

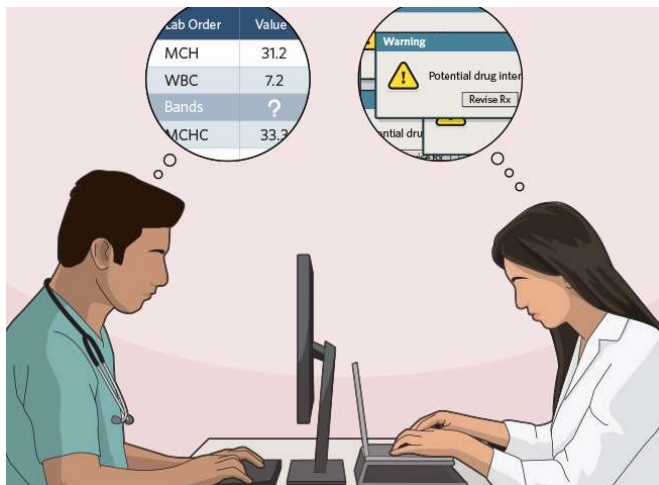
# Safety Assurance for Artificial Intelligence is Futile (and that's okay)

Drew Rae

Safety Science Innovation Lab





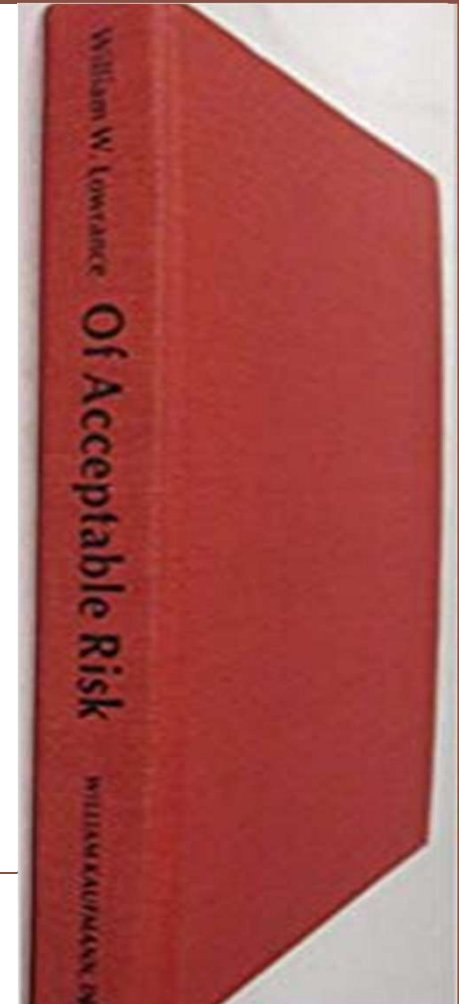


How do we worry about  
AI in safety?

---

“A thing is safe  
if the  
associated risk  
of harm is  
acceptably  
low”

---



# Why do we worry about AI?

## Knowledge

### Capability

- What can the system do?

### Predictability

- What will the system do?

## Control

### Comprehensibility

- Can insiders understand the system?

### Tractability

- Can we manage the system?

