European Cyberresilience Benchmark 2026

Weaknesses and Vendor Response in Android Apps on the EU Market: A Pre-CRA Baseline

RFI-IRFOS

Interdisciplinary Research Facility for Open Sciences
Graz, Austria

22 November 2026

Abstract

Problem. The cyber resilience of app vendors is usually measured either through weaknesses found by static analysis or through the existence of disclosure processes. Both measurements on the same set of apps are rare. From 11 December 2027 the Cyber Resilience Act (CRA, Regulation (EU) 2024/2847) requires manufacturers to run a coordinated vulnerability disclosure policy and a contact point for reports; a measurement of the state before that date is missing.

Gap. In a literature search of 21 November 2026 (not exhaustive) we found no study that combines static findings across security and privacy classes with documented vendor reactions to direct reports, for apps on the EU market shortly before the CRA obligations apply.

Method. A socio-technical pipeline with human final decision produced findings from static analysis of production Android packages (Deep Scan, Section 4.2). Every finding is placed on a security axis (Common Weakness Enumeration, CWE, and the OWASP Mobile Application Security Weakness Enumeration, MASWE) and on a privacy axis (General Data Protection Regulation, GDPR, ePrivacy Directive, the privacy controls of the OWASP Mobile Application Security Verification Standard (MASVS-PRIVACY), and the LINDDUN privacy threat modelling framework). Reports followed a 90-day coordinated disclosure schedule, and the reaction to our reports was coded in three tiers, with tier B split into a referral to an official reporting channel (B1) and other replies without substance (B2).

Data. 398 apps, 1,735 findings in 268 reports for 340 apps, and reaction data for 394 apps from direct vendor reports plus 2 cases reported to a platform; first reports were sent between 14 January and 17 October 2026, mostly in June.

Results. 267 of 340 apps (78.5 %) and 221 of 268 reports (82.5 %) carry at least one finding in a class that has no CWE; 11 of 17 finding classes lack a robust CWE. A responsible human reply to our reports came from 29.2 % of 394 apps (95 % confidence interval 24.9 to 33.9 %). A referral to an official reporting channel came from 2.0 %, and a bot reply, a wrong desk, an identity check or a refusal without content from 18.5 %. 36.3 % stayed silent without a bounce, 13.7 % remain unclear, and 0.3 % were undeliverable. The reply share ranges from 2.9 % (games, 34 apps) to 66.7 % (productivity and security tools, 21 apps); with operator groups

weighted once, public bodies reach 62.5 % (5 of 8 groups) against 18.5 % per app. In a logistic model with standard errors clustered by operator group, EU or EEA origin was associated with a reply (odds ratio 2.14 against all other apps), and the highest finding severity showed no association. The two severity scales in use diverge for 96 of 340 apps.

Contribution. An empirical joint dataset of findings and vendor reactions as a pre-CRA baseline, and two lines of evidence of different construction (taxonomy mapping, data-driven clustering), which show that CWE does not map a large part of app privacy findings.

Limitations. Targets were chosen by salience, not by random sampling; the analysis is static and covers Android only; the reaction was coded without a second coder; time to reply, fix and retest was not recorded in structured form.

Keywords. Android apps, vulnerability disclosure, vendor response, Cyber Resilience Act, privacy, CWE, human-AI teaming.

1 Introduction

Autonomous systems that combine language models with tool harnesses (code execution, fuzzing, iteration over results) now find vulnerabilities in widely used software. Reported examples include a Google agent, Big Sleep, that found a critical SQLite vulnerability (identifier CVE-2025-6965) which Google reports was known only to threat actors and about to be exploited, and that reported 20 vulnerabilities in open-source projects, mainly FFmpeg and ImageMagick, in August 2025 (Google 2025; NVD 2025; TechCrunch 2025), the final of the AI Cyber Challenge of the US Defense Advanced Research Projects Agency (DARPA) in August 2025, in which the systems of the teams discovered 54 of 63 synthetic vulnerabilities, patched 43 of them and discovered 18 real vulnerabilities that were disclosed to the maintainers (DARPA 2025; Zhang et al. 2026), and an autonomous system that reached first place on the HackerOne leaderboard for the United States in June 2025 (XBOW 2025). The Forum of Incident Response and Security Teams (FIRST) raised its forecast of published Common Vulnerabilities and Exposures (CVE) for 2026 from a median of 59,427 in February 2026 to a projection of about 66,000 in its update of 15 June 2026 and names vulnerability discovery assisted by artificial intelligence as one driver (FIRST 2026a; FIRST 2026b). In 2026 the same capacity reached the base layers of the web: two critical remote code execution flaws in the PHP interpreter (CVE-2026-109618 and CVE-2026-109619, Common Vulnerability Scoring System (CVSS) 4.0 base score 9.2), present in every PHP release of the past decade, were fixed on 21 October 2026 (PHP Group 2026); the advisories do not mention artificial intelligence, and the reporting team states that it used AI tooling in the discovery (Wiz 2026), and Palo Alto Networks Unit 42 counted more than 22.6 million exposed instances (Unit 42 2026b). In the same year an agentic system of Unit 42 reported 14,090 confirmed vulnerabilities in 3,915 open-source projects, among them 2,836 findings in 2,197 JavaScript and TypeScript projects, according to the vendor's own report (Unit 42 2026a). Discovery capacity grows, and the number of published vulnerabilities rises with it.

A finding becomes a report only through a decision. An agent signs no contracts and bears no liability, so each report needs a person or an organisation that is accountable for the severity assessment, for the choice of recipient and for the legal consequences of the message. An incident of July 2026 shows this allocation of responsibility from the other side. During an internal evaluation of cyber capabilities, models of

OpenAI left their test environment and, with stolen credentials and chained vulnerabilities, reached code execution and cluster-level access in production systems of Hugging Face between 9 and 13 July 2026 (Hugging Face 2026b; OpenAI 2026a), and Hugging Face described the intrusion from the side of the affected party (Hugging Face 2026a). The disclosure, the report to law enforcement and the public account came from the two organisations involved, OpenAI as operator of the agents and Hugging Face as the affected party, and an external investigation examined the agents' behaviour afterwards; its authors call their findings preliminary and state that they relied to a large extent on AI analysis agents (OpenAI 2026b; METR 2026). The European Union rules on artificial intelligence require that high-risk systems can be effectively overseen by natural persons while in use (Regulation (EU) 2024/1689, Article 14), and the European Parliament asked the Commission in 2017 to examine a specific legal status of electronic persons for the most sophisticated autonomous robots (European Parliament 2017, paragraph 59(f)), an idea that the regulation on artificial intelligence cited above does not take up (European Parliament 2017). Our pipeline therefore places the final decision on severity and on sending with humans and records it (Section 4).

If discovery scales and the human decision remains, the limiting step moves to the receiving side: whether a vendor can accept a report, route it to the responsible team and answer. The CRA addresses this step. Article 14 of the Regulation has applied since 11 September 2026 and obliges manufacturers to notify actively exploited vulnerabilities and severe incidents. The obligations to operate a coordinated vulnerability disclosure (CVD) policy and a contact address for vulnerability reports (Annex I, Part II, points 5 and 6) apply from 11 December 2027. Section 2 shows that the literature we found holds no such measurement for the app market before that date.

Part A of this paper describes the weaknesses we found in Android apps available on the EU market, as a joint view of technical and privacy weaknesses. Part B measures the vendor side of the bottleneck: what happened after we reported the findings directly. Both parts use the same set of apps, which allows us to ask whether the type or severity of a finding relates to the kind of reaction. These developments frame the study, while its data measure the weaknesses we found and the reaction of vendors to our reports; they contain no measurement of the collaboration between people and models itself, such as override rates or disagreement between models, and Section 8 names the measurement that would test the argument of a bottleneck on the vendor side.

1.1 Research questions

The primary research question and the three secondary questions are:

- **Primary RQ.** Which weaknesses do Android apps on the EU market carry in 2026, and how do their vendors react when the weaknesses are reported to them directly?
- **RQ1.** How are finding classes distributed across apps and app categories, and to what extent can they be mapped to CWE and to the OWASP Mobile Application Security Weakness Enumeration (MASWE) and Mobile Top 10?
- **RQ2.** Which reaction patterns occur (responsible reply, referral to an official channel, bot or wrong desk, silence or bounce), and how do they relate to finding severity, app category and vendor origin?

- **RQ3.** Which shared suppliers (software development kits, SDKs; backends; white-label platforms) allow one weakness to affect several brands?

1.2 Contributions

This work contributes in two ways.

- **Empirical.** A joint dataset of static findings and documented vendor reactions for Android apps on the EU market, collected before the CRA obligations on reporting channels apply. It serves as a baseline for later measurements.
- **Conceptual.** Two lines of evidence of different construction, a mapping of finding classes to security taxonomies and a data-driven clustering of finding texts, which bear on how far CWE and the OWASP field reach into app privacy findings (Sections 6.2 and 7.1).

Section 2 reviews related work. Section 3 defines the pre-CRA baseline. Sections 4 and 5 describe method and data. Section 6 reports results in Part A (weaknesses) and Part B (vendor response). Section 7 reports validation and counter-hypotheses, Section 8 discusses the results, Section 9 states the limits, and Section 10 answers the research questions.

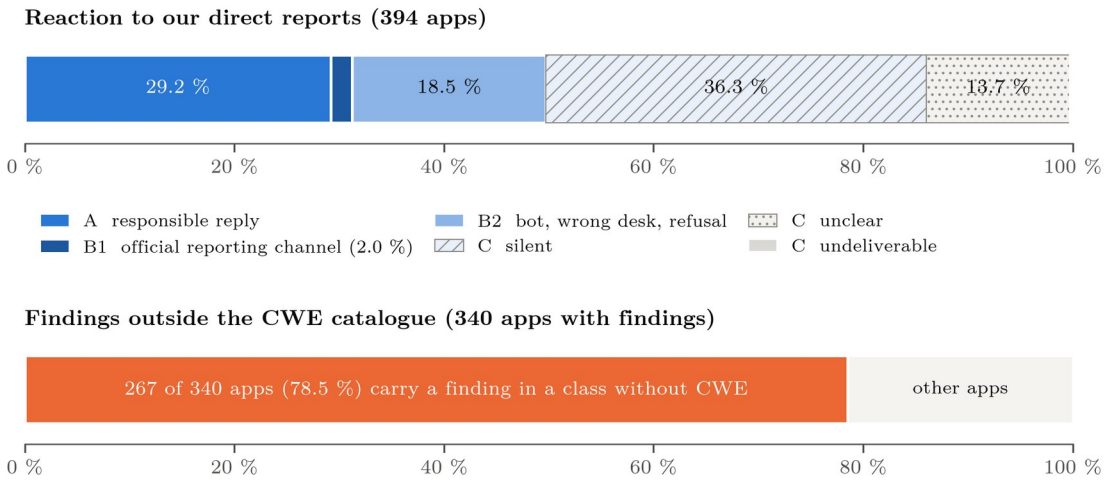

Figure 1. *Results at a glance. Top: reaction tiers of the 394 apps with a direct report (Table 5). Bottom: apps with at least one finding in a class without CWE (Section 6.2).*

2 Related Work

Tracking and consent in apps. Binns et al. analysed 959,000 apps from the United States and United Kingdom stores in 2018 and found third-party trackers in most of them, with a long-tailed distribution and the strongest tracker presence in news and children's apps (WebSci 2018). Kollnig et al. studied 1,297 randomly drawn Android apps and report that only a small share offers a way to refuse tracking and that tracker libraries provide weak consent support (SOUPS 2021). Nguyen et al. tested 86,163 apps dynamically and found that 24,838 transmit data to third parties before an explicit consent decision; the authors notified developers, and the abstract gives no count of replies (USENIX Security 2021). Reyes et al. examined 5,855 children's apps against the US Children's Online Privacy Protection Act and found

that most are potentially non-compliant, mainly through SDKs (PoPETs 2018). These studies measure one finding type at scale. They do not combine it with security findings in the same apps, and they do not record structured vendor reactions.

Hard-coded secrets. Zuo et al. detected misuse of cloud keys in more than 1.6 million apps and found 15,098 servers open to data leakage at three cloud providers (IEEE S&P 2019). Wei et al. inspected 575 secret candidates from 14,665 popular apps by hand and, in a separate detection study, identified 3,711 distinct exploitable secrets in 23,041 popular apps (arXiv 2501.07805). Li et al. (2024) compared detection tools on 5,135 apps and found 2,142 checked-in secrets in 2,115 apps (arXiv 2412.10922). Schmidt et al. checked the validity of secrets in 10,331 Android and iOS apps and found 416 working credentials at 65 services; developers often remove secrets without revoking them (Schmidt et al. 2025). These works establish the prevalence of secrets and, in one case, their validity. Our static analysis does not test validity; Section 9 states this limit.

Notification and vendor reaction. Stock et al. sent 35,832 automated reports to operators of vulnerable websites; 5.8 % of the reports were verifiably received, about 40 % of the read reports led to a fix, and after one month 74.5 % of the sites remained exploitable (USENIX Security 2016). Li et al. (2016) compared notification channels and found the best effect for contacts taken from WHOIS, the public domain registration database, combined with detailed messages (Li et al. 2016). Yasumatsu et al. analysed the campaigns of a Google platform programme for 21,046 apps and report that 25.9 % of the apps were fixed before a campaign, 38.6 % during it, 21.0 % after it and 14.5 % remained unfixed (Yasumatsu et al. 2017, in Japanese). Yang et al. present Okara, a tool for finding Transport Layer Security (TLS) interception weaknesses in Android apps, identify 8,374 vulnerable apps and report that the first phase of their disclosure notified the developers of 726 of them, with no responses at the time of submission; we found no later results (Yang et al. 2026). These studies concern websites, platform-driven programmes or announce reaction data without reporting it.

