Maskbreak Security Whitepaper

Sentinel Edge Networks LTD · Registered in England & Wales 17150600 · 134a West Hendon Broadway, London NW9 7AA · Version 2.15 · 5 November 2026 · canonical copy at maskbreak.com/security-whitepaper

Maskbreak is a real-time fraud detection API (maskbreak.com). This document states our security posture plainly for procurement and security review: the controls in place today, how data is handled, and — deliberately — what we do not yet have. It is written to be checked, not to impress: it describes what runs today, we can show evidence for any row on request, and we do not list controls we don't have.

1. Certifications & audits — the honest version

> Maskbreak is **pre-audit for SOC 2 Type II** and holds no SOC 2, ISO 27001, HIPAA, or PCI DSS attestation today. No third-party penetration test has been commissioned yet. We say this here rather than imply otherwise with badge walls; this section will be updated as each milestone lands, with the change dated on our changelog. A responsible-disclosure program is active (/responsible-disclosure, RFC 9116 `security.txt` published).

2. Technical controls in place today

Control	Implementation
Authentication	Email + bcrypt-hashed password (new passwords longer than bcrypt's 72-byte limit are refused rather than silently truncated), Google OAuth (RS256, JWKS verified), GitHub OAuth (since 5 November 2026 both redirect sign-ins bind the authorization code to the browser that started them with PKCE (S256), and Google's also with an OIDC nonce), TOTP 2FA with recovery codes, passkeys (WebAuthn), reviewable active-session list with per-device revocation. Sign-in is locked after 10 failed passwords in 30 minutes per account, and password re-checks before sensitive changes count toward the same lock; wrong two-factor codes have a per-account daily limit, and the account holder is emailed when a correct password is followed by wrong two-factor codes. Since 5 November 2026 the authenticator codes entered to change the second factor (turning 2FA off, replacing the authenticator, making new recovery codes, adding a passkey to an account with 2FA) have per-account limits of their own, 10 tries in 15 minutes and 30 wrong codes a day, kept apart from sign-in's so that they never lock the account holder out of signing in, and a wrong code for the authenticator in use emails the account holder (at most once a day); an organization's invitation of a member, or making a member an admin, asks for the password (or a recent Google or GitHub sign-in) first. Only genuine console-session tokens authenticate the console; other signed tokens we issue (OAuth state, admin and API access tokens) are refused there. Since 14 October 2026 the console session is held in an HttpOnly, Secure, SameSite=Strict cookie, which page scripts cannot read; the server accepts it only on requests carrying the console's own header, so other sites cannot use it. A browser that refuses cookies keeps the session token in page storage, as before.
API key management	Keys generated with `crypto.randomBytes` (prefix `sk_live_`); rotation and instant revocation via dashboard; key rotation and password change require re-authentication.
Password breach check	Have I Been Pwned k-anonymity (5-character SHA-1 prefix only) on signup, password reset and password change — the password itself never leaves our infrastructure.

Control	Implementation
Transport security	TLS 1.2+; HSTS for one year with includeSubDomains and the preload directive (the domain has not yet been submitted to browsers' preload lists); CORP same-origin, except where cross-origin use is the point (the public API, our embeddable scripts, brand images and machine-readable documentation); CSP with allowlists and live violation reporting on every page, and a no-script policy on HTML served under `/api/`; a strict first-party policy without inline script is enforced on the internal admin console, and a hash-based policy without `'unsafe-inline'` has been enforced on the customer console since 6 October 2026; X-Content-Type-Options nosniff; X-Frame-Options DENY (SAMEORIGIN for our own network-check frame and the live check shown without scripts); Permissions-Policy locks camera/mic/geolocation/payment.
Rate limiting	Per-account monthly allowances set by the plan (no account has an hourly cap since 29 October 2026; the live key and its predecessor share the allowance during a key rotation) and per-source-IP caps on the evaluate and lookup endpoints; fixed-window rate limits on authentication and form endpoints, with IPv6 addresses also counted per /64; per-account limits on signed-in console and API routes; a 30-minute account lockout after 10 failed password attempts; in-memory limiter state is bounded; every 429 response carries `Retry-After`. Since 5 November 2026 the fixed test tokens answer up to 1,000 times an hour per account, and the pretend visitors the console sends with them are stored up to 100 a day per account. The API-key IP allowlist, when configured, also applies to minting OAuth tokens (`/oauth/token`); `/v1/usage` stays readable from outside the list, as documented.