## Regulation

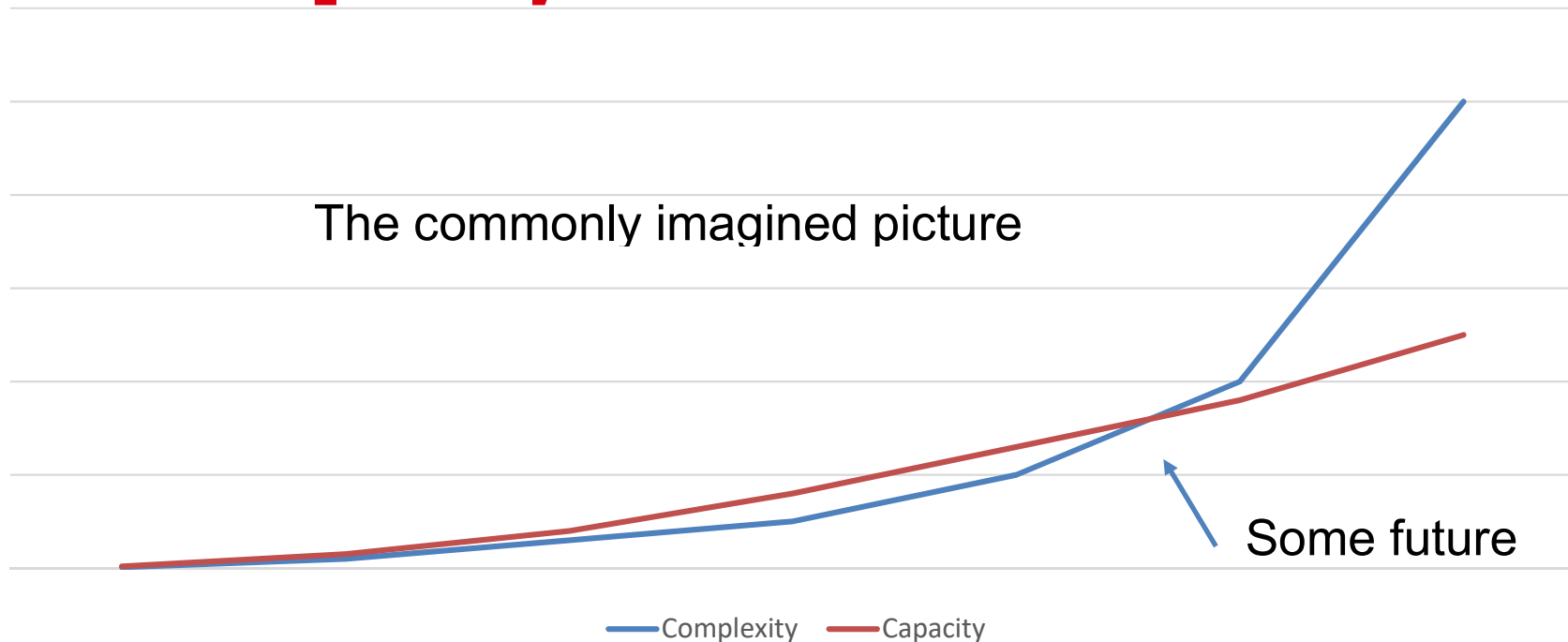
### Legibility

- Can outsiders understand the system?

