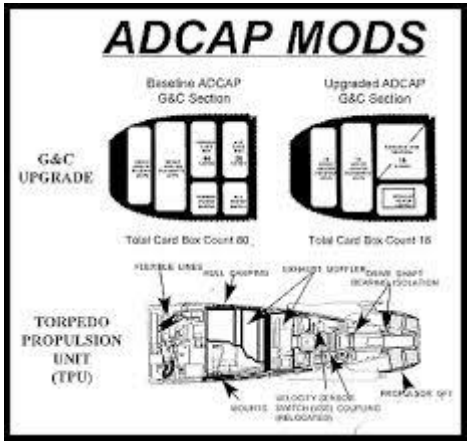


A Powerful New Approach to Safety Engineering

Prof. Nancy G. Leveson

Aeronautics and Astronautics
MIT




ADCAP MODS
Baseline ADCAP
G&C Section
Upgraded ADCAP
G&C Section
G&C
UPGRADE
Total Card Box Count 60
Total Card Box Count 16
TORPEDO
PROPULSION
UNIT
(TPU)

Bottom Line Up Front (BLUF)

- Complexity is reaching a new level (tipping point)
  - Old safety approaches becoming less effective
  - New causes of losses appearing (especially related to use of software and autonomy)

- Traditional analysis approaches do not provide the information necessary to prevent losses in these systems

- Need a paradigm change to a "systems approach"

  Change focus:

  ~~Increase component reliability (prevent failures)~~

  Enforce safe system behavior (constraints on system behavior)

BLUF (2)

- Allows creating new analysis and engineering approaches
  - More powerful and inclusive
  - Orders of magnitude less expensive
  - Work on extremely complex systems
  - Help to design safety, security, and other properties in from the beginning
- New paradigm works much better than old techniques:
  - Empirical evaluations and controlled studies show it finds more causal scenarios (the “unknown unknowns”)
  - Can be used before a detailed design exists to design safe and secure systems from the beginning

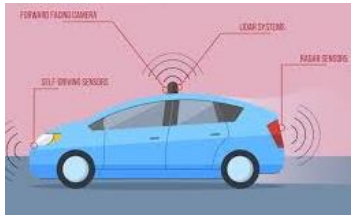
General Definition of “Safety”

Accident = Mishap : Any undesired and unplanned event that results in a loss

- Loss of human life or injury
- Property damage,
- Environmental pollution,
- Mission loss,
- Loss of protected information,
- Negative business impact (damage to reputation, etc.), etc.

Includes inadvertent and intentional losses: so security included

Our current safety tools are all 50-75 years old but our technology is very different today



1940 1950 1960 1970 1980 1990 2000 2010 2020

FMEA

FTA HAZOP ETA

STPA

Assume accidents caused by component failures (including human errors)

- **Introduction of computer control**
- **Exponential increases in complexity**
- **New technology**
- **Changes in human roles**

Traditional Approach to Safety Engineering

- Assumes accidents caused by chains of failure events
- Forms the basis for most safety engineering and reliability engineering analysis:

  FTA, PRA, FMEA/FMECA, Event Trees, FHA, etc.

- Evaluate reliability of components separately and later combine analysis results into a system reliability value
  - Assumptions: losses caused by component failure, independence, randomness---do software and humans behave this way?.
- Design (concentrate on dealing with component failure):
  - Redundancy and barriers (to prevent failure propagation),
  - High component integrity and overdesign,
  - Fail-safe design,
  - (humans) Operational procedures, checklists, training, ….

What Failed Here?

- Navy aircraft were ferrying missiles from one location to another.
- One pilot executed a planned test by aiming at aircraft in front and firing a dummy missile.
- Nobody involved knew that the software was designed to substitute a different missile if the one that was commanded to be fired was not in a good position.
- In this case, there was an antenna between the dummy missile and the target so the software decided to fire a live missile located in a different (better) position instead.

Warsaw A320 Accident

- Software protects against activating thrust reversers when airborne
- Hydroplaning and other factors made the software think the plane had not landed
- Pilots could not activate the thrust reversers and ran off end of runway into a small hill.

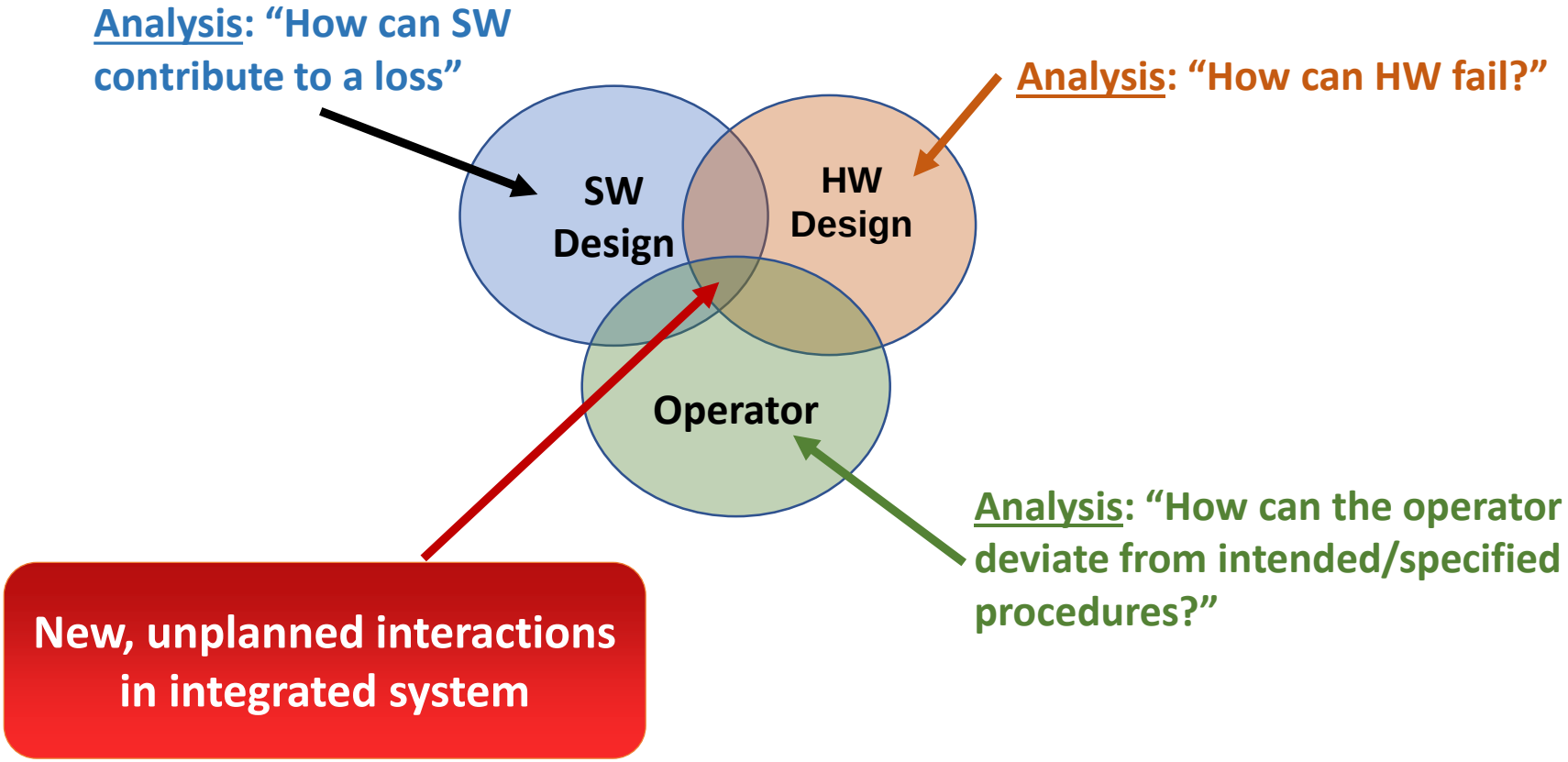
Human factors concentrates on the “screen out”

Hardware/software engineering concentrates on the “screen in”

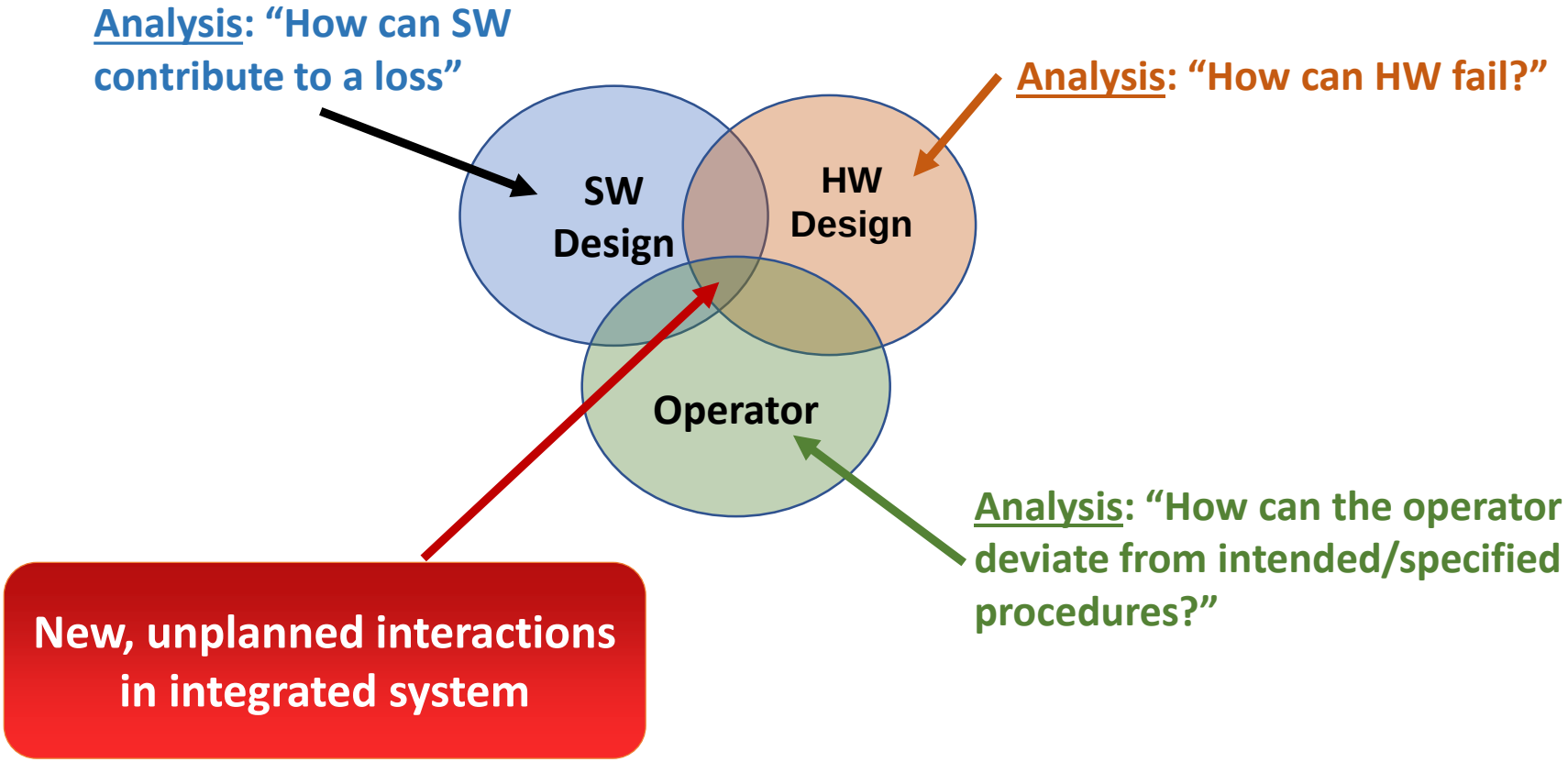
Not enough attention on integrated system as a whole

(e.g, mode confusion, situation awareness errors, inconsistent behavior, etc.

