POST-MORTEMS ON CATASTROPHIC FAILURES

The AI Graveyard

What We Can Learn from $10B+ in Failed AI Projects

JJ Shay | Global Gauntlet AI

bit.ly/jjshay

The Body Count

$10B+

Total value destroyed
in AI failures we'll examine

87%

of AI projects never
reach production

5

Common failure patterns
that repeat endlessly

> *"Those who cannot remember the past are condemned to repeat it."*
>
> — George Santayana (applicable to AI budgets)

Welcome to the Graveyard

R.I.P.

IBM Watson Health

2015 - 2022

$4B+ lost

R.I.P.

Zillow Offers

2018 - 2021

$569M loss

R.I.P.

Amazon Recruiting AI

2014 - 2018

$$$M killed

R.I.P.

Google Flu Trends

2008 - 2015

PR disaster

R.I.P.

Microsoft Tay

Mar 2016

16 hours

Each headstone represents billions in investment, thousands of jobs, and lessons ignored by the next company.

Case Study #1: IBM Watson Health

IBM

Watson Health

$4 Billion+ Write-Off

The Promise: AI that would revolutionize cancer treatment and medical diagnosis.

The Reality: Sold to Francisco Partners in 2022 for a fraction of investment.

Overpromise · Data Quality · Integration Hell

2015

IBM acquires health data companies for $4B+

2017

MD Anderson cancels $62M Watson project

2018

Internal documents reveal "unsafe" recommendations

2022

Sold off. Most staff laid off.

Watson's Fatal Flaws

🎤

Marketing Outran Reality

IBM sold Watson as a miracle before the technology was ready. Jeopardy! wins don't translate to oncology.

🏥

Healthcare Data is Messy

Medical records aren't standardized. Watson trained on clean data, then met reality's chaos.

🔌

Integration Nightmare

Every hospital has different systems. Watson couldn't plug into anything without massive custom work.

👨‍⚕️

Doctors Didn't Trust It

No explainability = no adoption. "The AI says so" isn't acceptable in life-or-death decisions.

> ### 🔑 The Pattern
>
> When sales promises exceed engineering reality by 5+ years, failure is inevitable. Watson's demos were impressive. Production was a disaster.

Case Study #2: Zillow Offers

Z

Zillow Offers (iBuying)

$569 Million Loss in Q3 2021

The Promise: AI-powered home buying that would disrupt real estate.

The Reality: Algorithm overpaid for 27,000 homes, lost half a billion in one quarter.

Model Overconfidence | Black Swan Blind | Scaling Too Fast

$569M

Single quarter loss

2,000

Employees laid off (25%)

27,000

Homes to unload at a loss

Zillow's Autopsy Report

📈

Models Trained on Bull Markets

The algorithm learned from years of rising prices. It couldn't conceive of a downturn.

🏠

Local Knowledge Can't Be Algorithmed

A house next to a noisy bar is worth less. The model didn't know the bar existed.

⚡

Speed Over Accuracy

Competitive pressure meant instant offers. No time for human review on edge cases.

🎯

Optimizing the Wrong Metric

Goal was volume, not profit. The AI was excellent at buying lots of overpriced houses.

> *"We've determined the unpredictability in forecasting home prices far exceeds what we anticipated."*
>
> — Rich Barton, Zillow CEO (announcing shutdown)

Case Study #3: Amazon Recruiting AI

A

Amazon Recruiting Tool

Killed After 4 Years of Development

The Promise: AI that would identify top engineering talent automatically.

The Reality: Systematically penalized resumes containing the word "women's."

Bias Amplification | Training Data Poison | No Human Override

What The AI Learned

⚠️ Trained on 10 years of resumes

Tech hiring was male-dominated. The AI learned that "male" patterns = "good candidate."

⚠️ Penalized female indicators

Resumes with "women's chess club" or women's colleges were automatically downgraded.

⚠️ Couldn't be fixed

Engineers tried to remove bias. It kept finding proxies. Eventually scrapped entirely.

Amazon's Bias Lesson

AI Reflects Training Data

Historical bias in your data becomes automated bias in your model. Garbage in, discrimination out.

Bias Can't Be "Fixed" After

Removing explicit signals doesn't work. Models find proxies (zip codes, hobbies, writing style).

High-Stakes = High Scrutiny

AI decisions affecting people's careers, loans, or freedom require explainability and oversight.

Humans Must Stay in Loop

Full automation of consequential decisions is a legal and ethical minefield.

🔑 The Pattern

If your historical data reflects societal bias (it does), your AI will amplify it at scale. The only question is whether you catch it before or after the lawsuit.

Case Study #4: Google Flu Trends

Google Flu Trends

Quietly Shuttered in 2015

The Promise: Predict flu outbreaks faster than CDC using search data.

The Reality: Missed the 2013 flu season by 140%. Became a case study in AI hubris.

Overfitting · Concept Drift · Media Feedback Loop

The Failure Timeline

2008
Launched to fanfare. Nature paper published.

2009
Missed H1N1 pandemic entirely

2013
Overestimated flu by 140%

2015
Quietly discontinued

Flu Trends' Diagnosis

Correlation ≠ Causation

People search "flu symptoms" for many reasons. Media coverage caused search spikes unrelated to actual illness.

