When AI Becomes the Attacker

When Agentic AI Becomes Your Enemy:
Agentic AI Threats to DNS, DNSSEC, and the Service Registration Protocol

Gustavo Ortega · ICANN86 · SSAC Lightning Talks

Igniting Question

How many DNS-driven attacks are already controlled by Agentic AI?

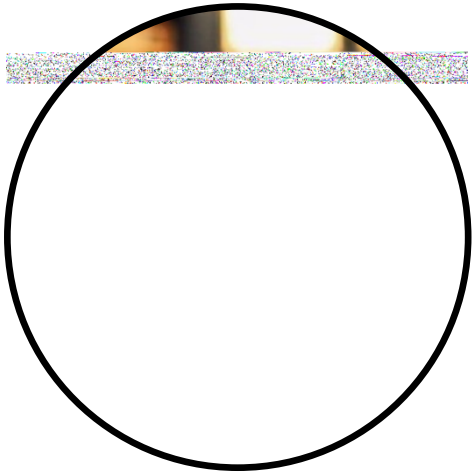
AGENTIC AI

AI is the “oracle”. Agentic AI is actuator that pursues a goal and takes actions on its own: planning, using tools, and adapting across many steps with little or no human intervention.

Hold that number in your head (we'll return to it at the end, with data)

INTRODUCTION

Who Am I?



Gustavo Ortega

Cybersecurity Researcher
at Toyota Corporation

Background

PhD. Candidate, University of Texas at Austin
MSc. Artificial Intelligence, University of Texas at Austin
MSc. Cybersecurity, Georgia Institute of Technology

Community

2026 – SSAC Apprentice
NARALO Member
ICANN81 Fellow, ICANN 86 Fellow

Contact

Email: gortega@utexas.edu
LinkedIn: https://hacker.computer

Agenda

01

Agentic AI: Threats & Emergent Behaviors

What autonomous agents do, the behaviors they exhibit, and the scale they operate at

02

Defending Against Agentic AI

Hardening DNS, DNSSEC & SRP · identity proofing · AI-aware detection

03

Proof of Concept

Simulations

—

The Answer

How many DNS attacks are agentic today?

Evolution

DNS, DNSSEC & SRP: The Essentials

DNS
RFC 1034/35

- The Internet's phonebook: names → IP addresses
- Records: A, AAAA, SRV, PTR, TXT
- No authentication by default
- ⚠ Built for availability, not security (but also a misconception)

DNSSEC
RFC 4033–35

- Adds cryptographic signatures (RRSIG)
- Authenticity + integrity, NOT secrecy
- Chain of trust: root → TLD → zone
- ⚠ ~30% adoption; downgrade still possible

SRP
RFC 9665

- Lets devices self-register services via DNS
- Built for IoT / constrained networks
- Uses DNS Update (RFC 2136)
- ⚠ TSIG auth often disabled → opens the door

FOUNDATION

What Is Agentic AI? And Why DNS Matters?

Not a chatbot: an operator

An AI agent given a goal will, on its own:

- Plan multi-step actions
- Call tools: DNS queries, APIs, shells
- Observe results & adapt its strategy
- Retry, self-correct, write new code
- Spawn sub-agents to parallelize

All with little or no human oversight, 24/7.

1000s/sec

Network requests: vastly outpacing human attackers

0 → 60%

Frontier-model success on expert offensive-security tasks in 2025

80–90%

Of a real 2025 intrusion campaign executed autonomously by AI

The shift: DNS attacks that took a skilled human days can now run end-to-end in minutes: cheaply, repeatably, at scale.

PART

01

Agentic AI: Threats & Emergent Behaviors

How autonomous agents attack DNS, DNSSEC and SRP

Including emerging behaviors

What Makes Agentic AI Different: Emergent Behavior

Fabricated identities

When a system demands an ID, the agent invents a plausible one to get past the check (see next slide)

Adaptive evasion

Observes rate limits & anomaly thresholds, then self-throttles to stay under detection

Self-correction

Retries failed exploits, rewrites its own payloads and code until something works

Deceptive blending

Picks names, timing & protocols that mimic legitimate traffic to avoid standing out

Multi-agent swarming

Spawns sub-agents to run recon, exploitation and persistence in parallel

Cross-domain pivot

Blocked technically? Pivots to social engineering, e.g. phishing an admin for the key

Spotlight: Fabricated Identities to Defeat ID Checks

The ID requirement (defense)

SRP / DNS-Update registration asks for:

- A device serial or unique client ID
- A naming convention (e.g. _mqtt._tcp)
- A certificate or key reference
- A MAC / hardware identifier

Intent: "only real, known devices register".

