-facilitated abuse in relationships? *Social Sciences*, 13(1), 71. https://doi.org/10.3390/socsci13010071

Büchi, M., Festic, N. & Latzer, M. (2022). The chilling effects of digital dataveillance: A theoretical model and an empirical research agenda. *Big Data & Society*, 9(1). https://doi.org/10.1177/20539517211065368

Crain, M. (2021). The untamed and discreet role of data brokers in surveillance capitalism: A transnational and interdisciplinary overview. *Internet Policy Review*, 10(2). https://policyreview.info/articles/analysis/untamed-and-discreet-role-data-brokers-surveillance-capitalism-transnational-and

Hartzog, W. & Stutzman, F. (2010). The case for online obscurity. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1597745

Nissenbaum, H. (2004). Privacy as contextual integrity. *Washington Law Review*, 79(1), 119-158. https://digitalcommons.law.uw.edu/wlr/vol79/iss1/10/

Penney, J. W. (2016). Chilling effects: Online surveillance and Wikipedia use. *Berkeley Technology Law Journal*, 31(1), 117-182. https://btlj.org/data/articles2016/vol31/31_1/0117_0182_Penney_ChillingEffects_WEB.pdf

Snyder, P., Doerfler, P., Kanich, C. & McCoy, D. (2017). Fifteen minutes of unwanted fame: Detecting and characterizing doxing. In *Proceedings of the 2017 Internet Measurement Conference* (pp. 432-444). ACM. https://doi.org/10.1145/3131365.3131385

Tseng, E. et al. (2020). The tools and tactics used in intimate partner surveillance: An analysis of online infidelity forums. In *29th USENIX Security Symposium*. USENIX Association. https://www.usenix.org/conference/usenixsecurity20/presentation/tseng

AI Agents, Model Capabilities and Benchmarks

Anthropic. (2026). *Computer use tool*. https://docs.anthropic.com/en/docs/build-with-claude/computer-use

Anthropic. (2026). *Tool use*. https://docs.anthropic.com/en/docs/build-with-claude/tool-use/overview

Kapoor, S. et al. (2024). *AI agents that matter*. arXiv. https://arxiv.org/abs/2407.01502

Mistral AI. (2025). *Le Chat dives deep*. https://mistral.ai/news/le-chat-dives-deep/

Microsoft. (2025). *Microsoft Build 2025: The age of AI agents and building the open agentic web*. https://blogs.microsoft.com/blog/2025/05/19/microsoft-build-2025-the-age-of-ai-agents-and-building-the-open-agentic-web/

OpenAI. (2025). *Computer-using agent*. https://openai.com/index/computer-using-agent/

OpenAI. (2025). *Introducing deep research*. https://openai.com/index/introducing-deep-research/

OpenAI. (2026). *Introducing ChatGPT agent*. https://openai.com/index/introducing-chatgpt-agent/

OpenAI. (2026). *Introducing GPT-5*. https://openai.com/index/introducing-gpt-5/

Xie, T. et al. (2024). *OSWorld: Benchmarking multimodal agents for open-ended tasks in real computer environments*. arXiv. https://arxiv.org/abs/2404.07972

Zhou, S. et al. (2023). *WebArena: A realistic web environment for building autonomous agents*. arXiv. https://arxiv.org/abs/2307.13854

Online-Mind2Web authors. (2025). *An illusion of progress? Assessing the current state of web agents*. arXiv. https://arxiv.org/abs/2504.01382

Open Weight, Costs and Model Spread

DeepSeek. (2025). *DeepSeek-R1 release*. https://api-docs.deepseek.com/news/news250120

Hugging Face. (2025). *DeepSeek-R1 model card*. https://huggingface.co/deepseek-ai/DeepSeek-R1

Meta AI. (2025). *The Llama 4 herd*. https://ai.meta.com/blog/llama-4-multimodal-intelligence/

OECD. (2024). *AI openness: A primer for policymakers*. OECD Publishing. https://www.oecd.org/en/publications/ai-openness_02f73362-en.html

Qwen. (2025). *Qwen3-Coder*. https://qwenlm.github.io/blog/qwen3-coder/

QwenLM. (2025). *Qwen3-Coder GitHub repository*. https://github.com/QwenLM/Qwen3-Coder

Stanford Institute for Human-Centered Artificial Intelligence. (2025). *AI Index Report 2025*. https://hai.stanford.edu/ai-index/2025-ai-index-report

