Human-Centered AI:
Reliable, Safe & Trustworthy

Ben Shneiderman   **@benbendc**

Founding Director (1983-2000), Human-Computer Interaction Lab
Professor, Department of Computer Science

Member, National Academy of Engineering

Photo: BK Adams

UNIVERSITY OF MARYLAND

Interdisciplinary research community
- Computer Science & Info Studies
- Psych, Socio, Educ, Jour & MITH

`hcil.umd.edu`
`vimeo.com/72440805`

Designing the User Interface

Design Theories

Direct manipulation

Menus, speech, search

Social Media

Information Visualization

`www.cs.umd.edu/hcil/DTUI6`

Web links

The **University of Maryland, College Park** (often referred to as the **University of Maryland**, **Maryland**, **UM**, **UMD**, **UMCP**, or **College Park**) is a public research university[10] located in the city of College Park in Prince George's County, Maryland, approximately 4 miles (6.4 km) from the northeast border of Washington, D.C. Founded in 1856, the university is the flagship institution of the University System of Maryland. With a fall 2010 enrollment of more than 37,000 students, over 100 undergraduate majors, and 120 graduate programs,
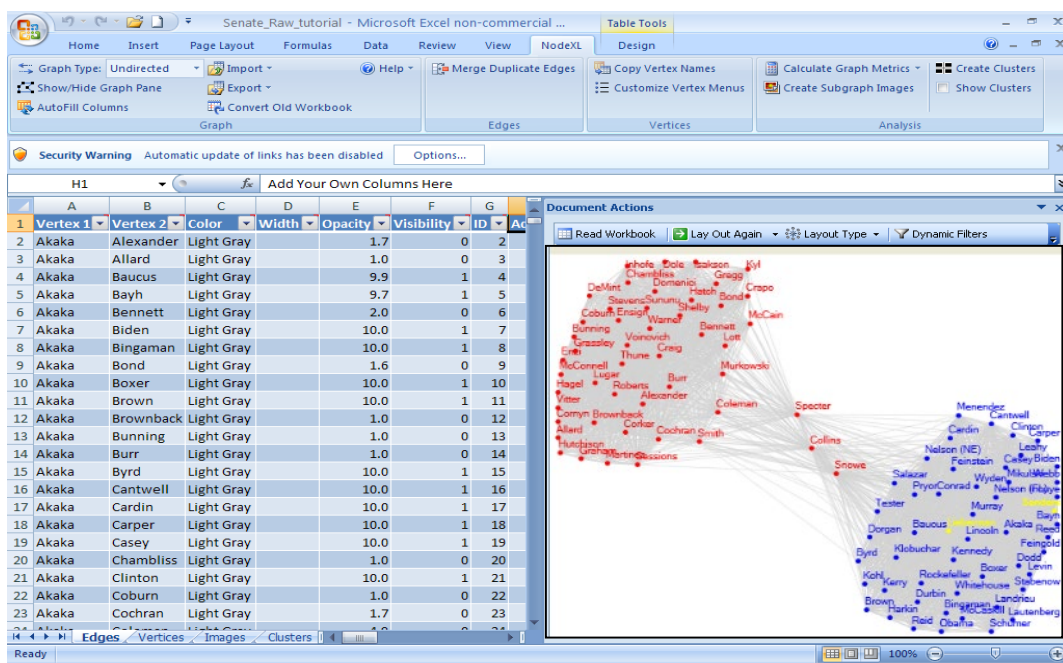
Tiny touchscreen keyboards
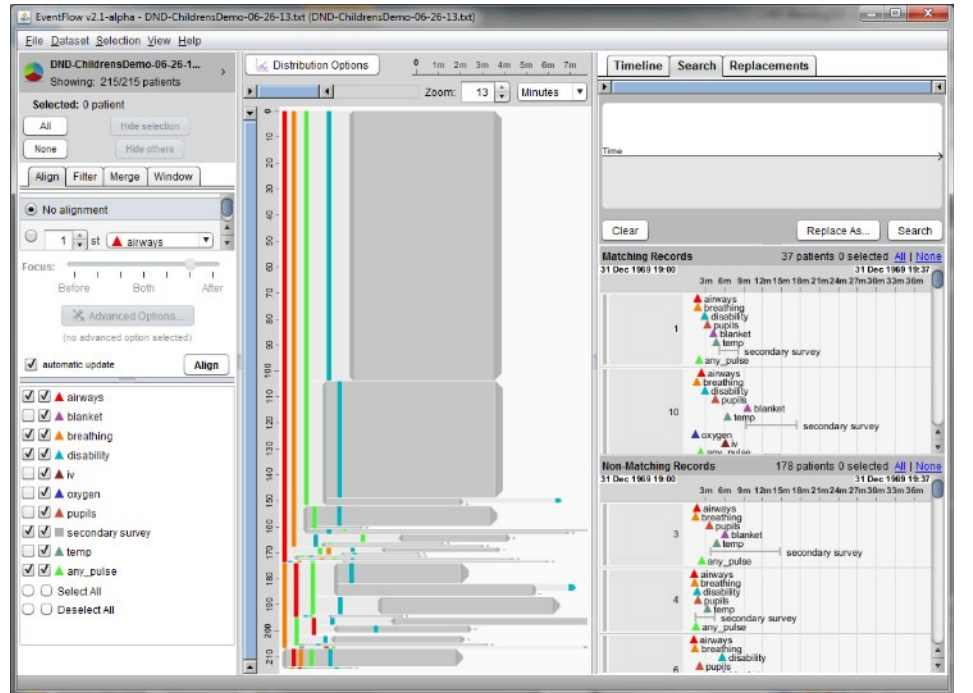