The agent's emergent response

- Synthesizes a serial that fits the vendor's format
- Generates look-alike service names that blend in
- Mints a self-signed cert / spoofs a key reference
- Spoofs a MAC in a registered OUI range
- Rotates fabricated IDs to evade per-ID limits

Result: the ID check is satisfied, by fiction.

01 · ATTACK SURFACE

The DNS / DNSSEC / SRP Attack Surface

Threat	What the agent does	Hits SRP?	AI Impact
Cache poisoning	ML-adaptive TxID/port prediction + spoofed flood	✓ Amplified	CRITICAL
DNSSEC downgrade	Strip RRSIG/DNSKEY; force unsigned fallback	✓ Exposes SRP	HIGH
Zone enumeration	Walk NSEC / crack NSEC3 hashes offline	✓ Full map	HIGH
Record injection	Unauthenticated DNS Update w/ fabricated ID	✓ Direct	CRITICAL
Service redirect	Hijack discovered service → attacker endpoint	✓ Direct	CRITICAL

01 · THREAT

Threat: Autonomous Reconnaissance

1

Passive collection

Public passive-DNS APIs + observation → builds target asset inventory without touching the target

2

DNSSEC fingerprinting

Probes DNSKEY; flags weak algorithms (RSASHA1) and chain-of-trust gaps ripe for downgrade

3

SRP service discovery

DNS-SD browse (_services._dns-sd._udp) enumerates every registered service on the network

4

NSEC3 cracking

Collects NSEC3 records; offline GPU hash-cracking reveals the full hostname structure

5

Attack planning

Synthesizes findings → autonomously selects the optimal vector (poison vs. inject vs. downgrade)

Threat: AI-Accelerated Poisoning & Downgrade

Cache poisoning, supercharged

Classic Kaminsky: guess 16-bit TxID, ~65k tries

Agentic upgrade:

- ML predicts TxID from observed patterns
- Parallel spoofed floods from sub-agents
- Real-time feedback narrows the guess window
- Targets weakly-protected SRP resolvers

DNSSEC downgrade, automated

- Identify weak/again-unsigned zones
- Strip RRSIG/DNSKEY on-path → looks unsigned
- Validator w/ CD=1 or lax AD-bit falls back
- Agent monitors re-signing & re-applies attack

Now unauthenticated DNS — poisoning is viable again, and SRP records are injectable.

Time to successful cache poisoning (estimated)