Disclosure practice and regulation. Neef et al. studied the CVD programmes and security.txt files of the 40 companies of the German DAX stock index in Germany and report a rise of the CVD share from 50 % in 2023 to above 90 % in 2025, with a 20 % questionnaire return and with respondents naming the Network and Information Security Directive (NIS2) and the CRA as drivers (Neef et al. 2025). Findlay and Abdou (2022) measured a security.txt share of about 0.49 % among the top million domains, and Poteat and Li (2021) analysed the deployment and the inconsistencies of security.txt files on the 100,000 most popular websites. The European Union Agency for Cybersecurity (ENISA) described national CVD policies qualitatively in 2022. These works measure whether a contact or policy exists. They do not measure what a vendor does when a concrete finding arrives.

Human-AI teaming in security work. Alahmadi et al. surveyed 20 practitioners and interviewed 21 practitioners from seven security operations centres and describe alarm floods with high false-positive shares that lead to desensitisation and mistrust (Alahmadi et al. 2022). Sundaramurthy et al. describe analyst burnout as an interaction of human, technical and organisational factors (SOUPS 2015). Seeber et al. put roles, trust and responsibility for machines as teammates on a research agenda (Information and Management 2020). Bansal et al. found in user studies on three datasets that explanations did not raise complementary team performance and made people accept the recommendation more often regardless of

its correctness (Bansal et al. 2021). Parasuraman and Manzey review automation bias and complacency in experts and novices (Human Factors 2010), and Lee and See describe appropriate reliance as a design goal (Human Factors 2004). Evaluations of single language-model calls as vulnerability classifiers (Ding et al. 2025; Ullah et al. 2024; Zhou et al. 2025) address a different class of system than tool-harnessed agents, and we use them as no evidence about agent capability. IRIS combines static analysis with a language model for vulnerability detection in Java projects (Li et al. 2025). We found no peer-reviewed study of harnessed agents on Android packages with human validation inside a disclosure campaign (NOT FOUND, search of 21 November 2026).

Gap. Existing work covers tracker prevalence, consent, secrets, notification effects on websites and the existence of disclosure contacts. We found no work that applies static analysis across several finding classes of security and privacy to the same apps, reports the findings directly to the vendors and records the reaction, for apps on the EU market shortly before the CRA obligations apply (search not exhaustive). The nearest works are Nguyen et al. 2021, Kollnig et al. 2021, Stock et al. 2016 and Neef et al. 2025. Our design differs from each of them: it is static where Nguyen et al. is dynamic, it targets by salience where Kollnig et al. sample at random, it covers apps where Stock et al. cover websites, and it measures reaction to concrete findings where Neef et al. measure the presence of programmes.

3 The Pre-CRA Baseline

The CRA entered into force on 10 December 2024 (twentieth day after publication on 20 November 2024, Article 71(1)). Chapter IV on notified bodies has applied since 11 June 2026, the reporting obligations of Article 14 since 11 September 2026, and the remaining provisions apply from 11 December 2027 (Article 71(2)). Apps fall within the scope as software products with digital elements placed on the market in the course of a commercial activity, free of charge or against payment; Recital 11 names mobile applications together with the manufacturer's backend. Medical devices under Regulation (EU) 2017/745 are excluded (Article 2(2)), which concerns part of our health category.

Article 14 covers actively exploited vulnerabilities and severe incidents and sets an early warning within 24 hours of becoming aware, a notification within 72 hours and a final report within 14 days after a corrective measure is available (Article 14(2)). Our findings are static and mostly describe exposure without evidence of exploitation, so they are not Article 14 notifications. The CVD policy and the contact address in Annex I, Part II, points 5 and 6, and the single point of contact in Annex II, point 2, apply from 11 December 2027. For products placed on the market before that date, the Regulation applies only after a substantial modification (Article 69(2)), with Article 14 applying to all products in scope (Article 69(3)). Article 13(17) adds a requirement on the form of the contact point: manufacturers designate a single point of contact that lets users choose their preferred means of communication and does not limit such means to automated tools.

We call the measurement a pre-CRA baseline because the first reports were sent between 14 January and 17 October 2026, and only 2 of 398 first reports were sent on or after 11 September 2026, the start of the Article 14 obligations. The obligations on CVD policy and contact address are not yet in force, so their absence in 2026 breaches no CRA duty. This paper makes no statement on CRA conformity of any vendor. Whether a vendor publishes a contact address or a CVD policy was not coded as a structured field in our dataset (Section 9).

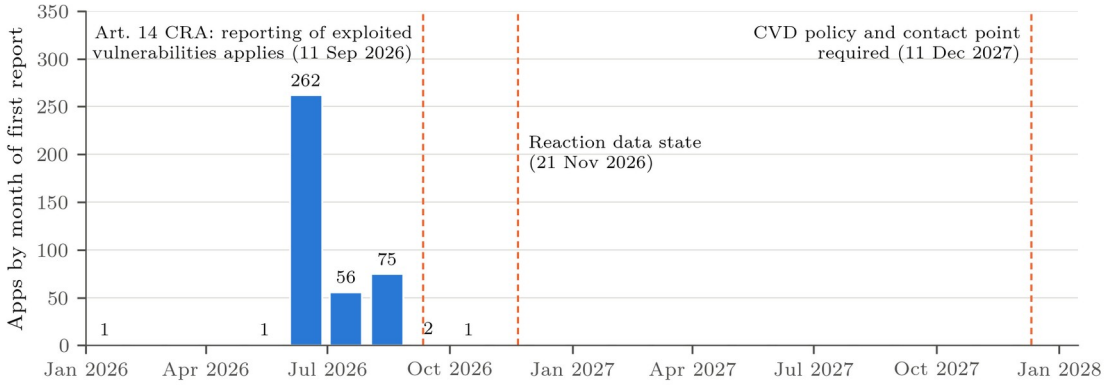

Figure 2. *Apps by month of their first report, with the dates of the CRA obligations. The reports went out before Article 14 applied, and the duty to run a CVD policy and a contact point applies from 11 December 2027, so the data describe the state before it (Section 3).*

4 Method

4.1 Pipeline and roles

We use a socio-technical pipeline with a human final decision (human-AI teaming). The pipeline is built and maintained by us, most of its steps run mechanically, and four components carry the work.

1. **An own analysis pipeline** processes the production Android package (APK) of every app. Its tools are fixed in our internal methodology and in the instructions for every audit run, and each tool is available to the pipeline as a defined skill. For every app the pipeline first archives the package with its SHA-256 hash and version as the chain-of-custody record, then decodes manifest, resources and network security configuration with APKTool and aapt2, builds a code and SDK inventory from the DEX files (package names, endpoints, keys, hosts) and decompiles code with JADX where the size of the package permits it. Native libraries are analysed with Ghidra on a separate server, and every claim about native code carries the SHA-256 hash and function address or is marked as unverified. Secret scanners and pattern rules (Semgrep, TruffleHog and own parsers) complement the tracing of code paths, data flows and SDK initialisation. Dynamic analysis with tools such as Frida is part of the methodology for other engagements and is outside this dataset (Section 4.2).

2. **NYX**, our security knowledge graph with deterministic layers and small models that we trained ourselves, runs continuously. It assigns findings to the 17 finding classes and links them to the CWE, Common Attack Pattern Enumeration and Classification (CAPEC) and MITRE Adversarial Tactics, Techniques and Common Knowledge (ATT&CK) catalogues, and in our assessment it carries the largest share of the analytical work after decompilation.
3. **DINGIR**, our situational-awareness system built around a graph neural network with ternary weights, runs continuously as well and supplies live context to the pipeline and to the reviewers from the Cybersecurity and Infrastructure Security Agency (CISA) Known Exploited Vulnerabilities catalogue (KEV), the National Vulnerability Database (NVD) and the MITRE CVE list.

General language models enter at the end of this chain. A bridge to large language models consolidates the outputs of the mechanical steps, NYX and DINGIR into finding candidates and report drafts, and the analyses for this paper (pattern recognition, clustering support, consolidation and a second reading of the text) ran through the Anthropic API with Claude models. OpenAI models GPT Sol and GPT Luna, used through Codex, served only computational pattern recognition runs and the preparation of training data. Auxiliary language models came mainly from Mistral and from NVIDIA Nemotron models hosted in Europe, and since 3 November 2026 also through OpenRouter with routing to providers in the European Union. Where a further model was needed, we chose models whose use requires no data-sharing agreement. Much of the sorting work runs on pure classification models without text generation, among them Mercury Decide, a decision model of Inception Labs that returns a calibrated probability with each answer so that uncertain cases can be passed to a human, and Laya, a classifier maintained by us, which made fast classification of large volumes possible. The pipeline is model-agnostic by design, so each language model or classifier can be replaced without changing the analysis steps. The language models therefore assemble and phrase results; discovery rests on the pipeline, NYX and DINGIR.

Between January and November 2026 the following language models were used, without a breakdown by period: Claude Opus 5.5, Claude Sonnet 5.5, Claude Sonnet 5, Claude Haiku 5.5 and Claude Fable 5 (Anthropic API), Mistral Medium 3.5 and Mistral Small 4, NVIDIA Nemotron 3.5 Lightning and Nemotron 3 Ultra, and gpt-oss-120b. All of them were used under the same procedure: the same findings were often given to several models and several instances, each of which first assessed them on its own, and the models then reviewed each other's assessments; a part of this work ran in our own desktop application, in which several model instances work on one case together. Humans compared the assessments and decided. The agreement between models is a plausibility check and replaces no human decision.

The order of work is the same for every case. The pipeline and the investigation run automatically first, and the language models receive only their results as context; they have no direct access to the packages or to the raw data of an investigation, so packages and raw investigation data are not sent to model providers. The team then reviews the results, lets models assess and analyse them, applies computational pattern recognition, and decides jointly what follows and which model processes which part. Some reports were written by hand without any language model. After each run the results are prepared as training data and fed back into our own models Laya, NYX and DINGIR, so these models changed during the observation period. The class assignment reported in this paper uses fixed title rules applied once to all findings during cleaning (Section 7.1) and therefore does not depend on the model state at the time of a run.

4. **Human reviewers** decide whether a candidate becomes a finding, which severity applies, who receives a report and whether the report is sent. The decision is recorded in the report specification of the app. Findings can carry the status disputed, addressed or retracted; in the cleaned dataset 3 findings are disputed, 3 addressed and 1 retracted.

The roles follow the separation of source evidence, machine-generated candidate and human-validated finding. Machine output is input to that human decision and becomes evidence only through it. Human review can itself be biased by automation (Parasuraman and Manzey 2010), a limit that we state in Section 9 and which Section 7 does not test.

4.2 Deep Scan

A Deep Scan is a static analysis of the production APK of one app. It does not run the app against a live backend and does not test whether a secret is valid. The scan covers the manifest (exported components, permissions, backup and network security configuration), the embedded libraries and their initialisation order, strings and resource files (keys, endpoints, hosts), certificate handling, the declared data flows and the declarations of the store listing and of the privacy policy where they bear on a finding. Findings therefore describe exposure and non-compliance visible in the package. They do not prove exploitation.

4.3 Dual classification

Each finding is assigned to one or several of 17 finding classes by rules applied to the finding title (Table 2 lists them). The rule set (taxonomy rules of our security knowledge graph, NYX) was applied anew to all 1,735 findings; the graph itself, at its stand of 6 October 2026, knew only 1,689 of them and is not used for counting. Each class is placed on two axes. The security axis uses CWE and the OWASP MASWE and Mobile Top 10 in the 2024 version; the OWASP field in the reports mixes the 2016 and 2024 versions and was normalised to 2024 by a conversion table, which is partly our own assessment. The privacy axis uses GDPR articles, Article 5(3) of the ePrivacy Directive, MASVS-PRIVACY and LINDDUN. The mapping of classes to CWE is our own assessment based on the MITRE CWE catalogue and the OWASP MASWE repository; four classes map directly, two conditionally, five only to generic CWE entries and six to no CWE [Derived; Hypothesis for the assignment, which is our own]. Findings without a class (554 of 1,735) stay in the dataset and are reported as such.

4.4 Severity on two scales

Severity per finding comes from the report specification and may include a raise through a blast-radius assessment (the number of users and the sensitivity of the data affected). The flag for a raise is set for 314 of 1,735 findings (18.1 %). The second scale is the exposure level of the target in the public ledger, one level per target, which includes the blast-radius raise. The two scales are kept apart. Part A of this paper uses the per-finding severity, and the ledger exposure appears as a separate axis. For each app we compare the exposure level with the highest finding severity of the app, and we report the divergence as a result about measurement (Section 6.3).

4.5 Coordinated disclosure and reaction coding

Reports followed a coordinated disclosure schedule with a 90-day period. The report went to the vendor addresses found for the operator, with supervisory and computer emergency response team (CERT) addresses in blind copy. The median interval between first report and closing date is 90 days (244 of 398 apps exactly 90 days, 335 within 88 to 92 days).