Photo tagging
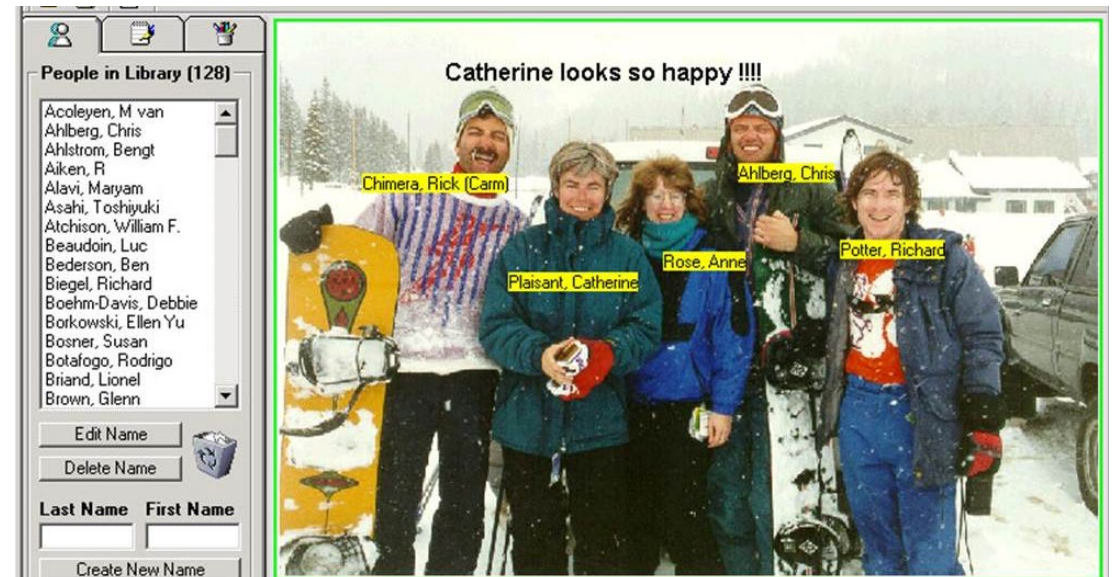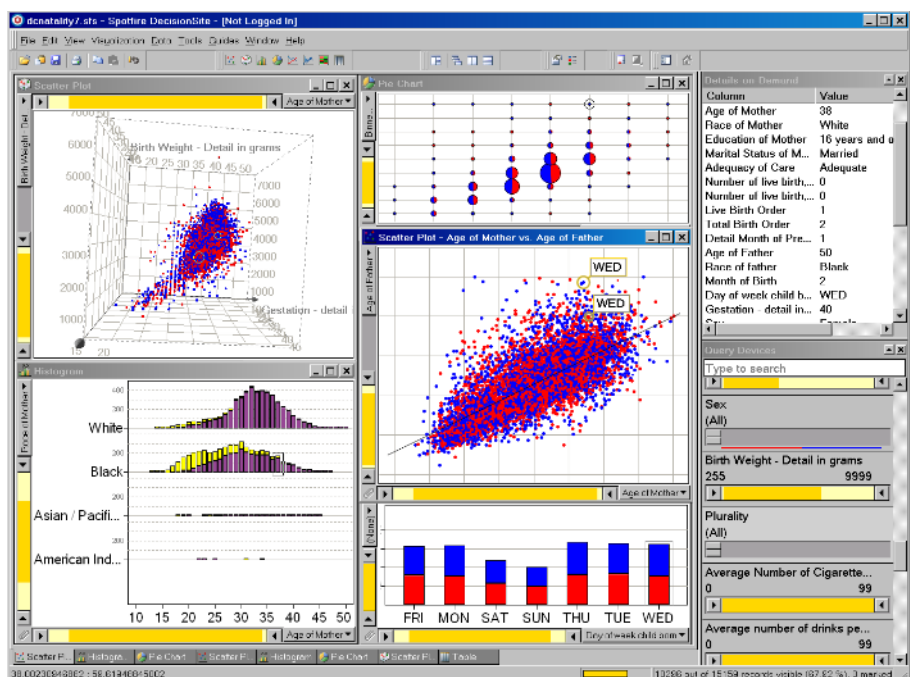
Spotfire

Treemaps
FinViz

NodeXL

EventFlow

What is Human-Centered AI?

What is Human-Centered AI?

Amplify, Augment, Empower & Enhance People

Human-Centered AI

Human Values
Rights, Justice & Dignity

Human-Centered AI

Human Values
Rights, Justice & Dignity

Individual Goals
Self-efficacy, Creativity, Responsibility & Social Connections

Human-Centered AI

Human Values
Rights, Justice & Dignity

Individual Goals
Self-efficacy, Creativity, Responsibility & Social Connections

Design Aspirations
Reliable, Safe & Trustworthy
Team, Organization, Industry & Government

Human-Centered AI

Stakeholders
- Researchers
- Developers
- Business Leaders
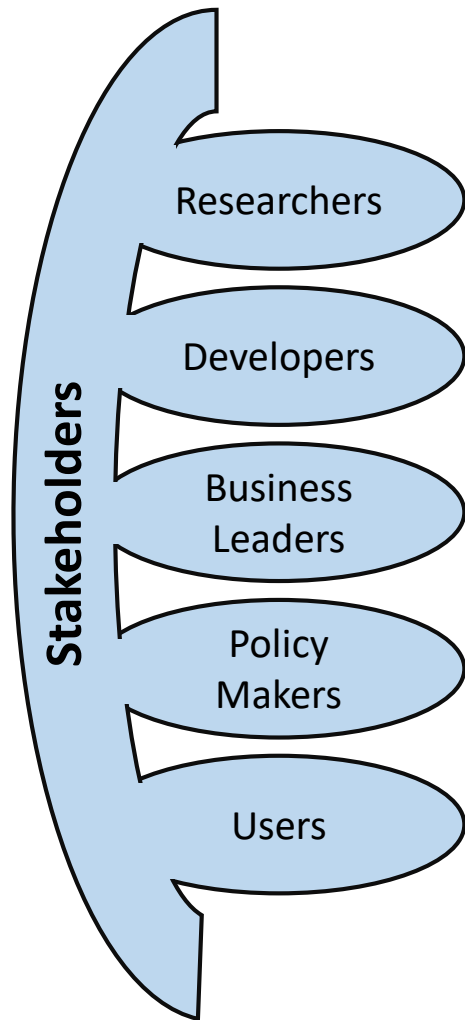- Policy Makers
- Users

