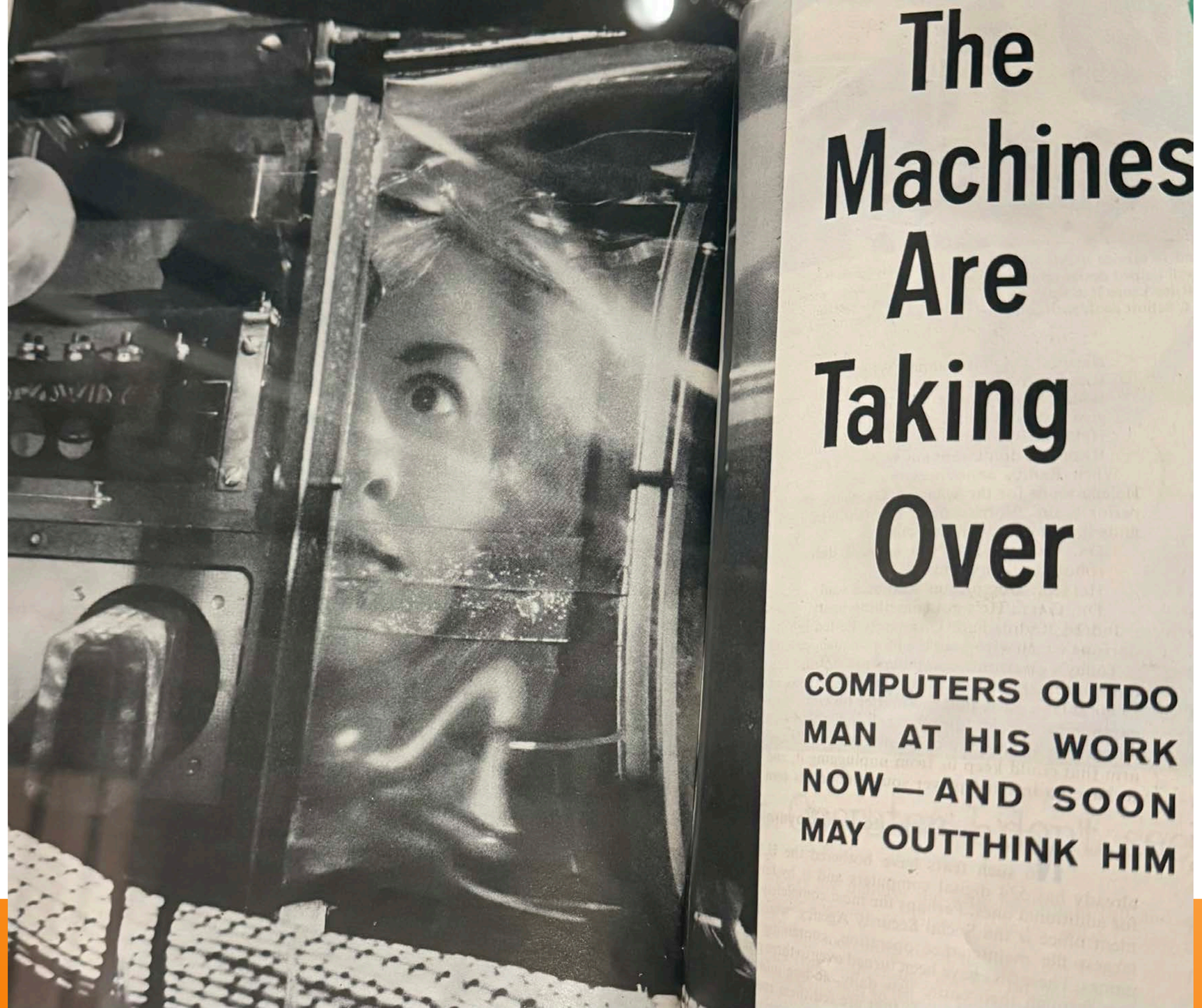


The Machines Are Taking Over

COMPUTERS OUTDO MAN AT HIS WORK NOW — AND SOON MAY OUTTHINK HIM

When AI Agents Misbehave and Fail to Coordinate: Architecting for Joint Activity

David D. Woods woods.2@osu.edu

How to design AI/software agents to be collaborative players in a joint activity space when disruptions arise?