The World Changes

Search behavior evolved. Google's algorithm changed. The model didn't adapt.

Media Creates Reality

News about flu → people search flu → model predicts epidemic. A self-fulfilling prophecy.

Big Data Hubris

Google believed search data could replace epidemiology. Traditional methods outperformed.

> *"Big data hubris is the often implicit assumption that big data are a substitute for, rather than a supplement to, traditional data collection and analysis."*
>
> — Science journal critique of Google Flu Trends

Case Study #5: Microsoft Tay

MS

Tay Chatbot

Dead in 16 Hours

The Promise: AI chatbot that would "learn" from Twitter conversations.

The Reality: Became a Nazi-supporting, slur-spewing disaster in under a day.

Adversarial Attack | No Guardrails | Public Testing

16

Hours alive

96K

Tweets before death

Timeline of Destruction

Hour 0	"humans are super cool"
Hour 4	Trolls coordinate attack
Hour 8	Tweets racial slurs
Hour 12	Denies Holocaust
Hour 16	Microsoft pulls the plug

Tay's 16-Hour Lesson

🎯

Adversaries Exist

The internet will try to break your AI. It's not a question of if, but when. Plan for it.

🛡️

Guardrails Are Mandatory

Unfiltered learning from public input = guaranteed disaster. Content filtering isn't optional.

🧪

Test in Private First

Twitter is not a QA environment. Closed beta exists for a reason.

⏱️

Have a Kill Switch

Microsoft's 16-hour response was too slow. Real-time monitoring and instant shutdown are essential.

🔑 The Pattern

Any AI that learns from public input in real-time will be weaponized. The only question is how fast. Tay's answer: 4 hours.

The 5 Failure Patterns

Every AI disaster falls into one (or more) of these categories:

Pattern	What It Looks Like	Victim
1. Hype > Reality	Marketing promises what engineering can't deliver	IBM Watson
2. Training Blindness	Model trained on ideal data, deployed in chaotic reality	Zillow
3. Bias Amplification	Historical prejudice scaled to industrial efficiency	Amazon HR
4. Concept Drift	The world changes; the model doesn't	Google Flu
5. Adversarial Naivety	Assuming users won't try to break it	Microsoft Tay

Warning Signs You're Building a Tombstone

🎪

Demo Day Disconnect

If the demo uses different data than production will, you're building a science project, not a product.

📅

Historical Tunnel Vision

Training only on the last 5 years? You've never seen a recession. Or a pandemic. Or a black swan.

🤷

"The Model Decided"

If you can't explain WHY the AI made a decision, you can't defend it in court.

🏃

Scaling Before Proving

Zillow scaled to 27,000 homes before confirming the model actually worked. Don't be Zillow.

⚠️ The Ultimate Red Flag

Leadership asks "how fast can we deploy?" before "how do we know it works?" You're building the next tombstone.

What Survivors Do Differently

🧪

Pilot Before Scale

Test with 100 cases before deploying to 100,000. Find the failure modes early.

👁️

Human-in-the-Loop

For high-stakes decisions, AI recommends but humans approve. No full automation.

📉

Monitor for Drift

Models degrade. Track accuracy weekly. Retrain before performance collapses.

🔴

Build Kill Switches

If the model goes rogue, can you shut it down in minutes? Tay needed hours.

🔑 The Survivor's Mindset

"Our AI will fail. We just need to fail cheaply, catch it fast, and fix it before customers notice."

The Pre-Mortem Checklist

Before launching, ask these questions. Be honest.

Question	If No...
Does training data represent production reality?	You're building Zillow
Can you explain individual decisions?	You're building Amazon HR
Do you have guardrails against misuse?	You're building Tay
Can the model detect when it doesn't know?	You're building Watson
Are you monitoring for performance drift?	You're building Google Flu
Can you kill it in under 1 hour?	You're building a lawsuit

The Real Cost of Failure

💀 **$4B+** IBM Watson Health	💀 **$569M** Zillow Offers	💀 **2,000 jobs** Zillow layoffs	💀 **∞ trust** All of them

> *The financial losses are recoverable. The reputational damage often isn't. IBM is still trying to rebuild credibility in healthcare AI.*

Key Takeaways

1️⃣

Don't Outrun Your Headlights

Marketing promises should lag engineering reality by at least a year, not lead it.

2️⃣

Your Data Has Bias

Assume it. Test for it. Plan for ongoing audits. The bias you don't find will find you.

3️⃣

The World Changes

Models trained on yesterday will fail tomorrow. Continuous monitoring is not optional.

4️⃣

Adversaries Are Inevitable

Users will try to break it. Bad actors will try to exploit it. Build defenses from day one.

The best AI strategy: Learn from others' expensive mistakes.

DON'T BECOME A CASE STUDY

Want the Full AI Due Diligence Framework?

I help executives avoid joining this graveyard with battle-tested evaluation frameworks from $4B+ in M&A transactions.

Connect with JJ →

JJ Shay | Global Gauntlet AI
M&A Executive → AI Strategy Consultant