AI Incident Response Guidebook: Recommendations for Practitioners

Andrea Brennen

SVP Applied Research, IQT

abrennen@iqt.org

IQT Again Named Most Active Investor in Top Defense and National Security Startups

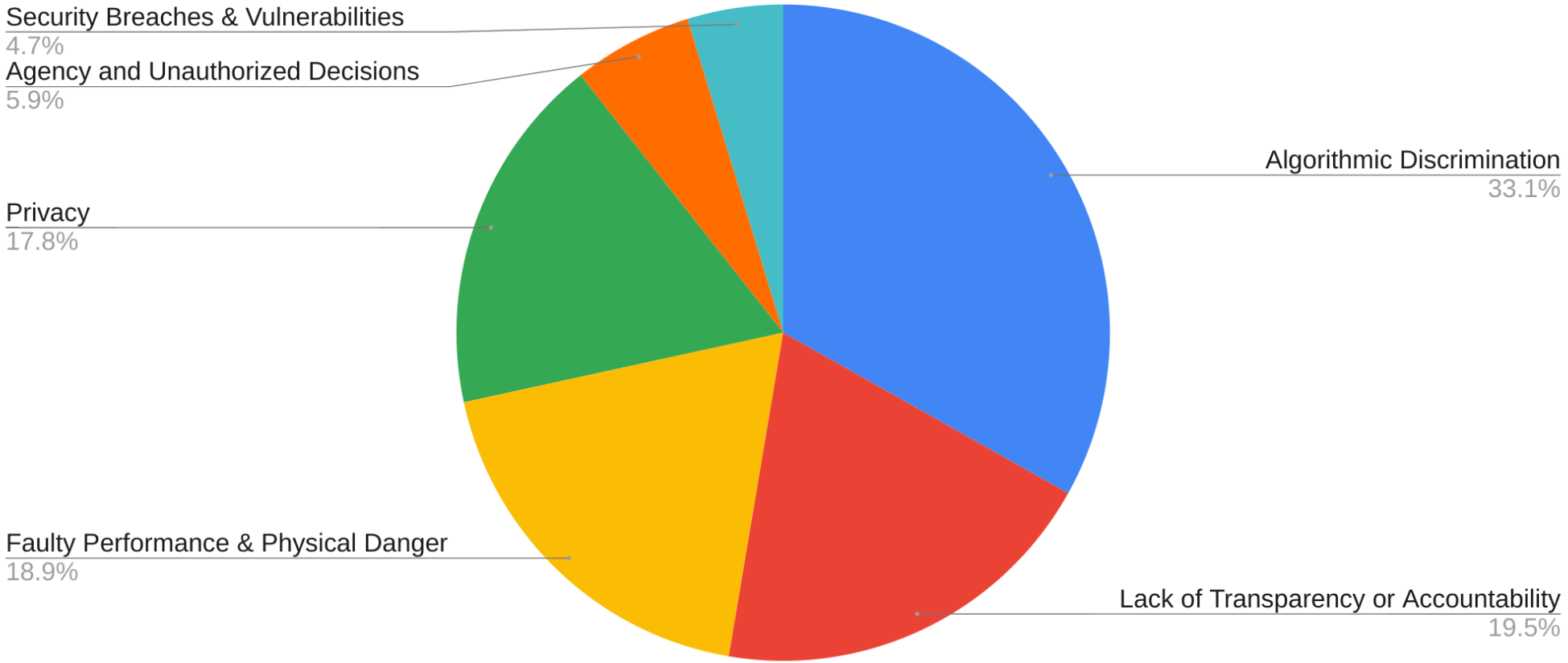
learn more

AI Assurance

Is your AI doing what you think it's doing?

2021 Analysis of 169* AI Incidents

95% were **unintentional** (i.e. they did not involve an attacker)



*** Incidents occurred between 1988 & 2021 and were publicly-reported in the public media.**
This work was done in collaboration with BNH.AI.

