Responsible Research and Innovation in System Design

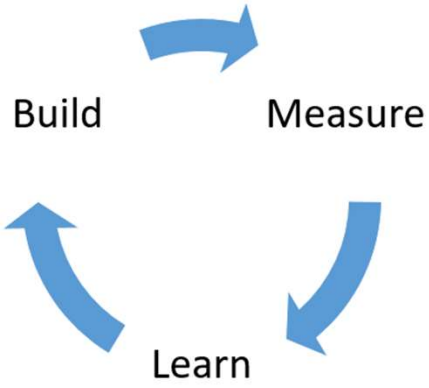
Dependable and Deployable AI for Robotics

Barbara Webb

- The Collingridge dilemma: *"attempting to control a technology is difficult…because during its early stages, when it can be controlled, not enough can be known about its harmful social consequences to warrant controlling its development; but by the time these consequences are apparent, control has become costly and slow"* (Collingridge, 1980: 19).
- Can we incorporate actions in our *current* design processes to try to minimise harmful (and increase beneficial) social consequences?

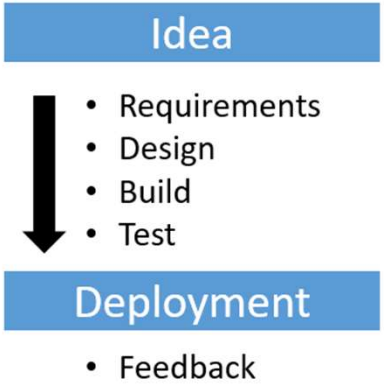
There are many alternative design approaches:

Agile

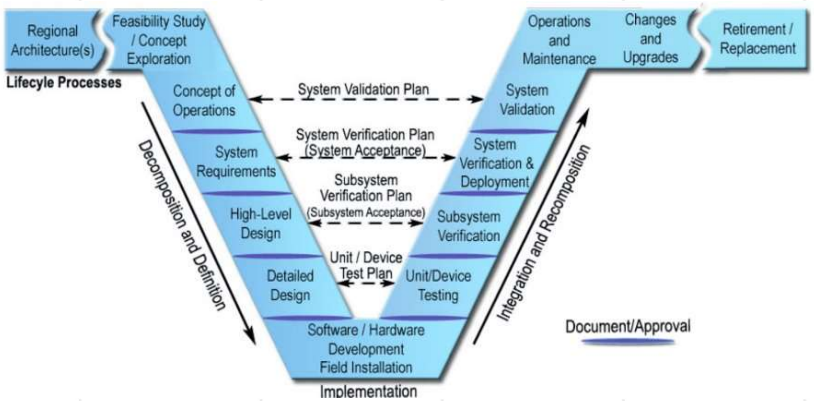


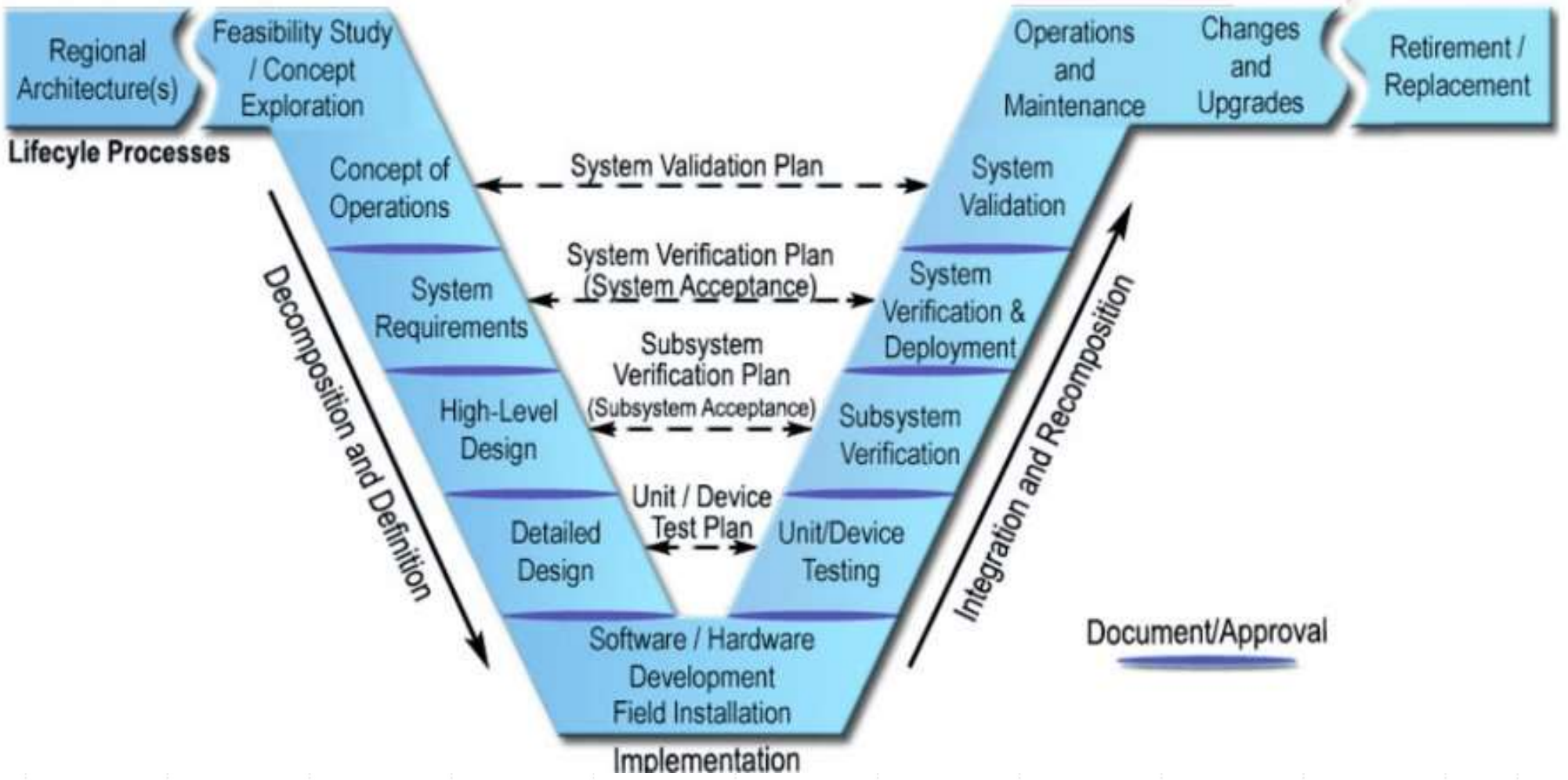
VS

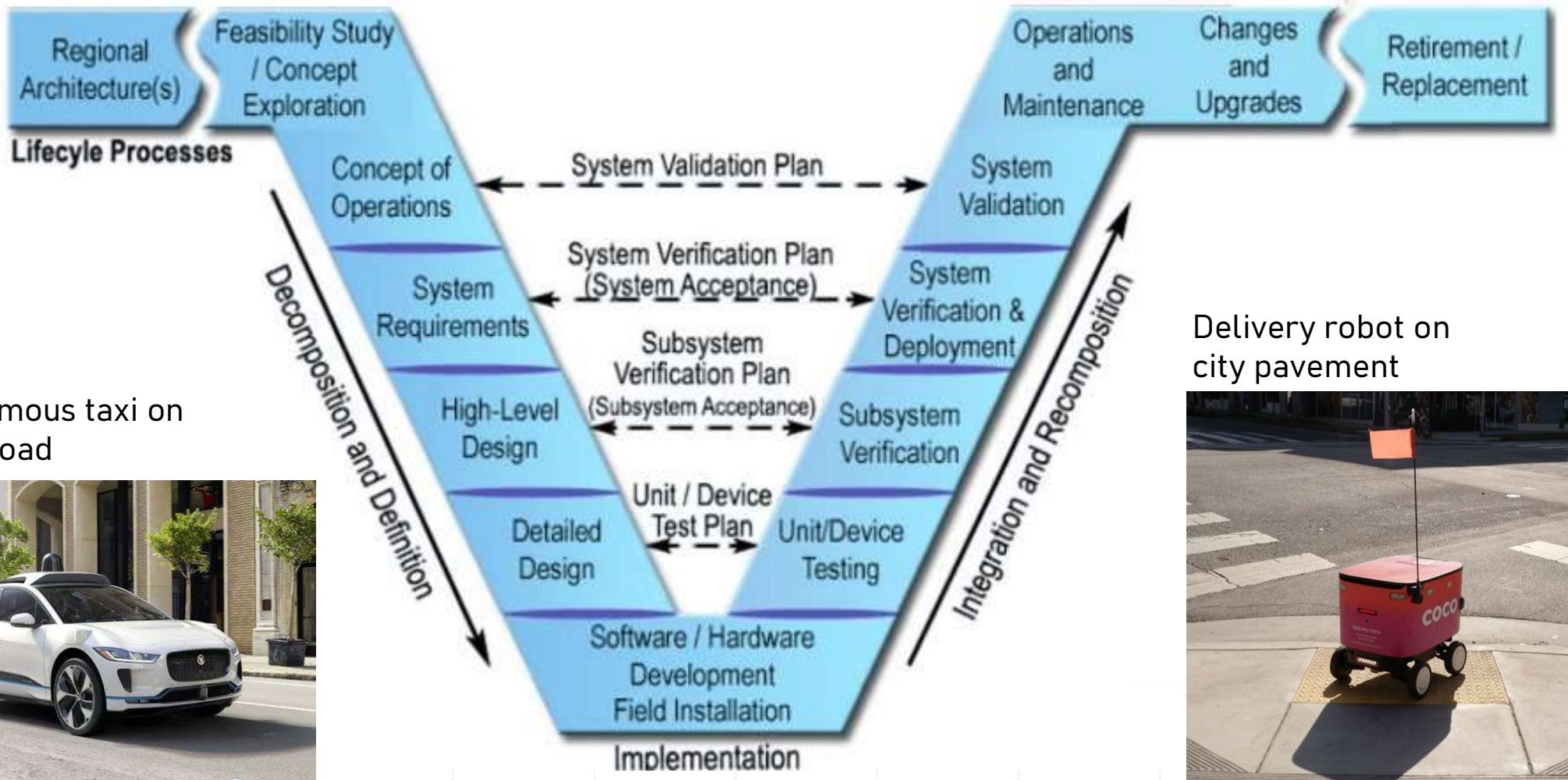
Waterfall



V-model



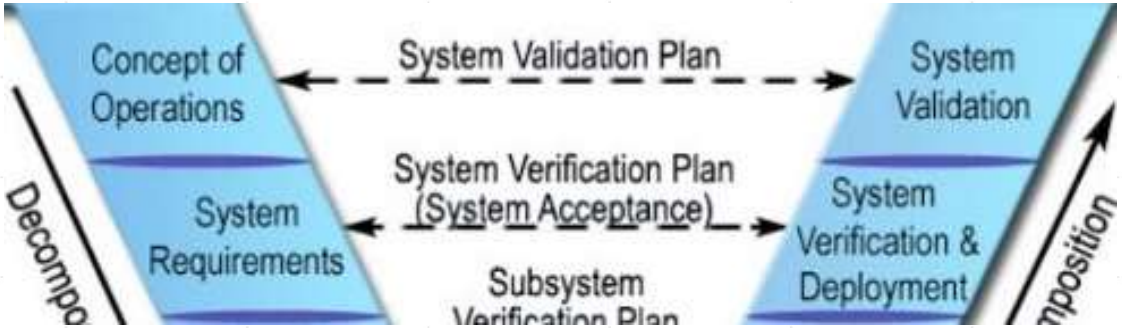

Regional Architecture(s)
Feasibility Study / Concept Exploration
Operations and Maintenance
Changes and Upgrades
Retirement / Replacement
Lifecyle Processes
Concept of Operations
System Validation Plan
System Validation
System Requirements
System Verification Plan (System Acceptance)
System Verification & Deployment
High-Level Design
Subsystem Verification Plan (Subsystem Acceptance)
Subsystem Verification
Detailed Design
Unit / Device Test Plan
Unit/Device Testing
Software / Hardware Development Field Installation
Implementation
Decomposition and Definition
Integration and Recomposition
Document/Approval