Control	Implementation
Admin access	The internal admin console signs in only through a single-sign-on provider with a phishing-resistant passkey (since 14 October 2026; Google sign-in and the admin key are switched off and come back only through a documented break-glass setting), sits behind an access gateway that admits only the owner (since 26 October 2026 the gateway no longer checks that the owner's device is enrolled), runs under an enforced Content-Security-Policy without inline script, and writes an audit line to the application log for every admin API request (operator type and email, route, target account id, status and operator IP); since 20 October 2026 the same record is also kept in an audit table for one year. Key rotations, two-factor removal, forced sign-outs and reinstatements made by an operator are emailed to the customer; admin sign-ins, key sign-ins and lockouts alert the owner.
Runtime	The server runs in a container as a non-root user on a pinned Node.js version, on Cloudflare Containers since 20 October 2026 (Railway before). The container has no public address: only our gateway Worker reaches it, and a new version takes over from the running one only once it answers its readiness check (a version that never gets ready leaves the old one serving). It refuses to start with known-weak secrets in production, and refuses to connect to the production database from outside the production hosting environment unless explicitly overridden. Outbound calls to providers have timeouts.
Session invalidation	Password change/reset or admin suspension immediately invalidates all sessions (token-epoch bump).
Database access	Since 19 October 2026 the primary database is Cloudflare D1, which the server reaches only through a Worker that requires a secret and runs one statement per request; the nightly backup job holds a separate secret that permits one read-only statement per request. That Worker keeps no request logs. Before, the database was Turso, deleted with our account the same night.

Control	Implementation
Encryption at rest	Managed encryption at rest on the primary database (Cloudflare D1, United States, Eastern North America, since 19 October 2026; Turso / libSQL, AWS us-east-1, before) and on the backups in Cloudflare R2 (EU jurisdiction, since 17 October 2026; the copies made up to 16 October 2026 in AWS S3, Stockholm, used S3 server-side encryption). Account IDs a customer sends are also encrypted in the database (AES-256-GCM, a key per customer derived from a key held outside the database), since 18 October 2026, and so are the GPU-evidence features of customers in that beta.
Error monitoring	Since 24 September 2026 the server reports unexpected errors to Sentry (EU region, Frankfurt, Germany) through `@sentry/node`: errors only — no performance tracing, profiling, log forwarding, metrics or session replay, and no Sentry code in the browser. Sentry's own data collection (user, cookies, headers, bodies, query strings, local variables) and trace propagation are switched off (Sentry SDK 11 since 21 October 2026, the same setup). Before sending, the server drops request headers, cookies, bodies, query strings, the user context and console breadcrumbs; replaces IP addresses, email addresses, UUIDs, purely numeric ids (account and record numbers), hexadecimal strings of 12+ characters and long token-like path segments with placeholders (`:ip`, `:email`, `:id`, `:token`, e.g. `/v1/lookup/:ip`); keeps only the method, status code and origin (scheme, host, port — no path or query) for up to 30 outgoing-call breadcrumbs, so a customer's webhook host name can appear; and scrubs API keys, webhook secrets, JWTs, bearer tokens, email addresses, IP addresses, database credentials and the exact values of our own secrets from error messages and the remaining event context (stack frames, which hold our own code, are sent as they are). This scrubbing is pattern-based: a strong safeguard, not a guarantee. In Sentry, IP-address storage is turned off, the server-side data scrubber is on as a second layer, and generative AI features (including Seer) are off. Without a configured DSN the SDK is not loaded at all. Retention follows Sentry's plan: 30 days on the free Developer plan; up to 90 days on paid plans and during a trial.