Easy to overlook the system problems when break up system analysis problem



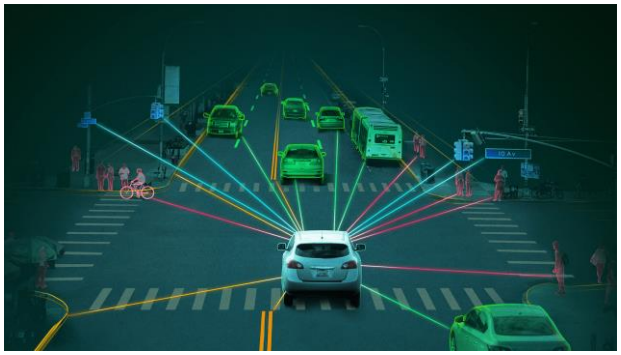
Easy to overlook the system problems when break up system analysis problem

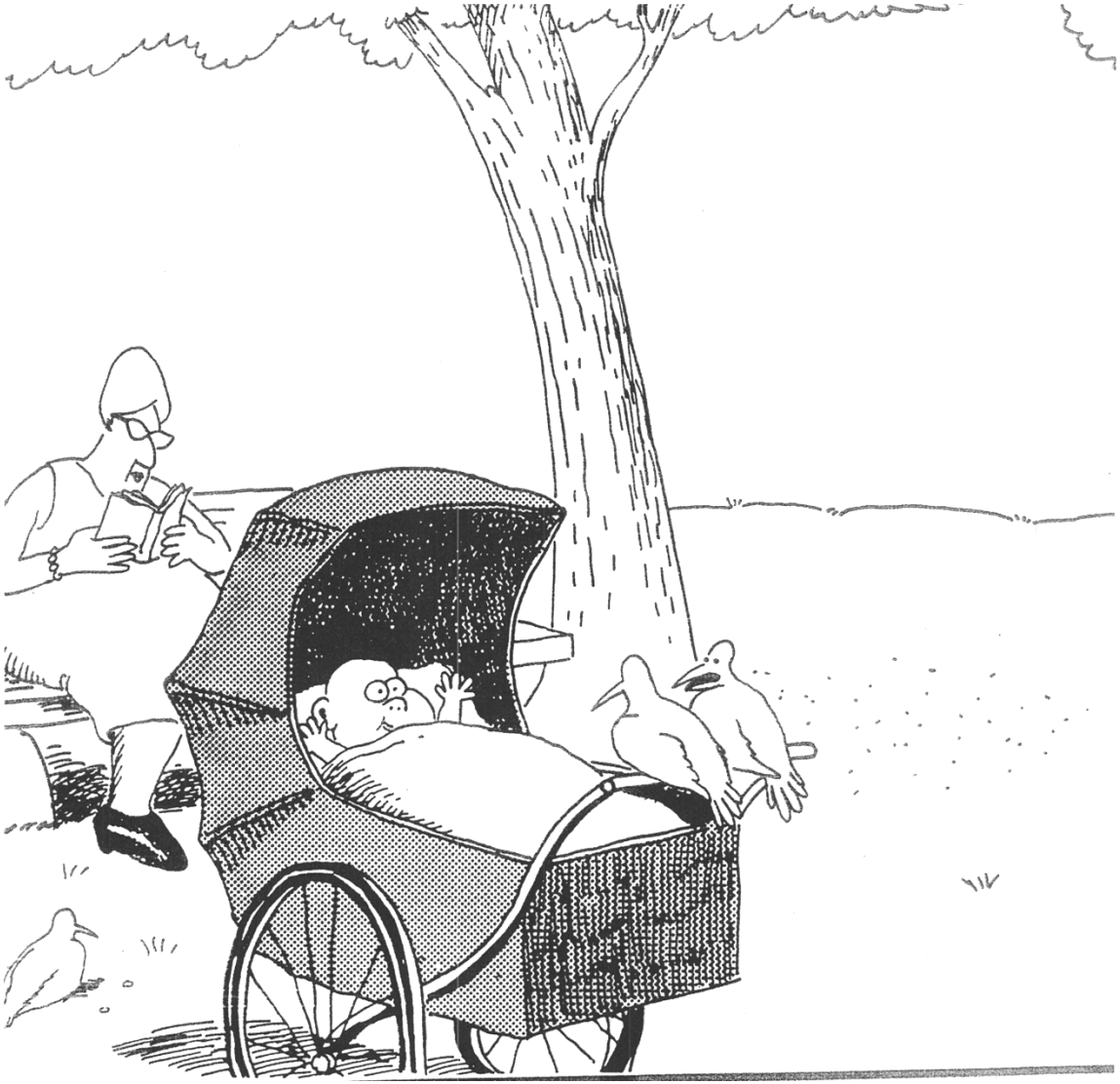


Need to look at integrated system as a whole

The Problem

- Traditional safety approaches do not work on today's systems
  - Don't handle complex systems, software, new roles for humans, management, social systems
  - Start too late – need a design first
  - Hardware, humans, software all treated separately
- No way to extend them as the underlying assumptions do not fit today's systems
- We need a paradigm change





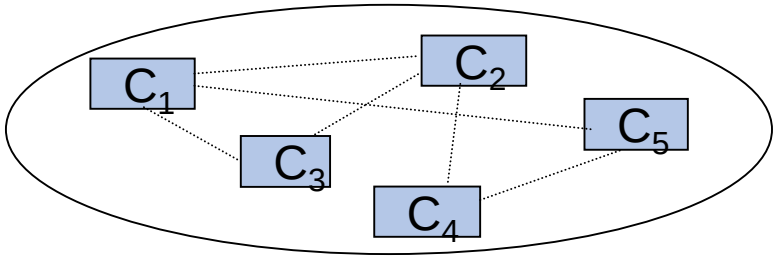
It’s still hungry … and I’ve been stuffing worms into it all day.

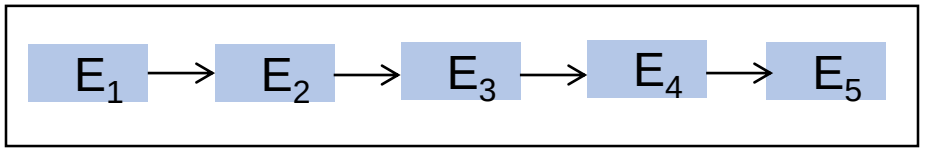
The Problem is Complexity

Ways to Cope with Complexity

- Analytic Decomposition
- Statistics
- Systems Theory

Analytic Decomposition





Analyze/examine pieces separately and combine results

- Assumes such separation does not distort phenomenon
  - ✓ Each component or subsystem operates independently
  - ✓ Components act the same when examined singly as when playing their part in the whole
  - ✓ Components/events not subject to feedback loops and non-linear interactions
  - ✓ Interactions can be examined pairwise

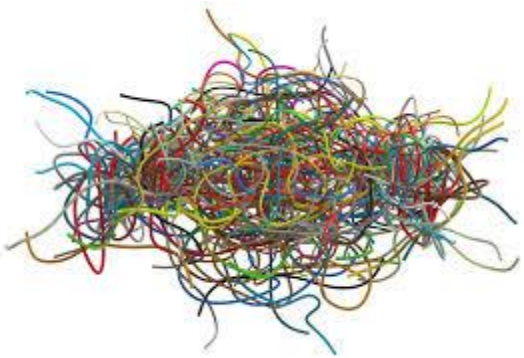
Limitations of Probabilistic Risk Assessment

- Failures of components must be independent
- Doesn't work for non-failure accidents (caused by system design errors and not component failures)
- Doesn't work for software or new technology or new designs
- Doesn't work for human errors in complex systems
- Unreliable results
  - Two scientific evaluations (1980s and 2002)
  - Both showed variance in results of 3-4 orders of magnitude
- Empirical results are terrible: All accidents I have seen had a PRA that showed they could not happen

The Problem

- The independence and simple interaction assumptions are no longer true in our
  - Tightly coupled
  - Software intensive
  - Highly automated
  - Connected

  engineered systems



- We need a paradigm change

Here comes the paradigm change!

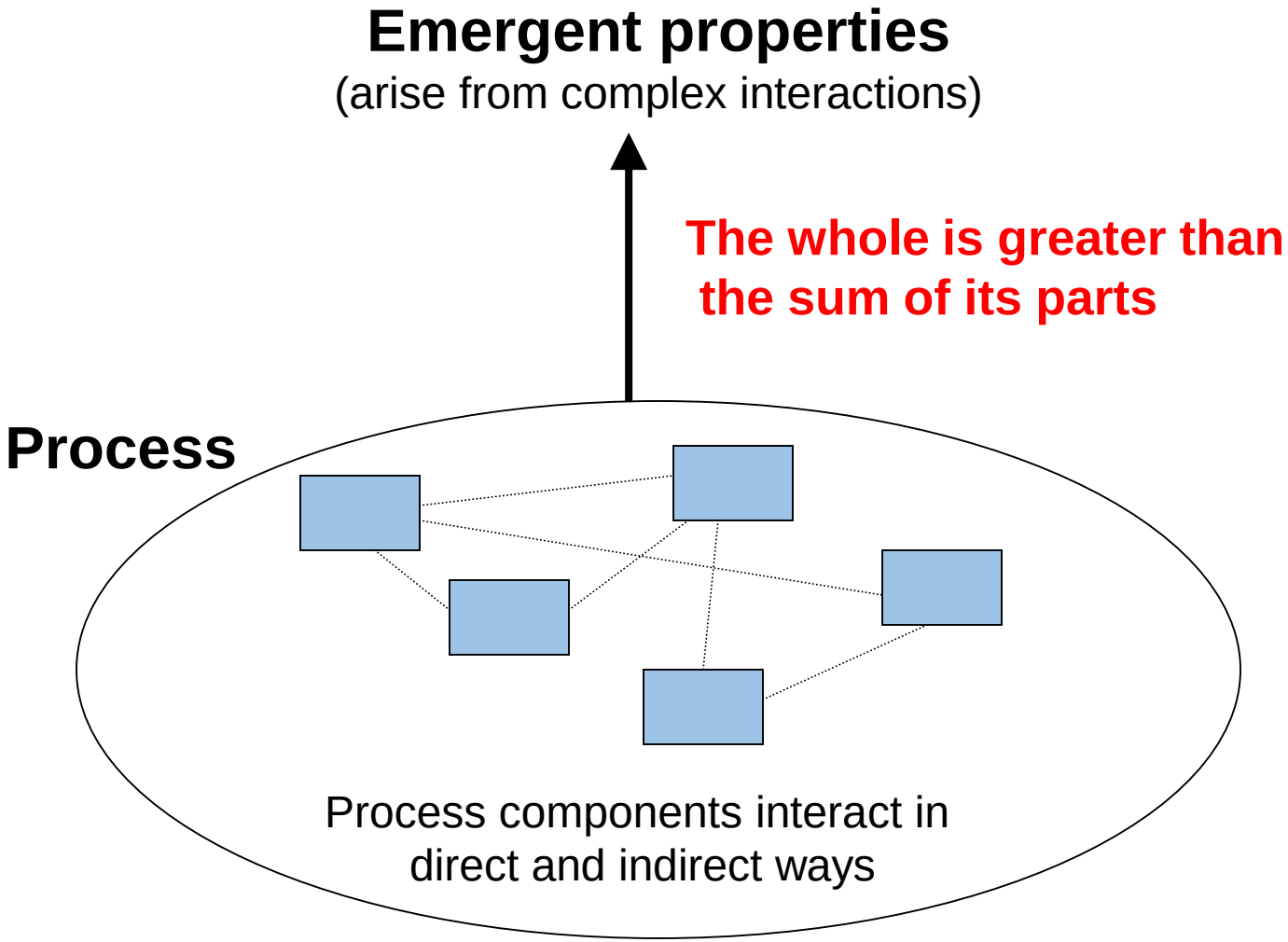
Prevent failures or errors

Enforce constraints on behavior:
- **components**
- **interactions among components**