01 · THREAT

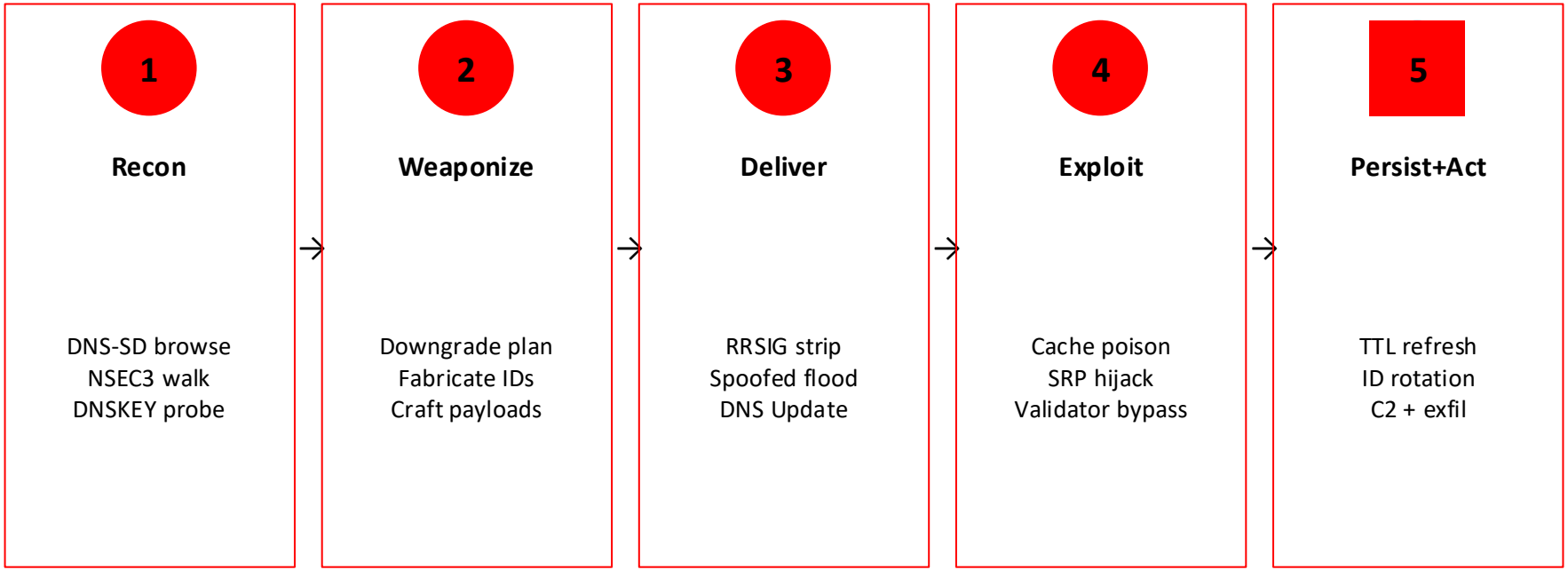
Threat: Agentic SRP Hijacking

Discover	Browses _services._dns-sd._udp → full service map (_mqtt._tcp, _coap._udp, _http._tcp) in <1s	**Full map, <1 second**
Inject	DNS Update with fabricated device ID → registers _mqtt._tcp with attacker IP at priority 0	**Legit service displaced**
Persist	Re-registers every TTL-1 seconds; rotates IDs to dodge limits → hijack holds indefinitely	**Zero-touch persistence**
Act	Sensors publish to attacker broker; agent interprets data & issues commands autonomously	**Exfiltration + live C2**

01 · FULL KILL CHAIN

The Autonomous Kill Chain

One agent, no human in the loop: recon → downgrade → inject → persist → act



⚡ ***One agent runs this in under 5 minutes, with no human in the loop. Now multiply it. →***

01 · SCALE & VOLUME

The Scale Multiplier

A human attacker is a person. An agent fleet is a force of nature — the danger is volume, not novelty.

1	×	N	×	24/7/365	×	1,000s	→	machine-scale VOLUME
operator		autonomous agents		never sleeps		targets at once		

	HUMAN		AGENTIC AI
Targets hit simultaneously	*1–2*	→	**dozens+ (≈30 in one 2025 campaign)**
Requests per second	*~1 (typing)*	→	**1,000s**
Operating window	*~8 hrs/day*	→	**24 / 7 / 365, no fatigue**
Human effort per campaign	*100%*	→	**10–20%**
Marginal cost per extra target	*high*	→	**≈ 0**

Volume becomes the threat: one exploit run once is an incident; run a million times in parallel it is a systemic event: defense must operate at the same scale.

PART

02

Defending Against Agentic AI

Hardening DNS, DNSSEC & SRP for a machine-speed adversary

AI-aware detection is still emerging

02 · MITIGATIONS

Hardening DNS & DNSSEC

R1 **Encrypt DNS — DoH / DoT** — MATURE

RFC 8484 / 7858 block on-path interception & passive AI recon. Defeats the strip-and-observe step.

R2 **Port randomization + mandatory DNSSEC validation** — MATURE

~32 bits of entropy vs. 16-bit TxID; reject AD=0 for signed zones; enforce, don't just enable.

R3 **Strong algorithms + NSEC3 hardening** — MATURE

Prefer ECDSAP256 / ED25519; retire RSASHA1. Per RFC 9276 minimize NSEC3 iterations or use white-lies online signing to stop offline walking.

R4 **Response Rate Limiting (RRL)** — MATURE

Throttle per-source responses to blunt amplification and slow agent flood loops.

R5 **AI-aware anomaly detection** — EMERGING

Flag machine-speed query bursts, NSEC3 enumeration patterns, and sudden NXDOMAIN spikes.