Disclaimer: No AI or other algorithms were used in the preparation these slides (except spellcheck - which made mistakes)

De-stabilizing forces at scale

multiple destabilizing forces are spawning crises
that demand adaptation across sector, national, and societal scales.

De-stabilizing forces at scale

multiple destabilizing forces are spawning crises
that demand adaptation across sector, national, and societal scales.

Benchmark failure
AWS outage October 20, 2025
wide swath outage in layered & tangled infrastructures

New risks

Wide Swath outages: 1 breakdown at 1 layer for 1 unit
produces secondary malfunctions & loss of critical services
for many "unrelated" others

The Future is already here

3 more in @ 1 month: AWS, then Azure, Cloudflare, again

De-stabilizing forces at scale

current trajectory is racing toward safety and systems engineering ***malpractice***

Context of Processes of Growth, Complexification & Adaptation
(GCA, not CAS)

https://resiliencefoundations.github.io/video-4-the-science-and-pragmatics-of-re-through-the-lens-of-complexification.html

How to design AI/software agents to be collaborative players
in a joint activity space when disruptions arise?

De-stabilizing forces at scale

multiple destabilizing forces are spawning crises
that demand adaptation across sector, national, and societal scales.

current trajectory is racing toward safety & systems engineering ***malpractice***

We have to adapt

but what *direction*?

Re-trenchment? or **Re-vitalization?**

Retreat to cope locally? or Re-prioritize/Reconfigure?

The Silicon Valley Way: Move fast and break…aviation safety?

By David Woods, Mike Rayo, Shawn Pruchnicki | May 29, 2025

A seemingly unending stream of incidents, close calls, and fatal accidents has challenged confidence in aviation safety. Image: ErsErg via Adobe Stock

CSEL

SNAFU CATCHERS

De-stabilizing forces at scale

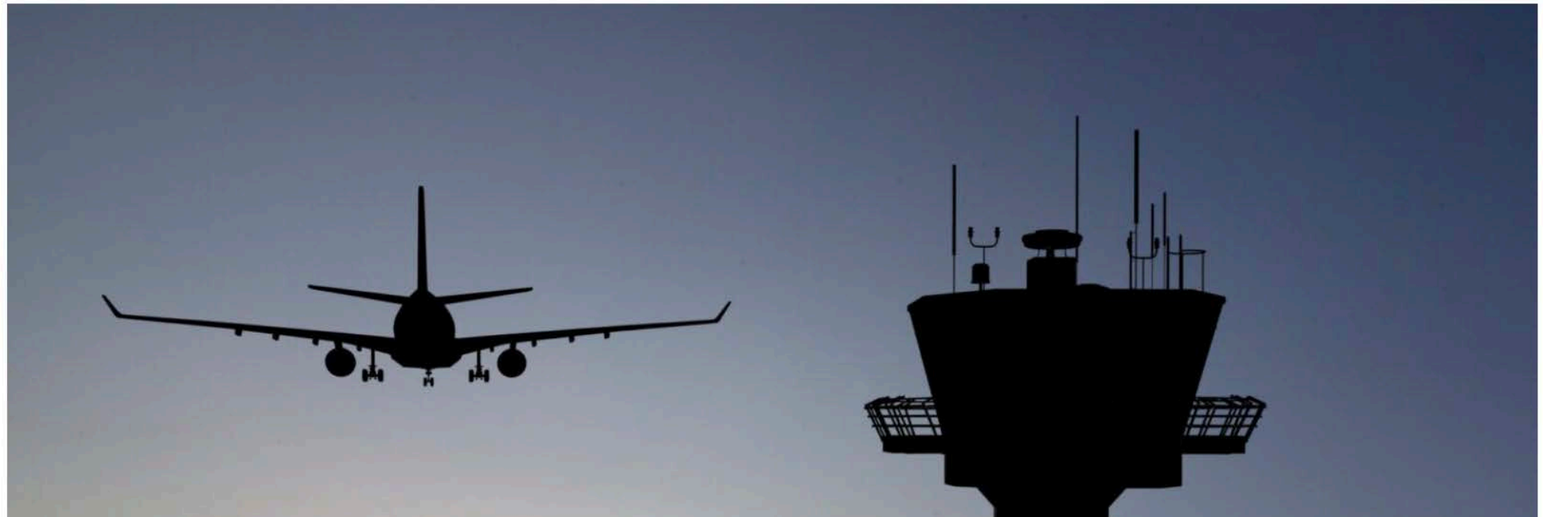
#1 mantra of Silicon Valley: Move Fast and Break Things (MFBT)