Control	Implementation
Customer script addresses	Since 20 October 2026 an account holder can serve our script from a subdomain of their own site (Cloudflare custom hostnames; Cloudflare issues the certificate for each name). On such a name only the script, its network-check redirect and its site report answer, with `noindex`; every other path is a 404 (the gateway answers `/robots.txt` there itself), and HSTS is not sent there, so the customer's own domain is never bound by our header. A name is refused when it is ours, an IP address, carries a port, or is a bare domain or its `www`. Since 5 November 2026 at most three names set up under one registrable domain are served, across accounts; and when an account adds a name whose DNS already leads to us, or whose DNS cannot be read, Cloudflare is not asked for it until a TXT record `_maskbreak.`<name> holds a value tied to that account (an HMAC of the account and the name under a key held outside the database), so a name set up for one account cannot be taken by another, and the account whose value is in the name's TXT record can take over a name another account holds. The server reads these records through Cloudflare's public DNS resolver (1.1.1.1, DNS over HTTPS), which receives only the names.
Uptime transparency	Self-hosted status page at /status with recorded probe samples (90 days) and incident history, plus independent scheduled checks from GitHub Actions (scheduled every 15 minutes; GitHub runs it every few hours in practice) and a Cloudflare schedule (every five minutes). Since 20 October 2026 that Cloudflare schedule emails the operator itself when two runs in a row fail, again if more checks fail, and on recovery; neither the application server nor the database is in its path, so an outage of either cannot silence it. Recorded samples are not a complete uptime measurement; see the service level objectives.

3. Data handling & retention

Data	Policy
Lookup IPs	Raw IPs retained 7 days, then replaced by a keyed one-way hash (per customer, key held outside the database) by a cleanup that runs every six hours (since 5 November 2026 a run that stops at its per-run limit is followed by another 15 minutes later). Records hashed before 28 September 2026 used an earlier, unkeyed hash; the same cleanup re-keys them with the secret key, and backup copies made before that keep the older form until they expire (60 days).
Application logs	Kept by our hosting provider (Cloudflare Containers, United States, since 20 October 2026) for 7 days, and since 22 October 2026 a copy for 30 days in Cloudflare R2 (EU jurisdiction), with keys, tokens and our own secrets removed from each line first; the logs Railway kept until then are deleted by Railway within 30 days. Lines identify accounts and records by number and API keys and tokens are not logged on purpose; some lines can still contain an email address or IP address (admin lines name the operator, a failed newsletter send names the recipient, and a request path or error message can contain one).

Data	Policy
Backups	A nightly GitHub Actions job (GitHub-hosted runner) reads the full database (through the database Worker, with its read-only secret), compresses it and uploads it to Cloudflare R2 in Cloudflare's EU jurisdiction, encrypted at rest, where a rule removes each copy after 40 days. The job holds no storage key: it writes through a Cloudflare Worker that admits only a GitHub-issued identity token for this repository's main branch, accepts only new copies and deletes nothing. Since 19 October 2026 each night's copy is also loaded into a temporary in-memory database on the runner, its row counts compared table by table and its integrity checked; a failure opens one issue for the operator. Since 22 October 2026 the job also encrypts the copy on the runner (a fresh AES-256-GCM key per copy, wrapped with RSA-4096 OAEP to a public key in our repository; the private key is held only by the operator, not by GitHub or Google) and uploads it to a shared drive in our Google Workspace, as a service account it reaches through Workload Identity Federation with its own GitHub-issued identity, so no Google key exists; copies there older than 25 days (30 days until 5 November 2026) are moved to the drive's trash, which Google empties 30 days later, so a copy is gone by day 56; this cleanup runs every night right after the job signs in to Google and before the upload, so a failed upload does not stop it. A failure of this second copy or of its cleanup opens one issue and never affects the R2 copy. Until 16 October 2026 the copies went to AWS S3 in Stockholm (eu-north-1), encrypted with server-side encryption; that account is suspended, so we can no longer check or delete those copies ourselves. See the retention-assurance note below.
Database history	Cloudflare D1 keeps a 30-day point-in-time history of the primary database (Time Travel). We would use it only to recover from a failure of the database itself; a rollback undoes the account deletions made after the point restored, so those are listed first (from our operations log or, if the database does not answer, the application logs) and carried out again before the service reopens.
Device sighting counter	A one-way hash of a device identifier with a visit count and first and last sighting, shared across customers (it answers `times_seen` and `first_seen`), with no site, account, IP address or user agent attached; deleted after 90 days without a sighting.

