Assuring Artificial Intelligence

Advanced T&E concepts for complex systems

Ben Luther

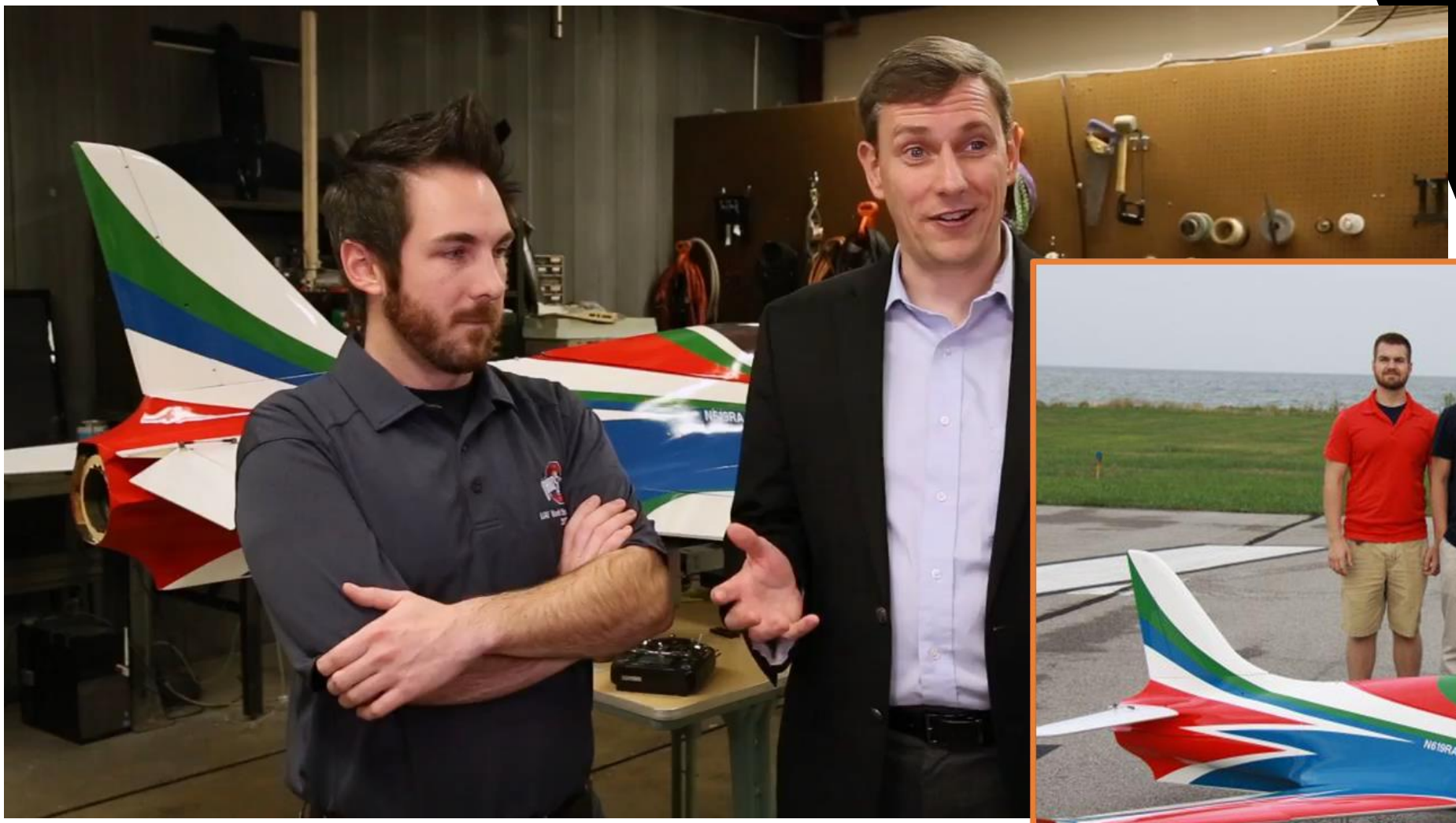
Content is the opinion of the author and
not necessarily the position of any organisation

Traditional assurance

https://theconversation.com/artificial-intelligence-carries-a-huge-upside-but-potential-harms-need-to-be-managed-173073

Artificial Intelligence

https://www.theguardian.com/technology/2014/oct/27/elon-musk-artificial-intelligence-ai-biggest-existential-threat



https://youtu.be/zccyN_z9WNk

https://engineering.osu.edu/news/2017/08/ohio-state-sets-drone-world-speed-record#

Difficulties in assuring Artificial Intelligence

Configuration - ?