versus

mantra of Proactive Safety: create foresight about the changing shape of risk before harm occurs

The Silicon Valley Way: Move fast and break…aviation safety?

By David Woods, Mike Rayo, Shawn Pruchnicki | May 29, 2025



CSEL

De-stabilizing forces at scale

multiple destabilizing forces are spawning crises
that demand adaptation across sector, national, and societal scales.

2 Pressure from aspiration for large scale benefits
if only we can deploy more autonomous systems

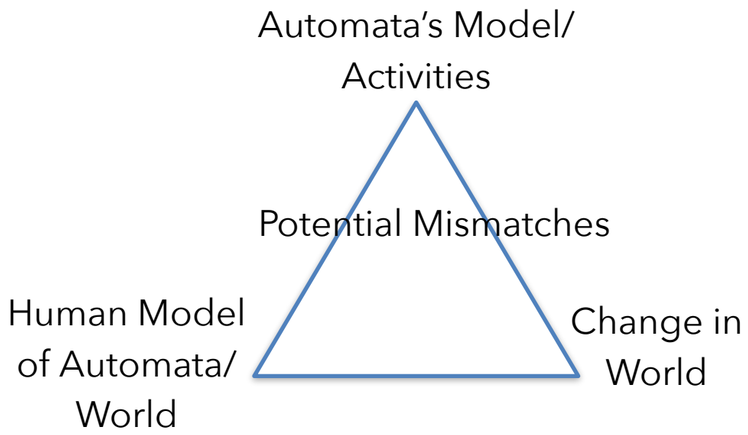
AI/software actors will *misbehave* as people delegate more authority to more capable machines/algorithms/software.

More capabilities mean larger consequences when software agents misbehave

"Strong and Wrong" Dekker and Woods 2023

risks of *literal-minded* automated agents—a system that can't tell if its model of the world is the world it is actually in

system will do the "right" thing—its actions are appropriate given its model of the world, but it is actually in a different world—producing unexpected/unintended behavior and potentially harmful effects.



Huge Premium on ***Re-framing***

De-stabilizing forces at scale

multiple destabilizing forces are spawning crises
that demand adaptation across sector, national, and societal scales.

3 all modern systems are distributed and layered,
thus all challenges are ones of coordination & synchronization
(Hochstein, 2024)

Architecting/designing/testing for ***joint activity*** in a space of interdependent roles,
multiple human and machine/software actors, multiple threads of activities
intertwined over time, across multiple tangled layers

De-stabilizing forces at scale

3 all modern systems are distributed and layered,

thus all challenges are ones of coordination & synchronization

(Hochstein, 2024)

The Future is Already Here

systems where breakdowns in joint, distributed activity lead to failures of critical services

& observe how a stream of incidents that threaten outages are well-handled

(see stella.report https://snafucatchers.github.io/ and https://www.youtube.com/watch?v-fbwDnpuys7w

Benchmark viability threatening failure

AWS outage October 20, 2025

wide swath outage in layered & tangled infrastructures

De-stabilizing forces at scale

current trajectory is racing toward safety and systems engineering *malpractice*

Context of Processes of Growth, Complexification & Adaptation

https://resiliencefoundations.github.io/video-4-the-science-and-pragmatics-of-re-through-the-lens-of-complexification.html

Re-vitalization Directions

How to design AI/software agents to be collaborative players
in a joint activity space when disruptions arise?