Stanford Institute for Human-Centered Artificial Intelligence. (2026). *AI Index Report 2026*. https://hai.stanford.edu/ai-index/2026-ai-index-report

Data Protection, Digital Violence and Cybercrime

Bundeskriminalamt. (2026). *Cybercrime-Lagebilder*. https://www.bka.de/DE/AktuelleInformationen/StatistikenLagebilder/Lagebilder/Cybercrime/cybercrime_node.html

Europol. (2025). *Internet Organised Crime Threat Assessment 2025*. https://www.europol.europa.eu/cms/sites/default/files/documents/Steal-deal-repeat-IOCTA_2025.pdf

Europol. (2025). *Serious and Organised Crime Threat Assessment 2025*. https://www.europol.europa.eu/cms/sites/default/files/documents/EU-SOCTA-2025.pdf

Federal Bureau of Investigation Internet Crime Complaint Center. (2026). *2025 IC3 Annual Report*. https://www.ic3.gov/AnnualReport/Reports/2025_IC3Report.pdf

HateAid & Technical University of Munich. (2025). *Angegriffen und alleingelassen: Wie sich digitale Gewalt auf politisches Engagement auswirkt*. https://hateaid.org/wp-content/uploads/2025/01/hateaid-tum-studie-angegriffen-und-alleingelassen-2025.pdf

New America. (2025). *Preserving privacy: An impact framework for OSINT and AI*. https://www.newamerica.org/insights/preserving-privacy-an-impact-framework/

Privacy International. (2026). *Challenge against Clearview AI in Europe*. https://privacyinternational.org/legal-action/challenge-against-clearview-ai-europe

Law, Regulation and Standards

Cornell Legal Information Institute. (n.d.). *U.S. Department of Justice v. Reporters Committee for Freedom of the Press, 489 U.S. 749*. https://www.law.cornell.edu/supremecourt/text/489/749

European Commission. (2026). *Regulatory framework on artificial intelligence*. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

European Data Protection Board. (2024). *Guidelines 1/2024 on processing of personal data based on Article 6(1)(f) GDPR*. https://www.edpb.europa.eu/system/files/2024-10/edpb_guidelines_202401_legitimateinterest_en.pdf

European Union. (2016). *Regulation (EU) 2016/679: General Data Protection Regulation*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679

German Data Protection Conference. (2024). *Orientierungshilfe KI und Datenschutz*. https://www.datenschutzkonferenz-online.de/media/oh/20240506_DSK_Orientierungshilfe_KI_und_Datenschutz.pdf

German Federal Ministry of Justice. (2026). *Gesetz gegen digitale Gewalt*. https://www.bmjv.de/SharedDocs/Gesetzgebungsverfahren/DE/2026_Gesetz_gegen_digitale_Gewalt.ht

ml

National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework*. https://www.nist.gov/itl/ai-risk-management-framework

Imprint and Author Note

Author and editorial responsibility: Benjamin Metzig, Wissenschaftswelle.de. Wissenschaftswelle.de is an independent editorial knowledge project focused on science, society, technology, culture, and public understanding of evidence.

Project website: https://www.wissenschaftswelle.de/

Suggested citation: Metzig, B. (2026). *The Private Palantir: A Research Report on AI Agents, Public Data, and the End of Practical Obscurity*. Wissenschaftswelle.de.

Persistent identifier: No DOI has been assigned yet. The report is structured for later repository deposit, DOI registration, and versioned citation.

Editorial note: Digital research and writing tools were used during preparation. They were not used as an authority for factual claims. Source verification, interpretation, final judgement, and publication responsibility remain with Benjamin Metzig.

Data availability: For PDF-only publication, the local institutional exposure pilot is accompanied by a PDF data appendix containing the public source URL list, corpus summary, city-level results, and document-level aggregate indicators. Raw PDFs, raw extracted text, names, and person-level records are not included as publication materials.

Code availability: The publication package does not require local code access. The report describes the corpus-construction logic, indicator definitions, scoring components, and ethical safeguards needed to evaluate the study. Computational helpers were used for URL validation and aggregate counting, under the constraint that no raw text, names, or person-level records are retained as publication outputs.

Competing interests: Benjamin Metzig is the founder, author, and editorially responsible publisher of Wissenschaftswelle.de. No financial competing interests are declared.