Autonomous taxi on urban road

Delivery robot on city pavement

Verification vs. Validation

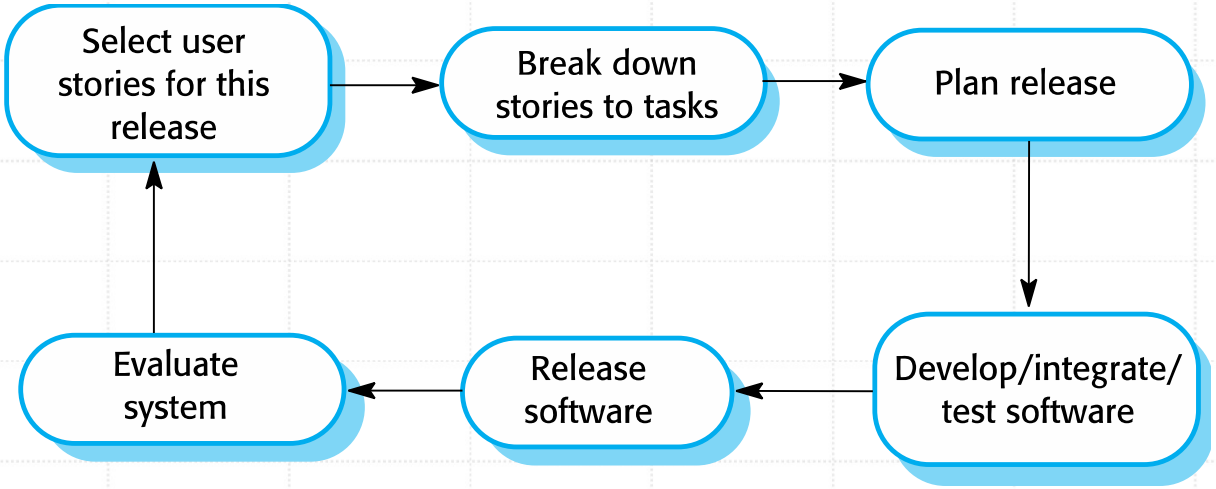


Verification: ensuring that the system meets its design requirements

Validation: ensuring the system meets the needs and requirements of intended users and environment

Was the specification itself the right one? Responsibility: did it include operational bounds aligned with values, needs and expectations of society?

The extreme/agile release cycle



Some principles of agile:

- Transparency: team responsibility to share information and team member responsibility to stay informed
- Inspection: both of what is produced, and of the process itself, to iteratively improve
- Adaptability: to the reality you are actually dealing with, rather than imagined scenario

Human-centred design

See also user-centred, participatory design, value-sensitive design

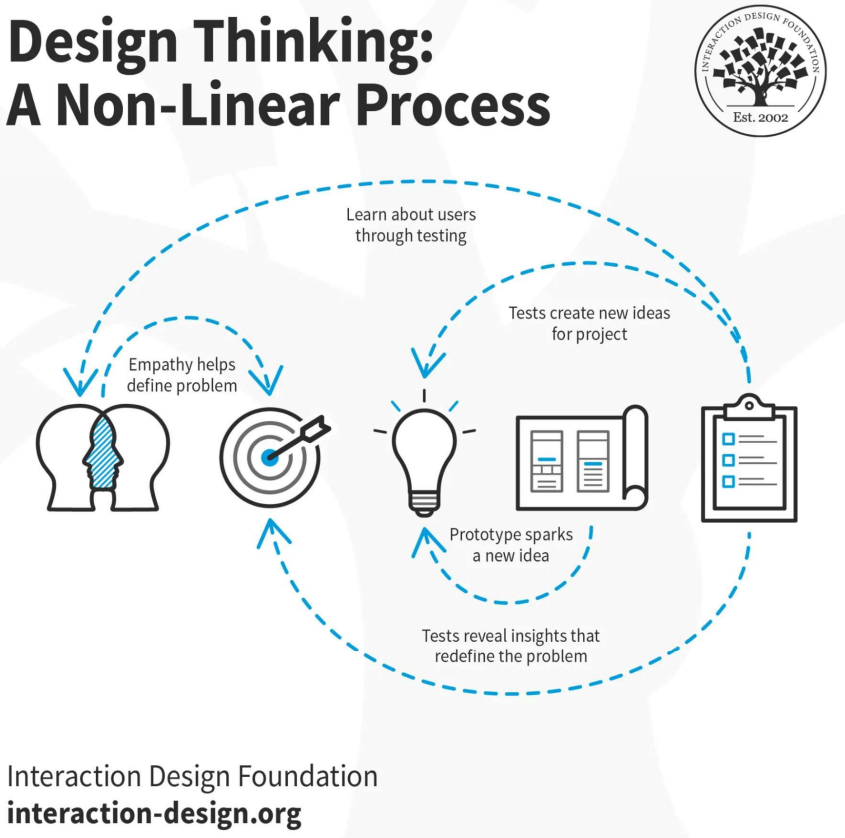
Four key principles:

- **People-centered**: Focus on people and their context to create appropriate solutions.
- **Understand and solve the right problems:** Understand the underlying fundamental issues.
- **Everything is a system:** Consider everything as a system of interconnected parts.
- **Small and simple interventions:** Do iterative work and get constant feedback.

Human-centred design

Five steps:

- **Empathize**: understand the user's point of view in depth; avoid assuming you know what the problem is.
- **Define**: analyze the core issues from step one to produce a clear problem statement..
- **Ideate**: Don't fixate on first solution but use brainstorming and other creativity tools to expand the solution space..
- **Prototype**: Create scaled down versions to test before sinking resource into full-scale implementation.
- **Test**: re-engage with the user



Critical design

- Creatively explore what might be
- Expressed as fictions, objects to think with, etc.
- Point is not useful or efficient engineering but to create a critical social context for the intended system
- For example, take goal of a project to its extreme, revealing how solving one problem might create a variety of others
- But sometimes reality is stranger than fiction…

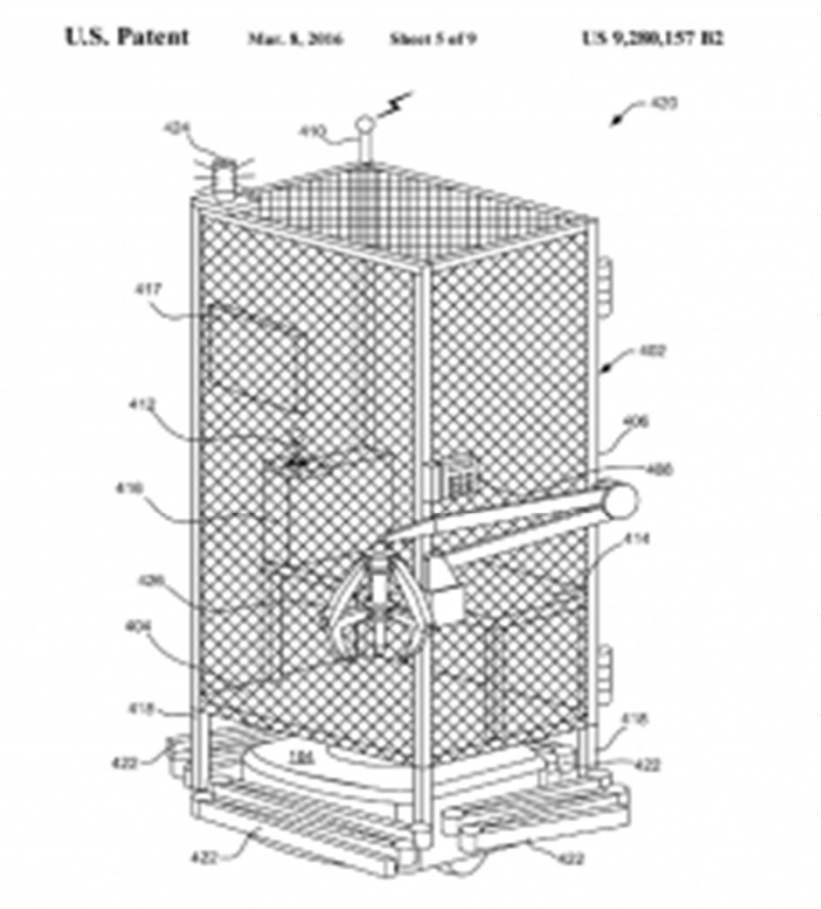