Re-vitalization Directions

Architecting/designing/testing for ***joint activity*** in a space of interdependent roles, multiple human and machine/software actors, multiple threads of activities intertwined over time, across multiple tangled layers

How to design AI/software agents to be collaborative players in a joint activity space when disruptions arise?

How to address the cross-scale dimensions of the challenges? Expanding to sector, national, and societal scales

Re-vitalization Directions

5 positive developments to architect highly adaptive layered networks of multiple human & software agents to overcome risk of miscoordination

1 Study: how did automation behaviors contribute to, hinder, or help during the evolution of incident/accident?

2 Managing a dynamic configuration of machine agent activities in pace with changing world, given risk, uncertainties

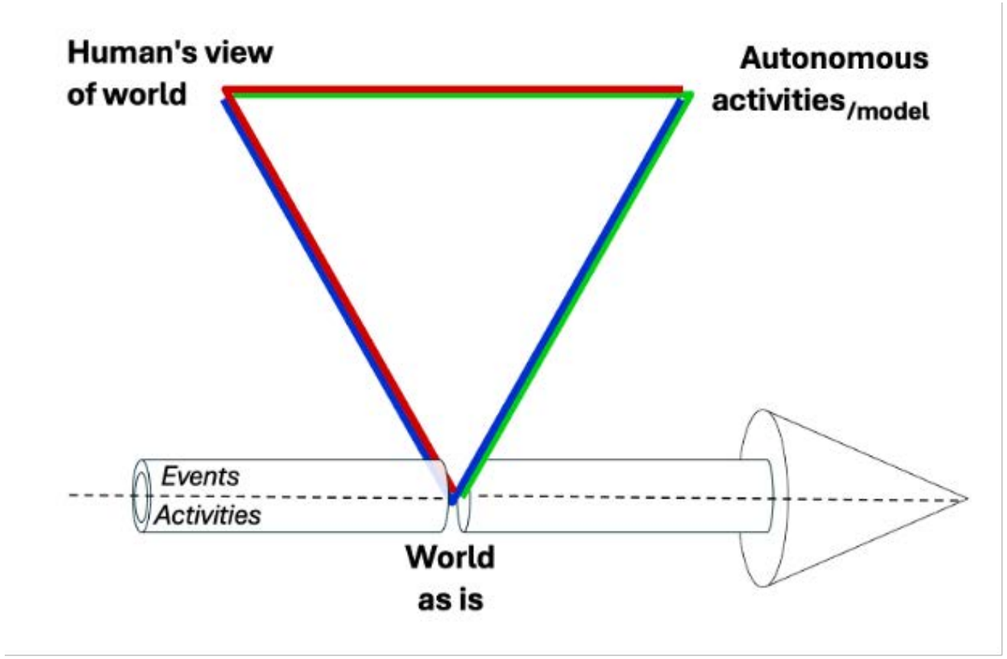
3 What's Next Diagrams—Looking ahead to track what software agents in a suite will do next

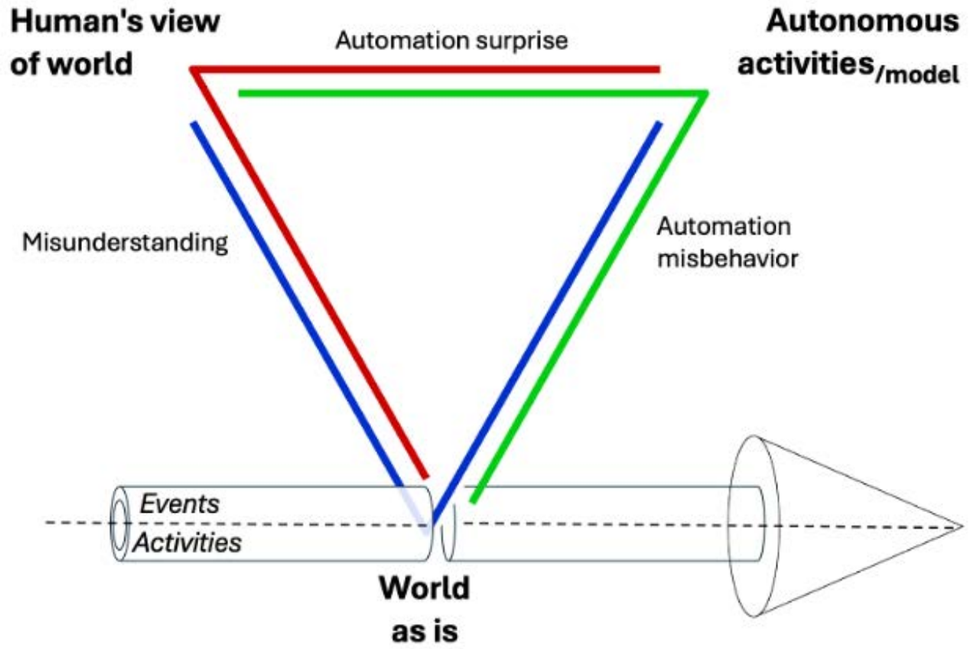
4 Capacity for Maneuver—Adapting ahead of approach to saturation

