MODULE 2 · SESSION 3

Defend and Recover AI Services

How to detect an attack, contain its impact, and rebuild a trustworthy account of what actually happened.

Theory session

Beginner

Running case: HarborPilot Health

ORIENTATION

Where this session sits

The last session of Module 2: when the system is attacked

MODULE 1
AI System Foundations

MODULE 2
Controlled AI Operations

MODULE 3
Strategic AI Operating Choices

SESSIONS IN THIS MODULE

2.1 Measure Reliability Before Scale
An evaluation loop that detects quality drift

2.2 Bound Autonomy With Human Escalation
What the system may decide alone, and what it must escalate

2.3 **Defend and Recover AI Services**
Detect an attack, contain it, and restore trust

YOU ARE HERE

SESSION OUTCOMES

What this session explains

Responding to an attack without making the second problem worse

01

AI systems have their own attack surface

Prompt injection, poisoning, extraction and abuse sit alongside ordinary security risks.

02

Restoration is not investigation

Service came back in eight hours; the scope took five months to settle.

03

Response follows a known sequence

Prepare, detect and analyse, contain and recover, then learn.

04

Premature certainty is expensive

A figure that moves from 345,000 to 3.76 million costs more credibility than saying “we do not yet know”.

CONCEPT 1 · THREATS

Four attack classes specific to AI systems

In addition to everything that already threatens software

EVASION	Crafted input that makes the model behave wrongly at the moment of use.
POISONING	Corrupted training or reference data, planted so the model learns the wrong thing.
EXTRACTION AND LEAKAGE	Pulling training data, credentials or the model itself out through its outputs.
ABUSE AND INJECTION	Instructions hidden in content the system reads, redirecting it to act for someone else.

NIST (2024). Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations. NIST AI 100-2.

THE ONE TO KNOW FIRST

Prompt injection

If your system reads untrusted content and can take actions, assume that content is trying to instruct it.

THE DEFENCE THAT GENERALISES

Least privilege

An AI component should hold the smallest set of permissions that lets it do its job, and no standing access.

THE UNCOMFORTABLE TRUTH

Data is part of the attack surface

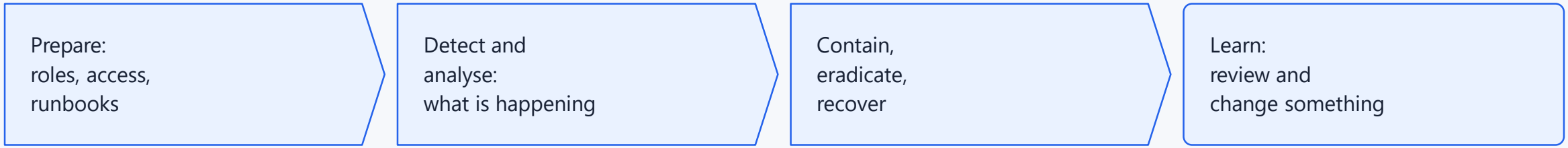
Session 1.2's provenance chain is also a security control — you cannot detect poisoning without it.

CONCEPT 2 · THE SEQUENCE

Four phases of incident response

The order matters; skipping the first makes the rest improvised

THE RESPONSE LIFECYCLE



WHY THE FIRST BOX IS THE LARGEST

Preparation is the only phase you can do calmly. Everything decided in advance — who leads, who speaks, who can revoke access at 3 a.m. — is a decision you do not have to make under pressure.

