Will AI Change Everything?

Predictions & recommendations for fighting internet abuse at scale

David Freeman | Principal Scientist, Delphi Research

CAMP 11th Annual Meeting

Seoul, Korea · 7 July 2026

What is “Cybersecurity”?

Wikipedia:

“Protecting software, systems, and networks from threats that lead to unauthorized disclosure, theft, or damage.”

My definition:

“Preventing people from using computers to cause harm.”


Financial loss


Denial of service


Data leakage


Physical harm

SETTING THE STAGE

What do I know about Cybersecurity?

(2012-17)

(2017-26)

14 years working on **high-volume adversarial problems:**

- Fraud & scams
- User authentication & impersonation
- Unauthorized data scraping
- and others...

3.5B	5B	Millions	Trillions
fake accounts disabled	malicious ads blocked	of advertisers verified	of scraping requests blocked
Meta Transparency Report	*Google Ads Safety Report*	*Meta / Google*	*Meta Press*

Incidence will never go to zero.
The goal is to get the problem under control.

Principles for Fighting Abuse at Scale

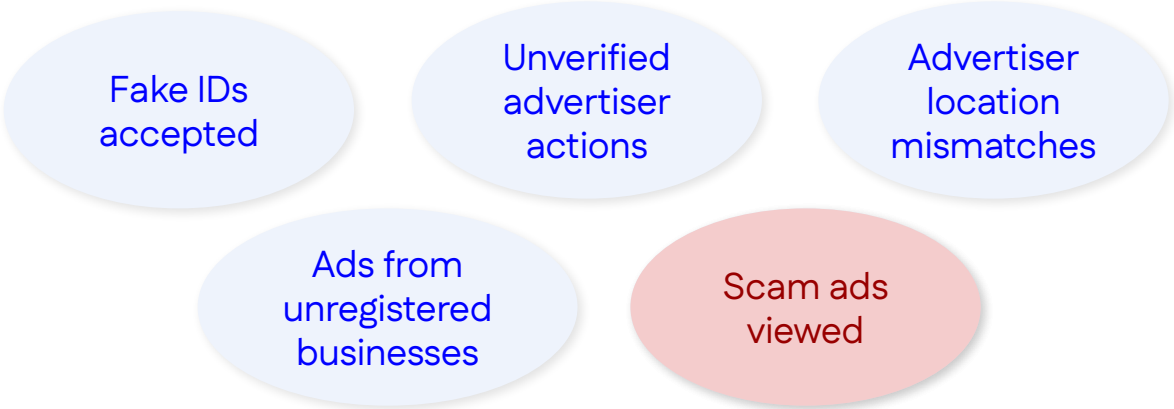
1. What gets measured gets fixed
2. Everything is a tradeoff
3. Change the economics
4. Focus on behavior

PRINCIPLE #1

What gets measured gets fixed

Example: Verifying advertisers

Which of these events should we count to size the problem?



These are all bad things. Only one reflects a bad user outcome.