Human Values
Rights, Justice & Dignity

Individual Goals
Self-efficacy, Creativity, Responsibility & Social Connections

Design Aspirations
Reliable, Safe & Trustworthy
Team, Organization, Industry & Government

Human-Centered AI

Stakeholders
- Researchers
- Developers
- Business Leaders
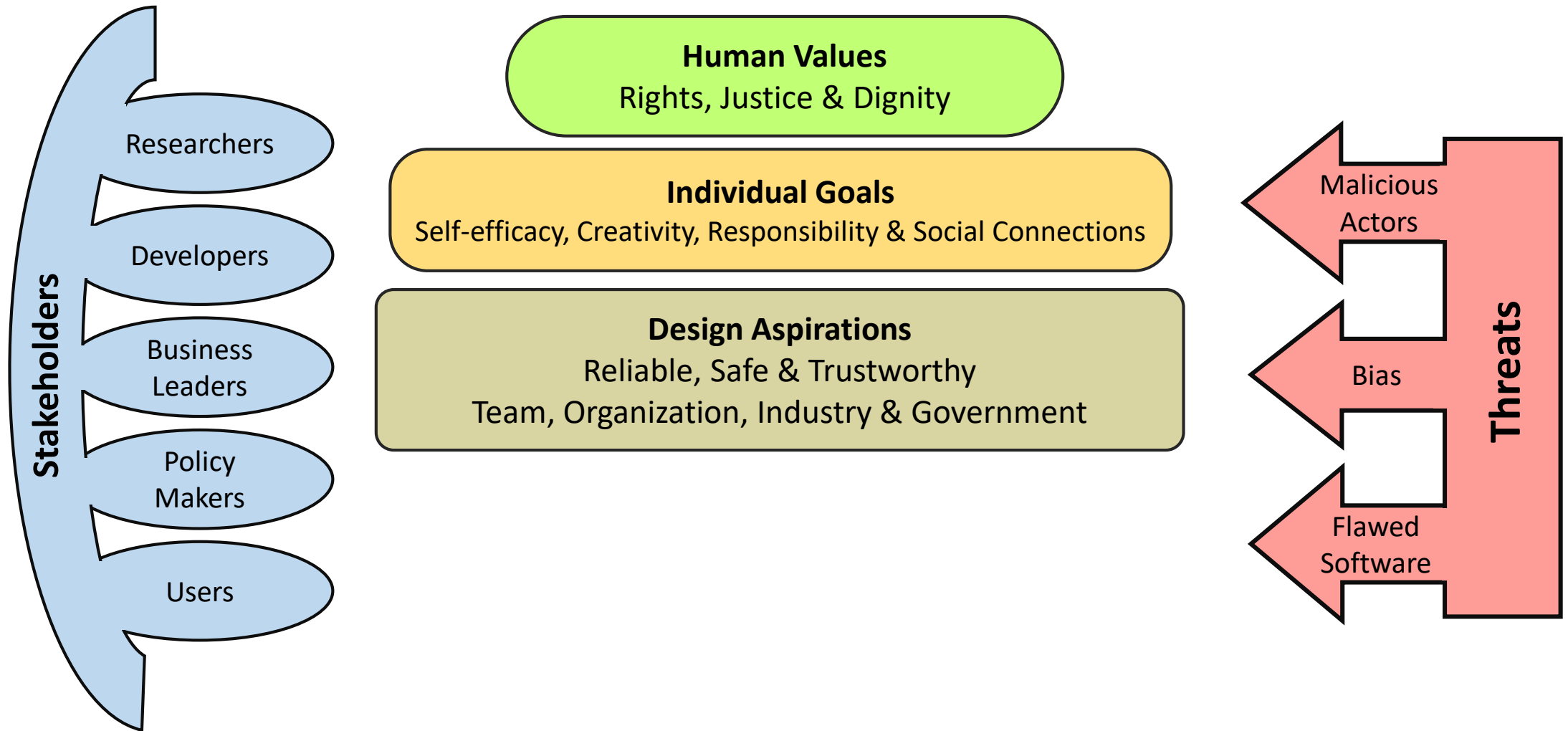- Policy Makers
- Users

Human Values
Rights, Justice & Dignity

Individual Goals
Self-efficacy, Creativity, Responsibility & Social Connections

Design Aspirations
Reliable, Safe & Trustworthy
Team, Organization, Industry & Government

Threats
- Malicious Actors
- Bias
- Flawed Software

Human-Centered AI

Stakeholders
- Researchers
- Developers
- Business Leaders
- Policy Makers
- Users