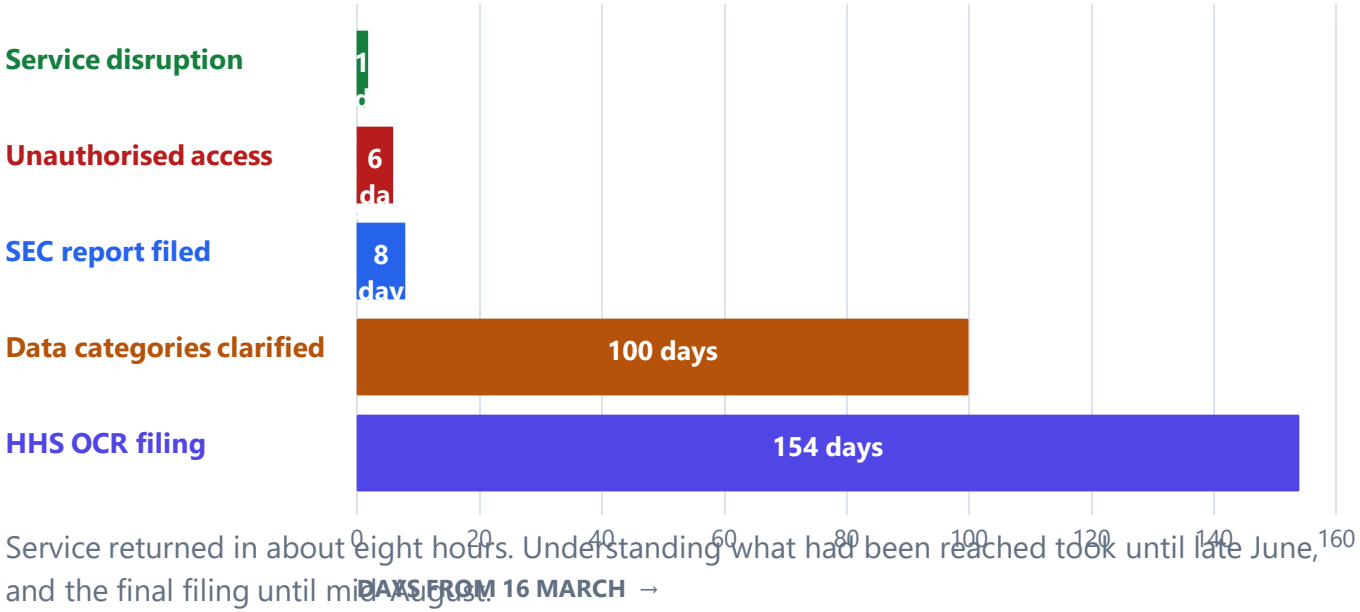
NIST (2012). Computer Security Incident Handling Guide. Special Publication 800-61 Revision 2.

CONCEPT 3 · TIMELINE

Two clocks run at different speeds

The CareCloud sequence, in days from first disruption

OPERATIONAL RECOVERY AND EVIDENTIARY RECOVERY



Service returned in about eight hours. Understanding what had been reached took until late June, and the final filing until mid-August.

THE OPERATIONAL CLOCK

Hours

Disruption lasted about eight hours on 16 March, and service was restored that evening.

THE EVIDENTIARY CLOCK

Months

The review had to clarify which categories of data were involved — settled by 24 June.

THE LESSON

Do not let one speak for the other

“Service is restored” says nothing about whether information was taken.

CONCEPT 4 · EVIDENCE

Three kinds of statement, never mixed

The discipline that keeps an evolving account credible

WORKING DEFINITION

In an incident, separate what is confirmed, what is provisional, and what is unknown — and label every public number as one of the three. A provisional figure reported as a fact becomes a correction later.

1

Confirmed

Verified from evidence you hold. "Access occurred between these dates."

2

Provisional

Current best estimate, expected to change. "Our present estimate is X; the review continues."

3

Unknown

Saying so is a stronger position than a number you will have to multiply by ten.

CONCEPT 5 · COMMUNICATION

Who needs to hear what, and when

The gap you leave is filled by someone else

Audience	What they need	The common error
Customers	Whether they are affected, what to do, and when the next update comes.	Waiting for a complete picture while they hear it elsewhere.
Regulators	Facts, within the required window, labelled by confidence.	A firm number that later moves by an order of magnitude.
Staff	What is known, what they may say, and who to direct questions to.	Silence internally, which produces speculation externally.
The public	A brief, accurate account and a named point of contact.	Near-silence, which becomes part of the story itself.

WHAT THE CASE SHOWS

CareCloud's early notifications referenced roughly 345,000 affected individuals; the later filing listed 3,756,469. Labelling the first as provisional would have cost nothing and preserved a great deal.

CONCEPT 6 · CONTAINMENT

How aggressively to contain

Containment and evidence pull in opposite directions

TOO CAUTIOUS

Watching to learn more while access continues. Every extra hour is more data reached.

BALANCED

Cut access immediately; preserve logs and images first so the investigation still has evidence.

TOO BLUNT

Wipe and rebuild at once. The intrusion stops and so does any chance of knowing what happened.