Data	Policy
Signup emails (API signal)	Checked transiently against disposable-domain feeds; never stored or logged.
Device and account linking	Per customer only, never linked across customers. Device links use one-way hashes. Since 18 October 2026 an `accountId` a customer sends is kept as a keyed hash for linking and, encrypted with that customer's own key, only so its console can show it; an account's network addresses are kept as keyed one-way hashes and never for shared exits (VPN, proxy, datacenter, Tor, Apple's iCloud Private Relay, mobile carriers' shared addresses). Since 5 November 2026 a login, second-factor check or password change the customer has not reported as successful is kept with the same hashes but marked as an attempt: it counts toward how many of the customer's accounts a device or an address tried, never as a device or an address the account used. Deleted 90 days after the account was last checked, and with the customer's account.
Email and phone check (console)	Since 18 October 2026: the one email address or phone number an account holder types is sent to an email- and phone-intelligence provider (United States), with a daily limit of checks per account. Neither the value nor the answer is stored; only the day's count of checks per account and, since 5 November 2026, of unanswered checks given back to it (kept 3 days).

Data	Policy
Login protection	Since 23 October 2026, on for every account (a customer can turn it off in Settings; nothing is sent until the customer's server reports a login or an account event with its own account ID, which is its instruction to use it): each login attempt the customer reports as a login is sent to Sift (United States) for an account-takeover score, with the account and session identifiers and our device key as keyed one-way hashes (a different key for every customer, derived from the key that pseudonymises stored IP addresses, so Sift never sees a customer's own identifiers and two customers' users never become one), whether the password was right, the end user's IP address and user-agent string, the email address when the customer sends one, the customer's website domain and our own verdict. One Sift account serves every customer that uses it, so its scores learn from all of their logins. Since the same day the customer can also report an account's signups, password changes, verifications, logouts and account-detail changes; each goes to Sift as that event with the same data and no score in return, so that Sift knows what is normal for the account. The stored evaluation record of a login keeps the score, the names of Sift's reasons (never a value), Sift's pseudonymous user ID for that account and the customer's own report of the login as the real user or a takeover (sent to Sift as a label on that pseudonymous user), with the record's normal retention; a high score also leaves the reason code `account_takeover_risk`. Our log line carries the account number and the outcome only. Each call waits at most 1.5 seconds and never fails the check. Since 5 November 2026 each account has a share of the calls: at most 10 scored logins a second and 80 open calls, and the Free accounts together at most 15 and 100, within the overall 25 and 200; a login over its share of scored calls is sent without asking for a score, and a login or account event over its share of open calls is not sent. Test keys, test tokens and console checks never reach Sift; without the server setting `SIFT_API_KEY` nothing does.

Data	Policy
Email and phone intelligence (API)	Since 23 October 2026, on for accounts created from that day and off for older accounts until they turn it on in Settings (any account can turn it off there; nothing is sent until an address or a number arrives with a check): the email address a customer sends with a visitor check, and the visitor's phone number sent as `phone`, go to the same email- and phone-intelligence provider (United States) as the console's check, which returns a risk score, its reason names and a few facts (whether the address accepts mail, when its domain was registered, on how many sites it was seen, in how many known data breaches it appears; a number's validity, line type and carrier). The stored evaluation record keeps the scores, the reason names and those facts, never the address or the number, which are not logged either; a score of 80 or more leaves the reason code `email_high_risk` or `phone_high_risk` and raises allow to review, never to block. Each address or number scored counts toward the plan's daily number of email and phone checks (shared with the console's check); one the provider was asked about and did not score is given back, since 5 November 2026 at most the plan's daily number of times a day, after which it uses one of the day's checks. Test keys, test tokens and console checks never reach the provider; without the server setting `SEON_API_KEY` nothing does.
Page check (console)	Since 20 October 2026: a browser run by Cloudflare (Browser Run) opens, once, a page address an account holder enters and reports what it saw of our script there. No end-user data; neither the address nor the result is stored. Addresses with an IP literal, a port or a user name, and our own sites, are refused; 20 checks an hour per account.