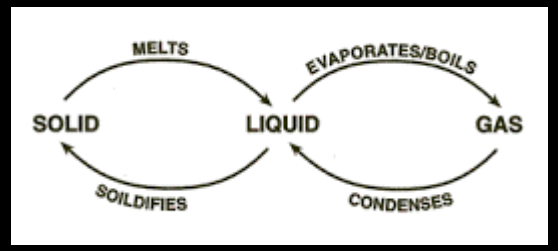
Performance - ?

Complexity

Complicated ≠ Complex

Complex is:
- Dynamic (Sterman)
- Unknowable *a priori* (Snowden)
- Emergent functions (Leveson)

State changes



Crisis leadership

Non-Statistical Approaches

Unique occurrences
Process assured

Statistical Approaches

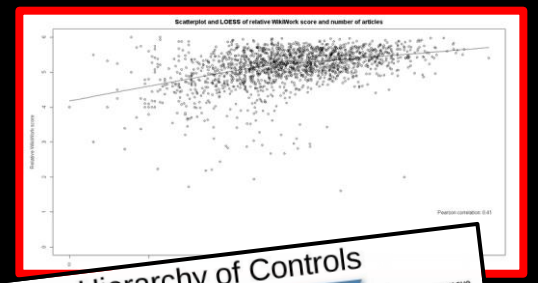
Process driven
Product assured

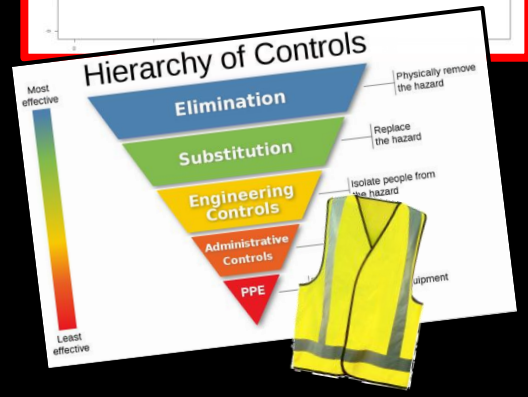
COMPLEX
Enabling constraints
Loosely coupled
probe-sense-respond
EMERGENT PRACTICE

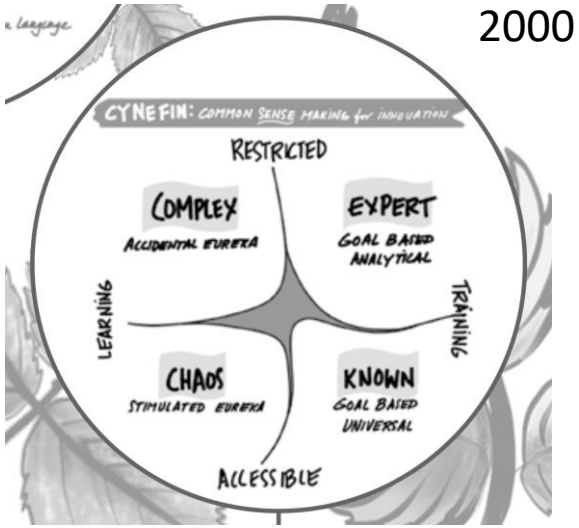
COMPLICATED
Governing constraints
Tightly coupled
sense-analyze-respond
GOOD PRACTICE

CHAOTIC
Lacking constraint
De-coupled
act-sense-respond
NOVEL PRACTICE

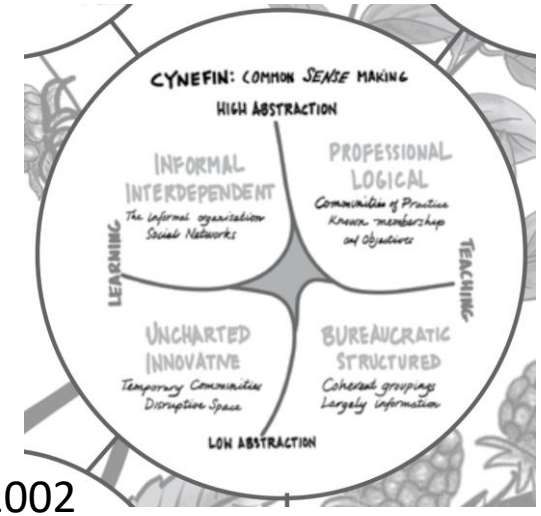
CLEAR
Tightly constrained
No degrees of freedom
sense-categorize-respond
BEST PRACTICE