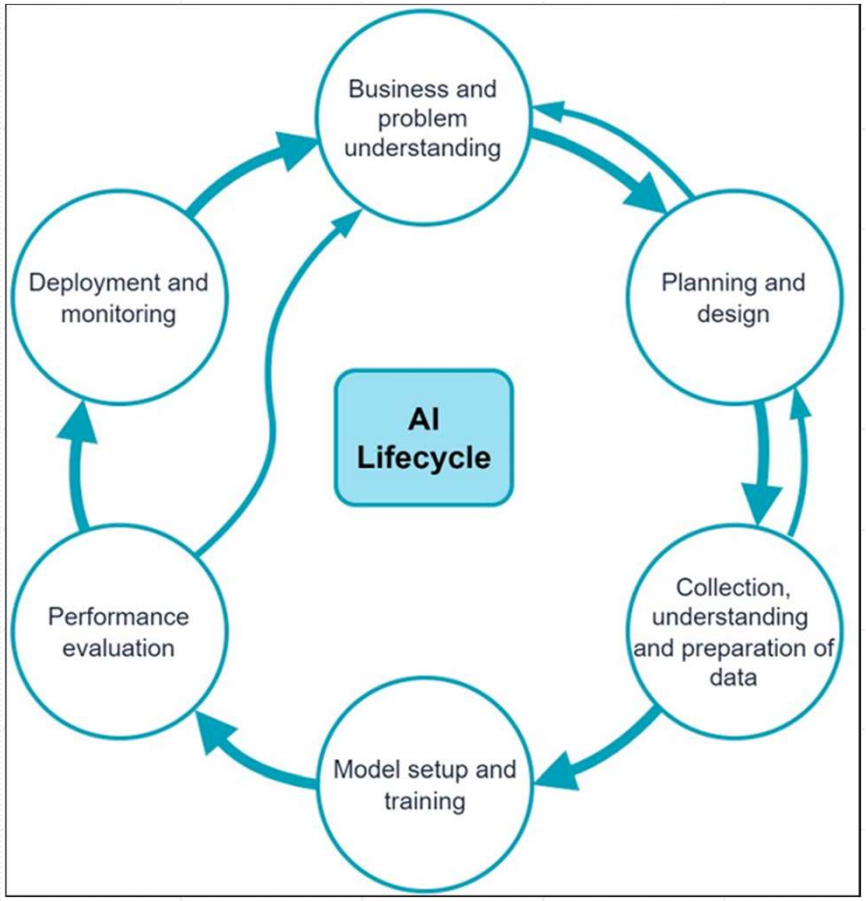


Illustration from a 2013 Amazon patent, “System and method for transporting personnel within an active workspace.”

Critical design

- Exercise:
- Can we use imaginative processes (fiction) to anticipate harmful (or beneficial) consequences?
- Utopia/dystopia: write a one paragraph story describing how the world might be different in 10/20/50 years if your colleague's research is massively successful...

RRI in the AI (ML) lifecycle



- Ortega-Bolaños, R., Bernal-Salcedo, J., Germán Ortiz, M. *et al.* Applying the ethics of AI: a systematic review of tools for developing and assessing AI-based systems. *Artif Intell Rev* **57**, 110 (2024). https://doi.org/10.1007/s10462-024-10740-3

Responsible ML engineering

"An ML system can satisfy latency, throughput, and aggregate accuracy requirements while reproducing historical discrimination, rewarding harmful engagement, consuming unjustified energy, or using data without adequate privacy and accountability controls. Such a system has not malfunctioned; it is efficiently optimizing an incomplete specification... Responsible engineering treats such outcomes as system failures to be diagnosed, measured, and mitigated with the same rigor as latency or accuracy regressions."