Data	Policy
Script address	The subdomain an account holder adds for our script is kept, with its set-up state, when it was added and first worked and, since 5 November 2026, when its TXT record was found, while it is in use, and deleted with the account; when Cloudflare does not confirm its removal at once, a job that runs every hour asks again until it does, keeping only the name and Cloudflare's reference until then. Requests to it pass through Cloudflare as they do for maskbreak.com. Since 5 November 2026 the name and `_maskbreak.`<name> are looked up through Cloudflare's public DNS resolver (1.1.1.1) when the name is added and while it waits for its TXT record.
Account deletion	GDPR Art. 17 self-service deletion, confirmed with the password (or, for a Google or GitHub account, a sign-in from the last 10 minutes), which first cancels a paid plan at Stripe and asks Cloudflare to remove the account's script address (an hourly job asks again until Cloudflare confirms it); live-system deletion, including a newsletter subscription under the same address, invitations addressed to the account and pending reset or signup records. Since 5 November 2026 the deletion's first database step marks the account suspended, the account-network, device-link, browser-consistency and Health-page records are written only while the account exists and is not suspended, and for a day after the deletion the server repeats the deletion of late records once an hour, so a record that a call already under way wrote after the deletion is removed too; when the deleted account was a member of another account's organization, that organization's test key is replaced and its owner is told. The backup-expiry commitment is 60 days; the retention-assurance note below says what the daily audit confirms and what stays open.

Data	Policy
Analytics & advertising	Fonts are self-hosted, and no advertising tag runs on our pages. No chat widget loads on any page: the Widgo chat was removed on 30 September 2026, and Widgo and the earlier provider Crisp keep legacy records only. Since 23 September 2026 PostHog product analytics records page views, clicks and sessions on public pages, `/signup` and the dashboard without a consent prompt (skipped under Global Privacy Control, Do Not Track and Save-Data; never on sign-in, reset, verification, invitation, settings, security, members, admin or result pages). Its library (posthog-js 1.434.11, byte-identical to the npm release) is served from maskbreak.com and the CSP grants PostHog's EU hosts `connect-src` only, never `script-src`, so PostHog cannot run code on our pages. Since 17 October 2026 it also receives page-speed measurements (Core Web Vitals) and, on public pages and `/signup`, clicks that changed nothing and the page's own script errors, with URLs, IP addresses, email addresses and key-shaped strings removed from their messages. Form fields, the visitor's own network readouts and all dashboard text are masked in the browser; console logs, network requests and canvas are not recorded; see Cookie Policy §2.3–2.5. Since 18 October 2026 the homepage and sign-up page also send one log-only browser-check line per page load (privacy notice, Section 3).

Retention assurance — review dated 17 September 2026, updated 5 November 2026

The existing 60-day backup-retention commitment remains unchanged. Until 16 October 2026 the repository audited every listed current and older backup version in AWS S3 and the storage expiry configuration, without a newest-snapshot exemption; a denied or incomplete inspection could not pass, and a successful upload alone does not establish retention compliance. The audit does not itself erase old versions or change the storage policy. On 10 October 2026 it passed for the first time: an enabled S3 lifecycle rule made every backup version eligible for deletion within 56 days of its upload, and no stored backup version was 60 days old or older. The audit runs daily and reads what is stored on the day it runs, so it is evidence, not a guarantee of future deletion, and a failed run reopens this item.

Since 17 October 2026 the backups go to Cloudflare R2 (EU jurisdiction), which keeps no older versions: a rule removes each copy after 40 days. The daily audit lists every stored copy through the backup receiver (names, sizes and dates, never contents) and fails if any copy is older than 43

days, so a missing or broken rule shows within days; it first passed on 17 October 2026. The second, encrypted copies in Google Drive (since 22 October 2026) are not part of this audit: the nightly backup job moves them to the trash itself after 25 days and opens an issue for the operator when it cannot, which leaves about four days to move them by hand before any copy would pass 60 days. The copies made up to 16 October 2026 remain in AWS S3 (Stockholm), where the audit had confirmed the deletion rule from 10 October 2026; that account was suspended between the backups of 16 and 17 October, we are not restoring access to it, and we can no longer check or delete those copies ourselves.