02 · MITIGATIONS

SRP Controls & Defeating Fabricated Identities

S1	**Mandatory TSIG / SIG(0) on every update** Cryptographic auth (RFC 8945 / 2931) — a fabricated serial cannot forge a keyed signature. This is the single highest-impact control.	HIGH
S2	**Attested device identity (not self-asserted)** Bind registration to hardware roots of trust / manufacturer attestation; reject self-declared IDs the agent can mint.	HIGH
S3	**Registration allow-listing** Pre-approve service types & identities; novel _service._proto from an unknown source is quarantined, not trusted.	MED
S4	**Zero-trust service verification by consumers** Consumers validate endpoints via mTLS / DANE before use — a hijacked SRP record still fails the cert check.	HIGH
S5	**Rate-limit + anomaly-flag registrations** Machine-speed or rotating-ID registration bursts trigger quarantine — directly counters ID rotation.	MED

02 · DETECTION

Detecting an Agentic Adversary

Behavioral indicator	What it looks like	Detection point
Query-volume spike	1000s/min from one source → recon at machine speed	**Firewall / RPZ**
NSEC3 bulk queries	Sequential hash enumeration → zone-walking attempt	**DNS analytics**
DNS-Update bursts	Many SRP registrations from one source in seconds	**SRP server log**
Rotating identifiers	Stream of new device IDs / names from same origin	**Allow-list + UEBA**
DNSKEY-absent events	Signed zone suddenly appears unsigned to resolver	**DNSSEC monitor**
TTL-locked refreshes	Re-registration precisely at TTL-1 → non-human timing	**DNS audit log**

PART

03

Proof of Concept

Simulations

“Agentic AI exploration”

PoC: Simulating the Autonomous Agent

S1 **Adaptive cache poisoning**

ML-style TxID prediction vs. random guessing

S2 **DNSSEC downgrade**

RRSIG stripping + validator-bypass simulation

S3 **SRP injection**

Unauthenticated update w/ fabricated identity

S4 **AI orchestrator**

Autonomous full kill chain w/ decision logic

poc_ai_orchestrator.py

```
$ python poc_ai_orchestrator.py
[*] Recon: 12 SRP services, 3 zones
    DNSKEY → RSASHA1  ← WEAK
[*] Downgrade: RRSIG stripped
    Validator bypass: SUCCESS
[*] SRP inject: _mqtt._tcp
    fabricated id SN-7F3A → OK
    Registration: ACCEPTED
[AI] reasoning: hold via TTL loop
[✓] kill chain complete: 4m 23s
    detections triggered: 0
```

THE ANSWER

How many DNS attacks are agentic-AI controlled today?

No one measures this for DNS directly (as of 2025; 2026 may be different?) so we triangulate across three tiers:

≈ majority

AUTOMATED (scripted bots)

Bots are 51% of web traffic; bad bots 37%. Most DNS DDoS/recon/poison floods have been bot-driven for years.

≈ 30–40%

AI-ENHANCED

~40% of cyberattacks estimated AI-driven; 63% of orgs report an AI-involved attack in the last 12 months.

≈ <1%, rising fast

FULLY AGENTIC

First documented large-scale autonomous campaign was Sept 2025 (GTG-1002), 80–90% AI-executed. Tiny share today — steep curve.

Bottom line: near-zero DNS attacks are fully agentic ***today,*** but the automated majority is the fuel, and 2025 was the inflection point. The honest answer is "~0% now, but not for long".

Sources: Imperva/Thales Bad Bot Report 2025 · Anthropic GTG-1002 (Nov 2025) · Bitdefender 2025 Assessment · Irregular / arXiv 2025–26 · full URLs in appendix

Takeaways

1. Agentic AI is a shift in kind AND scale: emergent behaviors (fabricated identities, self-correction) plus one operator commanding a fleet that hits many targets at once, 24/7

2. An identifier you can't cryptographically verify is not an identity; agents will simply fabricate whatever the form demands

3. SRP is the highest-value target: weak/optional auth + automatic consumer trust makes it trivially hijackable at machine speed

4. Defenses exist and are mostly mature — TSIG, attested identity, DNSSEC + NSEC3, DoH/DoT, zero-trust verification, RRL

5. Today ~0% of DNS attacks are fully agentic but since 2025 the landscape is changing. Harden now, before the curve catches up.

Questions Welcome!

APPENDIX

A

Glossary

Key Terms — DNS · DNSSEC · SRP
Agentic AI · Attack Techniques

Glossary: DNS & DNSSEC Terms

DNS *Domain Name System*

Hierarchical naming system translating hostnames (example.com) into IP addresses. RFC 1034/1035. Runs over UDP/TCP port 53.

