bit.ly/safer-operations
Complexity
Compare 2020 with 2010

Developers can do much more with much less
launch entire stack

few lines of code
“Do you think that kind of automation is easy? Or cheap?”
Findings from 2020 Emergence of Big Code Survey

Over half of developers surveyed are working with 100x the volume of code than 10 years ago
Findings from 2020 Emergence of Big Code Survey

Majority of developers report

- 92% Pressure to release changes more quickly
- 90% Software their teams deliver is more critical
- 50% Understanding dependencies is more difficult
- 74% Updating code is more dangerous
Why isn’t this leading to more company-ending outages?
What is driving this increase in complexity?

It’s success
Law of Stretched Systems
Capacities being stretched

+ Organizational workload
+ Pace of development cycle
+ Demands on expertise
+ Speed of technological innovation
Are We Getting Better Yet?
Progress toward safer operations
Indeed is currently unavailable

We should be back online in a few minutes. Thanks for your patience and good luck with your job search.
Network Engineer Sr. - Palmer, Antarctica (Austral Winter 2020)
GHG Corporation
Antarctica

- Strong ability to troubleshoot complex multi-vendor network issues in LAN and WAN networks.
- The Network Engineer Senior provides IT services at the National...
30+ days ago  - More...

Network Engineer, Sr. - McMurdo, Antarctica (Austral Winter 2020)
GHG Corporation
McMurdo Station

- Monitor network performance and reliability to proactively identify potential problems.
- Experience with the configuration and management of network firewalls is...
30+ days ago  - More...
We help people get jobs.
Prioritize a learn and adapt safety mode over a prevent and fix safety mode.
Prevent & Fix
Learn & Adapt
The Prevent & Fix cycle

Focus on safety

Time

Prevent

Fix

Tradeoffs

Surprise

© 2020 Alex Elman
The Learn & Adapt reinforcing loop

- Learn
- Adapt
- Surprise

Getting safer

Focus on safety

Time

© 2020 Alex Elman
Action items form a defensive strategy but **do not lead to learning**
A taxonomy of action-items

+ Emergency work
+ Planned work
  - eventually gets completed
  - never gets completed
+ Undiscovered critical work
Better-preventions
Better-fixes
Measuring progress
It is wrong to suppose that if you can’t measure it, you can’t manage it – a costly myth.

W. Edwards Deming
It is wrong to suppose that if you can’t measure it, you can’t manage it – a costly myth.”

W. Edwards Deming
“The most important figures that one needs for management are unknown or unknowable, but successful management must nevertheless take account of them.”

Lloyd S. Nelson
Can these be measured?

Richer metrics

+ Magnitude of psychological safety
+ Comfort in surfacing risk to leadership
+ Potential $$ losses from incidents that were avoided
+ Net cost/benefit of experts changing teams
+ Amount an organization is learning
Be data-driven
avoid being *driven by data*
“It is a difficult thing to look at a winking light on a board, or hear a peeping alarm … and … draw any sort of rational picture of something happening out in the vast plant, let alone meet it with anything but a mechanical response”

Three Mile Island commission report
Valid Metrics
Everybody has a story to tell
Metrics anchor the story and the story gives meaning to the metrics
Activities around creating and maintaining safety

Preparation

Reporting reliability

Assessing accountability

Incident analysis

Incident write-ups
Activities around creating and maintaining safety

Preparation
Preparation

Emphasizes only avoiding a recurrence

Emphasizes a more accurate and complete understanding
Barriers and guardrails are used to prevent people from repeating mistakes.
Examples of barriers and guardrails

- Turning MySQL Safe Mode on in Production
- Disallowing SSH access
- Capping instance capacity
- Preventing rollbacks without approval
"All practitioner actions are gambles."