Measure using multiple techniques

User reporting

✓ Reflects what users care about
✗ Noisy, biased, can be gamed

Best for: attack discovery

Expert labeling

✓ High quality, statistically robust
✗ Expensive to scale

Best for: assessing interventions

Proxy metrics

✓ Easy to scale
✗ Loosely coupled with the real goal

Best for: continuous monitoring

Red-teaming

✓ Uncovers complex vulnerabilities
✗ Doesn't reflect attacker

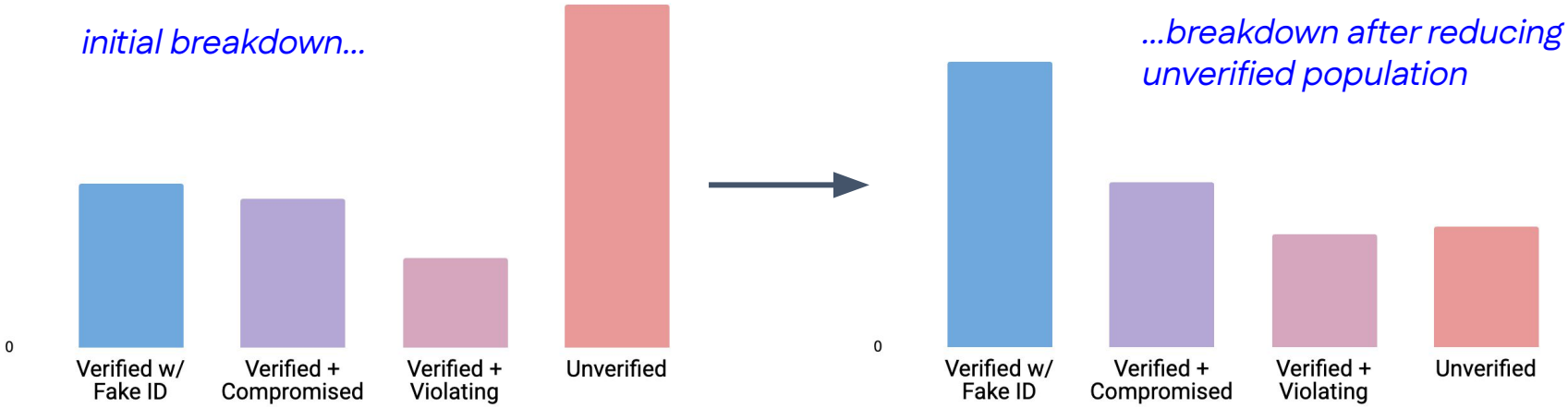
Best for: proactive defense

No single method is enough — monitor & set goals on as many as possible.

Segment to find the low-hanging fruit

Example: Verifying advertisers

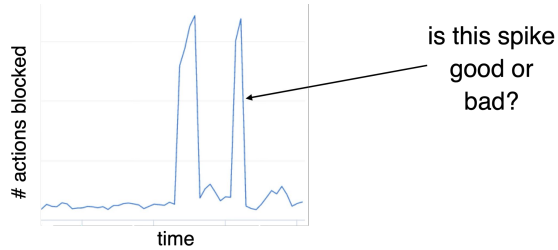
FICTIONAL breakdown of scam ads by advertiser segment:

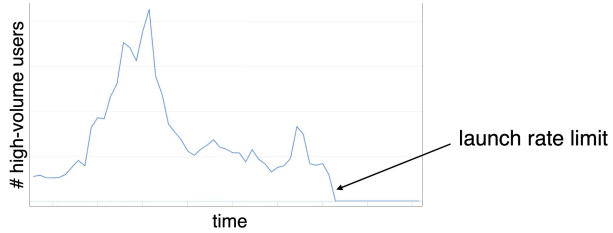


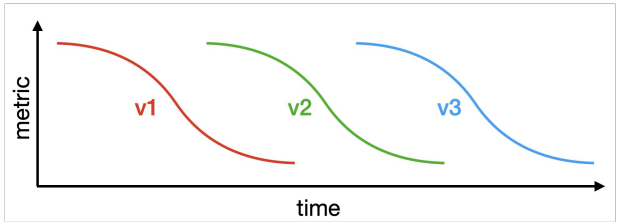
Solve one problem at a time. Focus on the segment with the biggest opportunity.

Avoid common measurement mistakes

- Don't count how many things you blocked.
- Don't "grade your own homework."
- Don't assume your metric quality is constant.







Takeaway: constantly re-measure, re-segment, re-prioritize.

PRINCIPLE #2

Everything is a tradeoff

High Security

Low Friction

- Every “bad user outcome” metric has a **“user friction” countermetric.**
  - accounts compromised ↔️ users locked out
  - sensitive data scraped ↔️ benign actions blocked
  - loss due to fraud ↔️ loss of subscription revenue
- Countermetrics are different from harm metrics:
  - ✓ Not adversarially controlled
  - ✗ Requires counterfactual assessment (holdout)

PRINCIPLE #3

Change the economics — Raise attack cost

Attackers are economically motivated.