DECIDE THESE BEFORE AN INCIDENT

Decision	Pre-agreed answer
Who can revoke access immediately?	Named roles with standing authority, reachable out of hours.
What is preserved before rebuilding?	Log retention and snapshot rules, set in advance and tested.
Who speaks externally?	One named voice, so the account stays consistent.

CONCEPT 7 · FAILURE MODES

Three responses that turn one incident into two

The second incident is always about credibility

Declaring the scope too early

Looks like A precise figure in the first notification.

Why it fails Corrections are read as concealment, however honest.

Fix Publish ranges and label them provisional until the review closes.

Treating restoration as resolution

Looks like "Service restored; the matter is closed."

Why it fails The operational clock and the evidentiary clock run at completely different speeds.

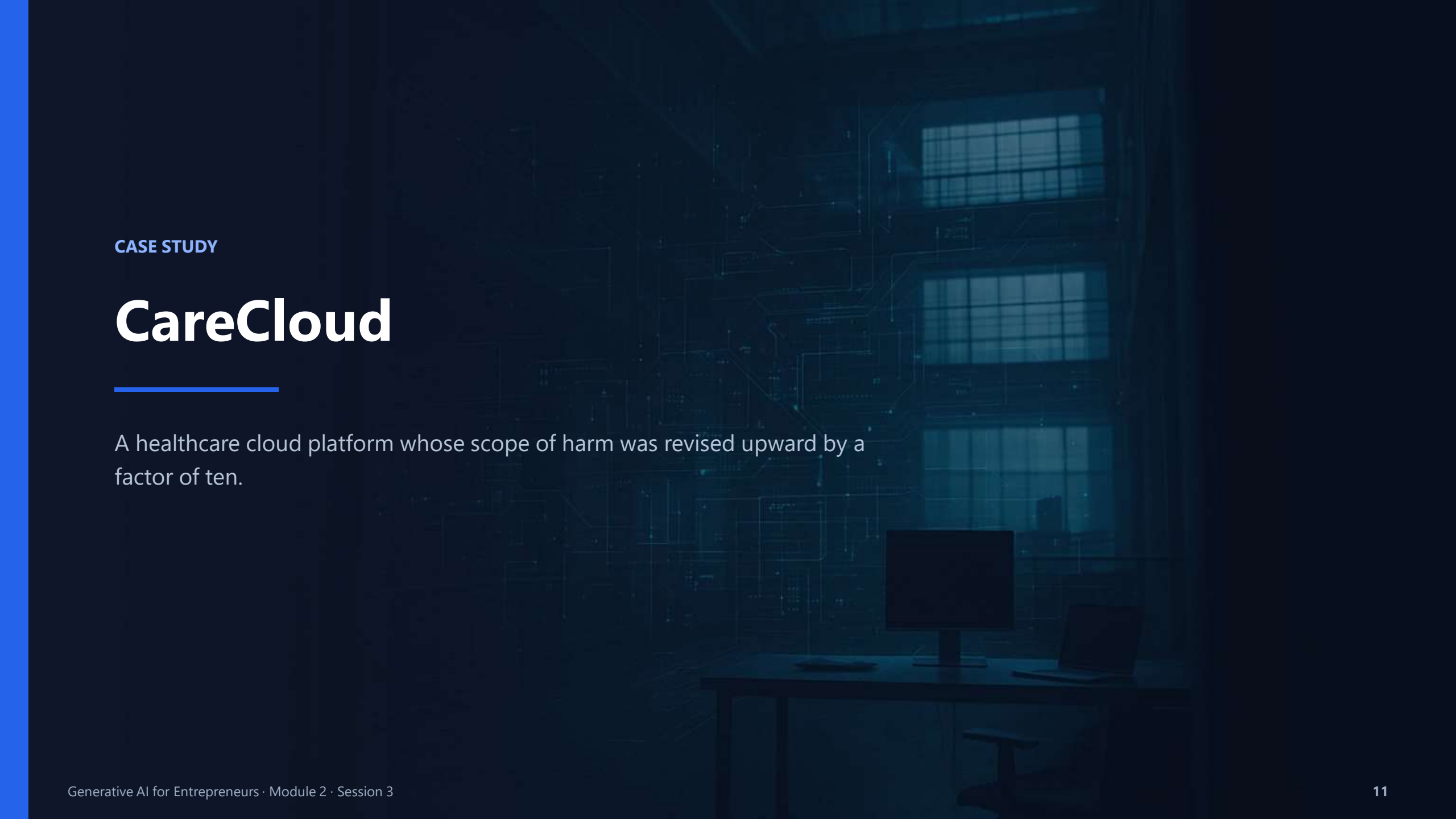
Fix Two separate workstreams with separate status reporting.