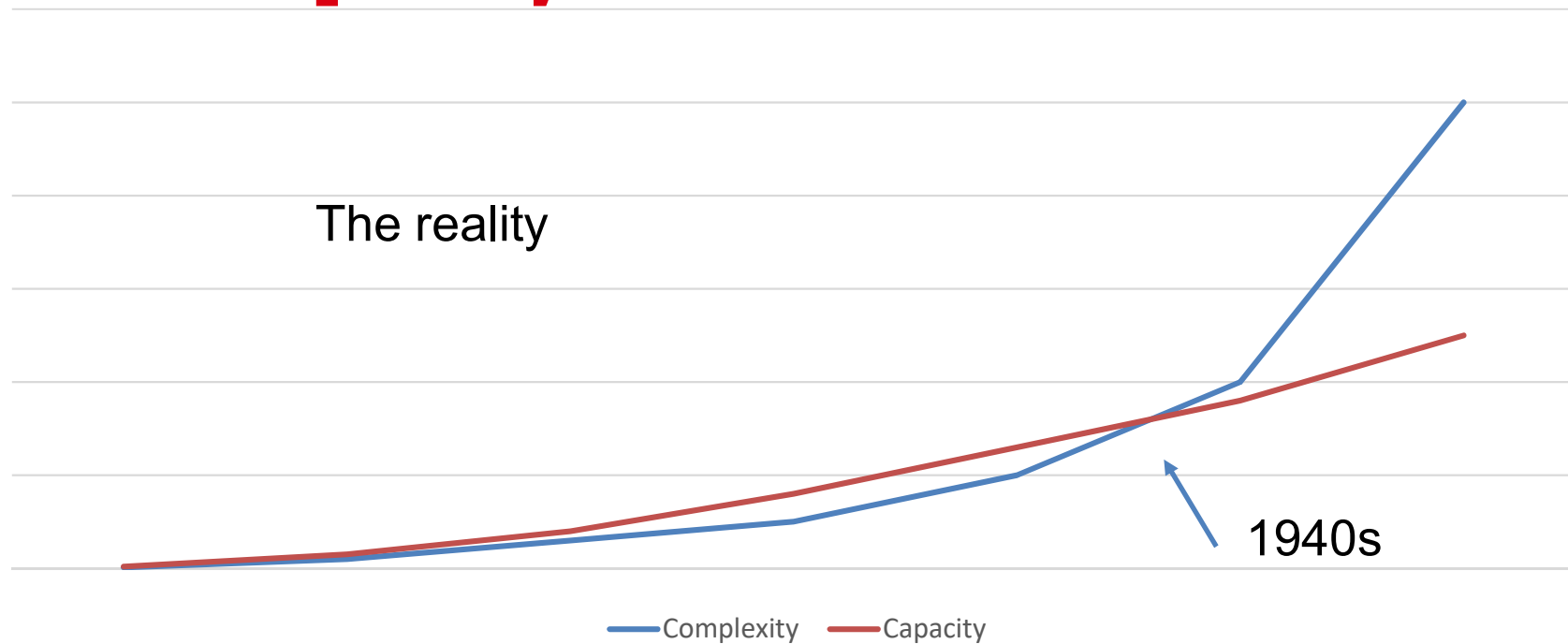
### Regulatability

- Can outsiders influence the safety?

# Can our methods and organisations cope with complexity?



# Can our methods and organisations cope with complexity?





# Capability

In most cases AI doesn't represent an increase in system capability to do harm





Northeast Blackout (1965)



# Predictability

Safety Science Innovation Lab

Safety assurance relies on systems being predictable.

We identify what can go wrong, and then provide assurance that it won't.



United Airlines Flight 608 (1947)