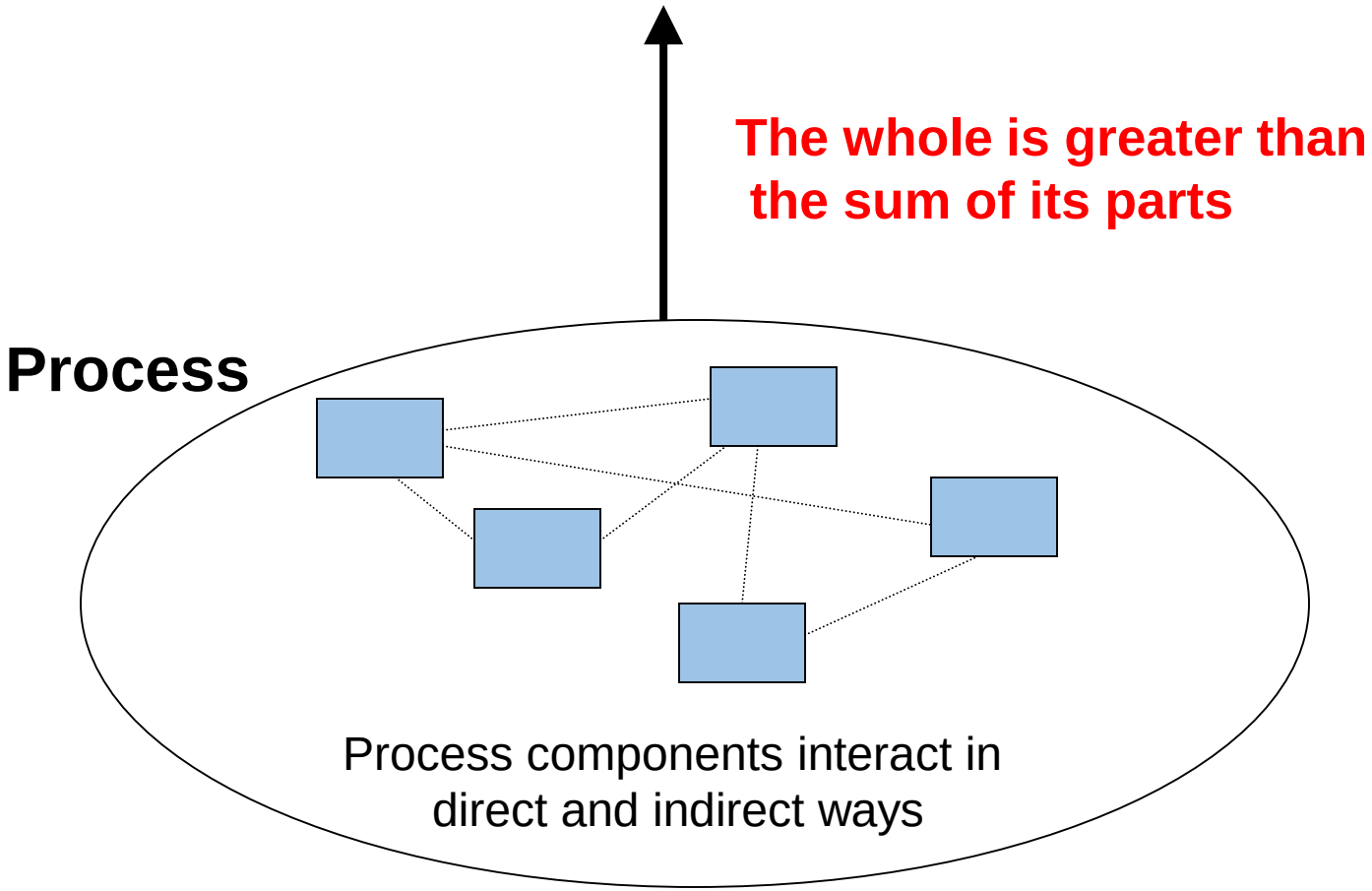
Treat Safety as a **Reliability** Problem

Treat Safety as a **Control** Problem

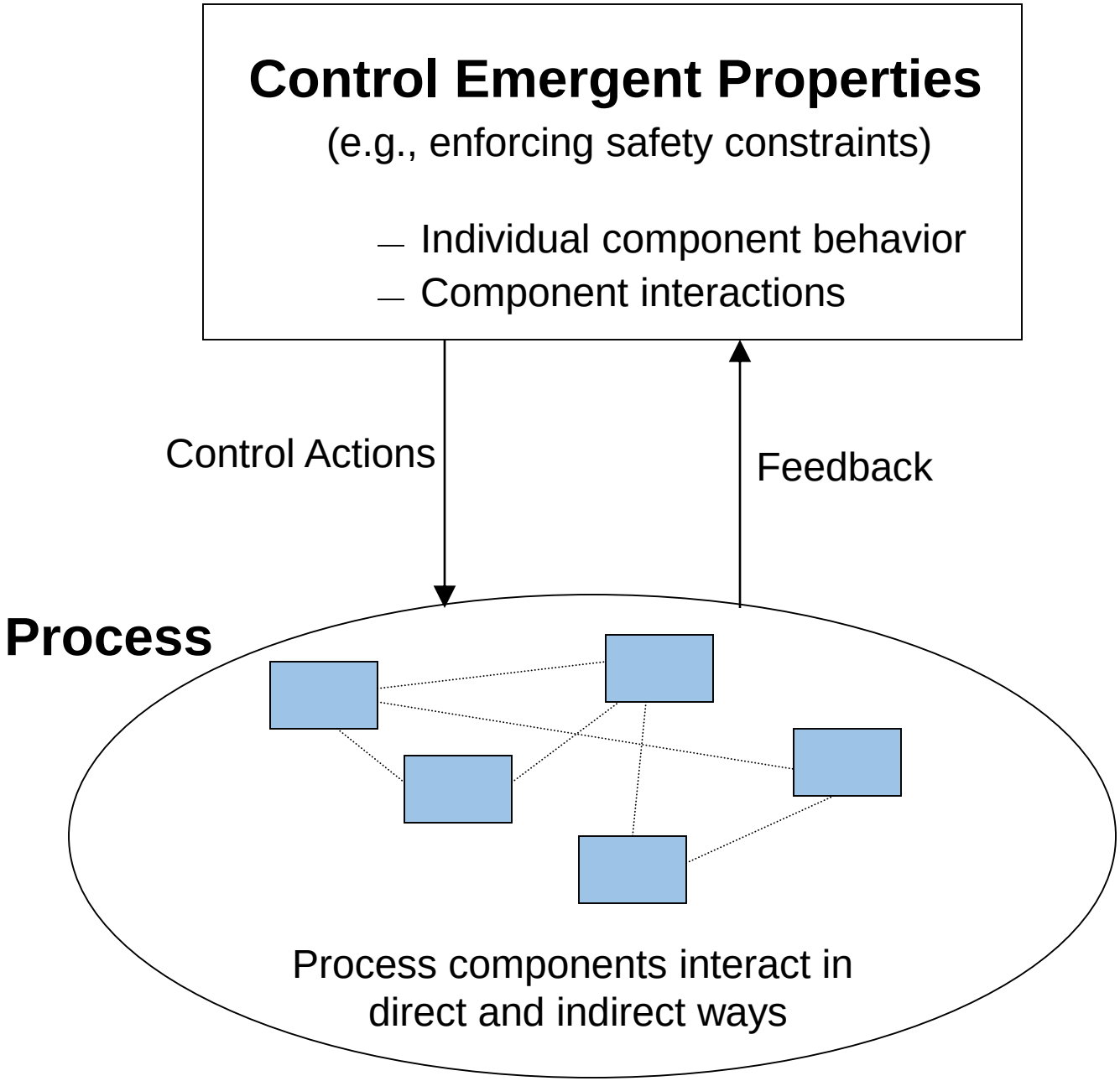
System Theory



Emergent properties
(arise from complex interactions)



Safety and security are emergent properties


Control Emergent Properties
(e.g., enforcing safety constraints)
– Individual component behavior
– Component interactions
Control Actions
Feedback
Process
Process components interact in
direct and indirect ways

A Broad View of "Control"

Component failures and unsafe interactions may be "controlled" through design

(e.g., redundancy, interlocks, fail-safe design, …)

or through process

- Manufacturing processes and procedures
- Maintenance processes
- Operational processes

or through social controls

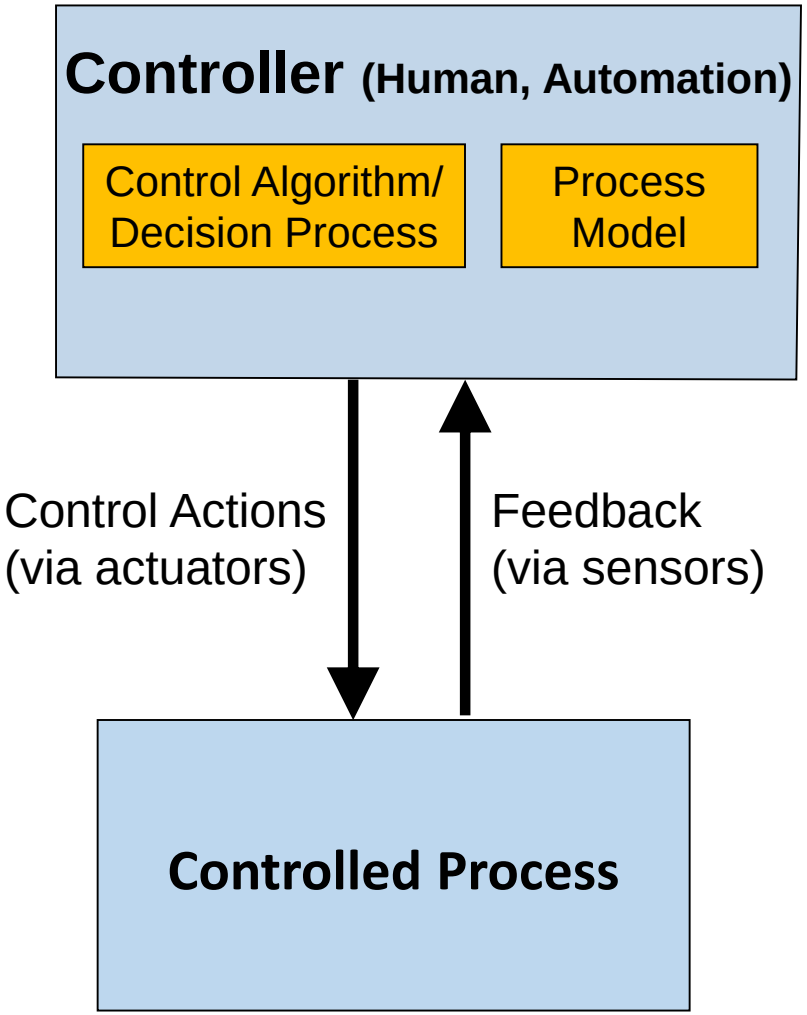
- Governmental or regulatory
- Culture
- Insurance
- Law and the courts
- Individual self-interest (incentive structure)