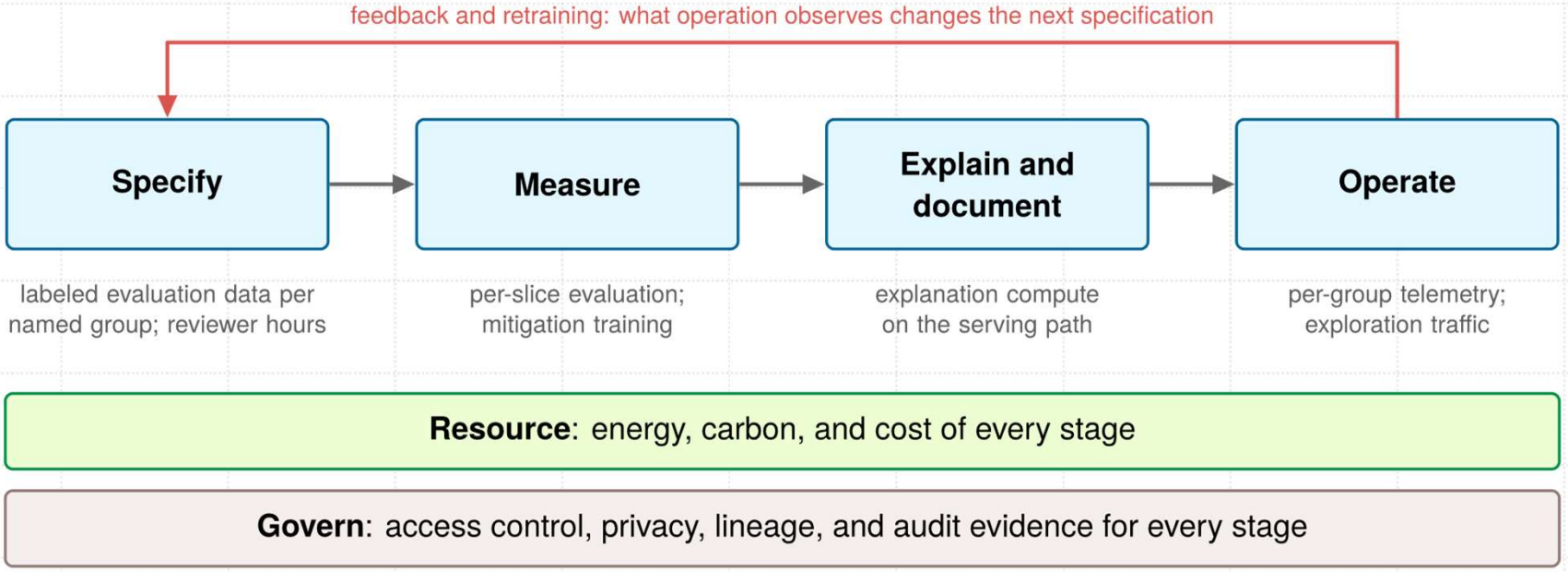
- Vijay Janapa Reddi Machine Learning Systems
- https://mlsysbook.ai/vol1/15_responsible_engr/15_responsible_engr.html

Responsible ML engineering

Some specific issues that may differ from conventional systems engineering:

- Lack of independent 'subsystems' that would allow potential issues to be isolated as clear subsystem specifications
- The source of problems (e.g. bias) can be hidden in the data, the loss function, the architecture, the training method, or subsequent optimisation: e.g. a compressed model may have accuracy loss that differs across demographic groups
- Resulting 'error' does not lead to failure to compile or run-time crash but can silently affect the whole system and take a long time to become measurable
- Feedback mechanisms can result in drift

Responsible ML engineering



- Vijay Janapa Reddi Machine Learning Systems
- https://mlsysbook.ai/vol1/15_responsible_engr/15_responsible_engr.html

For Friday's session

- In pairs, pick a current topic relevant to responsible research in AI for Robotics
- Research the current status or debate in this area and prepare a short presentation
- Some suggested topics (or suggest your own, discuss now):
  - Consequences of the introduction of foundation models in robotics
  - Application of robotics in warfare
  - Innovation outpacing regulation and standards
  - Privacy implications of social assistive robotics