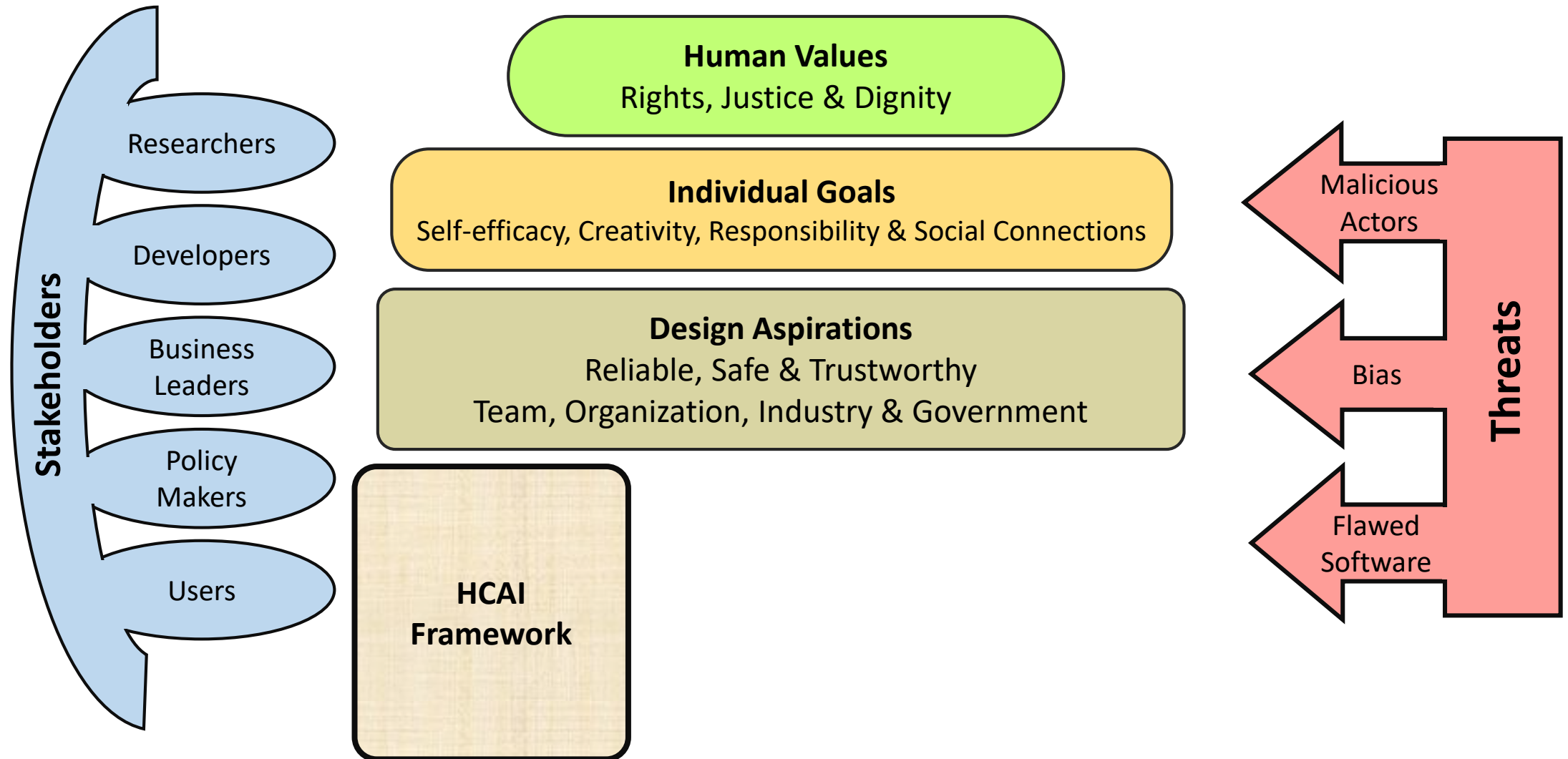**Human Values**
Rights, Justice & Dignity