Controls/Controllers Enforce Constraints

- Aircraft must maintain sufficient lift to remain airborne
- Vehicles must maintain minimum separation
- Public health system must prevent exposure of public to contaminated water, food products, and viruses
- Pressure in an offshore well must be controlled
- Integrity of hull must be maintained on a submarine
- Toxic chemicals/radiation must not be released from plant
- Workers must not be exposed to workplace hazards

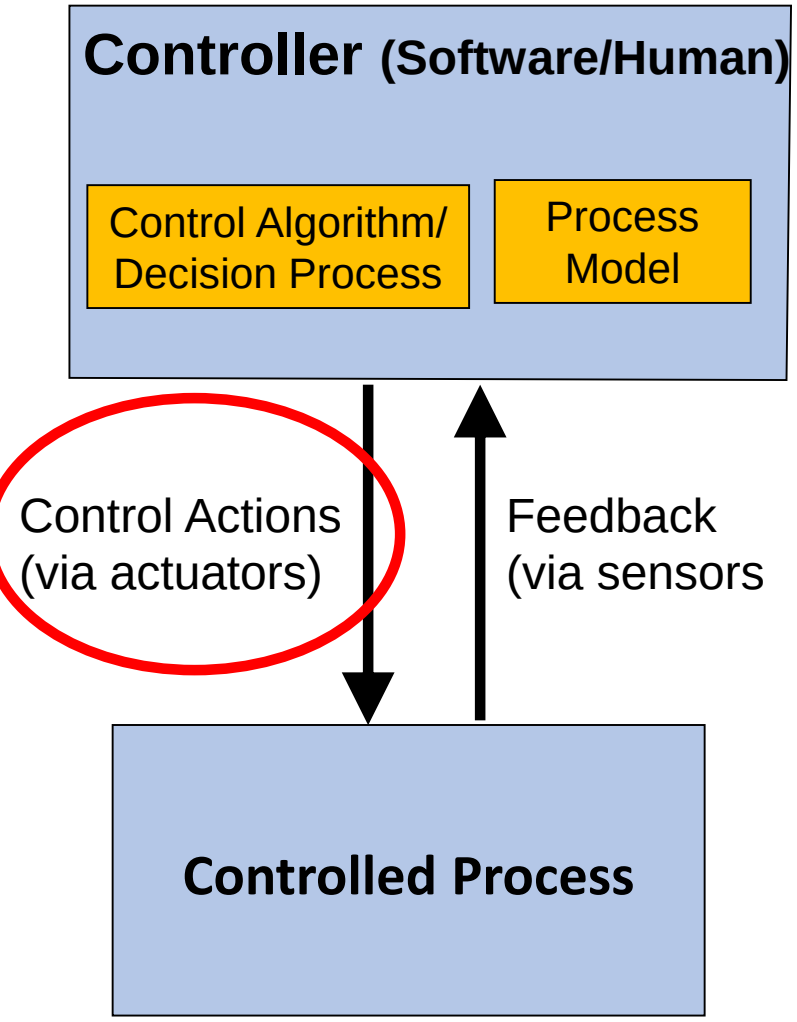
These represent the system-level requirements on the sociotechnical system

Safety as a Control Problem (vs. failure problem



- Controllers use a **process model** to determine control actions
- Software/human related accidents usually occur when the process model is incorrect (inconsistent with real state of process)
- Captures software errors, human errors, flawed requirements …

Hazard and Accident Analysis



Four types of unsafe control actions:

- Control commands required for safety are not given
- Unsafe ones are given
- Potentially safe commands given too early, too late
- Control stops too soon or applied too long

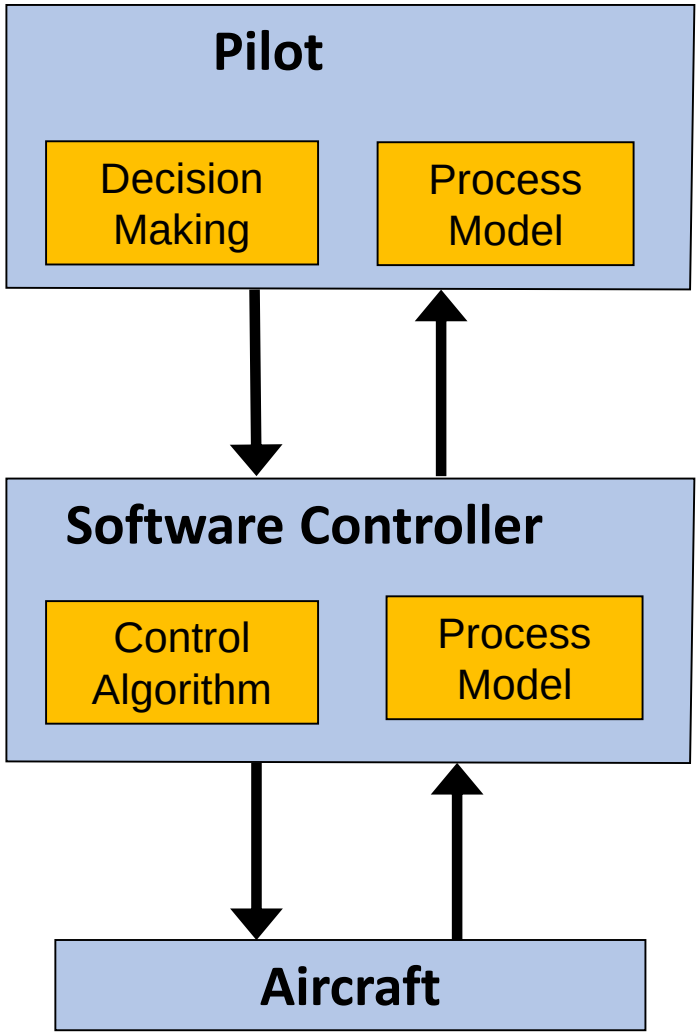
Analysis

- Identify potential unsafe control actions
- Identify why they might be (or were) given and then eliminate or mitigate them
- If safe ones provided, then why not followed?

(Leveson, 2003); (Leveson, 2011)

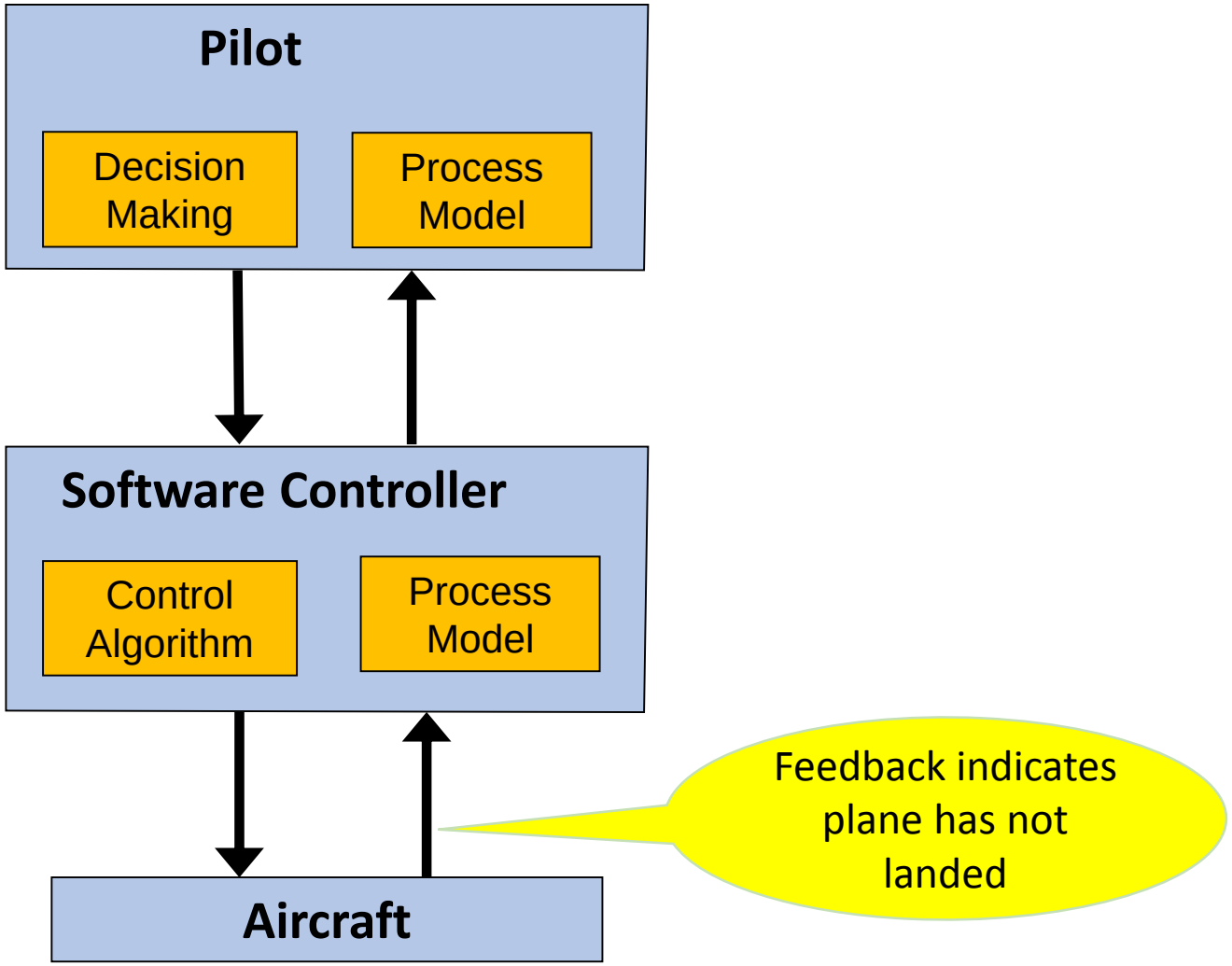
Warsaw (Reverse Thrusters)