Richard I. Cook
https://how.complexsystems.fail/#10
Performance variability
Ensure positive outcomes through activities like team practice and chaos experiments.
Chaos experiments as scrimmage
Practice: safe, predictable
Chaos: safe, unpredictable
Incidents: unsafe, unpredictable
Why are incidents so difficult?
Escape Rooms

Still from Schitt’s Creek (2020), Pop TV/Netflix
Hindsight
Mistakes are a feature not a bug
Stories

+ The tale of the well-choreographed incident response

+ A brand new on-call responders painful experience through the obstacle course to address a stuck deploy

+ The one where an accidental line of YAML burned $250,000 in cloud costs
Humans in the loop
Automation an opportunity to **enhance** or **enable** humans
Activities around creating and maintaining safety

Preparation

Reporting reliability
Reliability outcomes and human performance are predicted and controlled.

Reliability outcomes and human performance are monitored and influenced.
Incidents are a source of insights but not a good measure of reliability.
Reliability
Historically good performance

___

Robustness
Retains good performance within a threshold when challenged

___

Brittleness
Predictably poor performance when challenged
How fast can we safely go in a brittle system?
Service Level Objectives
Service Level Objectives

+ How **reliable** have we been?

+ How **fast** can we go?

+ How fast **should** we go?

+ What is the user **experiencing**?

+ Can we keep relying on FooService?
Control vs Influence
inputs → outputs → outcomes

Circumstances out of our control

circumstances

Things we control

Things we influence

What we want

@_pkill | bit.ly/safer-operations
circumstances

runbooks
training
procedures
fault tolerance

queue delays
timeouts
hypotheses
mitigations

Production stabilized

packet loss
traffic spikes
hardware failures
users
Outcomes over Outputs

Watch the inputs
Influence the outputs
Target the outcomes
Activities around creating and maintaining safety

Preparation

Assessing accountability

Reporting reliability
People who make mistakes are blamed. They are obligated to take responsibility.

People who make mistakes feel supported which inspires them to seek opportunities.
Attribution is important to learning but can also lead to blame.
Opportunity vs Obligation
Opportunities are taken, not given
Opportunities

+ Identifying
  ● Defined goals and rationale

+ Selecting
  ● Career growth
  ● Assumptions and risks

+ Realizing
  ● Definition of “done”
Activities around creating and maintaining safety

Preparation

Reporting reliability

Assessing accountability

Incident analysis
Incident analysis

1. Incidents result only in technical fixes
2. Incidents are investments in more capable organizations
"Focusing on reducing errors diverts energy and attention into ... narrowly targeted ‘fixes’ that treat symptoms but not the underlying problem"

Robert L. Wears
↑ Performance = ↓ errors + ↑ insight generation
Ignoring “above the line” misses at least 50% of the opportunities.

Line of representation

Local-only fixes “below the line”
Improving a process
What are we looking for during incident analysis?
Mickey Oct 21st, 9:24 am
Peiwen what's involved in launching the new instances? I can press the button but don't know how long they'll take to provision puppet. Also wondering if we want to just stop dbs1005 and change its size to i3 (or m4.8xL)

Peiwen Oct 21st, 9:26 am
I think the process is similar to what we do when launching ec2 in cmhqa. But if it's faster, we can change an existing instance. Maybe 1015, since we already have a change going on 1005

Note
Instance changes are faster than launching instances. Is this widely known?
Peiwen  Oct 21st, 9:20 am
while we are waiting, can we get 3 i3en.6xlarge instances going?

Mickey  Oct 21st, 9:22 am
Andrew is that something you want to do or want me to? I'm not sure if we need a proc ticket or can just go for it

Note
Unclear procedures on how to handle change management during an incident.

Andrew  Oct 21st, 9:22 am
i'm too tired to do that, tbh

Note
Notable that Andrew mentioned he stayed up all night earlier in #prod-on-call
Is this in a runbook?

How does the lag impact this ability?