Individual Goals
Self-efficacy, Creativity, Responsibility & Social Connections

Design Aspirations
Reliable, Safe & Trustworthy
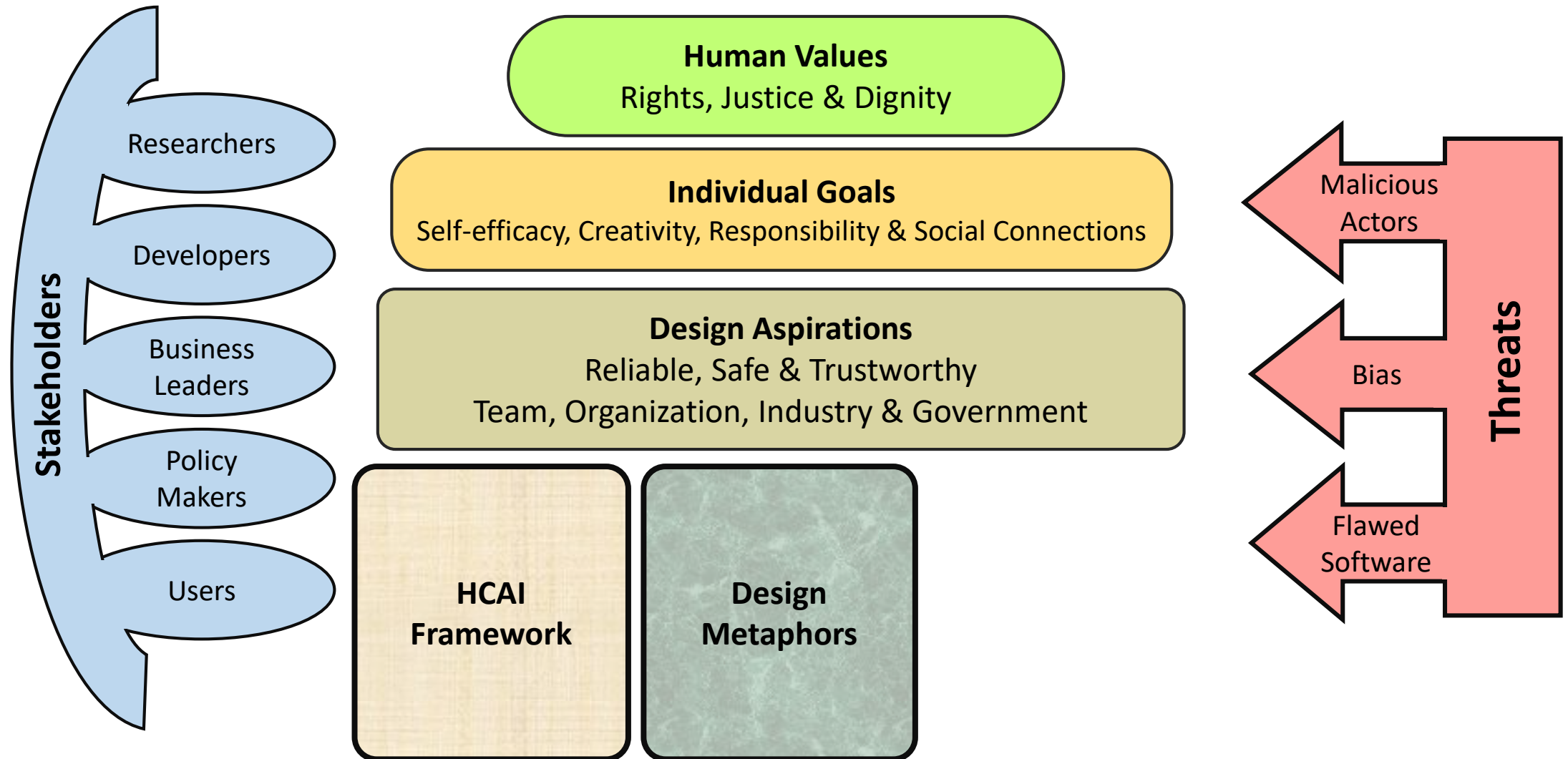Team, Organization, Industry & Government

HCAI Framework

Threats
- Malicious Actors
- Bias
- Flawed Software

Human-Centered AI

Stakeholders
- Researchers
- Developers
- Business Leaders
- Policy Makers
- Users

Human Values
Rights, Justice & Dignity

Individual Goals
Self-efficacy, Creativity, Responsibility & Social Connections

Design Aspirations
Reliable, Safe & Trustworthy
Team, Organization, Industry & Government