Make the attack more expensive to run. For example:

- Require login to access data.
- Require verification to perform risky actions.
- Adjust requirements based on probability of bad outcomes.
- Go after financial chokepoints (see Kondrashin et al, McCoy et al)
- Do **all of the above** (“defense in depth”)

Change the economics — Reduce attack value

Attackers are economically motivated.

If you can't make the attack more expensive, make it pay off less per success.

- Limit the number of actions a single account can perform.
- Limit the amount of data returned per request.
- Limit potential financial loss per user.

  (& vary limits across segments)

PRINCIPLE #4

Don’t play “Whack-a-mole”

PRINCIPLE #4

Focus on behavior, not content

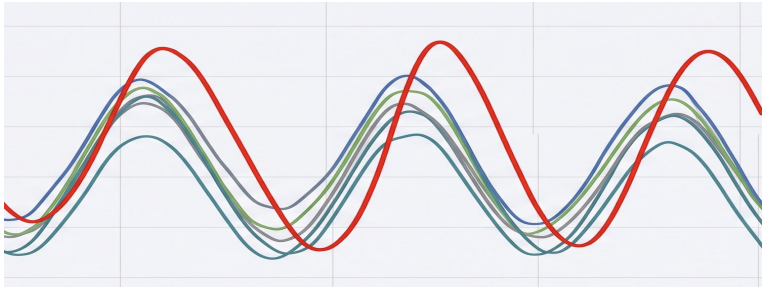
Do less of these things:	Do more of these things:
Invest in content classifiers to find violating **instances**.	Invest in **clustering algorithms** & cluster classification to find violating **groups**.
Block specific IP addresses, app versions, text strings, etc.	**Dynamically assess** IP activity, app usage, content generation, etc. vs. benign baseline.
Engage in an arms race to stop deep fakes.	Corroborate **multiple identity sources** against each other.

Attackers will always be anomalous in some dimension — the key is to collect enough data to identify the anomaly.

Use information asymmetry in your favor

- **Graph data**
  - users
  - devices
  - locations
  - ...

- **Baseline distributions**
  - action type
  - request latency
  - system parameters
  - ...



Now AI is Here!

What Changes?

1. What gets measured gets fixed?

2. Everything is a tradeoff?

3. Change the economics?

4. Focus on behavior?

Both sides get better at their jobs

Attackers can

- Produce realistic (fraudulent) images & video
- Engage in native-language (fraudulent) conversation
- Automate everything at low cost

malicious content

economic motivation

Defenders can

- Scale high-quality expert labeling (better visibility, faster iteration)
- Identify anomalous patterns in unstructured data
- Automate everything at low cost

measure everything

behavior analysis

But their jobs don't change!

Attacker's job: Look like a benign user.

Defender's job: Distinguish attack traffic from benign traffic.

What AI doesn’t change

- Attackers are still economically motivated.
- Attack patterns still differ from normal behavior.
- You still need objective, ground-truth observations to measure opportunity and impact.
- Attackers never sleep!

A cautionary tale

You still need experts to determine whether the AI is correct!

WRAPPING UP

Takeaways

1. **Measure the bad user outcome** — use many sources to get the complete picture.
2. **Everything is a tradeoff** — put numbers on both sides of the equation.
3. **Shift the attacker's cost/benefit calculation** — raise cost or cut value.
4. **Focus on behavior, not content** — it's expensive to fake consistently across all surfaces and dimensions.

AI increases speed and quality on both sides. It doesn't change the fundamentals.

Thank You / 감사합니다

Questions?

David Freeman · www.delphiresear.ch

Opinions expressed in this talk are my own and do not represent the views of any current or past employer.