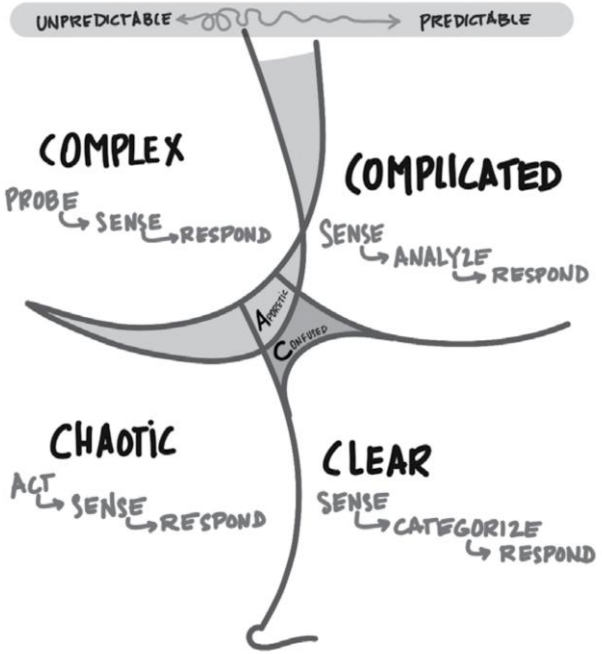




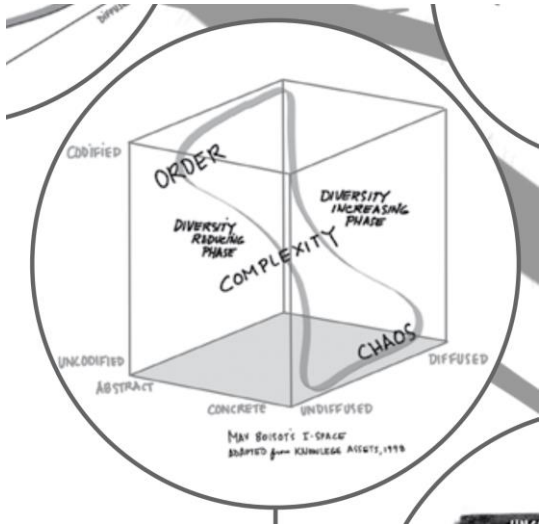
2000

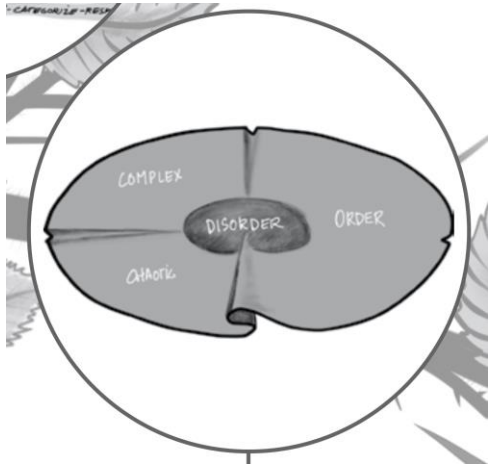


2002



2020





2010