HCAI Framework
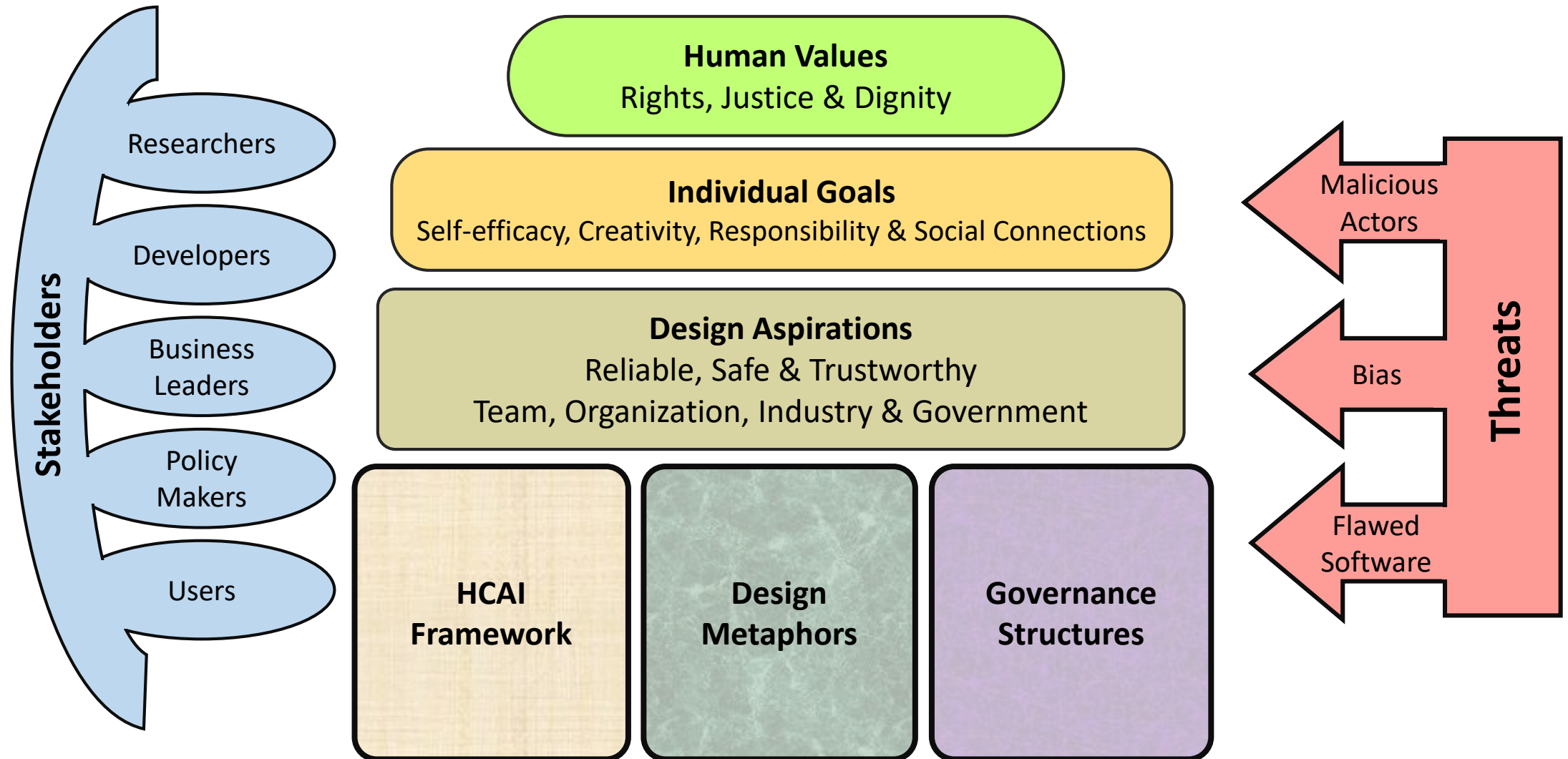
Design Metaphors

Threats
- Malicious Actors
- Bias
- Flawed Software

Human-Centered AI

Stakeholders
- Researchers
- Developers
- Business Leaders
- Policy Makers
- Users

Human Values
Rights, Justice & Dignity

Individual Goals
Self-efficacy, Creativity, Responsibility & Social Connections

Design Aspirations
Reliable, Safe & Trustworthy
Team, Organization, Industry & Government

HCAI Framework

Design Metaphors

Governance Structures

Threats
- Malicious Actors
- Bias
- Flawed Software

Oxford University Press (Early 2022) `https://hcil.umd.edu/human-centered-ai/`

HCAI Framework

Designing the User Interface

Balancing automation & human control

Designing the User Interface
Strategies for Effective Human-Computer Interaction
Ben Shneiderman

First Edition: 1986

Designing the User Interface

Balancing automation & human control

Human control

Computer automation

LEVELS OF DRIVING AUTOMATION

0	1	2	3	4	5
NO AUTOMATION	**DRIVER ASSISTANCE**	**PARTIAL AUTOMATION**	**CONDITIONAL AUTOMATION**	**HIGH AUTOMATION**	**FULL AUTOMATION**
Manual control. The human performs all driving tasks (steering, acceleration, braking, etc.).	The vehicle features a single automated system (e.g. it monitors speed through cruise control).	ADAS. The vehicle can perform steering and acceleration. The human still monitors all tasks and can take control at any time.	Environmental detection capabilities. The vehicle can perform most driving tasks, but human override is still required.	The vehicle performs all driving tasks under specific circumstances. Geofencing is required. Human override is still an option.	The vehicle performs all driving tasks under all conditions. Zero human attention or interaction is required.

THE HUMAN MONITORS THE DRIVING ENVIRONMENT | **THE AUTOMATED SYSTEM MONITORS THE DRIVING ENVIRONMENT**

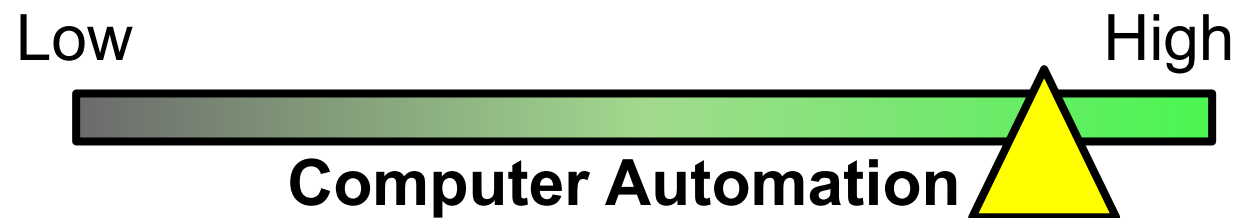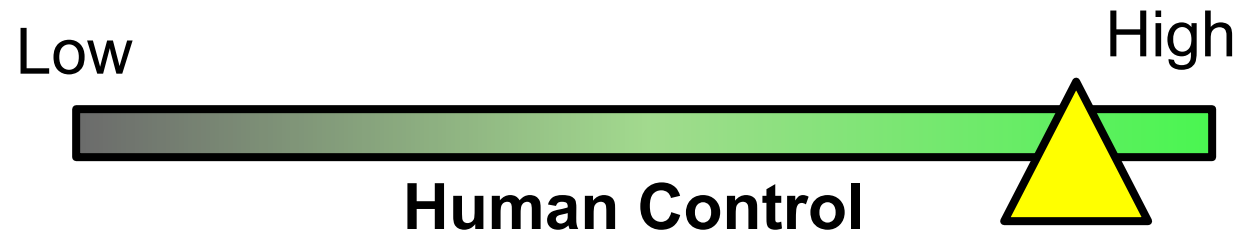(Society of Automotive Engineers, 2016)