Hazard: Inadequate aircraft deceleration after landing



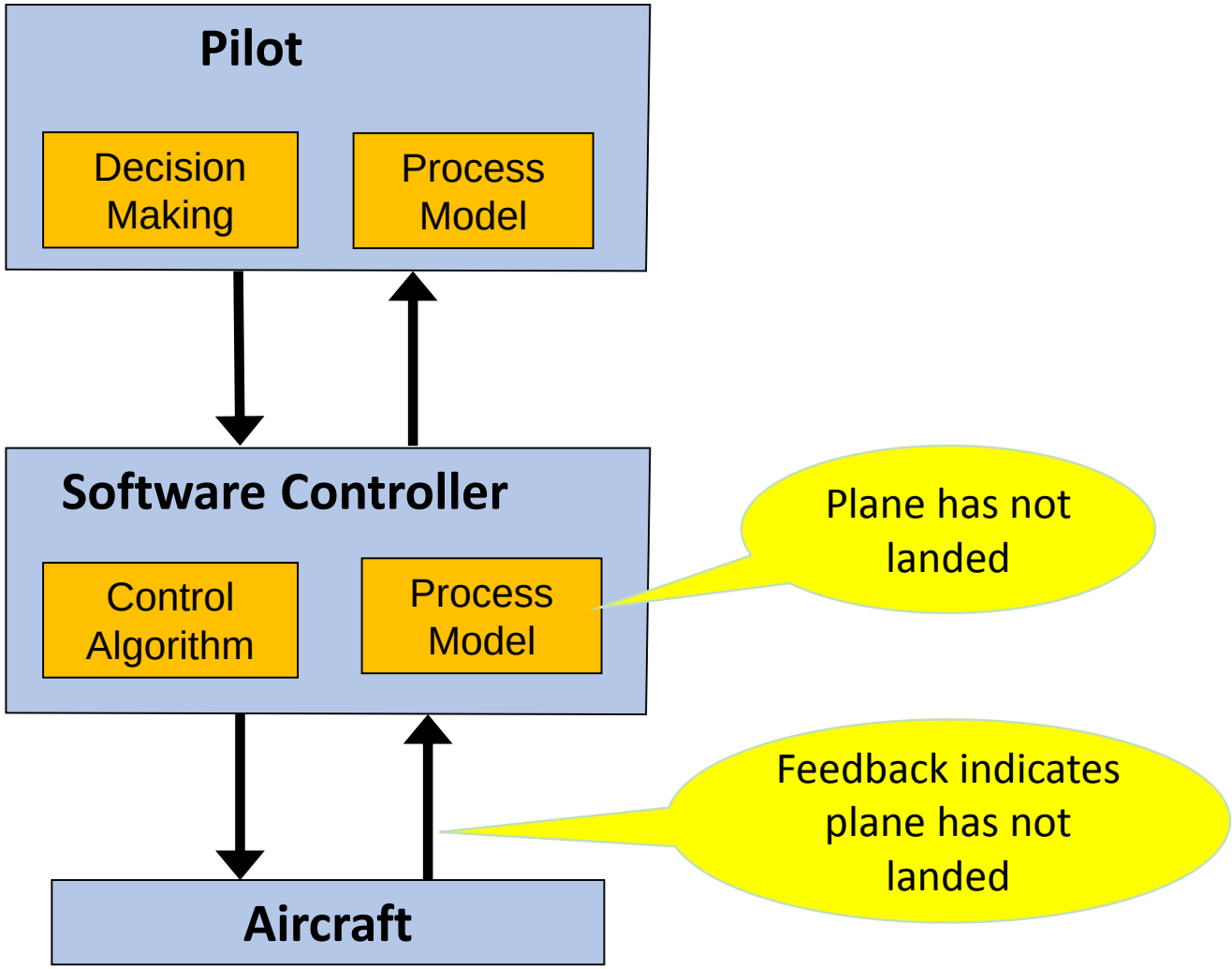
Warsaw (Reverse Thrusters)

Hazard: Inadequate aircraft deceleration after landing



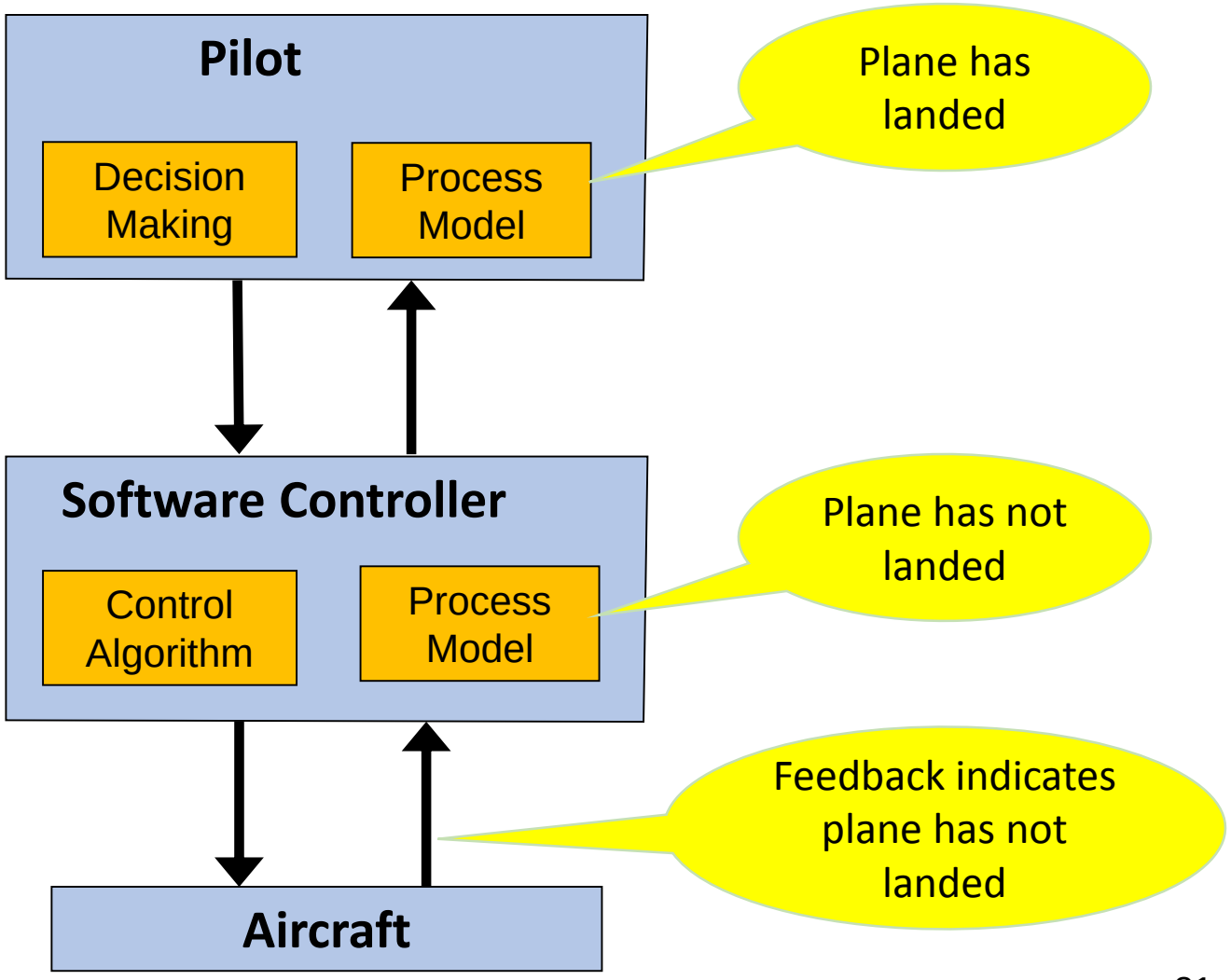
Warsaw (Reverse Thrusters)

Hazard: Inadequate aircraft deceleration after landing



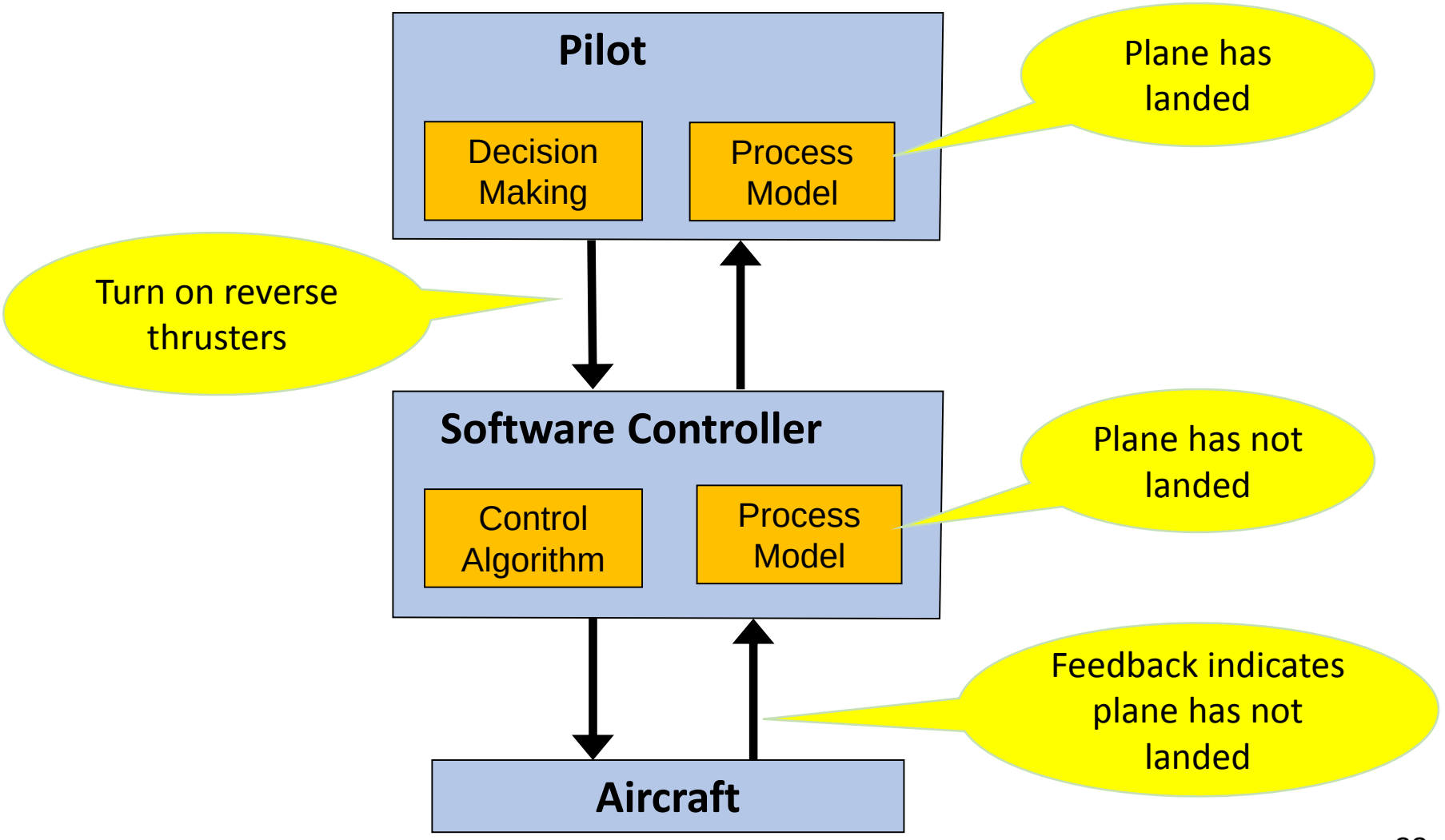
Warsaw (Reverse Thrusters)