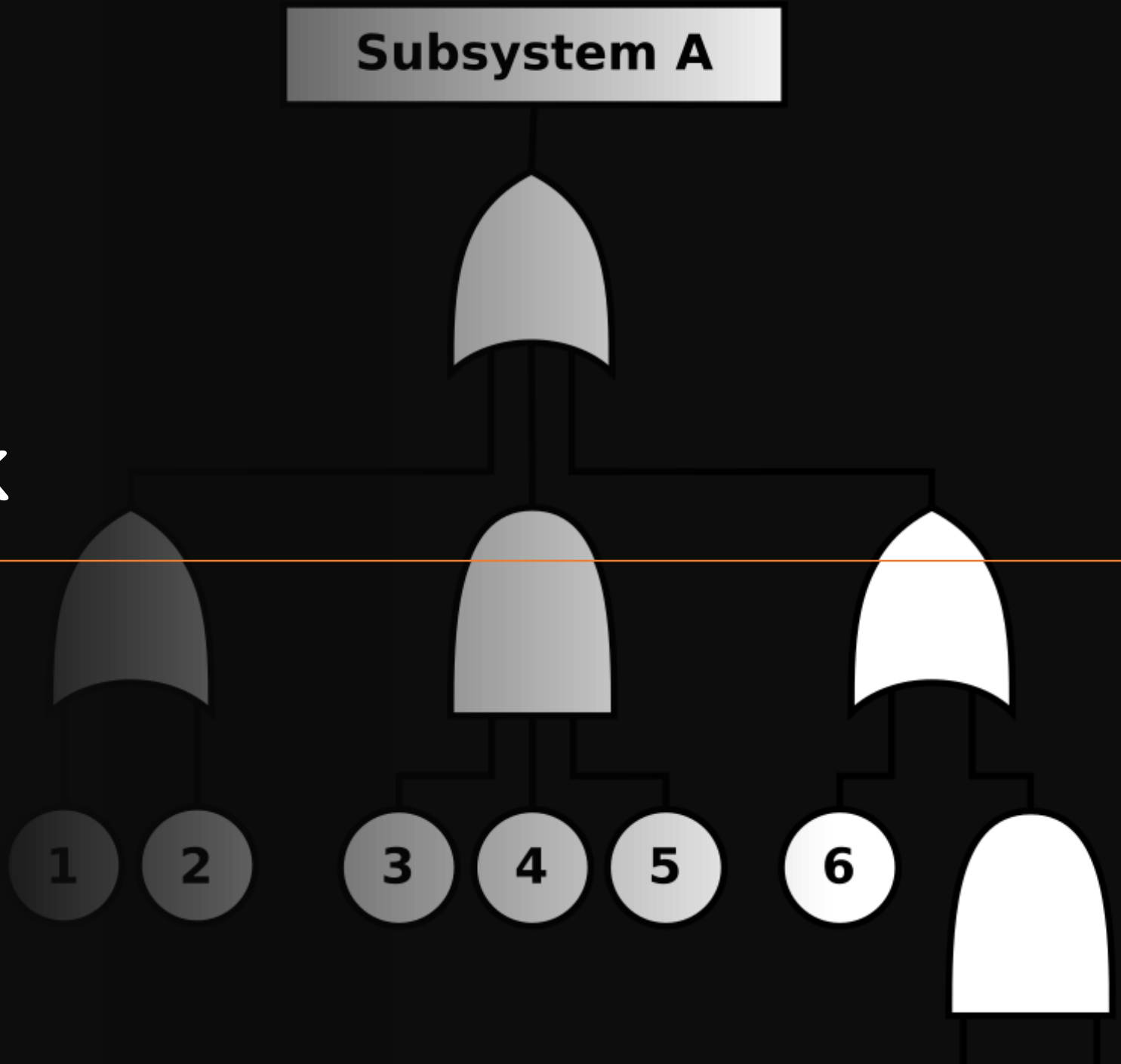
Snowden, D. et. al. (2021) Cynefin: weaving sense-making into the fabric of our world. Cognitive Edge. Singapore.