Designing the User Interface

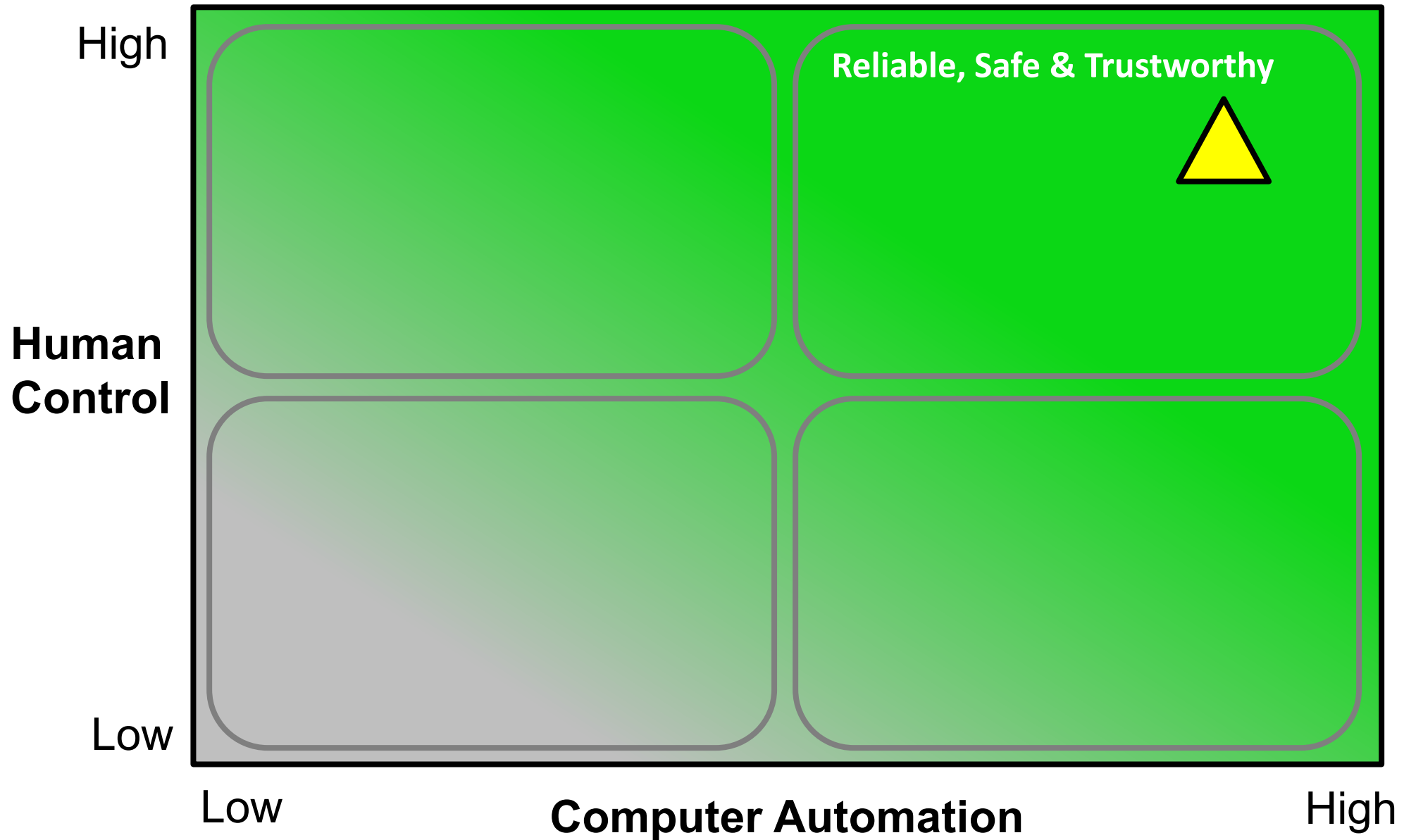Ensuring human control
while increasing automation