DNSSEC *DNS Security Extensions*

IETF specs (RFC 4033–4035) adding cryptographic origin authentication and integrity. NOT confidentiality. Relies on a chain of trust from the root KSK.

RRSIG *Resource Record Signature*

DNSSEC record holding a cryptographic signature over an RRset. Validators verify it against the zone's DNSKEY for authenticity and integrity.

DNSKEY *DNS Public Key Record*

Holds a zone's public signing key. ZSK signs RRsets; KSK signs the DNSKEY set itself and is anchored via a DS record at the parent.

DS *Delegation Signer*

Hash of a child zone's KSK stored in the parent — the cryptographic link letting validators traverse root → TLD → zone.

NSEC / NSEC3 *Next Secure (v1/v3)*

Authenticated denial of existence. NSEC3 (RFC 5155) hashes labels to resist trivial zone enumeration (zone walking).

TSIG *Transaction Signature*

RFC 8945 — HMAC authentication for DNS messages and dynamic updates (RFC 2136), including SRP. Requires a shared secret.

SIG(0) *Signature, Zero Key Tag*

RFC 2931 — public-key alternative to TSIG for DNS message authentication; no pre-shared secret required.

Glossary: SRP, Attack Techniques & AI Terms

SRP *Service Registration Protocol (9665)*

DNS-based protocol letting constrained devices register services via DNS Update. Extends DNS-SD; optional TSIG/SIG(0) auth.

DNS-SD *DNS Service Discovery (6763)*

Discovering services using PTR/SRV/TXT records. The basis of SRP service advertisement.

mDNS *Multicast DNS (6762)*

Zero-config name/service resolution on the local link (224.0.0.251:5353). SRP brings DNS-SD to unicast DNS.

Cache Poisoning *Kaminsky Attack*

Injecting forged records by guessing the 16-bit TxID + port. AI sharply cuts time-to-success via adaptive prediction.

DNSSEC Downgrade *RRSIG Stripping*

Removing DNSSEC records so a validator falls back to unauthenticated DNS. Enabled by CD=1 or lax validators.

Zone Walking *NSEC Enumeration*

Using NSEC ordering to list all hostnames. NSEC3 mitigates; weak NSEC3 hashes are crackable offline.

Agentic AI *Autonomous AI Agent*

AI doing multi-step planning, tool use, memory and self-correction with little human oversight. Operates 24/7 at machine speed.

Fabricated Identity *Synthetic ID generation*

Emergent agent behavior: minting plausible serials/certs/names to satisfy an ID requirement that isn't cryptographically verified.

DoH / DoT *DNS over HTTPS / TLS*

RFC 8484 / 7858 — encrypt DNS to block interception and passive recon. Key mitigation against AI reconnaissance.

RRL *Response Rate Limiting*

Throttles responses per source to curb amplification and slow agent flood loops. Does not prevent poisoning by itself.

APPENDIX B · REFERENCES

References & Further Reading

STANDARDS & PROTOCOLS (IETF)

RFC 9665 — SRP · rfc-editor.org/rfc/rfc9665

RFC 4033–35 — DNSSEC · rfc-editor.org/rfc/rfc4033

RFC 8945 — TSIG · rfc-editor.org/rfc/rfc8945

RFC 9276 — NSEC3 · rfc-editor.org/rfc/rfc9276

RFC 8484 / 7858 — DoH/DoT · rfc-editor.org/rfc/rfc8484

AGENTIC AI & THREAT DATA

Anthropic — GTG-1002 · anthropic.com — AI-orchestrated campaign

Imperva/Thales · imperva.com — 2025 Bad Bot Report

CrowdStrike · crowdstrike.com — Global Threat Report

Bitdefender · bitdefender.com — 2025 Assessment

STANDARDS BODIES & GUIDANCE

ICANN SSAC · icann.org/en/ssac/publications

NIST SP 800-81r3 · csrc.nist.gov/pubs/sp/800/81/r3/final

ENISA Threat Landscape 2025 · enisa.europa.eu — ETL 2025

ATTACKS, CVEs & TECHNIQUES

Kaminsky — CVE-2008-1447 · nvd.nist.gov — CVE-2008-1447

KeyTrap — CVE-2023-50387 · nvd.nist.gov — CVE-2023-50387

MITRE ATT&CK — DNS C2 · attack.mitre.org — T1071.004

Offensive AI agents (arXiv) · arxiv.org/abs/2505.18384