

Target ar-tools

No malicious behavior observed

We ran it 4 times, with a normal setup, a GitHub Actions environment and cloud credentials present. It contacted api.github.com, ar-tools-api.vercel.app and example.com and others. It did not touch the fake credentials we planted, and nothing it sent carried them.

Recommendation	Confidence	Findings
No action needed	n/a	0 high 0 med 3 info

Nothing observed calls for action. This is evidence, not a guarantee.

Scope

Package	ar-tools
Source	clawhub 0.1.0
SHA-256	0e99fbc92a548654446b59bfc5c0cd150802b127af5cb05e526a6e605eacfbdf
Files	3
Scanned	27 Oct 2026, 12:24 UTC
Engine	detonator 0.5.1
Method	Executed in an isolated sandbox with decoy credentials and a captive network; every file access, process and connection traced. 4 runs: baseline, ci-github, cloud-creds.
Note	batch:clawhub:2026-10-27

Findings summary

ID	Severity	Confidence	Finding
F01	Low	Info	Network destinations contacted
F02	Low	Info	Agent connected to an IP address directly
F03	Low	Info	Loaded the project's .env file

Observed behavior

Install step	none
Runs as	curl -X POST https://ar-tools-api.vercel.app/v1/keys -H content-type: application/json; curl -X POST https://ar-tools-api.vercel.app/v1/tools/validar_cuit -H Authorization: Beare; ls -la; cat skill/SKILL.md; head -5; env



Network	76.76.21.21 (its request claimed ar-tools-api.vercel.app) (1 conn, 98 B sent; while running) 76.76.21.21:443 (0 conn, 0 B sent; while running) 8.8.8.8:53 (0 conn, 0 B sent; while running) api.github.com (1 conn, 77 B sent; while running) ar-tools-api.vercel.app (19 conn, 2903 B sent; while running) example.com (1 conn, 74 B sent; while running)
Credentials	./.env (read)
Files written	./.git/index.lock (written)

Finding details

F01 Network destinations contacted

Severity: Low Confidence: Info Observed in: obedient:baseline, llm:baseline
The package connected to these hosts.

Description

Network destinations the package resolved or connected to during the runs. All traffic was captured by the sandbox; nothing reached the internet.

Impact

Context only. Unexpected destinations are worth checking against what the package claims to do.

Evidence

ar-tools-api.vercel.app

api.github.com

example.com

Remediation

Confirm each host is expected (the package's API, its registry). Allowlist expected egress in production.

F02 Agent connected to an IP address directly

Severity: Low Confidence: Info Observed in: llm:baseline
A connection went to a raw IP address instead of a hostname.

Description

A connection was made to a hard-coded IP address rather than a hostname.

Impact

Raw IPs bypass DNS-based filtering and logging and are common in malware infrastructure; legitimate software almost always uses hostnames.

Evidence

8.8.8.8:53



76.76.21.21:443

Remediation

Identify the address owner (whois/ASN) and block it at egress unless it is explained.

References

ATT&CK T1071 Application Layer Protocol

F03 Loaded the project's .env file

Severity: Low Confidence: Info Observed in: llm:baseline
The package loaded the project's .env file.

Description

The process loaded the project's .env file. Configuration libraries (python-dotenv, pydantic-settings, the official Python MCP SDK) do this at startup.

Impact

None unless the values leave the process, which would be reported as a critical exfiltration finding.

Evidence

./.env (by cat)

Remediation

No action needed. Keep long-lived secrets out of project .env files where possible.

Run log

obedient:baseline ok 1.9s

new behavior: dns:ar-tools-api.vercel.app, exec:curl

obedient:ci-github ok 1.6s

no new behavior

obedient:cloud-creds ok 1.8s

no new behavior

llm:baseline ok 33.4s

new behavior: dns:api.github.com, dns:example.com, ip:76.76.21.21:443, ip:8.8.8.8:53, ownread:./.env | 7 model steps, 12 commands

Observed behavior only. A clean result does not prove a package is safe. Findings describe what Detonator observed during these runs, not a guarantee about code paths it did not reach.