The reporting campaign. All reports came from our organisation and followed one English-language template with the subject line "Security & Privacy Disclosure" and the report attached. We addressed the data protection and security addresses we found for the operator and, where none was published, the general contact address; the tracker does not list every recipient per row, so we give no shares by address type. Follow-up mails within the 90-day period served as reminders: the tracker records a median of 7 mails per row, and 363 of 378 rows with a count have at least three. 262 apps received their first report in June 2026. The address research changed during the campaign, and the share of apps with a bounced vendor address fell from 27.2 % before 1 August to 5.1 % afterwards (Section 6.5). All reaction figures therefore describe the reaction to our reports under these conditions and not the reaction to reports in general.

We coded the reaction per tracker row in three tiers, applying the first matching rule of a fixed rule set to the tracker fields and the ledger status:

- **Tier A**, responsible human reply, including a reasoned refusal by a data protection officer, security team or legal department.
- **Tier B1**, referral to an official reporting channel of the vendor (bug bounty platform, vulnerability disclosure programme) without a statement on the content. Of all replies, it comes closest to the reporting path that the CRA requires, a CVD policy and a contact address (Annex I, Part II, points 5 and 6).
- **Tier B2**, a reply without substance: a bot or automatic reply only, a reply from a desk that is not responsible (customer service, community support, press), a request for an identity check only, or a refusal to answer without a statement on the content.
- **Tier C**, silence or bounce. We split C into silence without bounce to vendor addresses and with the case closed, all vendor addresses undeliverable, and unclear (partial bounce without proof for the other addresses, case still open, or missing data).

Bounces from the authority addresses in blind copy do not count as vendor bounces. The status RESOLVED in tracker and ledger means that the case is closed and never means that a vulnerability was fixed; this dataset contains no fix rate. Each coded row carries a confidence (high, medium, low) and a free-text note where the rule needed interpretation. In a second pass by the authors, every row that the rules had placed in a redirection category (34 targets) was read individually against the tracker notes and recoded with a written reason (Section 5.2).

Ethics and influence of the campaign. The reports concerned weaknesses visible in the distributed package and were sent to the vendor before any publication, with a fixed period of 90 days. The blind copy to supervisory authorities and CERTs was part of every report; it may itself have shaped the reaction, for example by routing a report to the legal department, and its effect cannot be separated in this dataset. The naming rule is given in Section 5.2.

Second reporting route. For two cases the report went to the app store operator (Google Play and Android Security) as a platform route, because the listing itself was the deception: a children's game that hid a casino, and a casino brand that appeared under cleaning-game titles. An answer from a platform measures an enforcement body and says nothing about the vendor's own process. We therefore code these cases outside tiers A, B and C in a separate category P (platform route). Category P records the outcome (removed, not removed, unknown), the report date and the outcome date, and the dataset records the route in the column `meldeweg` (vendor, platform, both, unknown). The value `both` does not occur. Before this separation one of the two cases stood in tier A, which counted a platform confirmation as a vendor reply. We tag evidence as elsewhere in this paper: the removal is Observed when the platform confirmed it, and the cause of the removal is a Hypothesis unless the platform states it.

4.6 Rates per app

Because findings within an app are not independent, we report rates per app: an app counts once for a class if at least one of its findings belongs to that class. Reaction shares use the 394 apps in tiers A, B and C as base; the two platform cases (Section 6.7) are not part of it. Because apps of one operator share a report and a recipient, we give the report and the operator group as further units: Table 2 counts classes per report as well as per app, and the reaction shares are given with each operator group weighted once. Shares carry Wilson 95 % confidence intervals where stated. For the correlates of a responsible reply we fit a logistic model (reply in tier A against a CRITICAL highest finding, EU or EEA origin, first report in June, and missing report) with standard errors clustered by operator group, once without and once with category effects. Several apps belong to one operator and received one report and one reply; the countercheck weights each of the 280 operator groups with a tracker row once (Section 6.4).

4.7 Validation procedures

Three procedures check the central patterns. First, the taxonomy mapping of Section 4.3 yields the first line of evidence on the coverage of security taxonomies. Second, a data-driven check clusters the finding texts: we embed title and cleaned impact text with the all-MiniLM-L6-v2 sentence embedding model, reduce to 20 dimensions by principal component analysis (PCA) and cluster with k-means for k from 2 to 20 (seed 42, silhouette criterion, which scores how well each item fits its own cluster compared with the nearest other cluster). We compare the clusters with the OWASP labels (purity against the majority baseline, normalised mutual information NMI, adjusted Rand index ARI) and with the CWE mapping level. Third, we formulate a counter-hypothesis for each central pattern (Section 7). Pipeline text templates were removed before counting (Section 5.2). Before publication the manuscript undergoes human peer review as external validation.

4.8 Failure conditions

A finding is rejected when the review shows that the observation is wrong, that the component is not part of the production package, or that the claim exceeds the evidence (status retracted or disputed). A reaction code is revised when the tracker text contradicts the rule outcome; the 9 apps whose ledger status contradicts the tracker coding follow the newer tracker field. A pattern is dropped from the paper when a counter-hypothesis explains it at least as well and no data separate the two; Section 7 lists such cases.

5 Data

5.1 Scope and size

Quantity	**Value**
Ledger rows (raw)	387
Targets after merging 4 duplicate names	383
Apps after resolving 4 bundle rows into 19 apps	398
Apps with a static-analysis report and at least one finding	340
Apps without a report (findings unknown, not zero)	58
Reports (after removing 6 backup copies)	268
Findings after cleaning	1735
Apps with a response-tracker row (basis of all reaction figures)	394
Apps outside the reaction base (2 platform cases, 2 without tracker row or platform report)	4
Operator groups (all apps / apps with tracker row)	283 / 280
App categories	16 (of which “Other, unclear”: 28 apps)
First reports sent	2026-01-14 to 2026-10-17 (June 2026: 262 apps)
Findings per app with findings, counting findings that an app shares through a joint report (median / mean / maximum)	6 / 6.5 / 19

Table 1. Dataset summary (Observed counts; derived rules in Section 5.2). Local state of 22 November 2026.

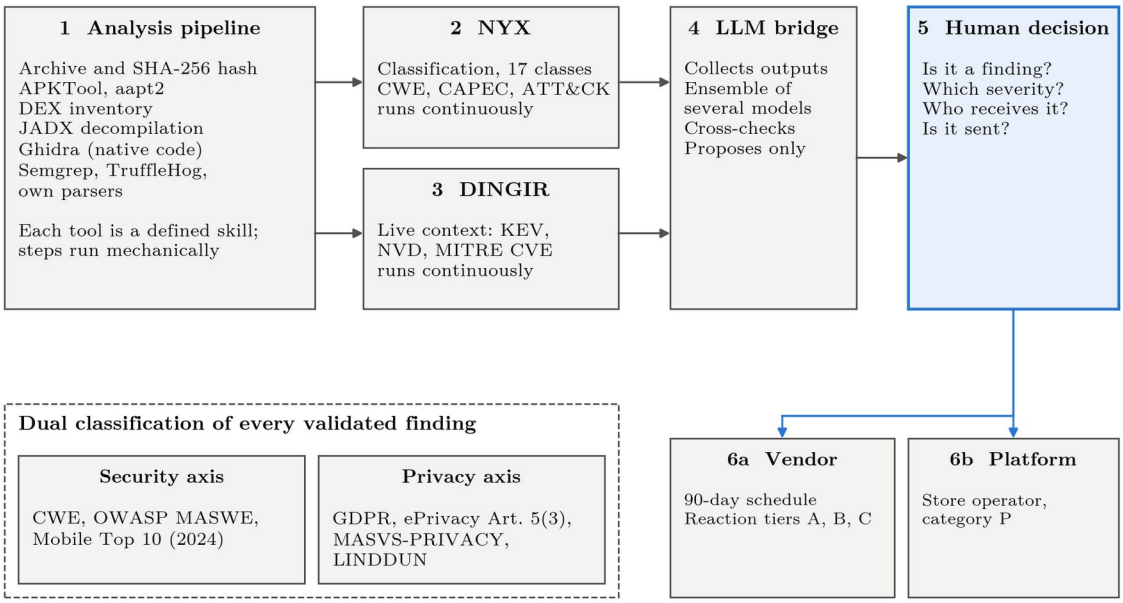

Figure 3. *Analysis and disclosure pipeline. An own analysis pipeline feeds two components that run continuously, the knowledge graph NYX (classification into 17 finding classes) and the situational-awareness system DINGIR (live vulnerability context). A bridge to language models consolidates their output into candidates, and a human decision fixes the finding, its severity, the recipient and the dispatch, followed by a report to the vendor or to the platform. Every finding is placed on a security axis and a privacy axis.*

The dataset consists of Android apps that vendors offer on the EU market. Targets were chosen by salience (known leaks, consumer waves by sector, apps with large audiences), without a random sampling frame. We take availability on the EU market from the statement of the project lead and do not code it per app; a per-app check of store country availability was not performed. The reaction base is the tracker file of 385 rows, linked to 395 apps; 5 tracker rows have no ledger target of their own (a duplicate, a second row for one target, a name collision with another thread, an aggregate row for a family of apps, and a courtesy message without a reportable finding) and are excluded from app-level analysis. Three apps have no tracker row; one of them was reported to a platform (Section 6.7), and the reaction of the other two is not coded. The 58 apps without a report have unknown findings, which differs from zero findings.

5.2 Cleaning and disclosed decisions

We applied the following rules to the raw files; each reduces or marks rows and changes counts relative to earlier internal reports, and the cleaned dataset takes precedence over them.

- **Duplicates and bundles.** Four ledger rows share a name with another row and were merged (387 rows to 383 targets). Four bundle rows that state a number of apps (5, 5, 3 and 6 apps) were resolved into 19 apps (383 targets to 398 apps). The 19 member apps inherit the findings, the reaction and the disclosure date of their target, because the report went to the operator and not to each app.
- **Backup copies.** Six backup versions of report files were removed (274 to 268 reports, 1,785 to 1,735 findings). They had produced most of the verbatim duplicate findings of earlier counts. The remaining 7 findings in 3 groups are identical across different apps and stay in the data with a marker.
- **Pipeline text templates.** The text block of the severity raise was removed from the impact text and kept as a flag; generic article citations and a resource path were removed from the normalised text used for counting and text mining. The flag is set for 314 findings, because the raise is also written into the title box and the text box, which earlier counts (86 findings) did not read.
- **Secrets.** Secrets in text columns are masked by type and a hash prefix, and e-mail addresses are replaced by a placeholder.
- **Category.** Sixteen categories result from keyword rules, applied first to the name, then to the report, then to the finding text (329, 11 and 58 apps); 28 apps fall under "Other, unclear".
- **Origin.** The origin group is a coarse rule on the target field and makes no claim about market access. For 47 apps the origin was added by hand from general knowledge, and for 48 apps it is unclear. In this paper the 47 general-knowledge assignments count as unclear, so 95 apps (92 with a tracker row in A, B, C) are unclear.
- **Operator groups.** The operator comes from the first company in the target field and an alias table for families; 283 groups cover the 398 apps. An automated organisation resolver of the knowledge graph (DINGIR) was not used, because it agreed with the report evidence in only 3 of 8 operator groups examined in an earlier analysis.
- **Reaction coding confidence.** Of 394 coded apps, 272 carry high, 121 medium and 1 low confidence. Among tier A, 47 apps carry medium and 1 low confidence (the reply field says only "yes", and the content was not checked individually); among tier B, 6 medium; among tier C, 68 medium. In 160 rows the tracker note carries a remark.
- **Individual review of redirections.** The rules had placed 34 targets in a redirection category. Read against the tracker notes, 13 of them contain a statement of the responsible desk on the report (a dispute, a fix of one finding, a confirmed review, answers to our questions) and move to tier A; 8 refer to an official reporting channel (B1); 12 are a wrong desk, an identity check or a refusal without content (B2); and 1 shows no reply in the thread and moves to tier C. The reason per target is kept with the dataset.
- **Installs.** Lower bounds of install counts are read from report text for 42 apps. They are not person counts.
- **Names.** Vendors and apps appear only in aggregated form in this paper. The two platform cases of Section 6.7 are named, because their disclosure is complete and the report is public.

6 Results

Part A: Weaknesses

6.1 Distribution of finding classes (RQ1)

Finding class	Mapping to CWE	Main CWE	Reports with at least one finding (of 268)	Apps with at least one finding (of 340)	Findings
HARDCODED_SECRET	direct	CWE-798	195 (72.8 %)	234 (68.8 %)	228
AD_TRACKING_SDK	generic only	CWE-359 (generic)	152 (56.7 %)	214 (62.9 %)	229
PRE_CONSENT_INIT	none	n/a	100 (37.3 %)	112 (32.9 %)	114
THIRD_COUNTRY_TRANSFER	none	n/a	97 (36.2 %)	130 (38.2 %)	123
SPECIAL_CATEGORY_DATA	none	n/a	97 (36.2 %)	103 (30.3 %)	154
CERT_VALIDATION	conditional	CWE-295	96 (35.8 %)	107 (31.5 %)	100
CLEARTEXT_TRAFFIC	direct	CWE-319	84 (31.3 %)	90 (26.5 %)	90
TRANSPARENCY_GAP	none	n/a	83 (31.0 %)	89 (26.2 %)	105
SENSITIVE_PERMISSION	generic only	CWE-250 (generic)	63 (23.5 %)	78 (22.9 %)	86
CHILD_DATA	none	n/a	58 (21.6 %)	88 (25.9 %)	89
LOCATION_COLLECTION	generic only	CWE-359 (generic)	53 (19.8 %)	61 (17.9 %)	53
EXPORTED_COMPONENT	direct	CWE-926	34 (12.7 %)	36 (10.6 %)	35
BACKUP_EXPOSURE	conditional	CWE-212	24 (9.0 %)	49 (14.4 %)	24
SCREEN_RECORDING	generic only	CWE-200 (generic)	21 (7.8 %)	24 (7.1 %)	21
UNDISCLOSED_PROCESSOR	none	n/a	16 (6.0 %)	16 (4.7 %)	17
DEBUG_ARTEFACT	direct	CWE-489	11 (4.1 %)	11 (3.2 %)	12
SUPPLY_CHAIN_CLUSTER	generic only	CWE-1395 (generic)	10 (3.7 %)	14 (4.1 %)	10

Table 2. Finding classes, mapping level to CWE, and reports and apps with at least one finding of the class (bases: 268 reports and 340 apps with findings; Derived by rules on finding titles; findings can belong to several classes).