5 Joint activity testing & modeling—how well can agents coordinate when ongoing organized activities are disrupted by events.

Coordination / miscoordination in network of multiple human & software agents

Study: Rayo et al., 2025: How Automation Behaviors Helped and Hindered Abnormal Operations: Re-analysis of Eighteen Aviation Incidents and Accidents





Automation centered process tracing
How mismatches develop & are resolved between human, machine roles & actual situation

Finding 1: Automated systems can interfere with the work of the flightcrew and/or other automated systems.
Finding 1.1: Automated systems can constrain, counteract, and/or overrule flightcrew attempts to resolve degrading situations.
Finding 1.2: Automated systems can obscure the misbehavior or misconfiguration of other automated systems.
Finding 1.3: Automated systems compensating for the misbehavior of other automated systems can progress aircraft to unusual states which are even more difficult for flightcrews to recover from.

Finding 2: Automated failure mitigations, especially those that mitigate the failure of other automation, can fail and complicate flightcrew responses.
Finding 2.1: Misses from automated failure mitigations can delay, mislead, and/or impede flightcrew diagnosis of dangerous situations.
Finding 2.2: False alerts from automated failure mitigations can induce inappropriate flightcrew actions.

Finding 3: Interdependencies between automated systems can propagate failures, aggravate misbehaviors, and introduce new potentials for failure.
Finding 3.1: Multiple (direct or indirect) paths can couple information automation to control automation
Finding 3.2: Interconnections between automated systems can propagate otherwise isolated failures to other automated systems.
Finding 3.3: The complexity of the automation suite can obfuscate couplings that render designed redundancies as if they were a common mode or single point of failure.

Finding 4: Diagnosing misbehaviors can be increasingly difficult for flightcrews with increasingly complex automation.
Finding 4.1: Cues available to flightcrews can be ambiguous, unreliable, or otherwise insufficient to diagnose automation misbehaviors.
Finding 4.2: Alert overload can complicate and/or obscure critical cues needed for diagnosing automation misbehaviors.

Finding 5: Managing the configuration of automation can be challenging and/or burdensome for the flightcrew to keep pace w/ events.
Finding 5.1: Who and/or what has authority for what functions of the aircraft in the current configuration of automation can be ambiguous to flightcrews.
Finding 5.2: Changes in the configuration of automation can be hidden or poorly communicated to flightcrews.
Finding 5.3: Difficulties in managing and changing the automation configuration can result in bumpy, large, and late transfers of control.

Finding 6: Flightcrew responses to automation misbehaviors are constrained by temporal factors.
Finding 6.1: Tempo can complicate flightcrew responses to automation misbehaviors.
Finding 6.2: Automation misbehaviors can escalate the tempo of situations.
Finding 6.3: Automation misbehaviors can unexpectedly occur at high-tempo phases of flight when time available to diagnose misbehaviors is less than expected.

Theme 1: Increasing use and addition of high-authority and high-autonomy automation exacerbates system interdependencies.

Theme 2: Erroneous sensor data and/or faulty automated system logic and automated system algorithms can produce automation misbehaviors that are difficult for the flightcrew to understand and resolve.