Social sciences

& ethics

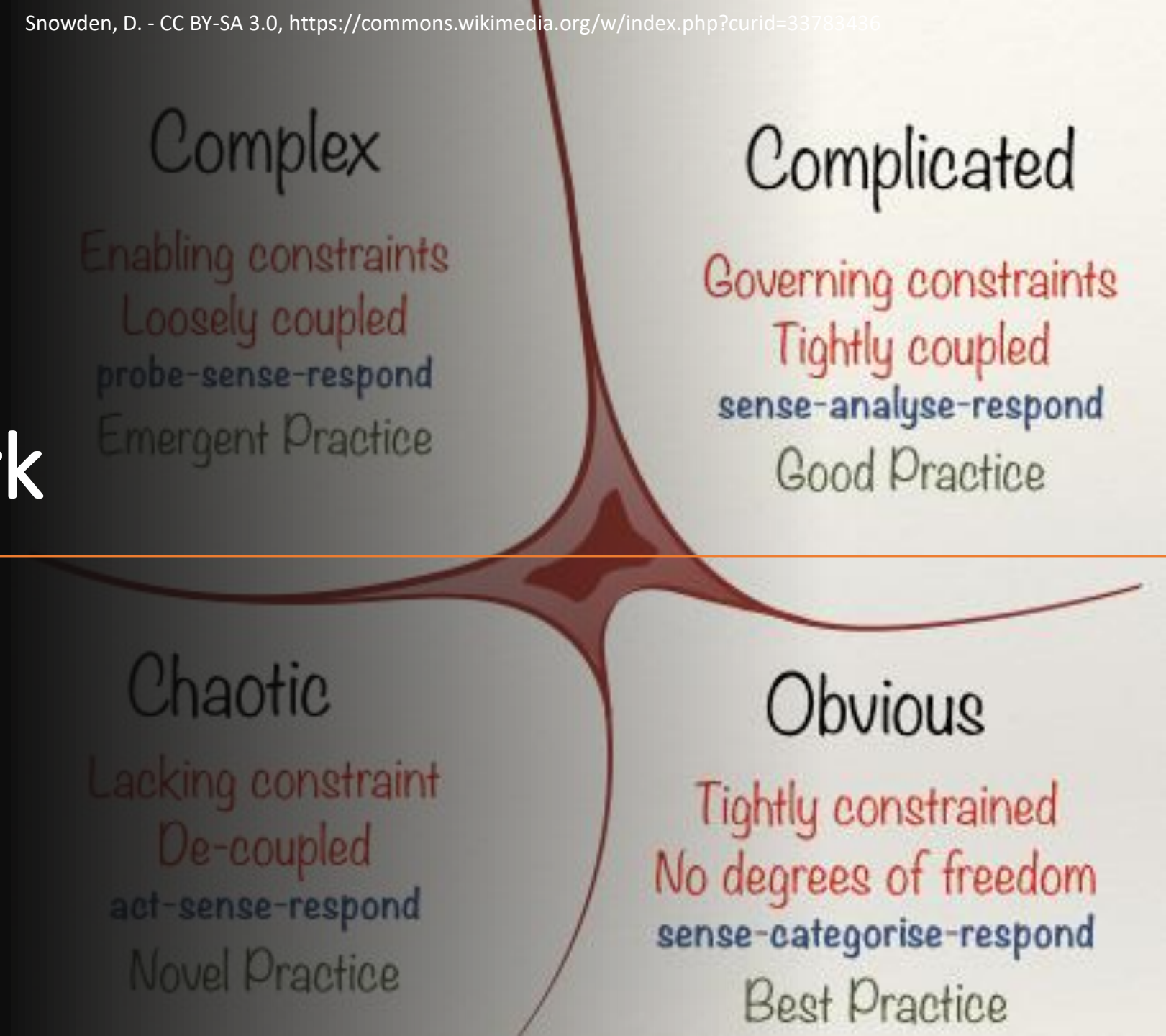
Tools that work

System Safety
- use first



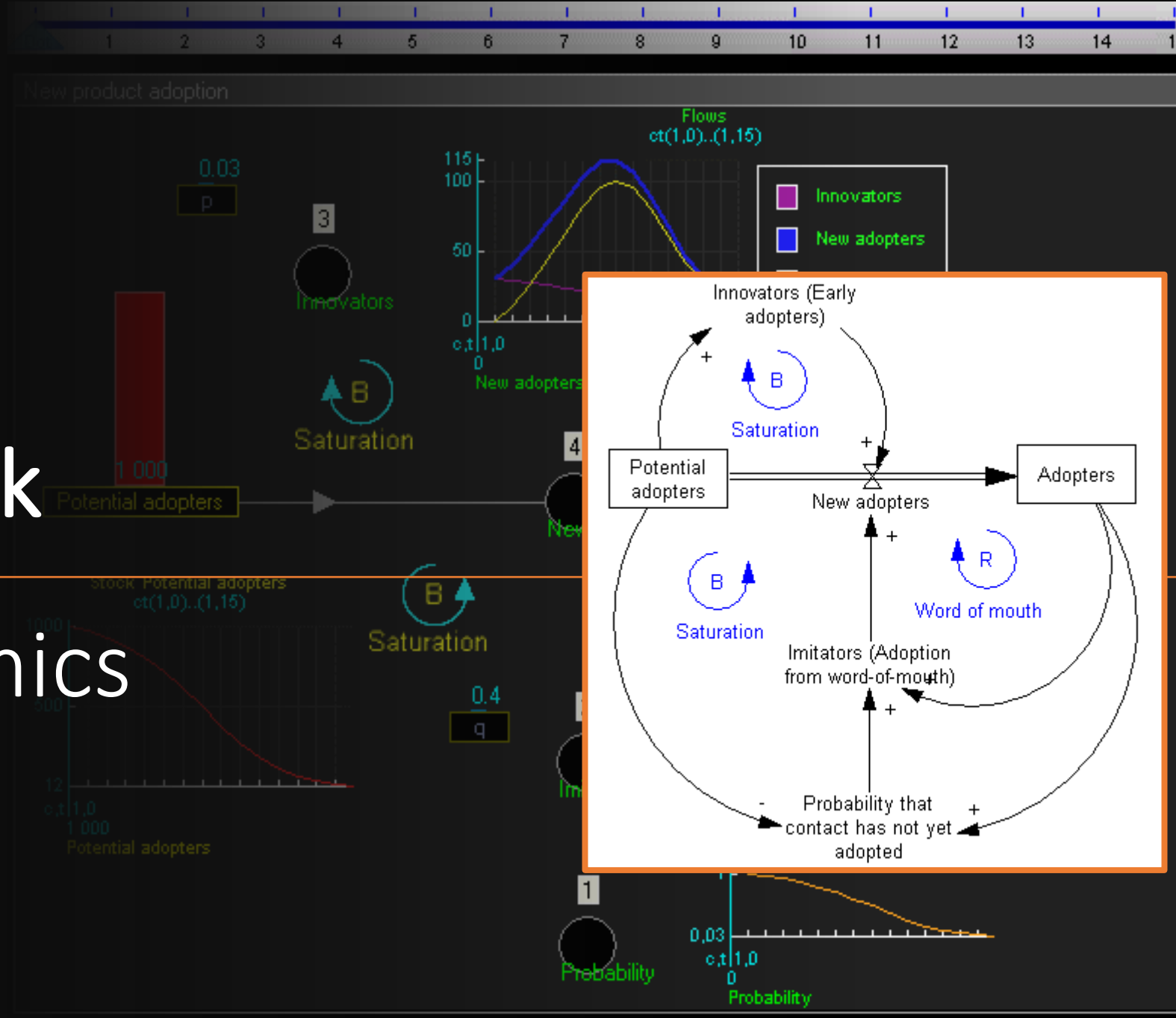