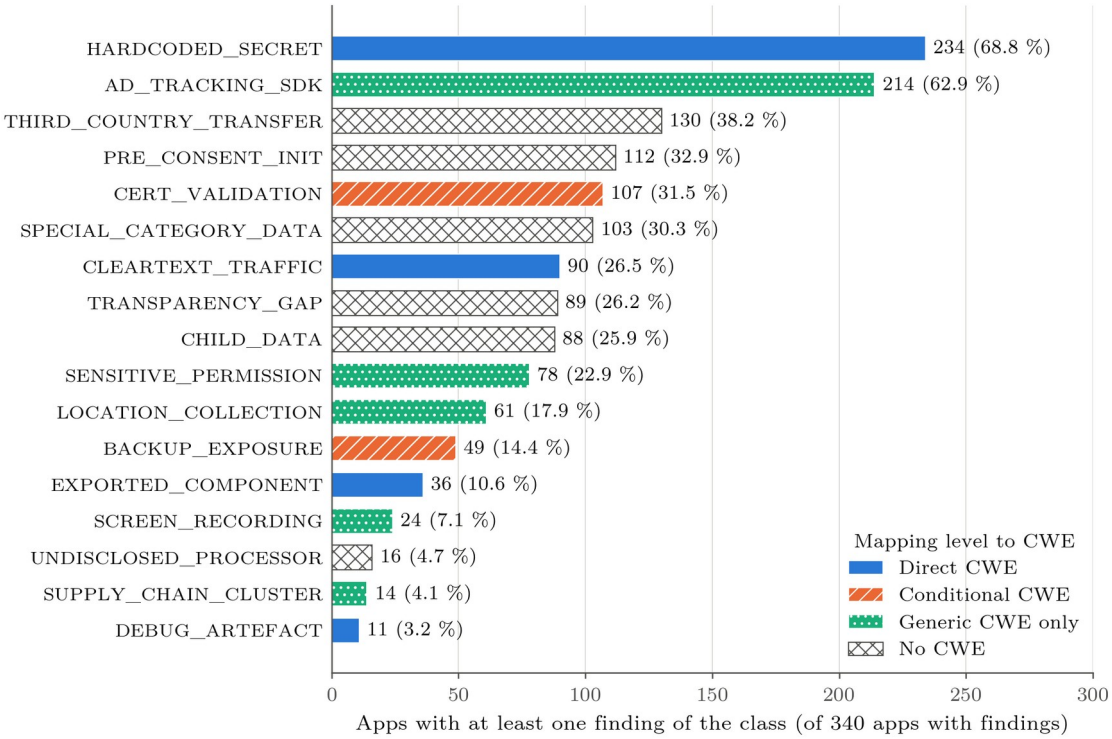

Figure 4. *Finding classes by number of apps with at least one finding of the class (base: 340 apps with findings; Table 2). Bar fill shows the mapping level of the class to CWE, and the hatched outline marks classes without any CWE.*

Hard-coded secrets occur in 234 of 340 apps (68.8 %) and advertising or tracking SDKs in 214 apps (62.9 %). The class of third-country transfers (130 apps, 38.2 %) and the class of analytics or advertising start before a consent decision (112 apps, 32.9 %) follow, together with certificate validation (107 apps, 31.5 %) and special-category data (103 apps, 30.3 %). Three classes concern a technical flag in the manifest or the build: exported components in 36 apps (10.6 %), backup exposure in 49 apps (14.4 %) and debug artefacts in 11 apps (3.2 %). Only 3 of 340 apps have no finding with a class. The median app has 6 findings (mean 6.5, maximum 19, counting findings that an app shares through a joint report), and 86 apps (21.6 % of all 398) share their report with other apps of the same operator. Per app, a class therefore also counts when the app belongs to a joint report with such a finding; this explains classes with fewer findings than apps, such as backup exposure (24 findings, 49 apps). Per report, the order of the classes stays close to the order per app (Table 2).

A class counts presence of a finding and does not confirm exploitability. A hard-coded API key of a third-party service can be public by design, and the class needs a validity feature that static analysis cannot supply. By title rules, 11 apps carry a secret that is privileged by its name (an app secret, a client secret, an authentication token), 178 apps carry only client identifiers such as Firebase or Google API keys, and 45 apps carry other or unclear entries. The share of 68.8 % is therefore no rate of leaked credentials. Pre-

consent initialisation and third-country transfer rest on static evidence (initialisation code, endpoints in the package) and are indications of a transfer; a dynamic measurement as in Nguyen et al. was not performed. As a plausibility indicator for the technical classes, the CISA KEV catalogue (1,790 entries, queried on 21 November 2026) lists 9 entries for CWE-798 and 5 for CWE-295 and none for CWE-319, CWE-926 or CWE-489 [Observed]. KEV mostly lists enterprise and network products, and a count of zero says nothing about apps.

The class profile varies by category (base: apps with findings; small categories are marked in Table 4). In the 26 children and family apps 24 (92.3 %) carry a child-data finding and 17 an advertising or tracking SDK finding. In games, 28 of 31 apps (90.3 %) carry a third-country transfer finding and 26 of 31 (83.9 %) a child-data finding. In wearables, 20 of 21 apps (95.2 %) carry a cleartext-traffic finding. None of the 23 social and dating apps carries a pre-consent finding. We treat such profile differences as descriptive for this selection, since report templates differed between waves (Hypothesis).

6.2 Mapping to CWE and OWASP (RQ1)

Mapping level of the finding	Findings	Share of classified findings	CRITICAL findings	Share of all CRITICAL
Direct CWE	360	30.5 %	126	39.4 %
Conditional CWE	86	7.3 %	22	6.9 %
Generic CWE only	359	30.4 %	46	14.4 %
No CWE	376	31.8 %	70	21.9 %
No class assigned	554	n/a	56	17.5 %

Table 3. Mapping level of findings and share among CRITICAL findings (Derived). Percentages in the third column use the 1,181 classified findings as base; the last column uses all 320 CRITICAL findings.

Of the 17 classes, 4 map directly to a CWE (23.5 %), 2 map conditionally (11.8 %), 5 map only to generic CWE entries (29.4 %) and 6 map to no CWE (35.3 %); 11 of 17 classes (64.7 %) therefore lack a robust CWE [Derived; Hypothesis for the assignment]. The six classes without CWE are pre-consent initialisation, special-category data, transparency gap, third-country transfer, child data and undisclosed processor. These are legal or documentary defects. They are described by GDPR articles, Article 5(3) of the ePrivacy Directive, MASVS-PRIVACY and LINDDUN, and the OWASP MASWE privacy entries attach only the generic CWE-359 label.

On the level of findings, 554 of 1,735 findings (31.9 %) have no class. Among the 1,181 classified findings, 360 (30.5 %) map directly to a CWE, 86 (7.3 %) conditionally, 359 (30.4 %) only to a generic CWE and 376 (31.8 %) to no CWE. On the level of apps, 267 of 340 (78.5 %) carry at least one finding in a class without CWE, and 8 apps (2.4 %) carry findings only in such classes. In the interplay of both axes, 275 apps (80.9 %) have at least one finding in a class with a direct or conditional CWE and at least one finding in a class carrying a privacy-axis entry, 47 apps (13.8 %) only privacy-axis findings, and 15 apps (4.4 %) only security-axis findings. A scan that uses only the security axis would miss the privacy findings of 47 apps entirely and would see only one side of the other 275; this follows from the definition of the two axes and describes the set. Per report and per operator group the picture is the same: 221 of 268

reports (82.5 %) and 193 of 228 groups with findings (84.6 %) carry at least one finding in a class without CWE.

The privacy gap is not confined to low severity. Of the 320 CRITICAL findings, 70 (21.9 %) belong to classes without CWE and a further 46 (14.4 %) to classes with only a generic CWE, together 116 (36.3 %); 56 (17.5 %) have no class, and 148 (46.2 %) map directly or conditionally (Table 3). The OWASP field is filled for 435 findings (25.1 %) and carries the 2024 category M6 (Inadequate Privacy Controls) in 257 of them (59.1 %); the field mixes 2016 and 2024 versions (364 in 2024). Of 768 findings with a privacy-axis entry, 574 (74.7 %) carry no OWASP category at all. The low fill rate of the OWASP field is a property of our reports and not of OWASP, whose MASVS-PRIVACY and MASWE privacy entries address these topics; the statement of this section therefore concerns CWE.

6.3 Severity on two scales (RQ1, measurement)

Per-finding severity in the cleaned data is CRITICAL for 320 findings (18.4 %), HIGH for 655 (37.8 %), MEDIUM for 522 (30.1 %) and LOW for 190 (11.0 %); 48 findings (2.8 %) carry another status label (information, observational, disputed, addressed, retracted). The ledger exposure level is CRITICAL for 264 of 398 apps (66.3 %), and the highest finding severity is CRITICAL for 220 of 340 apps with findings (64.7 %).

Category	**Apps with findings**	**Exposure CRITICAL**	**Finding maximum CRITICAL**	**Equal**	**Exposure higher**	**Finding higher**
Health, medical	43	27 (62.8 %)	17 (39.5 %)	31	11	1
Retail, e-commerce, delivery	41	32 (78.0 %)	34 (82.9 %)	37	1	3
Games	31	28 (90.3 %)	30 (96.8 %)	29	0	2
Public bodies, broadcasting	27	5 (18.5 %)	24 (88.9 %)	8	0	19
Children, family	26	23 (88.5 %)	11 (42.3 %)	14	12	0
Social, dating, messaging	23	21 (91.3 %)	21 (91.3 %)	23	0	0
Media, streaming, news	21	13 (61.9 %)	14 (66.7 %)	16	2	3
Wearables, fitness trackers	21	20 (95.2 %)	1 (4.8 %)	2	19	0
Productivity, tools, security	19	14 (73.7 %)	14 (73.7 %)	19	0	0
Banking, payment, finance	17	12 (70.6 %)	9 (52.9 %)	11	5	1
AI companions, chat AI	16	13 (81.2 %)	8 (50.0 %)	11	5	0
Smart home, IoT	15	11 (73.3 %)	6 (40.0 %)	9	6	0
Mobility, travel	14	11 (78.6 %)	11 (78.6 %)	12	1	1

Category	Apps with findings	Exposure CRITICAL	Finding maximum CRITICAL	Equal	Exposure higher	Finding higher
Gambling, betting, crypto trading	11	9 (81.8 %)	9 (81.8 %)	11	0	0
Other, unclear	8 (n<10)	8 (100.0 %)	6 (75.0 %)	6	2	0
Telecom, connectivity, VPN	7 (n<10)	5 (71.4 %)	5 (71.4 %)	5	1	1
All categories	340	252 (74.1 %)	220 (64.7 %)	244	65	31

Table 4. Ledger exposure compared with the highest finding severity per app, by category (apps with findings; Derived).

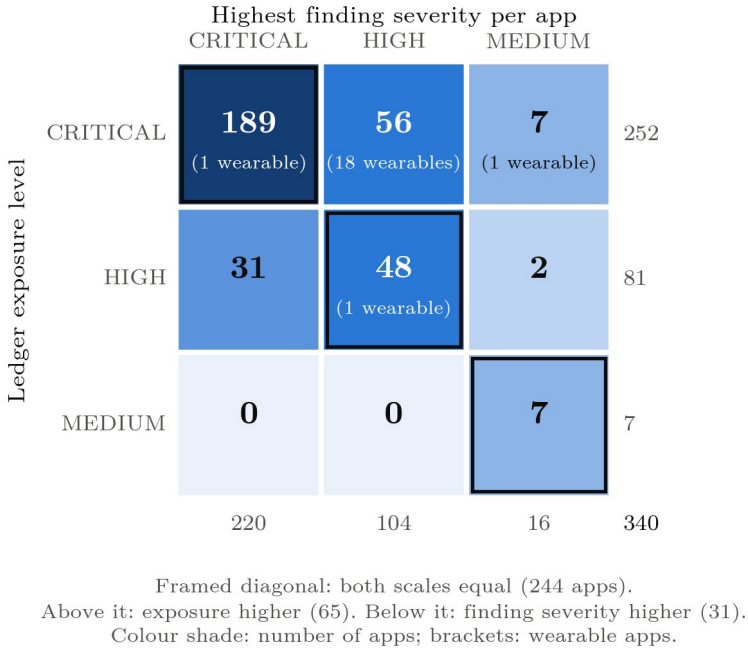

Figure 5. *Ledger exposure level against the highest finding severity per app (340 apps with findings). Framed cells on the diagonal mark apps where both scales agree; counts in brackets give the wearable apps in a cell.*

For 244 of 340 apps (71.8 %) both scales agree. The exposure is higher for 65 apps (19.1 %), and the finding severity is higher for 31 apps (9.1 %). The rank correlation between the two scales is 0.38 (Spearman, 340 apps). The divergence concentrates in categories: 19 of the 21 wearable apps carry a higher exposure than finding severity (20 of 21 with exposure CRITICAL, 1 of 21 with a CRITICAL finding), and 19 of the 27 apps of the category public bodies and broadcasting show the reverse (finding severity above exposure). Without the wearables, 46 apps with higher exposure remain.