What else can impact the website’s ability?
How are these responders performing?
Judging human performance with **metrics** applies conclusions without context
<table>
<thead>
<tr>
<th>Event Type</th>
<th>Absolute time (CST)</th>
<th>Delta time</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Event start</td>
<td>09/04 10:30:00</td>
<td>-</td>
<td>A/B test turned up to 1%</td>
</tr>
<tr>
<td>Time to Detect</td>
<td>10/22 03:00:00</td>
<td>1M 16d 16h 30m</td>
<td>Product Manager identifies a discrepancy</td>
</tr>
<tr>
<td>Time to Diagnose</td>
<td>10/26 04:30:00</td>
<td>4d 1h 30m</td>
<td>Software dev makes a diagnosis (bug)</td>
</tr>
<tr>
<td>Time to Correct</td>
<td>10/26 04:31:00</td>
<td>1m</td>
<td>A/B test turned off</td>
</tr>
<tr>
<td>Time to Recover</td>
<td>10/26 04:37:00</td>
<td>6m</td>
<td>Prod declared stabilized</td>
</tr>
</tbody>
</table>
Recording performance metrics promotes one perspective over others.
Incident analysis

Richer metrics

+ Number of
  ○ technical fixes inspired by incident analysis
  ○ insights generated per incident analyzed

+ Time spent
  ○ verbally coordinating actions
  ○ launching ec2 hosts
  ○ restarting many instances of a database
Activities around creating and maintaining safety

Preparation

Reporting reliability

Assessing accountability

Incident analysis

Incident write-ups

@_pkill | bit.ly/safer-operations
Incident write-ups favor a particular viewpoint above others.

Incident write-ups faithfully present multiple perspectives.
Interview Debriefing
“How did you become aware of the certificate expiration?”

“Oh, Edvaldo had some magic command that he ran...”
“I was kind of amazed that some of the certificates were pushed out through Puppet. Maybe there’s a gap there—that somebody knows how to push a certificate out but we weren’t aware that the Mongo [database] had to be restarted by a certain time and date.”
“Right from the start when we had to restart everything I knew this was going to be a ton of work ... helping to take care of the number of servers that we have but also to help verify everything.”
“I asked Shira and Dmitri to help split the workload.”
“We both made the mistake of assuming that users from India are served by the Hong Kong DC because that seemed to be the closest one, but it turns out it actually goes to London.”
and the number of errors if you don’t realize what you’re looking at.

Alexander Elman
06:29 So just to make sure that my understanding is the same as yours.

06:44 It was it’s either a rate of it’s the rate of change. It’s how the rate of change is changing over time versus just the instantaneous count at that time.

Eric
06:55 That’s how I understand that the rates is really a derivative, right. So,

Alexander Elman
07:01 So you thought it was instantaneous counts, but it was the rate of changing rate.

Eric
07:06 Right, yeah. So when you look at it, thinking it’s counts and it’s really the rate, like, oh, there’s nothing major. Here it’s, you know, there’s one very small spike at the beginning.

07:16 And then it’s just, you know, sort of just slightly increased and bouncing around and you know it didn’t look didn’t look drastic.

Alexander Elman
07:23 Got it. That makes sense. But yeah, I think that’s a very common, very common thing that people interpret.

Eric
07:31 So, um, other than that. Um, let’s see. I mean I think that’s kind of it, but that’s the thinking perspective.
Debriefing Facilitation Guide

Leading Groups at Etsy to Learn From Accidents

Authors: John Allspaw, Morgan Evans, Daniel Schuenberg

Etsy
Three Traps
In accident investigation

Johan Bergström
Reader, Lund University

https://youtu.be/TqaFT-0cY7U
Traps to avoid

+ Counterfactual reasoning
  ● “She should have waited before restarting...”

+ Normative language
  ● “He lacked an understanding of...”

+ Mechanistic reasoning
  ● “Maintenance would be less risky if we automated this.”
Incident write-ups

Richer metrics

+ Number of
  - distinct write-up document opens
  - attendees to review meetings
  - distinct perspectives represented
  - employees trained using write-up

+ Qualitative survey feedback
  - How was the write-up useful?
Ask deeper questions
Our system has been very reliable over the past few quarters. Why?
How close to the **safety boundary** is the pod autoscaler pushing my infrastructure?
Are my cloud provider’s staff a team player in my sociotechnical system?
Recap
+ Deeper understanding leads to better fixes and enduring prevention
+ Reliability is reported using SLOs not incidents metrics
+ Nobody has control over how an incident unfolds
+ Incidents are an opportunity to improve the accuracy of mental models
+ At least half of incident analysis should focus on human factors
+ Comparative storytelling enhances learning
bit.ly/safer-operations