Tools that work

Cynefin



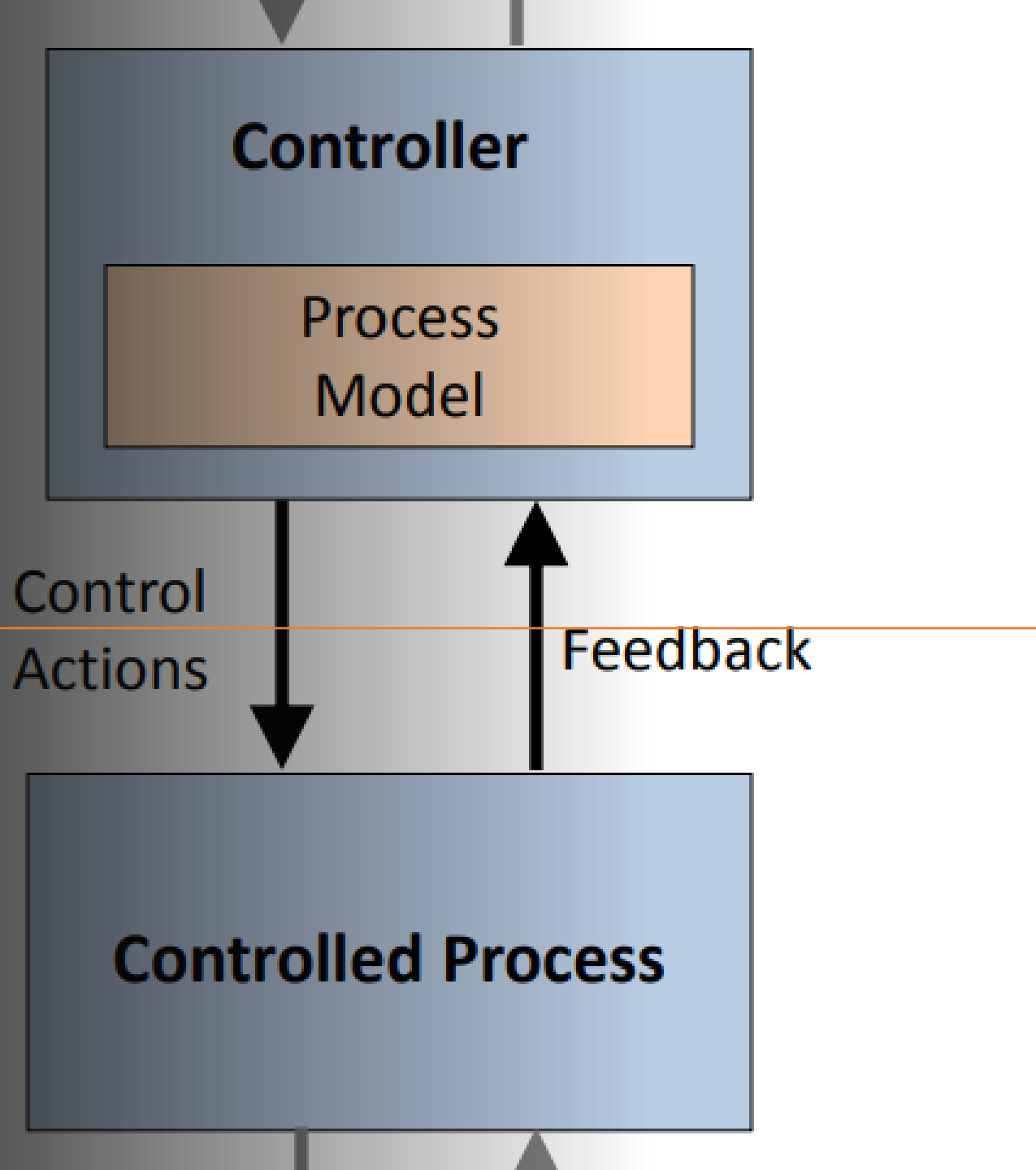
Tools that work

System Dynamics



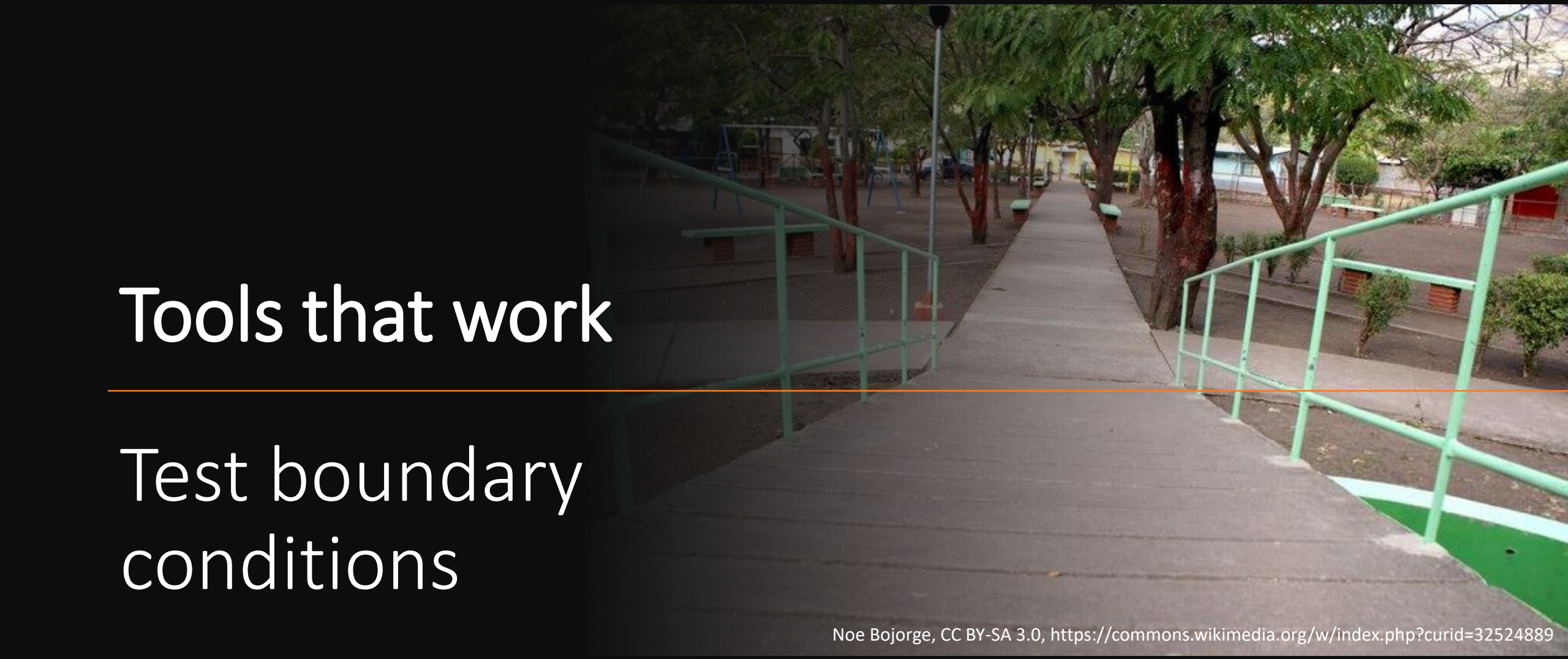
Tools that work

STPA