Hazard: Inadequate aircraft deceleration after landing



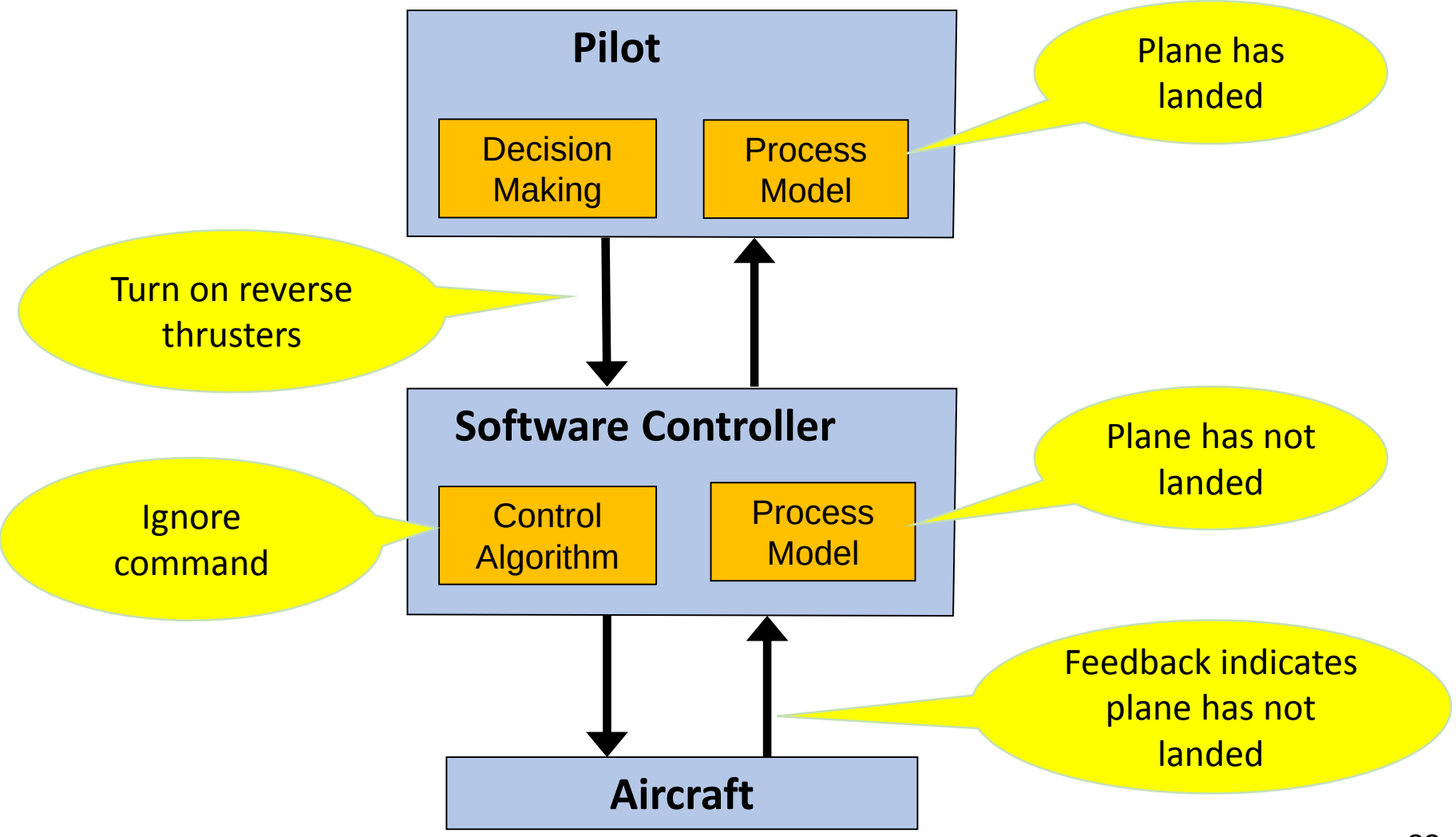
Warsaw (Reverse Thrusters)

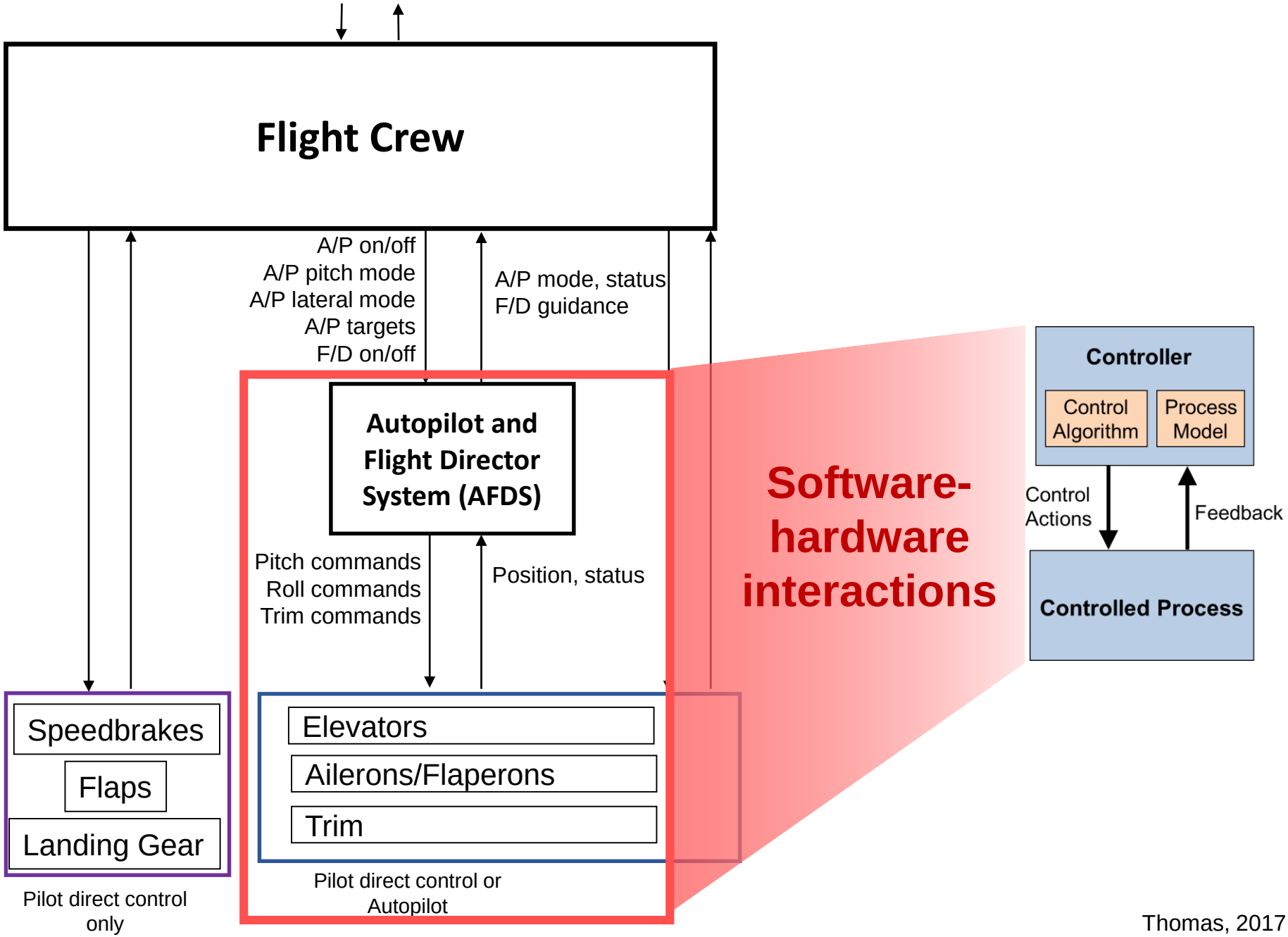
Hazard: Inadequate aircraft deceleration after landing



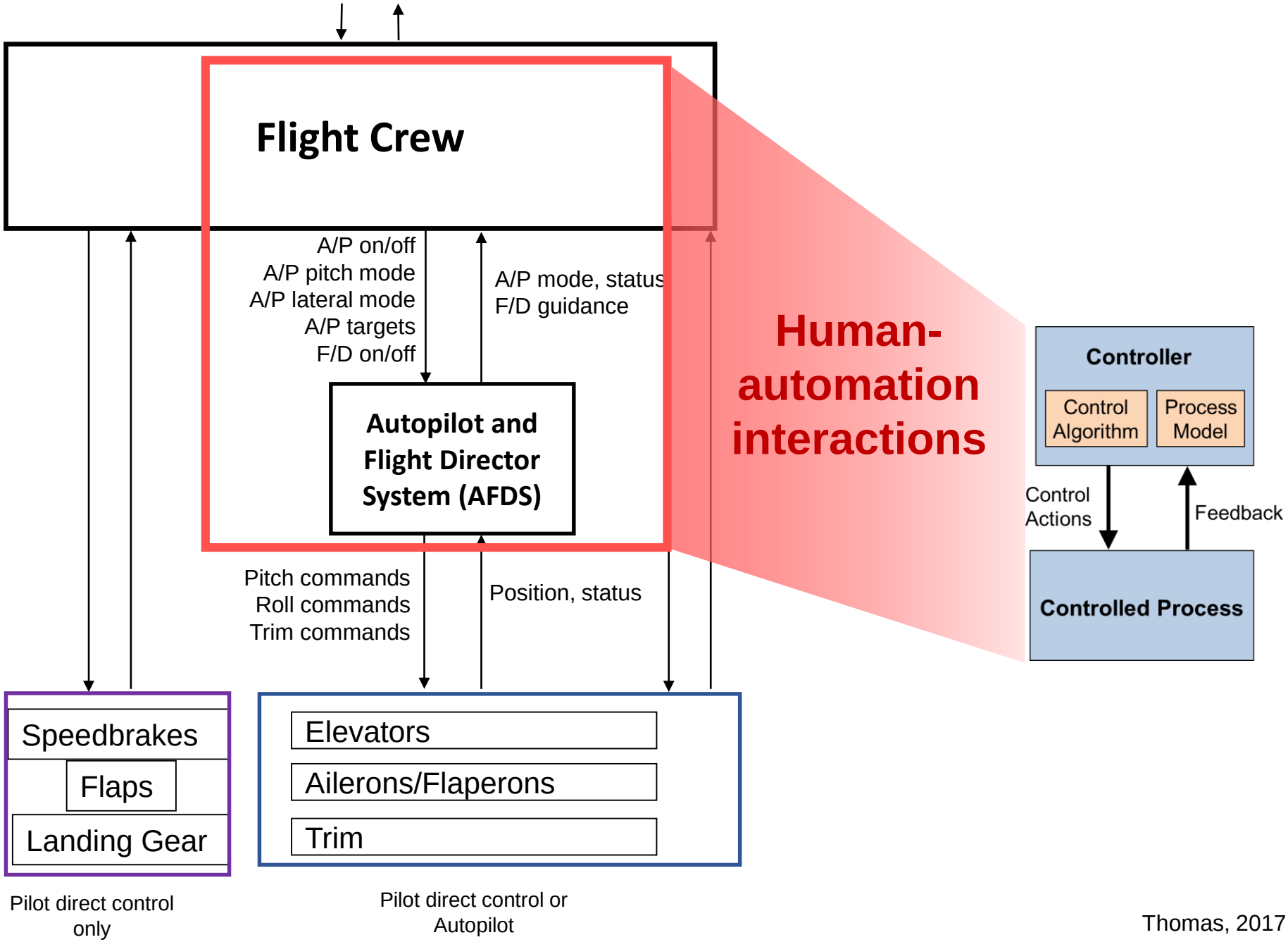
Warsaw (Reverse Thrusters)