The repository restore tool blocks downloads and restore instructions for the backup copies because an authoritative record of deletions after each snapshot has not yet been established. The database's own 30-day history (Cloudflare D1 Time Travel, above) is the recovery path in use; a rollback with it is followed by carrying out again the account deletions made after the point restored. It permits a metadata-only inventory. Reopening recovery requires a reviewed deletion-reconciliation process and an isolated restore drill showing that erased accounts and records remain absent. This is a safeguard, not a completed recovery solution. These assurance items remain open, and do not authorise longer retention or restoration of deleted accounts.

4. Sub-processors

Cloudflare (CDN/edge and Workers; since 17 October 2026 the encrypted nightly DB backups in R2, EU jurisdiction, each copy removed after 40 days within the 60-day commitment a daily audit checks; since 19 October 2026 the primary database, D1, United States, Eastern North America, including keyed abuse-prevention counters, and email delivery, Email Service; since 20 October 2026 application hosting, Containers, United States, Eastern North America, Browser Run for the console's page check, and custom hostnames for customers' own script addresses; since 5 November 2026 its public DNS resolver, 1.1.1.1, for the DNS records of those addresses, which receives the names only), Stripe (sells our paid plans as the merchant of record through its Managed Payments service, shown as Link, on its own hosted pages; account holders' billing data only, never customer end-user data), PostHog (product analytics and session recordings, EU Cloud, Frankfurt), Sentry (server error reports, EU region, Frankfurt; designed not to receive personal data, through pattern-based scrubbing that is a strong safeguard rather than a guarantee), GitHub (GitHub Actions for monitoring and the nightly backup job, which reads the full database; and GitHub Sign-In), Google (optional Sign-In, including One Tap on the log-in page; Google Workspace for the support mailbox and, since 22 October 2026, the Google Chat space for our operations alerts and Google Drive for the second, encrypted copy of the nightly DB backups, each copy moved to the trash after 25 days and deleted by Google 30 days later), flagcdn.com (flag images in the console's test panel), Have I Been Pwned (k-anonymity check), plus network- and device-intelligence providers (United States) and, for the console's email and phone check since 18 October 2026, an email- and phone-intelligence provider (United States; the console's email and phone check since 18 October 2026 and, since 23 October 2026, email and phone intelligence on the visitor checks of customers that have it on, the default for accounts created from that day), named to customers under DPA, and Sift (login protection since 23

October 2026, United States; on for every account, for the logins and the account events customers' servers send, and off for a customer that turns it off). Former providers: Railway (hosting until 20 October 2026), Turso (database until 19 October 2026, deleted with our account), Resend (email until 19 October 2026, account closed), Amazon Web Services / S3 (the encrypted nightly DB backups made up to 16 October 2026, EU Stockholm eu-north-1; the account is suspended, so we can no longer check or delete them), and Widgo and Crisp (former chat providers; legacy records only). The live list with regions and a dated change log is public at /sub-processors; new customer-data processing sub-processors are announced in that change log at least 30 days in advance, with right to object; the email- and phone-intelligence provider was the exception, switched on the day of its entry (18 October 2026), and the 30-day right to object still applies to it. The moves of the database, email delivery and hosting to Cloudflare, already listed, were announced on the days they took effect (19 and 20 October 2026), with a right to object until 18 and 19 November 2026. Sift (23 October 2026) processes a customer's data only after that customer turns login protection on; the right to object runs until 22 November 2026.

5. Compliance framework

UK GDPR / EU GDPR aligned (processor role for evaluation traffic; DPA published self-serve at /dpa for every customer, with the Standard Contractual Clauses and the UK IDTA incorporated for transfers (copies on request) — countersigned copies via support@maskbreak.com). Primary processing is in the United States; EU-only processing is not offered. Section 2 of this document is incorporated into Annex B of the DPA by reference. CCPA honored. Security questionnaires (SIG-Lite, CAIQ, bespoke) completed within 5 business days of request. MSA available where click-through Terms can't be accepted.

This document is a plain-language summary, not a contract, except that section 2 is incorporated into Annex B of the DPA; contractual commitments live in the Terms, DPA, and (for enterprise) MSA/SLA. Verify anything here against the live pages: /trust, /privacy, /status, /sub-processors. Questions: support@maskbreak.com.