Silence as a strategy

Looks like Saying nothing until everything is known.

Why it fails The absence of explanation becomes the explanation.

Fix A fixed update cadence, even when the update is that nothing has changed.



CASE STUDY

CareCloud

A healthcare cloud platform whose scope of harm was revised upward by a factor of ten.

CASE STUDY · CARECLOUD

Restoring service was only the first question

Six days of access, and five months to a settled account

THE INCIDENT

Access, disruption, exfiltration

THE FOOTPRINT

CareCloud serves over 45,000 healthcare providers with EHR, practice management and revenue cycle tools.

THE INTRUSION

An unauthorised party held access to an AWS environment for six days in March 2026.

THE DISRUPTION

Initial internal disruption lasted about eight hours on 16 March; service was restored that evening.

THE ACCOUNT

The scope moved by 10×

FIRST FIGURE

Roughly 345,000 affected individuals, used in early-August notifications to state attorneys general.

REVISED FIGURE

3,756,469 in the HHS OCR filing of 17–18 August — approximately 3.76 million patients.

THE SILENCE

No threat actor claimed responsibility, and CareCloud remained near-silent publicly — drawing its own silence into the story.

READING THE CASE

Restoring service answered the operational question. It did not answer what had been reached. A credible response has to separate confirmed facts from provisional estimates.

APPLICATION

Bringing the theory to HarborPilot Health

Preparing the response you cannot improvise

Step	What you put in place	HarborPilot example
Name the roles now	Lead, communicator, technical owner, legal contact.	One page, reviewed quarterly, with out-of-hours contacts.
Reduce standing access	Least privilege for every AI component.	The forecasting agents read one dataset and cannot reach patient records at all.
Treat untrusted input as hostile	Anything the system reads may be trying to instruct it.	Supplier documents parsed in a sandbox with no tool access.
Set retention before you need it	Logs and snapshots preserved automatically.	Ninety days of access logs, immutable, outside the production account.
Write the disclosure template	With confidence labels built in.	Confirmed / provisional / unknown sections, and a stated next-update time.
Run one exercise a year	A rehearsal, not a document review.	Two hours, one scenario, with the real on-call people.

WHAT COMES NEXT

Module 3 moves from control to strategy: reliability, cost and the suppliers you depend on.

SUMMARY

What to carry out of this session

1. **AI systems add four attack classes.**
   Evasion, poisoning, extraction and injection, on top of ordinary security risk.

2. **Restoration is not investigation.**
   Eight hours to restore service; months to establish what was reached.

3. **Label every number: confirmed, provisional or unknown.**
   345,000 becoming 3,756,469 costs credibility that “we do not yet know” would not have.

4. **Contain fast, but preserve evidence first.**
   Wiping immediately stops the intrusion and destroys the account of it.

5. **Silence becomes part of the story.**
   A fixed update cadence is cheaper than the interpretation people supply themselves.

FURTHER READING

Sources for this session

Primary references behind each concept

ADVERSARIAL ML	NIST (2024). *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations.* NIST AI 100-2 E2023.
INCIDENT HANDLING	NIST (2012). *Computer Security Incident Handling Guide.* SP 800-61 Rev. 2.
APPLICATION RISKS	OWASP (2025). *Top 10 for Large Language Model Applications.*
RISK FRAMEWORK	NIST (2023). *AI Risk Management Framework (AI RMF 1.0).*
POISONING	Biggio, B. & Roli, F. (2018). Wild patterns: ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84, 317–331.
CRISIS COMMUNICATION	Coombs, W. T. (2007). Protecting organization reputations during a crisis. *Corporate Reputation Review*, 10(3), 163–176.
LEARNING FROM FAILURE	Allspaw, J. (2012). Blameless postmortems and a just culture. *Code as Craft, Etsy.*