Theme 3: Interdependencies increase the potential impact of erroneous sensor data and/or faulty automated system logic and automated system algorithms.

Theme 4: Flightcrew play a central role in resolving automation misbehaviors by reconfiguring the suite of automation.

Theme 5: The capabilities and interconnectedness of new technologies can blur or modify previous engineering distinctions.

Theme 6: People monitoring automation that is monitoring automation creates new observability demands.

Theme 7: Configuration management of automation suite is central to the role of the flightcrew.

Architect highly adaptive layered networks

Directions come from the scientific developments that underpin Resilience Engineering

Foundational theorems / Theory of Graceful Extensibility (Woods 2018)

https://resiliencefoundations.github.io/overview.html

https://resiliencefoundations.github.io/video-4-the-science-and-pragmatics-of-re-through-the-lens-of-complexification.html

Processes of growth, complexification, adaptation (GCA) play out in lawful patterns.

Messiness is "conserved" — a la the No Free Lunch & Robust Yet Fragile & more theorems — over changes that aspire to 'improve' systems.

The Messy 9 is heuristic to map the formal results to pragmatic action in specific areas.

"Messy 9"heuristic

Congestion, Cascades, Conflict

Saturation, Lag, Friction

Tempos, Surprises, Tangles

2 Managing a dynamic configuration of machine agent & human activities in pace w/ changing world, given risk & uncertainty

current state; high demand on human roles; minimal support
see Rayo et al., 2025 study

Managing the configuration of automata to match changing situations
is itself Joint & Distributed Activity

Matters most when the Messy 9 are in play

Architecting highly adaptive layered networks

3 What's Next Diagram—Looking ahead to track what software agents will do next, given other ongoing events & activities

- Necessary to keep pace with changing tempo of events
- Representation plus computation techniques