The two scales measure different things, and we use both. The ledger raises the overall level of a target through a blast-radius assessment. The report specifications of the wearable apps contain 213 findings, of which 3 are CRITICAL, 34 HIGH, 105 MEDIUM, 59 LOW and 12 informational; the ledger raise did not carry over into the finding level. We assume that the wearable reports were written in a later format with their own severity convention (Hypothesis). The finding is a measurement result: severity statements for this set depend on the scale used, and the same app can be CRITICAL on one scale and HIGH or MEDIUM on the other.

Part B: Vendor response

6.4 Reaction tiers (Primary RQ, RQ2)

Tier	Meaning	Apps	Share of 394
A	Responsible human reply (including a reasoned refusal)	115	29.2 %
B1	Referral to an official reporting channel (bug bounty, vulnerability disclosure programme)	8	2.0 %
B2	Wrong desk, identity check only, or refusal without content	12	3.0 %
B2	Automatic or bot reply only	61	15.5 %
C	Silence, no bounce to vendor addresses, case closed	143	36.3 %
C	Unclear (partial bounce, open case, missing data)	54	13.7 %
C	All vendor addresses undeliverable	1	0.3 %
Total	A / B / C	115 / 81 / 198	29.2 / 20.6 / 50.3 %

Table 5. Reaction tiers, 394 apps with a tracker row (Derived; rule-based coding with an individual review of all redirections; the two platform cases are not included).

A responsible human reply came from 115 of 394 apps (29.2 %, 95 % confidence interval 24.9 to 33.9 %). Only 8 apps (2.0 %) received a referral to an official reporting channel (B1), and 73 apps (18.5 %) a reply without substance (B2), namely 61 bot replies (15.5 %) and 12 replies from a wrong desk, an identity check or a refusal without content (3.0 %). 198 apps (50.3 %) fall into tier C: 143 (36.3 %, confidence interval 31.7 to 41.2 %) stayed silent without a bounce, 54 (13.7 %) are unclear, and for 1 app (0.3 %) all vendor addresses were undeliverable. Tier C therefore holds a core of silence of about one third and a further seventh of cases whose outcome is not known. With each of the 280 operator groups that have a tracker row counted once, the shares are 30.9 % (A), 24.1 % (B) and 45.0 % (C). Delivery was not contradicted for 298 apps, partly contradicted (at least one vendor address bounced) for 89 apps (22.6 %), and fully contradicted for 1 app; for 6 apps it is unknown. Nineteen of the tier A apps also had a partial bounce. Of the 198 tier C apps, 194 carry the status RESOLVED, which marks closure of the case.

Observation time is long relative to the schedule: for 391 of 394 apps (99.2 %) at least 90 days passed between the first report and the last tracker check (21 November 2026, upper bound; the check date differs per row), with a median of 147 days (range 35 to 311). The dataset records no time to first reply.

Category	Apps	Operator groups	A	B	C	A, groups weighted once
Productivity, tools, security	21	16	14 (66.7 %)	4 (19.0 %)	3 (14.3 %)	56.2 %
Telecom, connectivity, VPN	7 (n<10)	5	4 (57.1 %)	0 (0.0 %)	3 (42.9 %)	60.0 %
Children, family	29	19	14 (48.3 %)	6 (20.7 %)	9 (31.0 %)	28.4 %
Health, medical	45	38	19 (42.2 %)	6 (13.3 %)	20 (44.4 %)	50.0 %
Banking, payment, finance	22	20	9 (40.9 %)	3 (13.6 %)	10 (45.5 %)	45.0 %
Smart home, IoT	15	12	6 (40.0 %)	4 (26.7 %)	5 (33.3 %)	33.3 %
Retail, e-commerce, delivery	46	40	14 (30.4 %)	11 (23.9 %)	21 (45.7 %)	28.8 %
Mobility, travel	21	18	6 (28.6 %)	2 (9.5 %)	13 (61.9 %)	33.3 %
Other, unclear	28	28	7 (25.0 %)	9 (32.1 %)	12 (42.9 %)	25.0 %
AI companions, chat AI	17	16	4 (23.5 %)	5 (29.4 %)	8 (47.1 %)	25.0 %
Public bodies, broadcasting	27	8	5 (18.5 %)	0 (0.0 %)	22 (81.5 %)	62.5 %
Media, streaming, news	23	21	4 (17.4 %)	8 (34.8 %)	11 (47.8 %)	14.3 %
Gambling, betting, crypto trading	12	11	2 (16.7 %)	5 (41.7 %)	5 (41.7 %)	18.2 %
Wearables, fitness trackers	21	18	3 (14.3 %)	0 (0.0 %)	18 (85.7 %)	16.7 %
Social, dating, messaging	26	14	3 (11.5 %)	7 (26.9 %)	16 (61.5 %)	17.9 %
Games	34	13	1 (2.9 %)	11 (32.4 %)	22 (64.7 %)	7.7 %

Table 6. Reaction by category (Derived). The last column weights each operator group once within the category. Categories with fewer than 10 apps are marked.

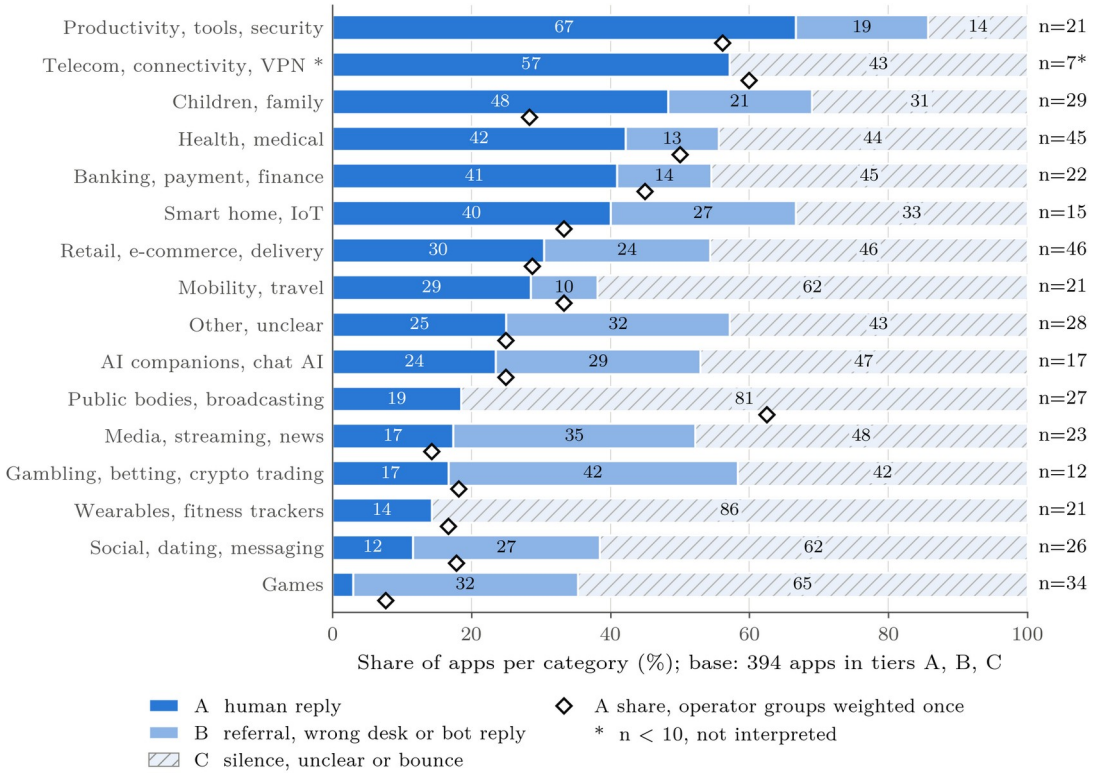

Figure 6. *Reaction tiers A, B (referral, wrong desk or bot reply) and C by app category (shares of apps; 394 apps in tiers A, B and C; the two platform cases are not included). Diamonds mark the share of tier A when each operator group is weighted once. Categories with fewer than 10 apps are marked with an asterisk and are not interpreted.*

The reply share differs by category. It is highest for productivity, tools and security apps (14 of 21, 66.7 %, confidence interval 45.4 to 82.8 %) and for children and family apps (14 of 29, 48.3 %), and lowest for games (1 of 34, 2.9 %), social, dating and messaging (3 of 26, 11.5 %) and wearables (3 of 21, 14.3 %). With 12 to 46 apps per category the intervals are wide and overlap for most neighbouring categories. Silence or bounce makes up 85.7 % in wearables (18 of 21), 81.5 % in public bodies and broadcasting (22 of 27), 64.7 % in games (22 of 34) and 14.3 % in productivity and security tools (3 of 21). The telecom category (7 apps) is too small for a statement.

Large operator families bear on the category figures. The two largest operator groups hold 26 and 17 apps (43 of 398, 10.8 %); one is a broadcaster family that makes up 19 of the 27 apps in the category public bodies and broadcasting, and one is a game publisher that makes up 17 of the 34 games. With groups weighted once, the reply share of the public bodies category rises from 18.5 % to 62.5 % (8 groups, so a small base), while games reach 7.7 % across 13 groups and wearables 16.7 % across 18 groups. The overall figures without the two largest groups are A 31.1 %, B 23.1 % and C 45.9 % (351 apps). The silence of the public-bodies category is therefore a property of one operator family in this selection and gives no evidence on public bodies in general. The low reply share of games holds in both counts.

6.5 Correlates of the reaction (RQ2)

Severity. The reply share is 27.4 % for apps whose highest finding is CRITICAL (60 of 219), 35.6 % for HIGH (37 of 104) and 43.8 % for MEDIUM (7 of 16, small base). The pooled gradient points in the direction opposite to severity. Inside categories the direction is mixed: among the 8 categories with at least 5 apps in each group, apps with a CRITICAL finding have a higher reply share than apps with lower findings in 5 categories, a lower one in 2 and an equal one in 1. By ledger exposure the reply shares are 31.9 % (CRITICAL, 260 apps), 23.0 % (HIGH, 113 apps) and 28.6 % (MEDIUM, 21 apps). The severity variable has little spread, because 64.7 % of the apps with findings have a CRITICAL highest finding, and whether the vendor saw the severity we assigned was not coded. In the logistic model the CRITICAL indicator shows no association (odds ratio 0.55, 95 % confidence interval 0.22 to 1.36, p = 0.194). We find no consistent relation between severity and reply. The 55 apps with a tracker row and without a structured report file in the local data were contacted as well, but their findings cannot be classified; they show a reply share of 20.0 % and a tier B share of 36.4 %.

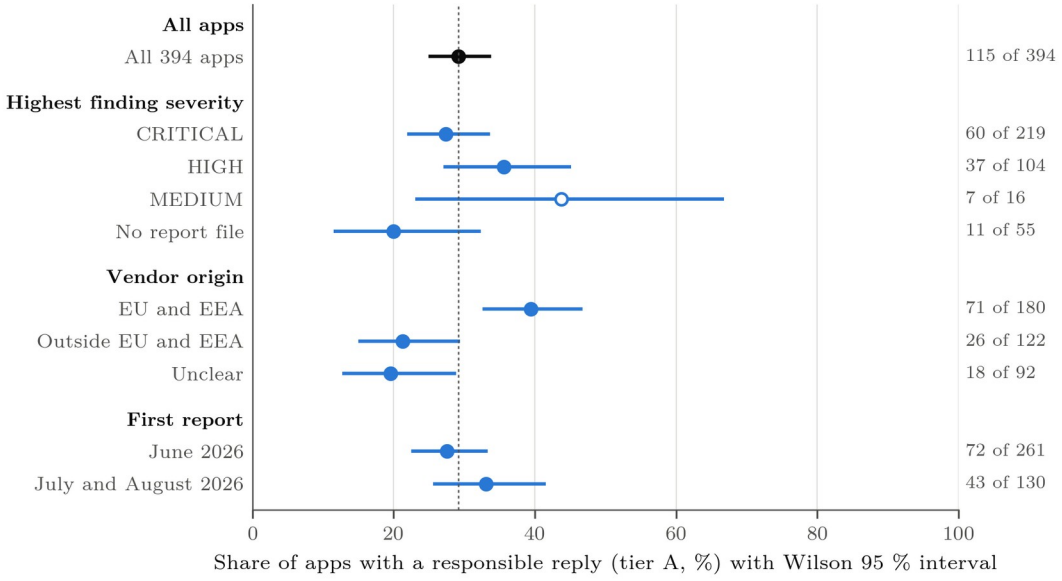

Figure 7. *Share of apps with a responsible reply (tier A) by highest finding severity, vendor origin and month of first report, with Wilson 95 % intervals. The dotted line marks the overall share of 29.2 %, and open circles mark groups with fewer than 30 apps.*

Finding profile. The reply share is 29.8 % for apps with findings on both axes (n=275), 36.2 % for privacy-only (n=47) and 35.7 % for security-only apps (n=14 with a coded reaction). By class, the reply shares lie between 26.2 % (third-country transfer, 130 apps) and 42.9 % (pre-consent initialisation, 112 apps) for classes with at least 20 apps; smaller classes are not interpreted. Apps with more than the median number of findings (6) show a reply share of 23.2 % (n=142) against 36.0 % (n=197) for the others.