Hazard: Inadequate aircraft deceleration after landing

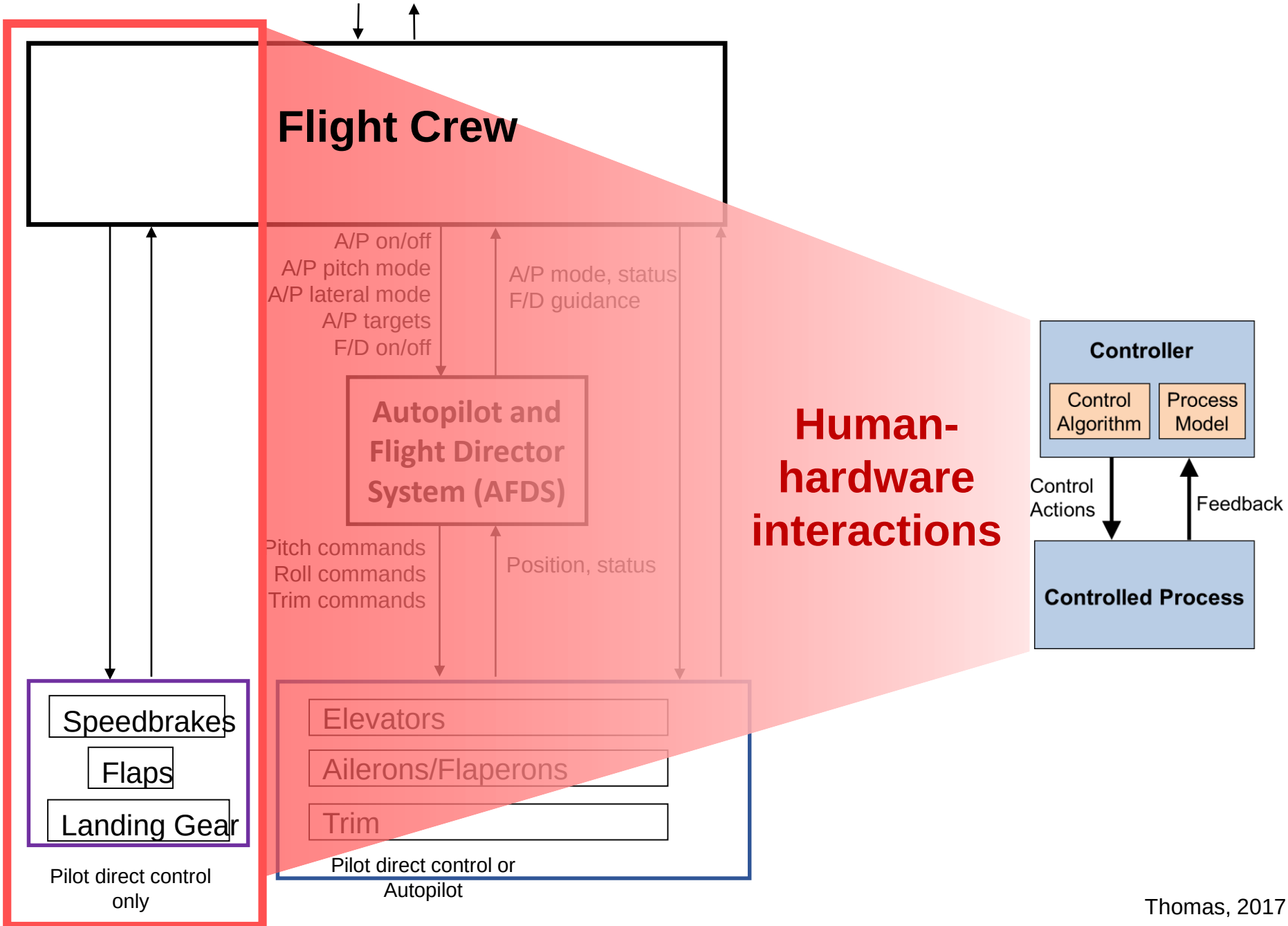



Flight Crew
A/P on/off
A/P pitch mode
A/P lateral mode
A/P targets
F/D on/off
A/P mode, status
F/D guidance
Autopilot and Flight Director System (AFDS)
Pitch commands
Roll commands
Trim commands
Position, status
Speedbrakes
Flaps
Landing Gear
Elevators
Ailerons/Flaperons
Trim
Pilot direct control only
Pilot direct control or Autopilot
Software-hardware interactions
Controller
Control Algorithm
Process Model
Control Actions
Feedback
Controlled Process

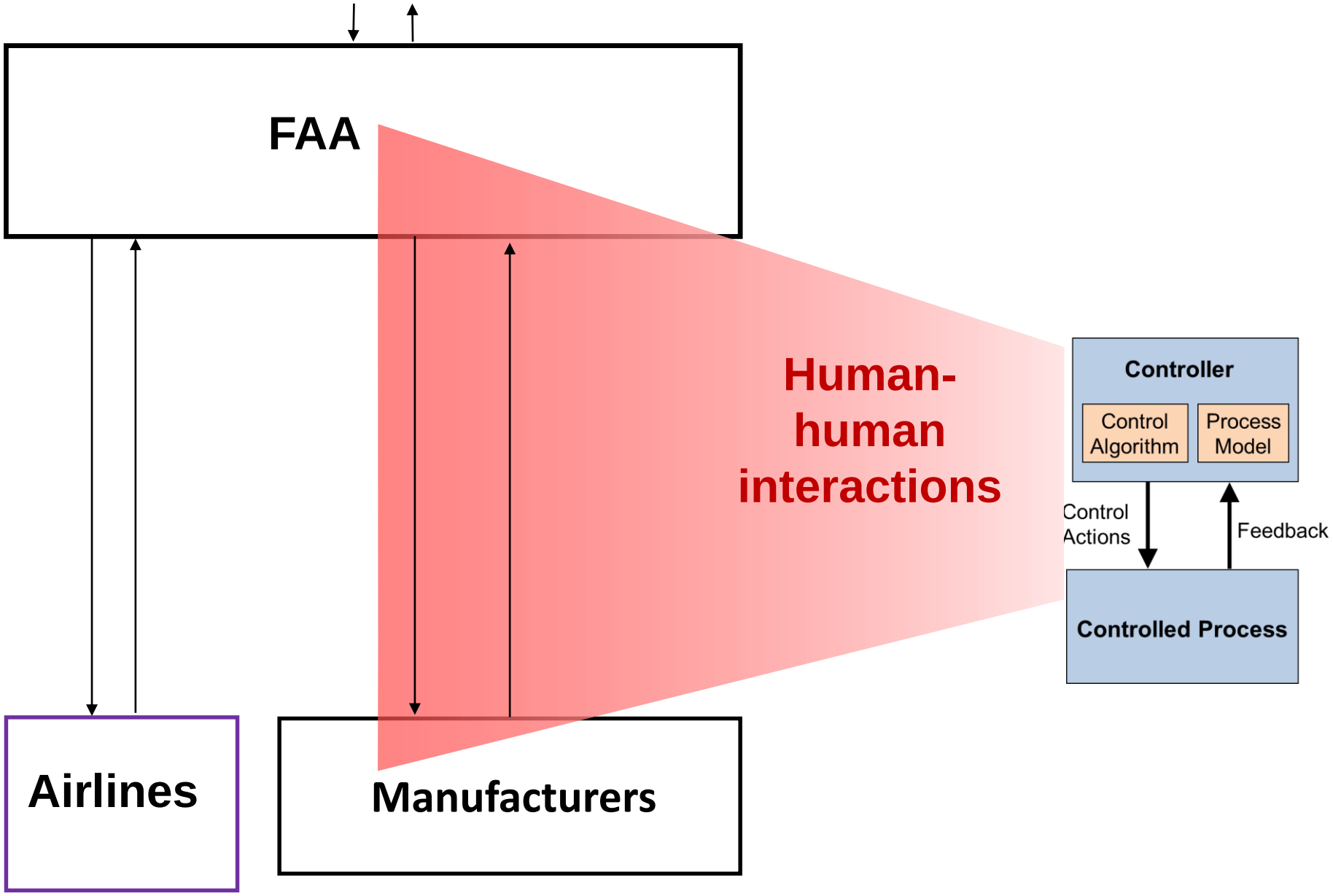
Thomas, 2017


Flight Crew
A/P on/off
A/P pitch mode
A/P lateral mode
A/P targets
F/D on/off
A/P mode, status
F/D guidance
Autopilot and Flight Director System (AFDS)
Pitch commands
Roll commands
Trim commands
Position, status
Speedbrakes
Flaps
Landing Gear
Elevators
Ailerons/Flaperons
Trim
Pilot direct control only
Pilot direct control or Autopilot
Human-automation interactions
Controller
Control Algorithm
Process Model
Control Actions
Feedback
Controlled Process

Thomas, 2017

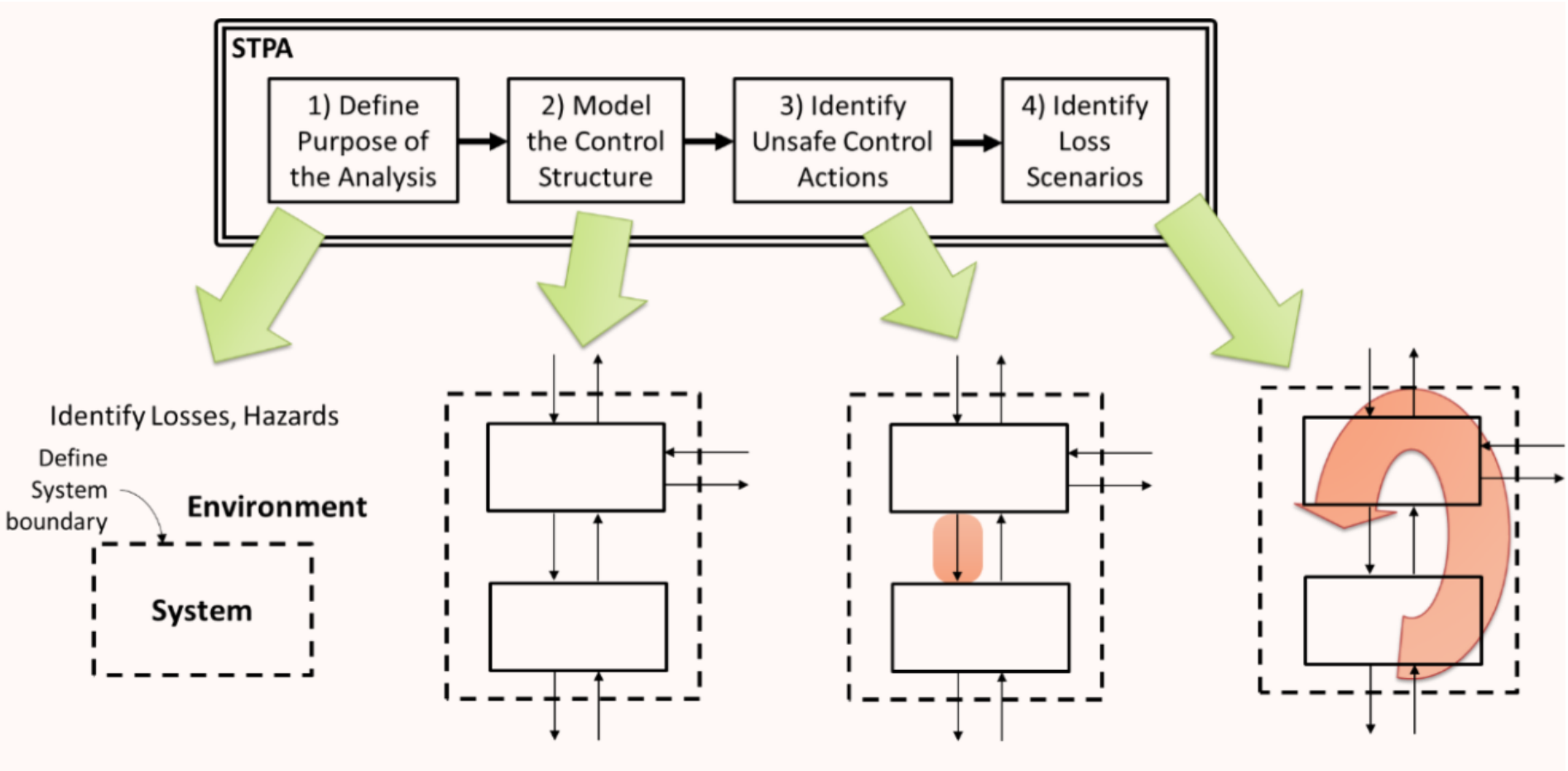

Flight Crew
A/P on/off
A/P pitch mode
A/P lateral mode
A/P targets
F/D on/off
A/P mode, status
F/D guidance
Autopilot and Flight Director System (AFDS)
Pitch commands
Roll commands
Trim commands
Position, status
Speedbrakes
Flaps
Landing Gear
Pilot direct control only
Elevators
Ailerons/Flaperons
Trim
Pilot direct control or Autopilot
Human-hardware interactions
Controller
Control Algorithm
Process Model
Control Actions
Feedback
Controlled Process