Local-out perspectives

Regional-around perspectives

Dynamic Field of Interest in time

Custom examples but no tooling/tools in architectures that scale

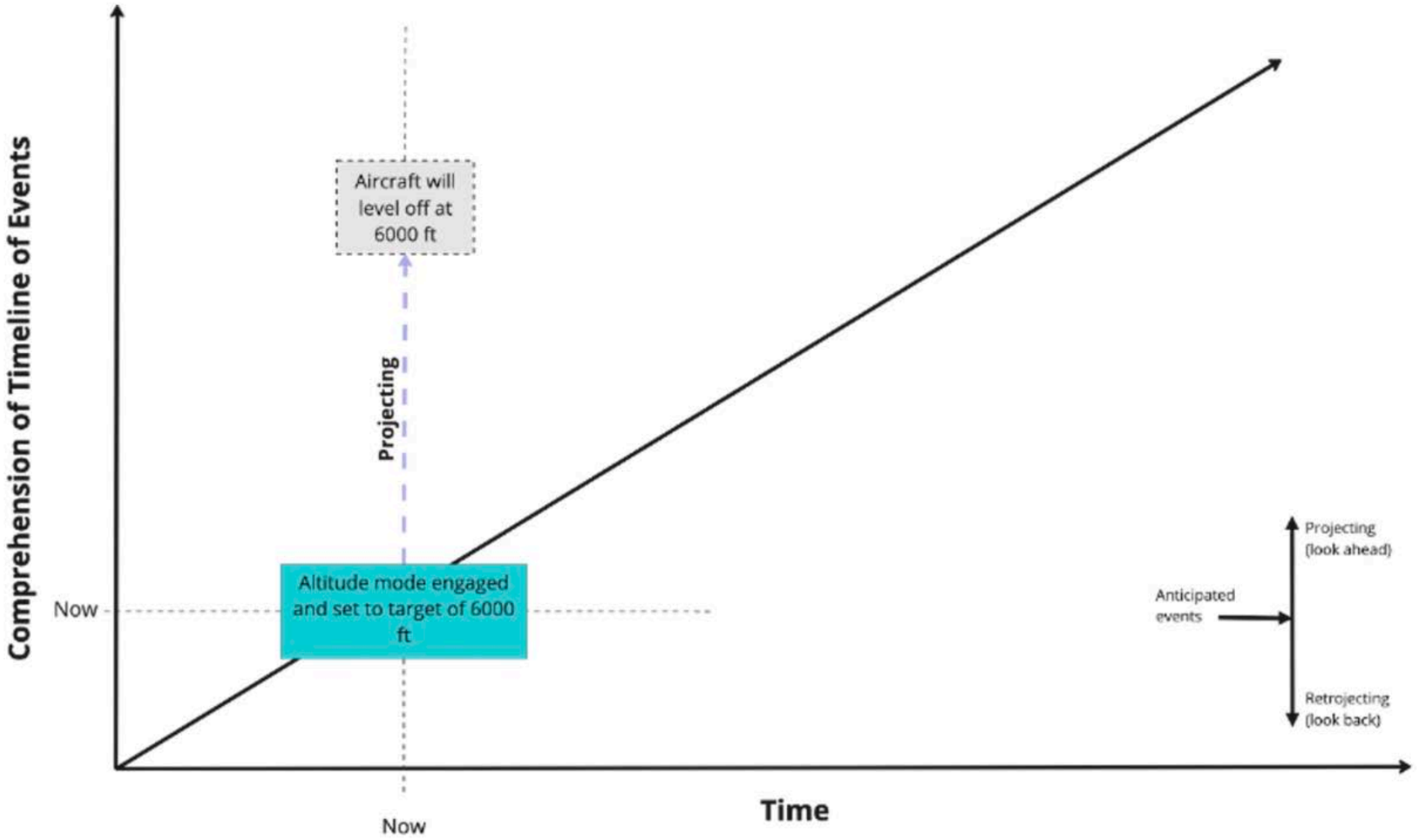


Fig. 4 Flightcrew projection of what will happen next based on previous events

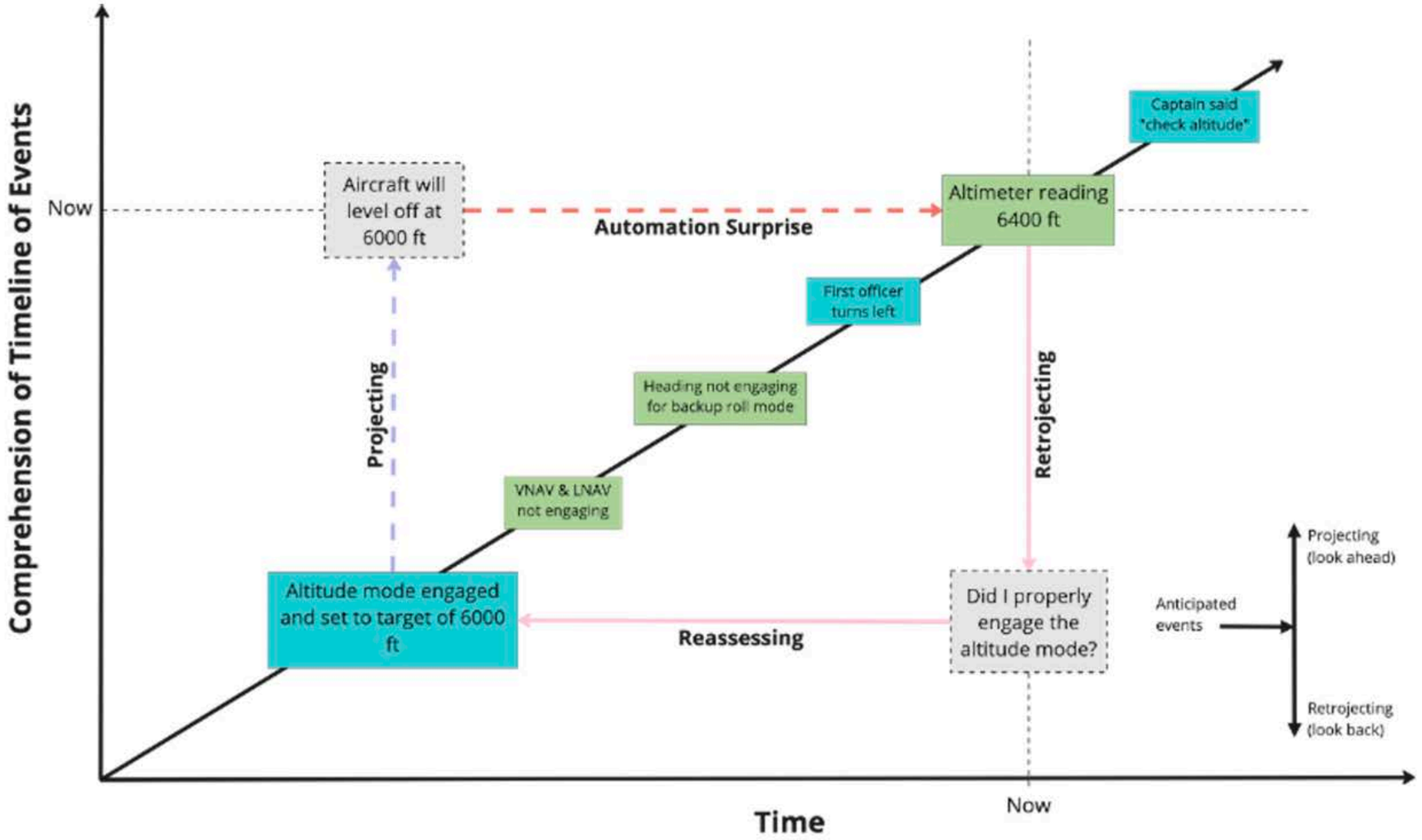


Fig. 5 Retrojective behavior of flightcrew reflecting back to uncover why the automation surprise occurred

Architecting highly adaptive layered networks

4 Adapting ahead of approach to saturation

mitigates the risk of brittleness in all agents at all scales

Capacity for Maneuver from the Theory Graceful Extensibility (Woods, 2018).

Extensibility & Reciprocity in adaptive layered networks

mathematical & realistic demonstrations at 2 scales

- flight deck automation (Farjadian, et al., 2021),
- layered networks of critical digital services https://www.youtube.com/watch?v-fbwDnpuys7w

4 Joint activity modeling & testing

assessing how well can agents coordinate
when ongoing organized activities are disrupted by events
(and then resume despite lingering effects).

Work Models that Compute, IJtsma
Joint Activity Design Heuristics Rayo & Morey

https://ai-frontiers.org/articles/how-ai-can-degrade-human-performance-in-high-stakes-settings

AI Frontiers

How AI Can Degrade Human Performance in High-Stakes Settings | AI Frontiers

Dane A. Morey, Jul 15, 2025 — Across disciplines, bad AI predictions have a surprising tendency to make human experts perform worse.