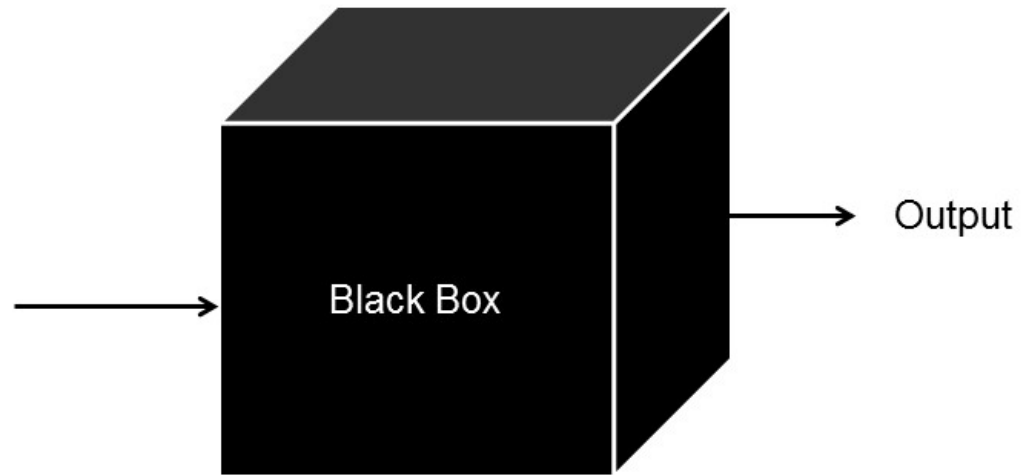




## UA965 B767 Galley Chiller Incident



Safety Science 2019, 9, 10



*Internal behavior of the code is unknown*

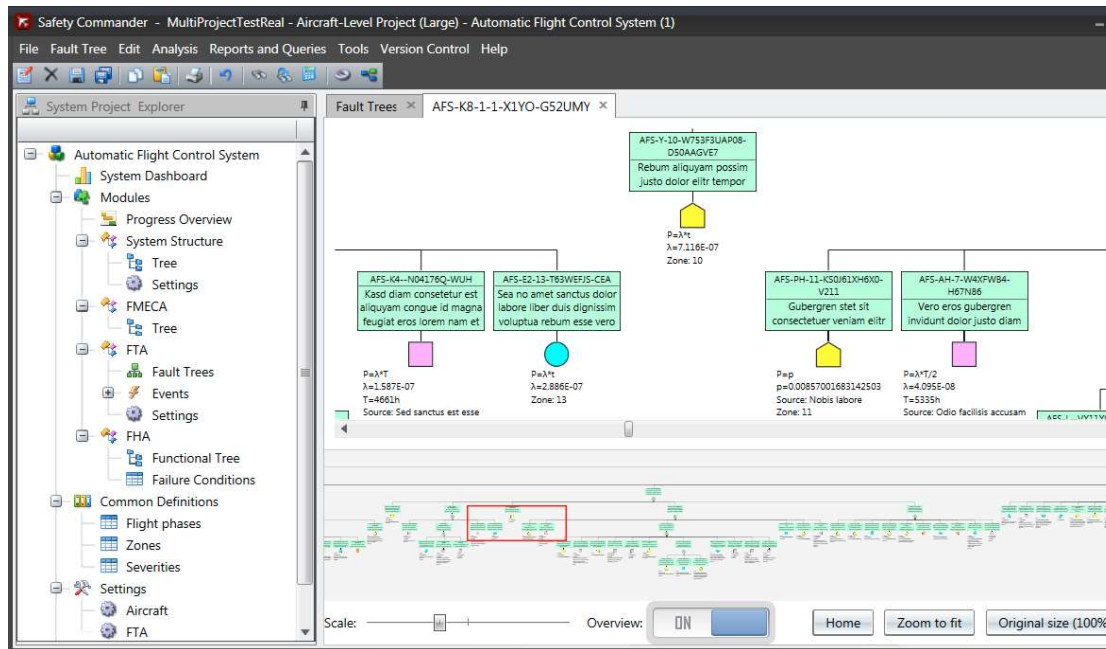
# Comprehensibility

Safety assurance depends on systems being analysable. We can't say that something is safe if we don't even know how it works.



AF447 “This maneuver is totally incomprehensible”



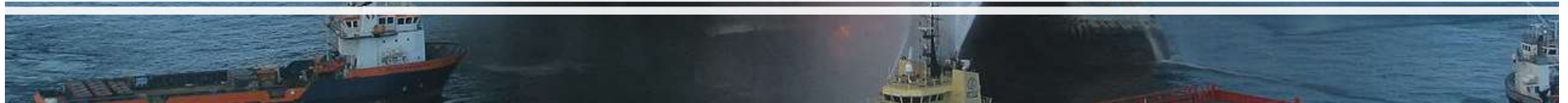


Safety assurance depends on systems being analysable. We can't say that something is safe if we can't determine how it will and won't behave.