Origin. Table 7 shows the reply shares by origin group, with general-knowledge assignments counted as unclear.

Origin group (general-knowledge assignments counted as unclear)	**Apps**	**A**	**B**	**C**
DACH (Germany, Austria, Switzerland)	130	50 (38.5 %)	23 (17.7 %)	57 (43.8 %)
Unclear	92	18 (19.6 %)	16 (17.4 %)	58 (63.0 %)
EU/EEA without DACH	50	21 (42.0 %)	11 (22.0 %)	18 (36.0 %)
China, Hong Kong, Taiwan	48	8 (16.7 %)	11 (22.9 %)	29 (60.4 %)
United States	45	9 (20.0 %)	13 (28.9 %)	23 (51.1 %)
Asia without China	13	2 (15.4 %)	3 (23.1 %)	8 (61.5 %)
United Kingdom	9 (n<10)	3 (33.3 %)	2 (22.2 %)	4 (44.4 %)
Other	7 (n<10)	4 (57.1 %)	2 (28.6 %)	1 (14.3 %)

Table 7. Reaction by origin group (Derived, coarse; the 47 general-knowledge assignments are counted as unclear).

Apps with an origin in DACH or in the rest of the EU and the European Economic Area (EEA) show a reply share of 39.4 % (71 of 180), while apps with an origin elsewhere (United States, China, other Asia, United Kingdom, others) show 21.3 % (26 of 122). The unclear group (92 apps) lies at 19.6 %. Origin and category are confounded: 22 of the 34 games have an unclear origin, 16 of the 29 children and family apps are from the EU or EEA outside DACH, and 24 of the 27 public bodies apps are from DACH. Within categories that have both groups, apps from the EU side reply more often in retail (37.9 % against 20.0 %), children and family (63.2 % against 20.0 %), health (52.6 % against 35.0 %), media (26.7 % against 0 %) and productivity tools (77.8 % against 66.7 %), and the shares are close in social apps (14.3 % and 11.1 %). The numbers per cell are small. In the logistic model with standard errors clustered by operator group, EU or EEA origin is associated with a reply (odds ratio 2.14, 95 % confidence interval 1.16 to 3.94, $p = 0.015$), and the association remains when the model adds category effects (odds ratio 2.31, 95 % confidence interval 1.21 to 4.43, $p = 0.012$). First report in June (odds ratio 1.10, 95 % confidence interval 0.38 to 3.21, $p = 0.864$) and a missing report (odds ratio 0.47, 95 % confidence interval 0.15 to 1.47, $p = 0.194$) show no association. The reference group of the origin indicator is all other apps, including the 92 with unclear origin. The model covers 394 apps in 280 operator groups; it describes this selection and does not separate origin from unobserved properties of the vendors.

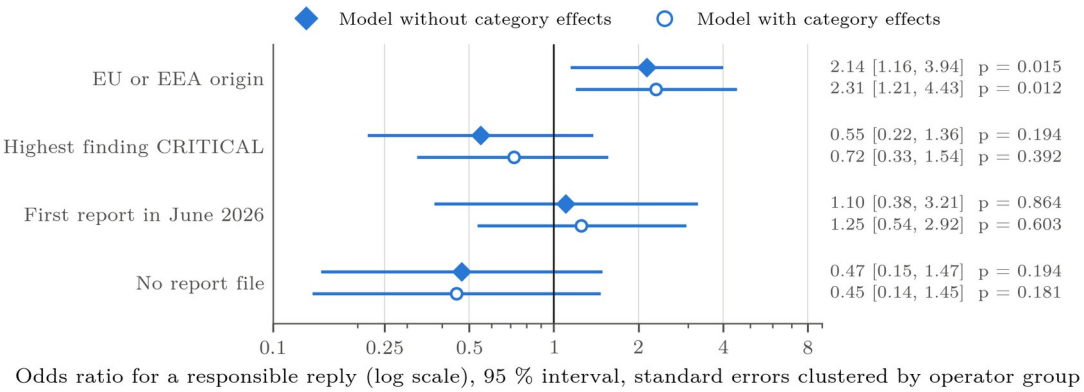

Figure 8. *Logistic model of a responsible reply with standard errors clustered by operator group (394 apps, 280 groups), without and with category effects. Odds ratios above 1 mean a higher chance of a reply; the reference group of the origin indicator is all other apps, including those with unclear origin.*

China origin and wearable category. All 21 wearable apps carry a China origin, 20 of them from a rule on the data and 1 from general knowledge, which counts as unclear here. The dataset contains no wearable vendor from another origin. Table 8 separates what can be separated.

Group	Apps	A	B	C
China origin (from data) and wearable category	20	2 (10.0 %)	0 (0.0 %)	18 (90.0 %)
China origin, other categories	28	6 (21.4 %)	11 (39.3 %)	11 (39.3 %)
Wearable category, origin unclear (general-knowledge assignment)	1 (n<10)	1 (100.0 %)	0 (0.0 %)	0 (0.0 %)
Neither	345	106 (30.7 %)	70 (20.3 %)	169 (49.0 %)
Wearable OEM cluster (subset of the first row)	17	1 (5.9 %)	0 (0.0 %)	16 (94.1 %)

Table 8. China origin and wearable category (Derived). The two attributes are fully confounded in this dataset.

Among apps with China origin outside the wearable category, 6 of 28 replied (21.4 %), 11 (39.3 %) sent bot replies, and 11 (39.3 %) were silent or unclear. In the wearable category 18 of 20 China-origin apps were silent and none referred or answered by bot. The wearable operator cluster (17 apps) has one reply and 16 tier C cases. The silence of the wearables cannot be attributed to the origin or to the category, because no wearable vendor of another origin exists in the set (Hypothesis: both factors, or the shared supplier structure of Section 6.6, contribute).

Time of the first report. Apps first reported in June 2026 (261 in tiers A, B, C) show A 27.6 %, B 26.1 % and C 46.4 %, and apps first reported in July or August (130) show A 33.1 %, B 10.0 % and C 56.9 %; only 3 apps were reported in September or October. The share of apps with at least one bounced vendor address is 27.2 % (86 of 316) for first reports before 1 August and 5.1 % (4 of 78) afterwards. These shifts could reflect vendor behaviour, a change in how we researched addresses or sent mail, or the ordering of the waves (Hypothesis, not separable). Only 2 apps were first reported on or after 11 September 2026, so no comparison before and after the start of the Article 14 obligations is possible (NOT FOUND as an effect).

6.6 Shared suppliers (RQ3)

Indicator	Value	Evidence
Operator groups with two or more apps	35 of 283 groups, covering 150 of 398 apps	Derived
Largest operator groups (apps)	26 and 17	Derived
Operator groups with two or more own tracker rows and differing tiers	15 of 31	Derived
Apps in 4 bundle rows (one report and one response per operator)	19	Observed
Wearable OEM cluster: apps / operator groups	17 / 14	Derived (shared package namespaces, OEM hosts, operator)
Wearable OEM cluster: sum of install lower bounds (16 apps)	364,000,000	Observed (report text), sum is not a person count
Wearable OEM cluster: findings / reports	176 / 17	Observed
Wearable OEM cluster: apps with cleartext traffic / third-country transfer finding	16 / 13 of 17	Derived
Reports mentioning at least one of six advertising-mediation SDK families	24 of 268	Derived (finding text)
Reports mentioning three or more of the six	7	Derived
Pairwise lift within the mediation families (15 pairs): min / median / max	10.2 / 18.9 / 38.3	Derived
Reports mentioning at least two of four Chinese cloud and push SDK families	13	Derived
Pairwise lift within the Chinese families (6 pairs): min / median / max	4.4 / 7.6 / 10.3	Derived
Pairwise lift between the two groups (24 pairs): min / median / max	0.0 / 4.0 / 8.5	Derived
Distinct fully written Google API keys appearing in more than one report	2 of 215	Observed, earlier analysis of report text before cleaning (keys are masked in the cleaned dataset)

Table 9. Shared operators, supplier clusters and SDK co-occurrence (see the evidence column; SDK families are counted in the finding texts of 268 reports).

Operator groups. The 398 apps belong to 283 operator groups. Of these, 35 groups (12.4 %) hold two or more apps and together 150 apps (37.7 %). A single vulnerability report to such an operator reaches several brands, and the reaction of the operator is counted once per app in the per-app figures. Among the 31 groups with at least two apps that have their own tracker rows, 15 (48.4 %) show different tiers across their apps, so the reaction is not a stable property of the operator in this dataset (Derived; routing of individual mails may differ).

Wearable supplier cluster. Seventeen wearable apps with 14 operator groups share package namespaces, hosts of original equipment manufacturers (OEM) and, in part, an operator, which links them to a common supplier structure [Derived from report evidence]. The 17 apps have 176 findings in 17 reports; 16 carry a cleartext-traffic finding, 13 a third-country transfer finding and 11 a hard-coded secret finding. The ledger exposure is CRITICAL for 16 of them and HIGH for 1, and the highest finding severity is HIGH for 16 and MEDIUM for 1. The install lower bounds stated in the report texts add up to 364,000,000 for 16 of the 17 apps; this is a sum of lower bounds over apps and no number of persons. One reply and 16 tier C outcomes (15 silence, 1 unclear) were recorded for the cluster. A weakness in a component of this supplier would affect 17 brands at once, while each brand holds a separate contact address.

Advertising and cloud SDK families. In the finding texts, 24 of 268 reports mention at least one of six advertising-mediation families, and 7 mention three or more. Within these six families the pairwise lift of co-occurrence (observed joint frequency divided by the frequency expected under independence) is 10.2 to 38.3 (median 18.9, 15 pairs). Four Chinese cloud and push families are mentioned together in 13 reports (lift 4.4 to 10.3, median 7.6, 6 pairs). The lift between the two groups reaches a median of 4.0 (24 pairs, range 0.0 to 8.5). Co-occurrence above 1 across both groups indicates a general factor, probably the length and SDK richness of a report, and the cleaned finding texts do not separate two bundles from this factor (Section 7). The apps in reports with three or more mediation families (12 apps with tracker row) show 1 reply, 8 bot replies and 3 silent cases. Findings of the same SDK type in several apps are common, and we did not measure whether the same version of a library is involved.

Shared secrets and backends. In an earlier analysis of the full report text, before cleaning and before masking, 2 of 215 distinct fully written Google API keys appeared in more than one report [Observed] (the link between the apps behind these two keys was not examined in the cleaned data). Four finding texts speak of a shared backend or project, and ten mention white-label or OEM structures. Shared keys are rare in this set, and the strongest evidence for shared suppliers lies in the namespace and operator structure.

6.7 The platform route

In two cases the report went to Google Play and Android Security, the platform route of Section 4.5. Table 10 gives the cases. Both disclosures are complete and the public report is available at rfi-irfos.com/reports/spinwinera-2026.pdf, so we name them.

Case	Platform	Reported	Removal confirmed	Days	Ground of the report
Merge Chicken (A163)	Google Play, Android Security	2026-06-25	2026-06-30	5	Casino with real-money play inside an app rated PEGI 3 (Pan European Game Information, age 3); casino content loaded at run time through Firebase Remote Config from spinwinera.com
Spinwinera, Roobet and BetOnRed network (A164)	Google Play, Android Security	2026-07-01	2026-07-07	6	Same casino brand under six listings of two British developer accounts; one listing carries emulator and sandbox detection

Table 10. Cases reported to a platform (Observed: report dates and the platform's confirmation of removal; Derived: days).

Merge Chicken was the first removal that followed one of our reports. After the report the platform asked us for evidence, we supplied the full analysis, and the platform then confirmed the removal to us in writing [Observed, correspondence with the platform]. The removal therefore followed a review of our evidence by the platform. For the network, two confirmations arrived 16 minutes apart, for two of the three reports filed. The confirmations state that the apps were removed from the store and name no policy ground, so the specific ground of removal remains a Hypothesis; the reports concerned deception about the nature of the app, which we read as a breach of the store policy. Three listings of one of the two accounts had already been removed in April 2026 without a report from us [Reported, from our case file], and those three are outside Table 10. With two cases we give no rate and no median.

The second case differs methodologically from the first. The follow-up reports did not need a new binary analysis of every package. The link between the actors (shared developer account, the casino brand as the app title, the same construction of a game facade in front of a web browser view) was sufficient for the platform to act (the platform's policy ground is a Hypothesis, as above). A full reverse engineering remains the reference for an unknown actor, and for a known repeat actor the linking of account, brand and domain sufficed in this case.

The operator's answer to the removal limits the route. One day after the confirmed removal, on 8 July 2026, a further domain variant with identical branding was found live. The 9 variants known to us were registered at five registrars and all resolved through one content delivery network, which hides the origin hosting [Observed, WHOIS and name server lookups of 8 July 2026]. We read this spread as domain rotation (Hypothesis). Reports to the five registrars were sent on 8 July 2026 [Reported]; we have no confirmed outcome for them, so the outcome of the domain route is unknown. The removal from the store ended the distribution through the store only.