Thomas, 2017


FAA
Human-
human
interactions
Controller
Control
Algorithm
Process
Model
Control
Actions
Feedback
Controlled Process
Airlines
Manufacturers

Thomas, 2017

STPA Method Overview



Processes

System Engineering | Risk Management | Organizational Design (SMS)

Operations | Certification and Acquisition | Regulation

Tools

Accident Analysis **CAST** | Hazard Analysis **STPA** | Security Analysis **STPA-Sec**

Organizational/Cultural Risk Analysis | Identifying Leading Indicators | MBSE **SpecTRM & …**

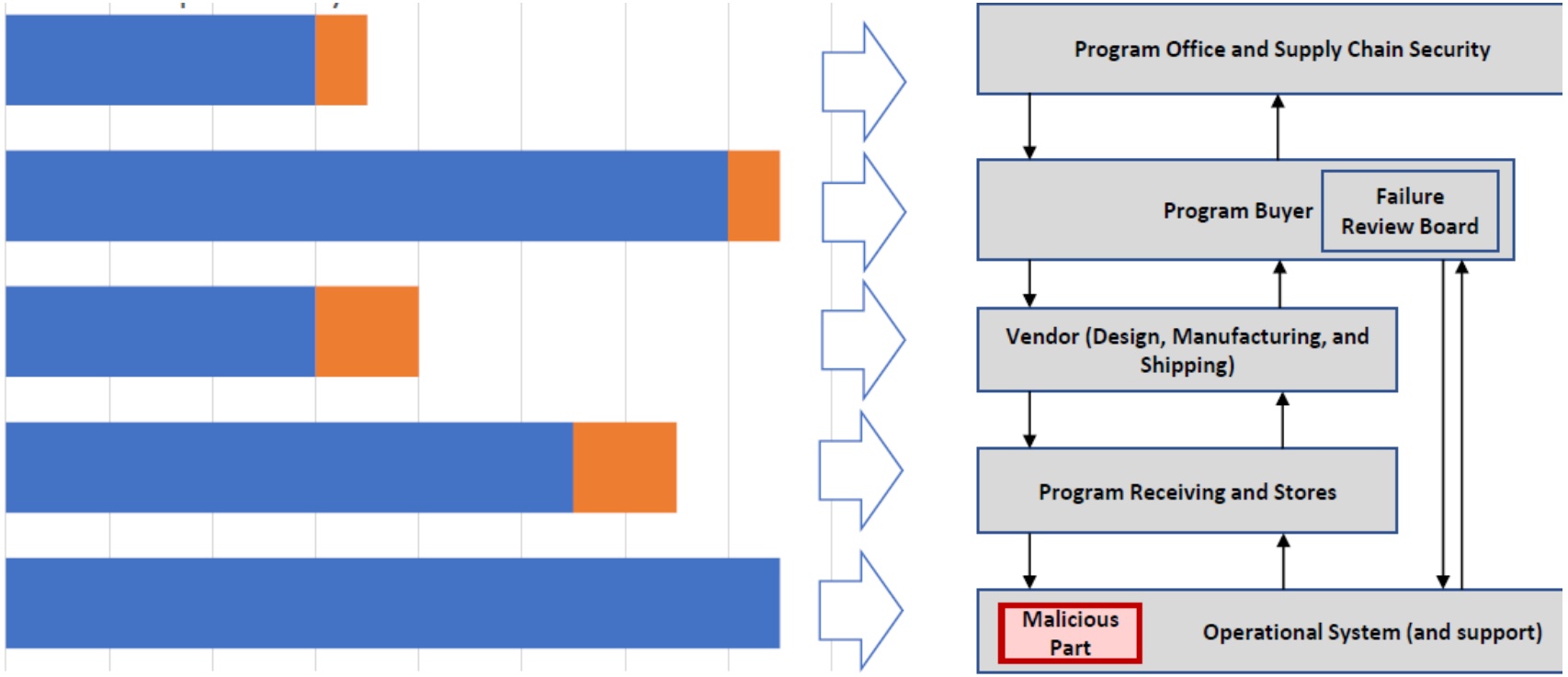
STAMP: Theoretical Causality Model

Example: STPA applied to one DoD program before SolarWinds attack

Michael Bear (BAE), John Thomas (MIT), Col. William Young (USAF)

- Program that used STPA was protected from SolarWinds
- Vulnerabilities found by STPA

Later exploited by SolarWinds attackers
Not exploited by SolarWinds attackers



New Research

- Top-down design from concept development: create requirements, architecture, and design using STPA analysis to guide tradeoffs

  Build safety into the system from the beginning

- Sophisticated teaming between humans and automation
- Leading indicators of increasing risk
- New approaches to risk assurance
- Extensions of STPA to new types of technology (e.g, ML)
- Applicability to other emergent properties (mission assurance, security, serviceability, quality, producibility, …)

Is The Approach Practical?

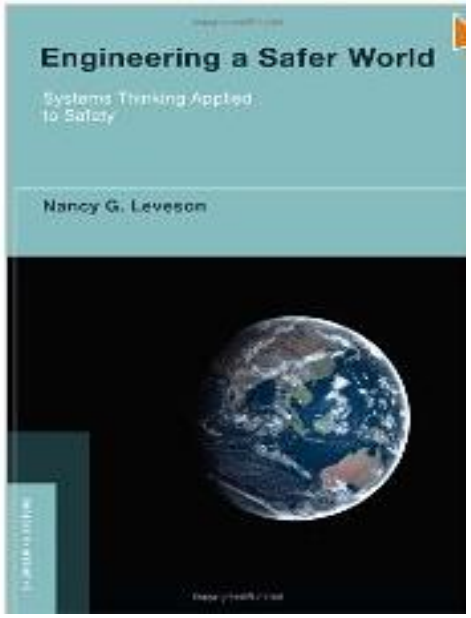
- Has been or is being used in a large variety of industries
  - Automobiles (>80% use)
  - Aircraft and Spacecraft (extensive use and growing)
  - Defense systems (UAVs, AF GBSD, Army FVL, etc.)
  - Ships/Marine
  - Air Traffic Control
  - Medical Devices and Hospital Safety
  - Chemical plants
  - Oil and Gas
  - Nuclear and Electric Power
  - Robotic Manufacturing / Workplace Safety
- Over 2,000 registrants (91 countries) for virtual STAMP Workshops
- New STPA international standards created or in development.

Evaluations and Estimates of ROI

- Hundreds of evaluations and comparison with traditional approaches used now
  - Controlled scientific and empirical (in industry)
  - All show STPA is better (identifies more critical requirements or design flaws)
  - All (that measured) show STPA requires orders of magnitude fewer resources than traditional techniques
- ROI estimates only beginning but one large defense industry contractor claims they are seeing 15-20% return on investment when using STPA

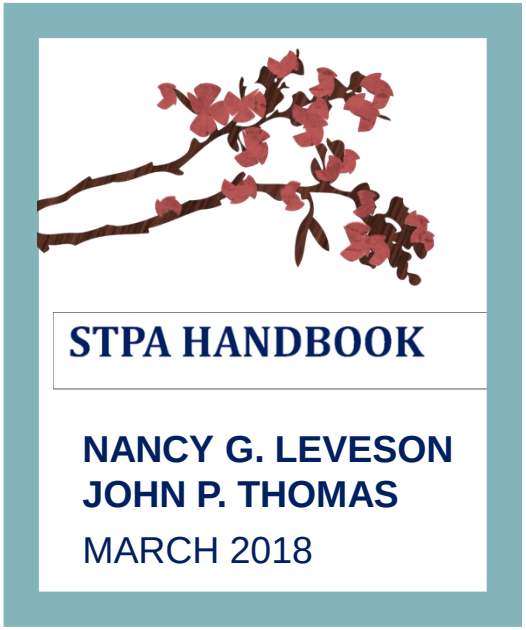
More Information

- http://psas.scripts.mit.edu (papers, presentations, tutorial slides, tutorials, examples, etc.)

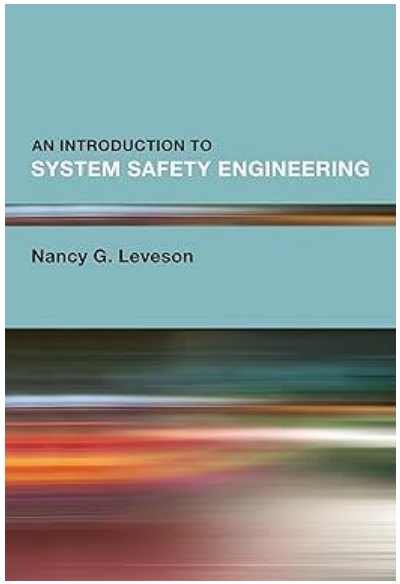


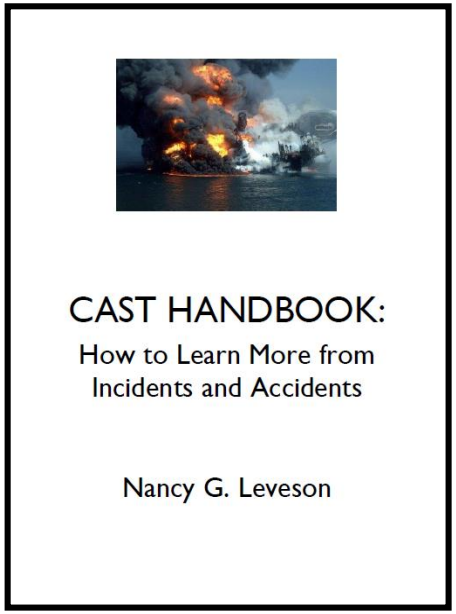
Free download:
http://mitpress.mit.edu/books/engineering-safer-world

In Japanese available 2024



Free download:
http://psas.scripts.mit.edu
(**400,000+** downloads since 2018. Japanese, Chinese, and Korean versions)





Free download:
http://sunnyday.mit.edu/CAST-Handbook.pdf
(Korean, Japanese versions)