2021 WHITE PAPER

When AI Fails

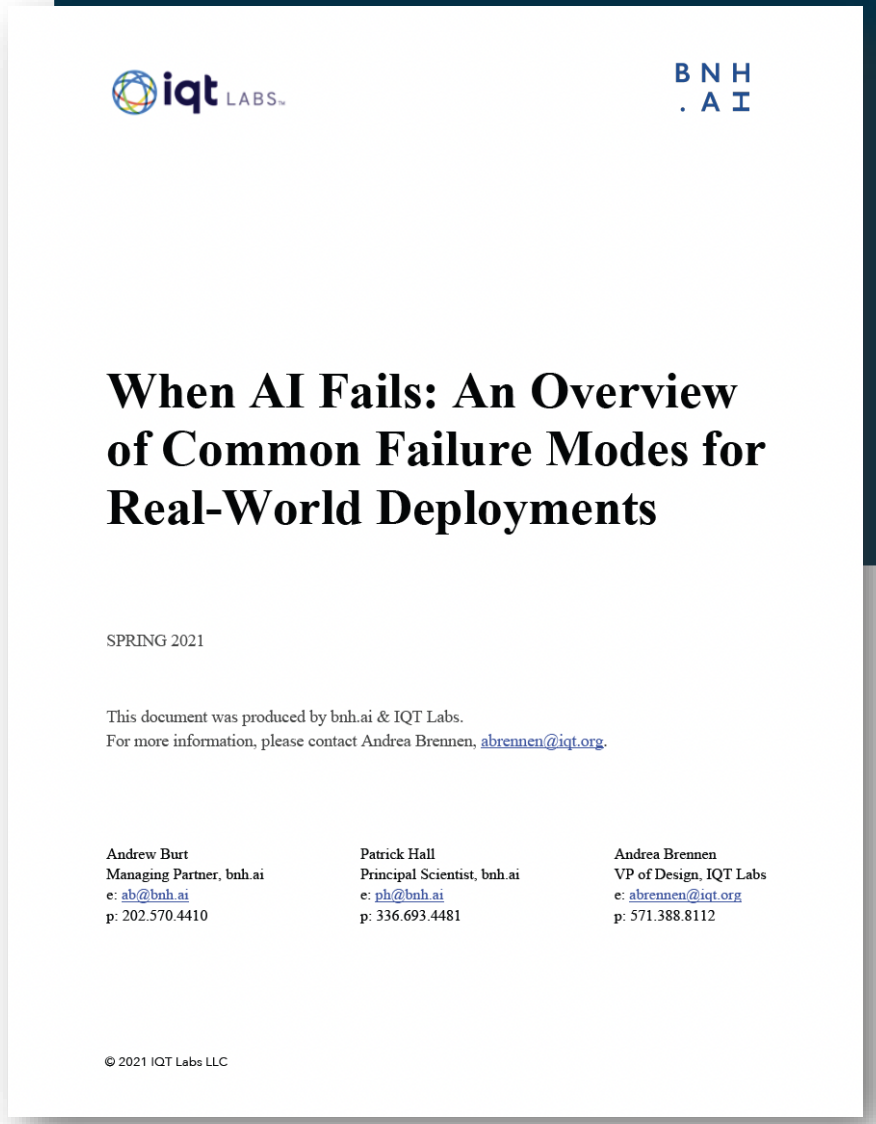
"AI Incident"

Unintentional

- Unauthorized decisions
- Data drift
- Discrimination
- Opaqueness
- User error
- ….

Intentional

- Backdoors
- Data poisoning
- Model extraction
- Membership inference
- Trojans
- ….


iqt LABS

BNH .AI

When AI Fails: An Overview of Common Failure Modes for Real-World Deployments

SPRING 2021

This document was produced by bnh.ai & IQT Labs.
For more information, please contact Andrea Brennen, abrennen@iqt.org.

Andrew Burt
Managing Partner, bnh.ai
e: ab@bnh.ai
p: 202.570.4410

Patrick Hall
Principal Scientist, bnh.ai
e: ph@bnh.ai
p: 336.693.4481

Andrea Brennen
VP of Design, IQT Labs
e: abrennen@iqt.org
p: 571.388.8112

iqt

AI Red-Teaming

- **Time-bound evaluation of risks posed by AI systems.**

- **3 Case Studies**
  - Deepfake detection tool
  - Pretrained LLM
  - Hardware at the Edge

Public reports show how to evaluate AI & what to expect from an AI audit/red-team

AI Tabletop Exercises
Serious Play About AI Risks

In an engaging 60-minute role-play session, participants confront decision-making challenges concerning a (fictional) AI system deployed in a high-stakes scenario.

IQT's Incident Response Tabletop Exercises highlight the importance of preparing for AI failures before they occur.

 |

A data scientist uncovers bias in a deepfake detection tool used to assess the veracity of a video with grave geopolitical consequences….

IQT's Incident Response Tabletop Exercises highlight the importance of preparing for AI failures before they occur.

2024

AI Incident Response Guidebook

- Few organizations are prepared to respond to AI failures.
- IQT's *AI Incident Response Guidebook* helps organizations bridge the gap between traditional (InfoSec) incident response and AI incident response.

We are working with partners across the USG to promote adoption of best practices for AI Incident Response.

What is an AI Incident?

An AI incident is an event caused by undesirable or unexpected behavior of an AI system that directly or indirectly causes immediate or potential harm to people or property.

AI incidents can be **intentional** or **unintentional**.

Prior to deployment, every AI system needs a Short-Term Containment Plan

AI systems are so complicated that it can take weeks to understand the cause when something goes wrong.

A Short-Term Containment Plan describes what to do with the system during this time.

9 Topics an AI Incident Response Policy should address:

1) Who is involved in response activities?
2) Who is responsible for decision-making during an incident?
3) Guidance on AI incident triage
4) SOPs for reproducing and verifying incidents
5) Guidance on surveillance vs. interdiction
6) Chain of custody instructions for affected systems and data
7) Guidance on communications during and after incident
8) Who is accountable for response activities?
9) Who will revise policies & procedures over time?

Andrea Brennen
abrennen@iqt.org