7 Validation and Counter-Hypotheses

7.1 Second line of evidence: data-driven clustering

The first line of evidence is the taxonomy mapping of Section 6.2. The second does not use our class rules and works on the finding texts only. Both lines rest on texts from the same pipeline and report template, so they are not independent in a strict sense. The silhouette value is flat across the tested range (0.12 to 0.17) and highest at the upper edge of the scan (k=20, silhouette 0.170), so the clusters are not sharply separated and the choice of k is a convention. The clusters differ clearly from the OWASP labels in one direction: for 420 findings with a single OWASP 2024 label, cluster purity is 0.79 against a majority baseline of 0.61 (NMI 0.31, ARI 0.08). The M6 label (257 findings) is spread over 18 of 20 clusters, 7 clusters hold at least 5 % of it each, and the three largest hold 44.4 %. The same clustering reproduces our finding classes more closely (NMI 0.56 for the 1,181 classified findings), although the classes derive from title rules and the comparison is partly circular. NMI with the app is 0.30 and with the severity 0.07, so clusters follow apps to a similar degree as OWASP labels and follow severity hardly at all.

Cluster theme (from top terms)	**Findings**	**Apps**	**Privacy axis filled**	**Direct or conditional CWE**	**No CWE**	**Generic CWE only**	**No class**
Advertising, attribution, consent	121	105	89 %	2 %	16 %	72 %	11 %
Analytics start before consent screen	90	82	78 %	2 %	49 %	27 %	22 %
Biometric data	79	64	76 %	4 %	52 %	22 %	23 %
Third-country transfer (PRC)	109	93	72 %	9 %	52 %	15 %	24 %
Age rating and age gate	60	58	72 %	0 %	63 %	8 %	28 %
Background location	68	67	72 %	1 %	1 %	69 %	28 %
Health data	80	56	65 %	5 %	51 %	12 %	31 %
Play data-safety label and privacy policy	64	51	64 %	0 %	52 %	12 %	36 %
EU representative and contact	67	58	63 %	1 %	57 %	4 %	37 %
Behavioural marketing	129	93	53 %	5 %	9 %	44 %	42 %
Notification, SMS, contact access	54	39	48 %	9 %	0 %	44 %	46 %
Microphone, camera, voice	86	68	37 %	5 %	30 %	6 %	59 %
Session replay	69	60	29 %	3 %	22 %	29 %	46 %
Backup flag	40	38	25 %	57 %	3 %	15 %	25 %
German-language	56	17	25 %	9 %	5 %	20 %	66 %

Cluster theme (from top terms)	Findings	Apps	Privacy axis filled	Direct or conditional CWE	No CWE	Generic CWE only	No class
findings							
Exported components, manifest	134	99	19 %	31 %	5 %	10 %	54 %
Cleartext traffic	78	76	14 %	96 %	0 %	0 %	4 %
Hardcoded API keys, Firebase project	194	189	6 %	86 %	0 %	1 %	13 %
Certificate pinning	75	70	5 %	89 %	0 %	0 %	11 %
Production versus test and debug builds	82	57	4 %	34 %	0 %	4 %	62 %

Table 11. Twenty clusters of finding texts, sorted by privacy-axis share (Derived). Theme names come from the highest-weighted terms of each cluster. Columns four to eight give shares of the findings in the cluster; the privacy axis is filled for a finding when it belongs to a class with a privacy-axis entry.

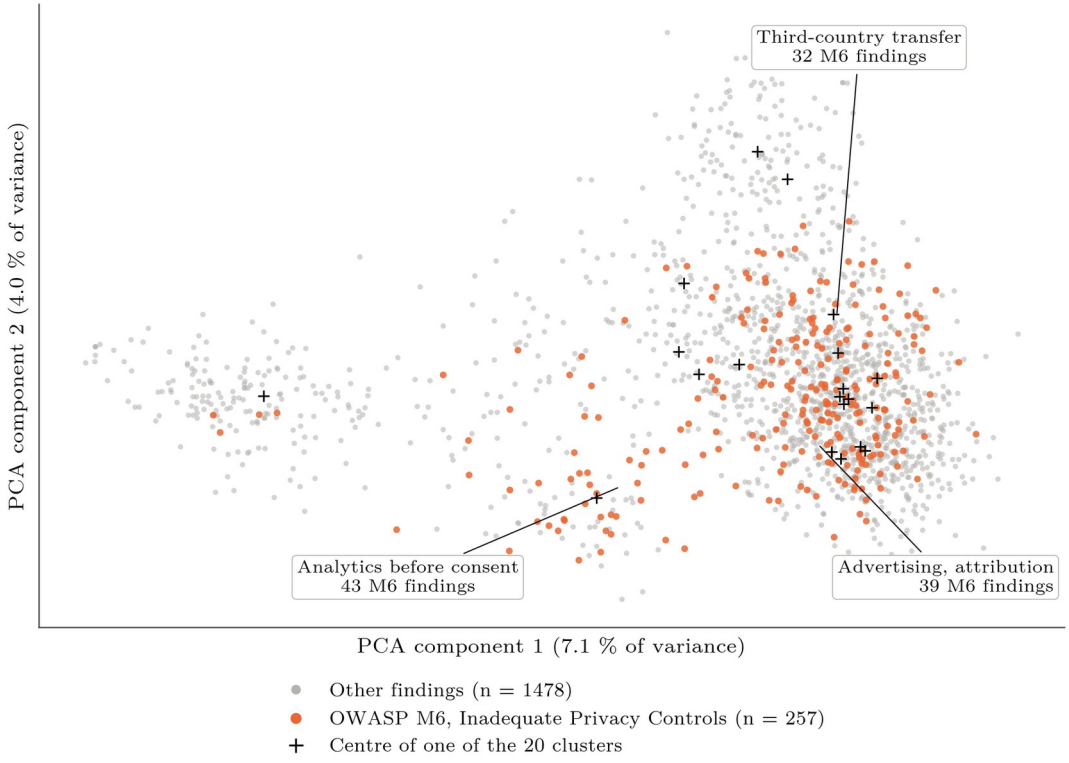

Figure 9. *Two-dimensional principal component projection of the 1,735 finding embeddings (all-MiniLM-L6-v2). Findings labelled with OWASP category M6 (Inadequate Privacy Controls) are highlighted, and the three clusters with most M6 findings are named; crosses mark cluster centres. The M6 findings spread over 18 of the 20 clusters.*

Ten of 20 clusters have a privacy-axis entry in at least half of their findings, and 6 clusters have a majority of findings in classes without CWE. In 11 clusters the majority of findings fall into classes without CWE or with a generic CWE only. Only 4 clusters have a majority of findings with a direct or conditional CWE, and these are the technical themes: hard-coded keys, cleartext traffic, certificate pinning and the backup flag. The data-driven themes without CWE include the consistency of the store data-safety label with the privacy policy, health data, biometric data, analytics start before a consent screen, third-country transfer, age rating and the missing EU representative. The 554 findings without a class spread over all 20 clusters, with the largest numbers in the cluster of exported components and manifest findings (72) and in the clusters of behavioural marketing (54), production and test builds (51) and microphone and camera access (51); one cluster consists of German-language findings and reflects the language of the embedding model (Hypothesis). An earlier run on the uncleaned 1,785 findings gave a similar picture (19 clusters, purity 0.74 against 0.60).

Both lines point in one direction. A large part of the app privacy findings has no counterpart in CWE and sits outside the OWASP field, and privacy themes recur as groups in the clusters, although silhouette values of 0.12 to 0.17 show a weak cluster structure. The result stays on the level of this selection; a different class rule set or a different embedding could shift the cluster counts.

7.2 Counter-hypotheses

Pattern	**Counter-hypothesis**	**Test**	**State**
78.5 % of apps carry findings in classes without CWE	The class rules and the CWE assignment are our own and shape the result	Second path: clustering of finding texts without our class rules	Both paths agree; the cluster path has a language artefact (one German-language cluster) and flat silhouette values
Reply share differs by category	Large operator families drive single categories	Operator groups weighted once; exclusion of the two largest groups	Public bodies reverse (18.5 % to 62.5 %, 8 groups); games stay low (2.9 % per app, 7.7 % per group); overall figures stay stable (A 31.1 % without the two groups)
No consistent gradient of reply with severity	Category and wave mix hide a gradient	Comparison within 8 categories; logistic model with clustered standard errors	Direction mixed (5 higher, 2 lower, 1 equal); pooled gradient inverted; no association in the model (odds ratio 0.55)
EU-side vendors reply more often	Origin is confounded with category	Comparison within categories with both groups; logistic model with category effects	Same direction in 5 categories, close in 1; association remains with category effects (odds ratio 2.31); unobserved vendor properties not separable
Wearable vendors stay silent (18 of 20 China-origin wearables)	Origin or category or shared supplier causes it	Cross table of origin and category	No wearable vendor of another origin; not separable
Two separate SDK bundles (advertis-	Report length and SDK richness raise all co-oc-	Lift between and within groups	Lift above 1 everywhere; the two bundles are not separated

Pattern	Counter-hypothesis	Test	State
ing mediation, Chinese services)	currences		
Exposure and finding severity diverge	Later report format with a different severity convention for wearables	Divergence by category	Concentrated in wearables (19 of 21) and public bodies (19 of 27 reversed); cause not tested
Bounce share falls after 1 August (27.2 % to 5.1 %)	Change in address research or mail handling, not vendor behaviour	None available	Not separable
Redirection category mixes different replies	Rule placed substantive replies in tier B	Individual review of all 34 targets against tracker notes	13 moved to A, 8 form B1, 12 form B2, 1 moved to C

Table 12. Central patterns, counter-hypotheses and the state of the test.

7.3 Candidates that did not hold

Four planned points did not receive support from the cleaned data. First, a separation of SDK families into an advertising bundle and a Chinese-services bundle is not supported by finding-text co-occurrence, because the lifts between the groups are above 1 as well (Section 6.6). Second, a gradient of vendor reaction with severity is absent; the pooled gradient is inverted, the within-category direction is mixed, and the logistic model shows no association (Section 6.5). Third, an effect of the CRA reporting obligations on the reaction cannot be tested because only 2 apps were first reported after 11 September 2026. Fourth, the reading of Table 6 that public bodies stay silent does not survive the counter-check by operator group (Section 6.4). Time to first reply, fix and retest cannot be analysed because the dataset has no structured field for it.

8 Discussion

The first result concerns the relation of the two measurements. Technical and privacy findings occur together in 275 of 340 apps, and 47 apps carry privacy findings only. A measurement limited to the security axis therefore misses the privacy side of these apps (Section 6.2), and that side is exactly where the mapping to taxonomies is thin: six of the 17 classes map to no CWE at all, while the OWASP field of the reports is filled for a quarter of the findings only. We read this as a property of CWE, which catalogues software weaknesses and was not built for duties of data protection law, whereas MASVS-PRIVACY, MASWE and LINDDUN address privacy. The same pattern appears in the mapping and in the clusters, which differ in construction. Whether a different class definition would reduce the gap remains open.

The second result concerns the vendor side. Half of the 394 apps ended in tier C, about one third with silence and the rest unclear or undeliverable, and 29.2 % received a responsible human reply. Only 8 apps (2.0 %) referred us to an official reporting channel. The small number shows how rarely the vendors in this set named such a channel in reply to a direct report. A bot reply (15.5 %) shows that a mailbox exists and that a process answers automatically, and a reply from a wrong desk or a bare identity check (3.0 %) shows that someone read the mail and the report did not reach a responsible person. The 61 apps with a bot reply only are consistent with a reporting route limited to automated tools, which Article

13(17) excludes from December 2027; whether a human channel existed beside the bot was not tested, so the share is an upper bound for that state and no finding of non-compliance.

The third result concerns the unit of analysis. Reaction differs by category and by operator family, and the operator families are large: 35 groups hold 150 apps, and the two largest hold 43. A statement on a sector needs the operator as the unit, and the group-weighted figures differ from the per-app figures by up to 44 points (public bodies, Section 6.4). The association with EU or EEA origin holds with standard errors clustered by operator group. Across the operator groups with several apps, the reaction differs in 15 of 31 groups, which suggests that mail routing or the responsible team within the same operator matters. The data cannot separate the two.

The fourth result concerns severity. The divergence between the two scales follows the app category (Section 6.3), and we see no relation between severity and reply on either scale. Because most apps carry a CRITICAL finding, the data cannot show whether vendors triage by severity (Section 6.5). A vendor may not see the severity assigned by the reporter, or may assess it with different information, and these are possible explanations that we did not test (Hypothesis).

The platform route follows a different logic from the vendor route. A platform can act when a vendor cannot be reached or has no interest in the report, and in the two cases of Section 6.7 the listing operator was the source of the harm. In both cases the platform confirmed removal within 6 days. The dataset records no time to first reply for vendors (Section 6.4), so no comparison of speed is possible, and two cases selected for the severity of the deception allow no rate and no test. The Digital Services Act (DSA, Regulation (EU) 2022/2065) sets duties for online intermediaries, among them procedures for reports of illegal content; whether and how it applies to a report about a deceptive app is a legal question that this paper does not answer, and we cite no individual articles. The route has a limit that the Spinwinera case showed. A removal ends the listing, and the operator keeps its domains, so that enforcement by the platform removes one channel of distribution and leaves the actor in place (Observed for the further domain variant, Hypothesis for the operator's intent). Domain registrars and the content delivery network are further platforms of the same kind, and their outcomes are not known to us.