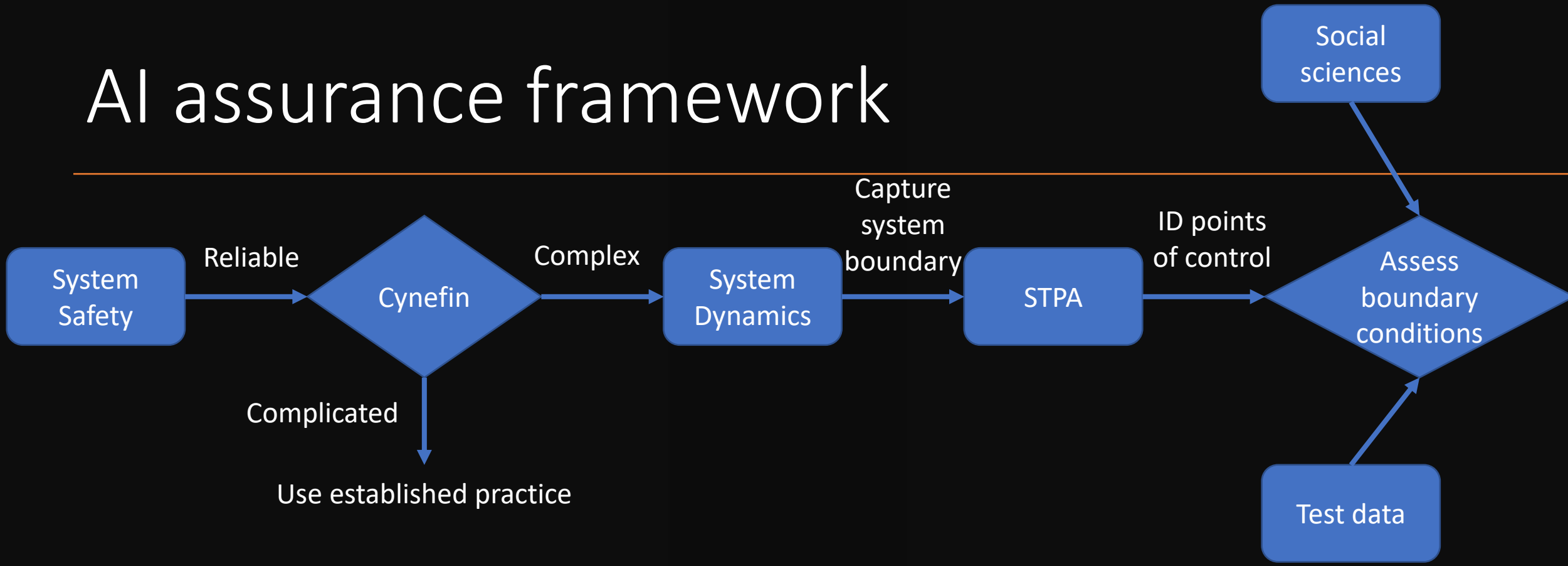


Tools that work

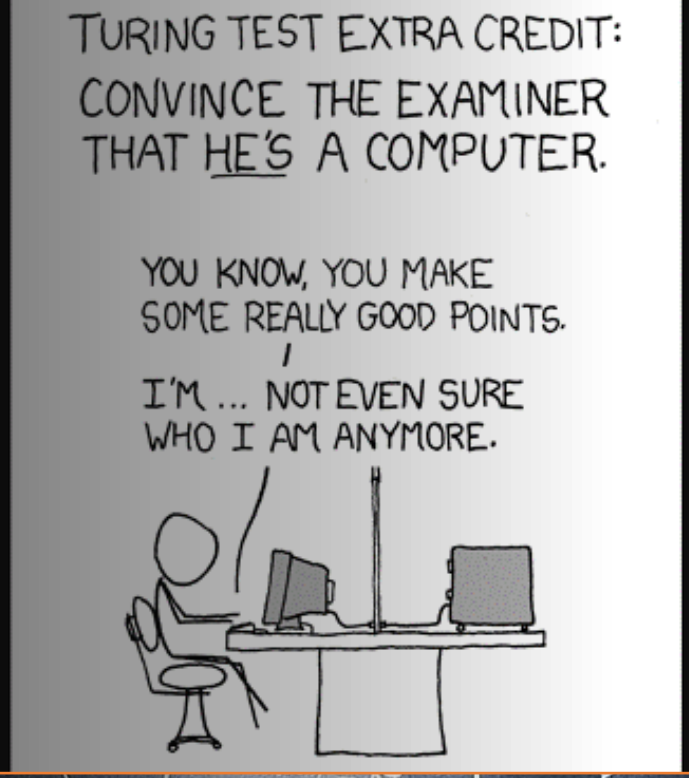
Test boundary conditions



AI assurance framework



Discussion


TURING TEST EXTRA CREDIT:
CONVINCE THE EXAMINER
THAT HE'S A COMPUTER.
YOU KNOW, YOU MAKE
SOME REALLY GOOD POINTS.
I'M ... NOT EVEN SURE
WHO I AM ANYMORE.