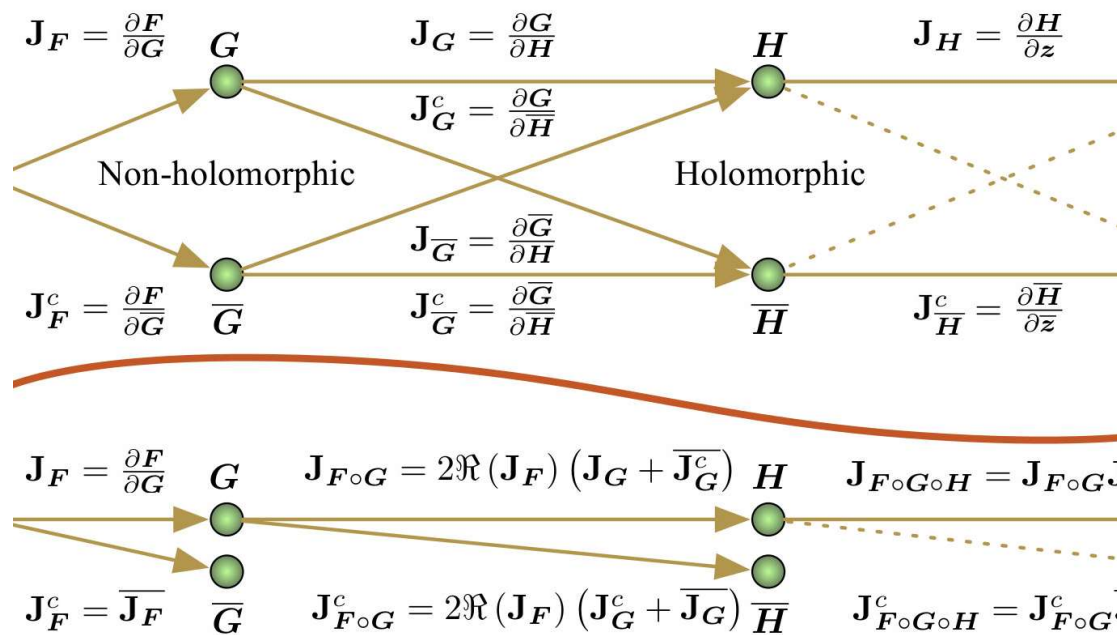
# Tractability



Deepwater Horizon (2010)







- “regulators of 'high' technologies face an inevitable epistemic barrier when making technological assessments, which forces them to delegate technical questions to people with more tacit knowledge”

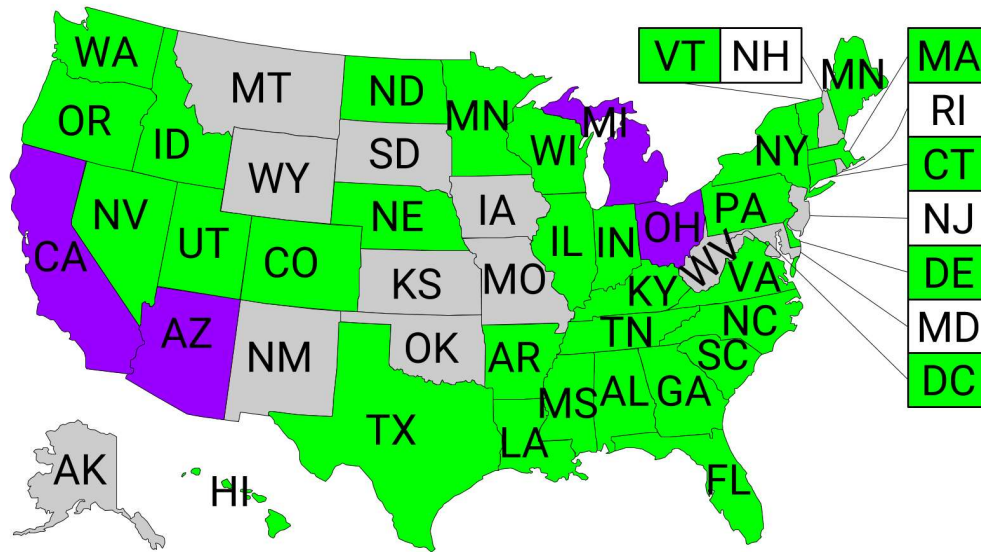
- Downer (2009)

# Legibility



Eastern Air Lines Flight 375 (1960)





Legend

# Regulatability

- “regulators of 'high' technologies face an inevitable epistemic barrier when making technological assessments, which forces them to delegate technical questions to people with more tacit knowledge”

- Downer (2009)



## Does AI actually make any of these worse?

### Knowledge

#### Capability

- What can the system do?

#### Predictability

- What will the system do?

### Control

#### Comprehensibility

- Can insiders understand the system?

#### Tractability

- Can we manage the system?

### Regulation

#### Legibility

- Can outsiders understand the system?

#### Regulatability

- Can outsiders influence the safety?



**DON'T  
PANIC**

The image features the text "DON'T PANIC" in a large, bold, black, sans-serif font. The letters are arranged in two lines: "DON'T" on top and "PANIC" on the bottom. Each letter is filled with a black background that contains various white stars of different sizes. Additionally, several colorful celestial bodies are integrated into the letters. In the "O" of "DON'T", there is a small gray moon and a small orange planet. In the "N" of "DON'T", there is a large orange planet with a white ring system. In the "P" of "PANIC", there is a small teal planet. In the "A" of "PANIC", there is a small orange planet with a white ring system. In the "N" of "PANIC", there is a small blue planet. In the "I" of "PANIC", there is a small red planet. In the "C" of "PANIC", there is a small blue planet. The entire graphic is set against a solid gray background.

For more information or to provide private feedback



d.rae@griffith.edu.au