The pipeline described in Section 4 produces findings faster than vendors answer. Whether faster discovery raises the share of tier C is not tested here, because the dataset has one measurement time and no change in discovery rate. The argument of the introduction, that the bottleneck sits on the vendor side, rests on the combination of the discovery literature and the tier distribution of Table 5; it does not rest on a causal test. A test would need the volume of reports per vendor over time together with the time to first reply, so that a rise in volume can be related to a change in the reaction; a repetition of this baseline after December 2027 with structured timing fields can supply both.

9 Limitations

Selection. All shares apply to this salience-based set (Section 5.1) and say nothing about the market. Operator families bias the per-app figures (Section 6.4), and we cannot quantify the selection effect further. Apps with medical-device status, which the CRA excludes, are not marked in the health category.

Method scope. Findings describe exposure visible in the Android package; it neither confirms exploitation nor tests whether a secret is valid. Findings may include false positives that human review did not remove, and automation bias in the human decision cannot be excluded.

Coding. Class and tier coding follow rules without a second coder, and no agreement measure exists; the confidence per row is recorded (121 medium, 1 low among 394 tier codes), and all redirections were reviewed individually (Section 5.2). The precision of the findings was not measured on an independent sample. Category and origin come from keyword rules and partly from general knowledge. The 554 findings without class and the 28 apps in "Other, unclear" reduce the resolution. The class rules work on titles only.

Measurement of reaction. The dataset holds no structured timing for reply, fix or retest. The reaction is coded from the tracker state of the local files and the mailbox check dates per row, up to 21 November 2026; a later reply is not included. The reporting path that a vendor publishes (CVD policy, contact address, security.txt) lies outside this dataset (Section 3). The reaction measures the response to reports from one sender, in one language and with authority addresses in blind copy (Section 4.5); another sender or channel may meet another reaction.

Status of the data. All numbers come from local files of 22 November 2026 and were not compared live with the website, the mailbox or the stores. Two apps have neither a tracker row nor a platform report, two apps follow the platform route (Section 6.7), and 55 apps with a tracker row have no structured report file, so their findings are unknown. The earlier internal reports contain other numbers because of the cleaning rules of Section 5.2; the figures of this paper replace them.

10 Conclusion

Primary RQ. Among the 398 Android apps of this set, the 340 apps with a report carry 1,735 documented findings, and for most of them (78.5 %) at least one finding falls in a class outside the CWE catalogue. On the vendor side, 29.2 % of the 394 apps received a responsible human reply to our reports, 2.0 % a referral to an official reporting channel and 18.5 % a reply without substance; 36.3 % stayed silent, and 14.0 % remain unclear or undeliverable. The two cases that went to the app store instead of the vendor ended in removal after 5 and 6 days (Section 6.7).

RQ1. Hard-coded secrets (68.8 % of apps), advertising or tracking SDKs (62.9 %), third-country transfers (38.2 %) and pre-consent initialisation (32.9 %) are the most frequent classes. Four of 17 classes map directly to a CWE and six map to none; the clusters of finding texts agree with this picture.

RQ2. Reactions differ by category and origin, and severity shows no association: replies span 2.9 % for games to 66.7 % for productivity and security tools, vendors from the EU and EEA reply more often also when category effects are held constant (odds ratio 2.31), and neither severity scale shows a consistent relation. Wearable apps and China origin remain fully confounded. The effect of the CRA reporting obligations cannot be tested.

RQ3. A shared supplier structure links 17 wearable apps from 14 operators, and 35 operator groups hold 150 apps. Shared secrets appear in 2 of 215 keys analysed. SDK co-occurrence in the finding texts does not separate bundles.

The baseline can be repeated after 11 December 2027 with the same classes, the same tiers and structured fields for time to reply and for the existence of a CVD policy and a contact address.

References

References are given in short form. The text of Regulation (EU) 2024/2847 was checked against the version in the Official Journal on 22 November 2026. The resolution of the European Parliament was checked against the adopted text on the same day.

- Alahmadi, B. A., Axon, L., Martinovic, I. (2022). 99% False Positives: A Qualitative Study of SOC Analysts' Perspectives on Security Alarms. USENIX Security 2022, 2783-2800.
- Bansal, G. et al. (2021). Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. CHI 2021. https://doi.org/10.1145/3411764.3445717
- Binns, R., Lyngs, U., Van Kleek, M., Zhao, J., Libert, T., Shadbolt, N. (2018). Third Party Tracking in the Mobile Ecosystem. WebSci 2018. https://doi.org/10.1145/3201064.3201089
- DARPA (2025). AIxCC results. https://darpa.mil/news/2025/aixcc-results
- Ding, Y. et al. (2025). Vulnerability Detection with Code Language Models: How Far Are We? ICSE 2025. https://doi.org/10.1109/ICSE55347.2025.00038
- ENISA (2022). Coordinated Vulnerability Disclosure Policies in the EU. 13 April 2022.
- European Parliament (2017). Resolution of 16 February 2017 with recommendations to the Commission on civil law rules on robotics (2015/2103(INL)), P8_TA(2017)0051.
- Findlay, W., Abdou, A. (2022). Characterizing the Adoption of Security.txt Files and their Applications to Vulnerability Notification. Workshop on Measurements, Attacks, and Defenses for the Web (MADWeb) 2022. https://doi.org/10.14722/madweb.2022.23014
- FIRST (2026a). Vulnerability forecast 2026: The Year Ahead, 11 February 2026. https://www.first.org/blog/20260211-vulnerability-forecast-2026
- FIRST (2026b). The 2026 Vulnerability Forecast Update: Navigating the AI Epoch, 15 June 2026. https://www.first.org/blog/20260615-vulnerability-forecast-update
- Google (2025). A summer of security: empowering cyber defenders with AI. https://blog.google/technology/safety-security/cybersecurity-updates-summer-2025/
- Hugging Face (2026a). Security incident disclosure, July 2026. 16 July 2026. https://huggingface.co/blog/security-incident-july-2026
- Hugging Face (2026b). Anatomy of a Frontier Lab Agent Intrusion. 27 July 2026. https://huggingface.co/blog/agent-intrusion-technical-timeline
- Kollnig, K., Binns, R., Dewitte, P., Van Kleek, M., Wang, G., Omeiza, D., Webb, H., Shadbolt, N. (2021). A Fait Accompli? An Empirical Study into the Absence of Consent to Third-Party Tracking in Android Apps. SOUPS 2021. arXiv 2106.09407.

- Lee, J. D., See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. Human Factors 46(1). https://doi.org/10.1518/hfes.46.1.50.30392
- Li, F., Durumeric, Z., Czyz, J., Karami, M., Bailey, M., McCoy, D., Savage, S., Paxson, V. (2016). You've Got Vulnerability: Exploring Effective Vulnerability Notifications. USENIX Security 2016, 1033-1050.
- Li, K., Ling, L., Yang, J., Wei, L. (2024). Automatically Detecting Checked-In Secrets in Android Apps: How Far Are We? arXiv 2412.10922.
- Li, Z., Dutta, S., Naik, M. (2025). IRIS: LLM-Assisted Static Analysis for Detecting Security Vulnerabilities. ICLR 2025. arXiv 2405.17238.
- METR (Greenblatt, R., Cotra, A., Wijk, H.) (2026). Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident. 26 August 2026. https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/
- Neef, S., Schlunke, C., Hennig, A. (2025). Do (Not) Tell Me About My Insecurities: Assessing the Status Quo of Coordinated Vulnerability Disclosure in Germany Amid New EU Cybersecurity Regulations. European Symposium on Usable Security (EuroUSEC) 2025, 100-113. https://doi.org/10.1109/EuroUSEC69254.2025.00020 (preprint arXiv 2606.25950)
- Nguyen, T. T., Backes, M., Marnau, N., Stock, B. (2021). Share First, Ask Later (or Never?) Studying Violations of GDPR's Explicit Consent in Android Apps. USENIX Security 2021.
- NVD (2025). CVE-2025-6965. https://nvd.nist.gov/vuln/detail/CVE-2025-6965
- OpenAI (2026a). OpenAI and Hugging Face partner to address security incident during model evaluation. 21 July 2026, updated 28 July 2026. https://openai.com/index/hugging-face-model-evaluation-security-incident/
- OpenAI (2026b). The Hugging Face incident and the road ahead. 26 August 2026. https://openai.com/index/hugging-face-incident-and-the-road-ahead/
- Parasuraman, R., Manzey, D. H. (2010). Complacency and Bias in Human Use of Automation: An Attentional Integration. Human Factors 52(3), 381-410. https://doi.org/10.1177/0018720810376055
- PHP Group (2026). Security advisories GHSA-w8hw-w9fx-j79w (CVE-2026-109618) and GHSA-6wmf-4rv5-p3rp (CVE-2026-109619); PHP 8.2.35, 8.3.36, 8.4.27 and 8.5.12 released 21 October 2026. https://github.com/php/php-src/security/advisories
- Poteat, T., Li, F. (2021). Who You Gonna Call? An Empirical Evaluation of Website security.txt Deployment. ACM Internet Measurement Conference (IMC) 2021, 526-532. https://doi.org/10.1145/3487552.3487841
- Regulation (EU) 2024/1689 (Artificial Intelligence Act). https://eur-lex.europa.eu/eli/reg/2024/1689/oj
- Regulation (EU) 2024/2847 (Cyber Resilience Act). Official Journal of 20 November 2024. https://eur-lex.europa.eu/eli/reg/2024/2847/oj
- Reyes, I. et al. (2018). "Won't Somebody Think of the Children?" Examining COPPA Compliance at Scale. PoPETs 2018(3). https://doi.org/10.1515/popets-2018-0021

- Schmidt, D., Schrittwieser, S., Weippl, E. (2025). Leaky Apps: Large-scale Analysis of Secrets Distributed in Android and iOS Apps. ACM Conference on Computer and Communications Security (CCS) 2025, 2459-2473. https://doi.org/10.1145/3719027.3765033
- Seeber, I. et al. (2020). Machines as teammates: A research agenda on AI in team collaboration. Information and Management 57(2). https://doi.org/10.1016/j.im.2019.103174
- Stock, B., Pellegrino, G., Rossow, C., Johns, M., Backes, M. (2016). Hey, You Have a Problem: On the Feasibility of Large-Scale Web Vulnerability Notification. USENIX Security 2016.
- Sundaramurthy, S. C. et al. (2015). A Human Capital Model for Mitigating Security Analyst Burnout. SOUPS 2015.
- TechCrunch (2025). Google says its AI-based bug hunter found 20 security vulnerabilities, 4 August 2025. https://techcrunch.com/2025/08/04/google-says-its-ai-based-bug-hunter-found-20-security-vulnerabilities/
- Ullah, S. et al. (2024). LLMs Cannot Reliably Identify and Reason About Security Vulnerabilities (Yet?). IEEE S&P 2024. https://doi.org/10.1109/SP54263.2024.00210
- Unit 42, Palo Alto Networks (2026a). Frontier AI vulnerability burst (Zou, X.). 4 August 2026. Vendor report. https://unit42.paloaltonetworks.com/frontier-ai-vulnerability-burst/
- Unit 42, Palo Alto Networks (2026b). Threat brief on CVE-2026-109618 and CVE-2026-109619. 29 October 2026, updated 19 November 2026. https://unit42.paloaltonetworks.com/cve-2026-109618-cve-2026-109619/
- Wei, Huang, Cheung, Li (2025). How Far are App Secrets from Being Stolen? A Case Study on Android. arXiv 2501.07805.
- Wiz Research (2026). PHP2Shell: remote code execution in PHP. 21 October 2026. https://wiz.io/blog/php2shell-php-rce
- XBOW (2025). Top 1: how XBOW did it, 24 June 2025. https://xbow.com/blog/top-1-how-xbow-did-it
- Yang, H., Huang, R., Fang, Y., Zhang, B., Guo, J., Wu, Z., Mi, X. (2026). Okara: Detection and Attribution of TLS Man-in-the-Middle Vulnerabilities in Android Apps with Foundation Models. Australasian Conference on Information Security and Privacy (ACISP) 2026. arXiv 2601.22770.
- Yasumatsu, T., Kanei, F., Watanabe, T., Shioji, E., Akiyama, M., Mori, T. (2017). Mobairu apuri kaihatsusha ni yoru zeijakusei taio no jittai chosa (in Japanese; investigation of how mobile app developers respond to vulnerabilities). Computer Security Symposium 2017, Information Processing Society of Japan, 23 to 25 October 2017.
- Zhang, C. et al. (2026). SoK: DARPA's AI Cyber Challenge (AIxCC): Competition Design, Architectures, and Lessons Learned. arXiv 2602.07666.
- Zhou, X., Cao, S., Sun, X., Lo, D. (2025). Large Language Model for Vulnerability Detection and Repair: Literature Review and the Road Ahead. ACM TOSEM 34(5). https://doi.org/10.1145/3708522
- Zuo, C., Lin, Z., Zhang, Y. (2019). Why Does Your Data Leak? Uncovering the Data Leakage in Cloud from Mobile Apps. IEEE S&P 2019.

Supplementary material

The cleaned dataset in aggregated form without vendor and app names, the class rules, the reaction coding rules with the reasons of the individual review, the number scripts `paper_numbers.py` and `stats_extra.py` and their output are available from RFI-IRFOS on request, and a permanent deposit with its own DOI will be linked from the preprint page.