(257 kB)

4 Joint activity modeling & testing

AI developers over-estimate algorithm competence &
underestimate challenges of anomalies, exceptions, surprises in real world

Study Findings:

- mix of helpful and harmful impacts of AI augmentation as problem difficulty varied
- AI explanation did not improve joint performance.

https://ai-frontiers.org/articles/how-ai-can-degrade-human-performance-in-high-stakes-settings

AI Frontiers

How AI Can Degrade Human Performance in High-Stakes Settings | AI Frontiers

Dane A. Morey, Jul 15, 2025 — Across disciplines, bad AI predictions have a surprising tendency to make human experts perform worse.
(257 kB)

The real stories of technology change are human stories
about *growth, complexification, adaptation* in human systems.

Stories of technology change capture or envision the new forms of
congestion, cascade, conflict, …, that arise when apparent benefits get hijacked.

The forces at work are producing ***instabilities*** across multiple scales

From societal scale to individual roles struggling under pressure

Steps toward Architecture for highly adaptive joint, distributed, layered systems

Challenges ahead for us

Layered

Tangled interdependencies

Cross-scale effects

Circular dependencies / Strange Loops (see stella.report)

Tools

Tooling

Sector / National /Global scales

2 comprehensive formal theories

Doyle et al. Diversity Enabled Sweet Spots DeSS

Theory of Graceful Extensibility TGE