SHNEIDERMAN • PLAISANT • COHEN • JACOBS • ELMQVIST

SIXTH EDITION

DESIGNING THE USER INTERFACE
STRATEGIES FOR EFFECTIVE
HUMAN-COMPUTER INTERACTION

Sixth Edition: 2016

Designing the User Interface

Ensuring human control
while increasing automation

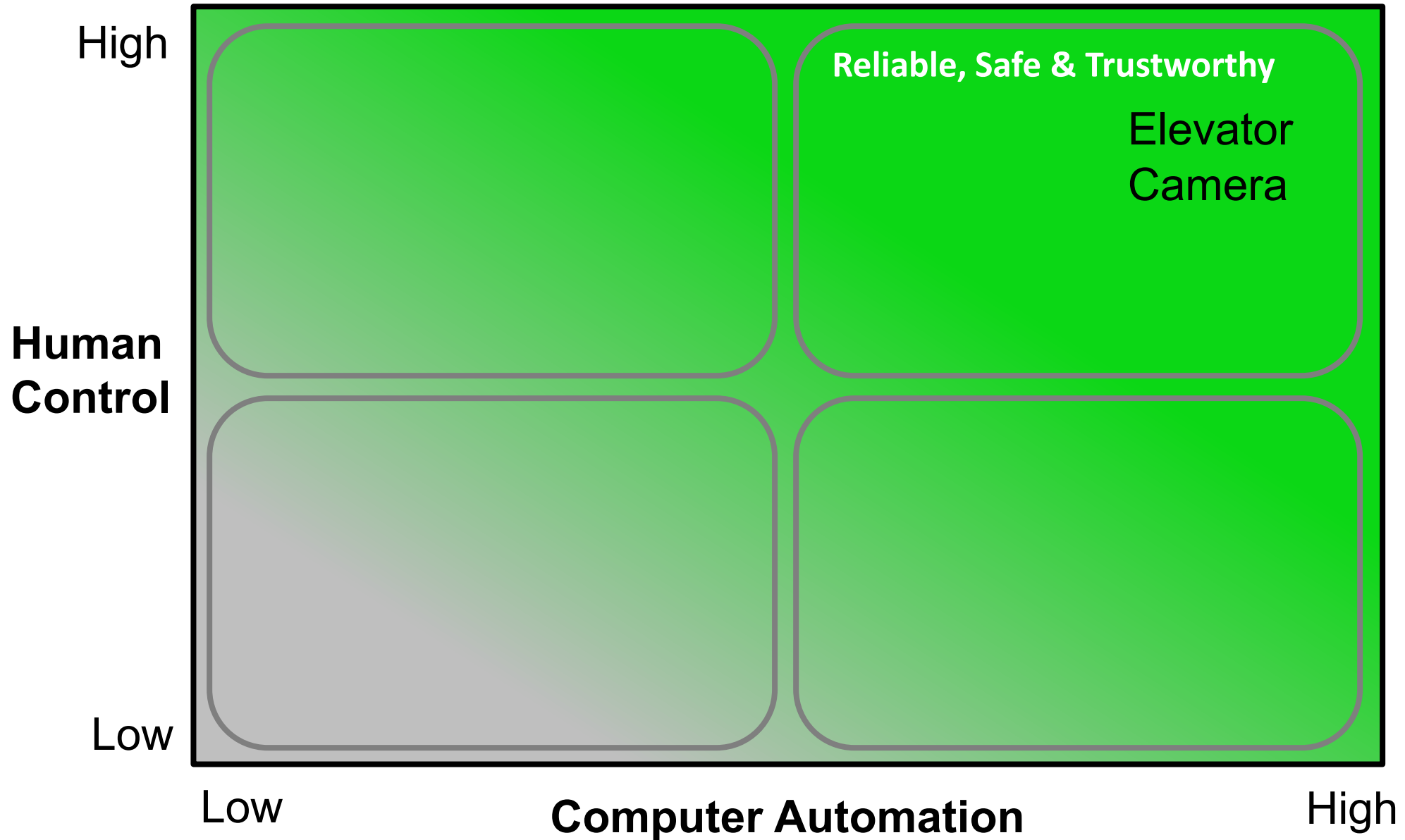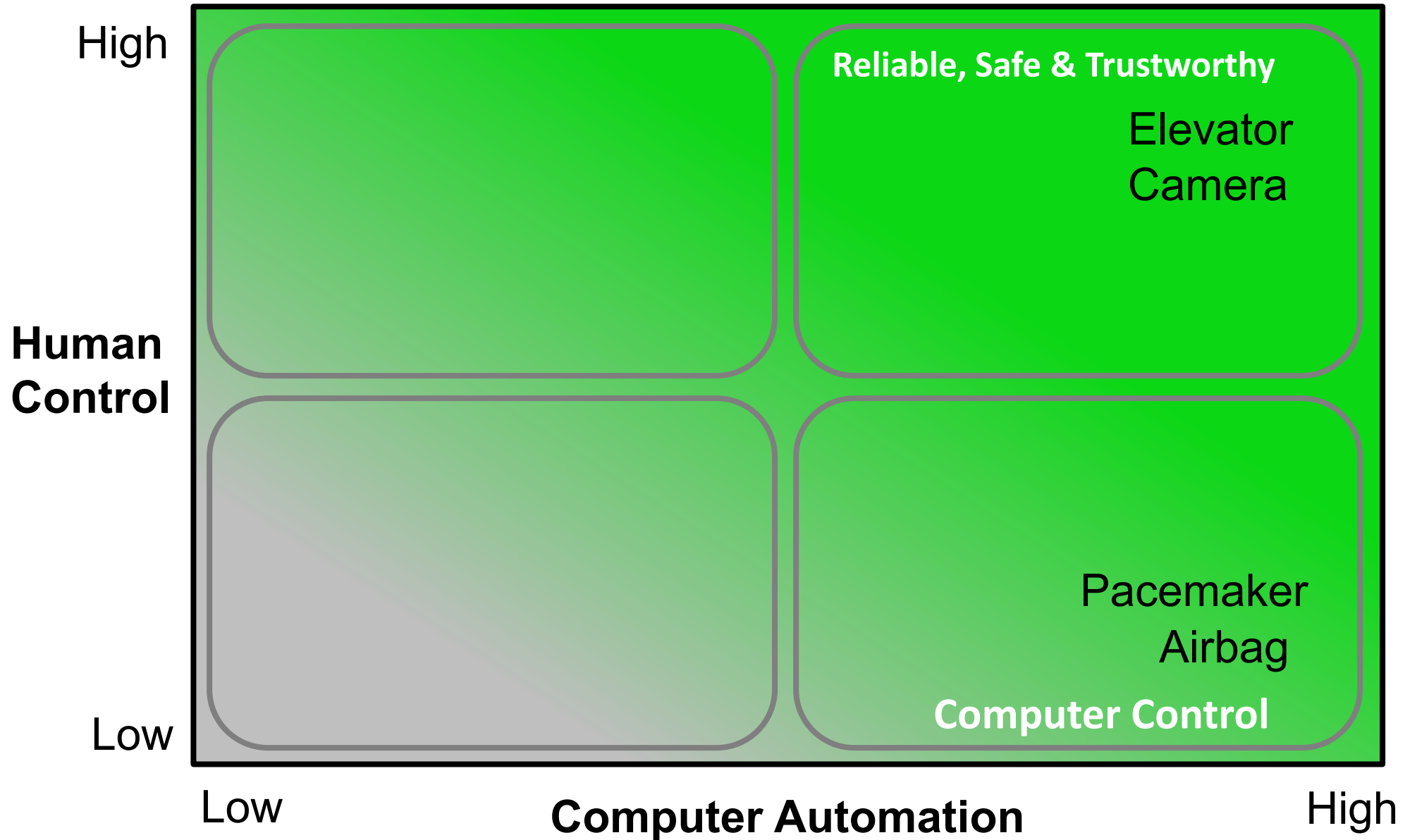Low   High

Human Control

Low   High

Computer Automation

IJHCI (2020) `https://doi.org/10.1080/10447318.2020.1741118`

SHNEIDERMAN • PLAISANT • COHEN • JACOBS • ELMQVIST

SIXTH EDITION

DESIGNING THE USER INTERFACE
STRATEGIES FOR EFFECTIVE
HUMAN-COMPUTER INTERACTION

Sixth Edition: 2016

Human-Centered AI

High

Human Control

Low

Computer Automation

Low   High

Reliable, Safe & Trustworthy
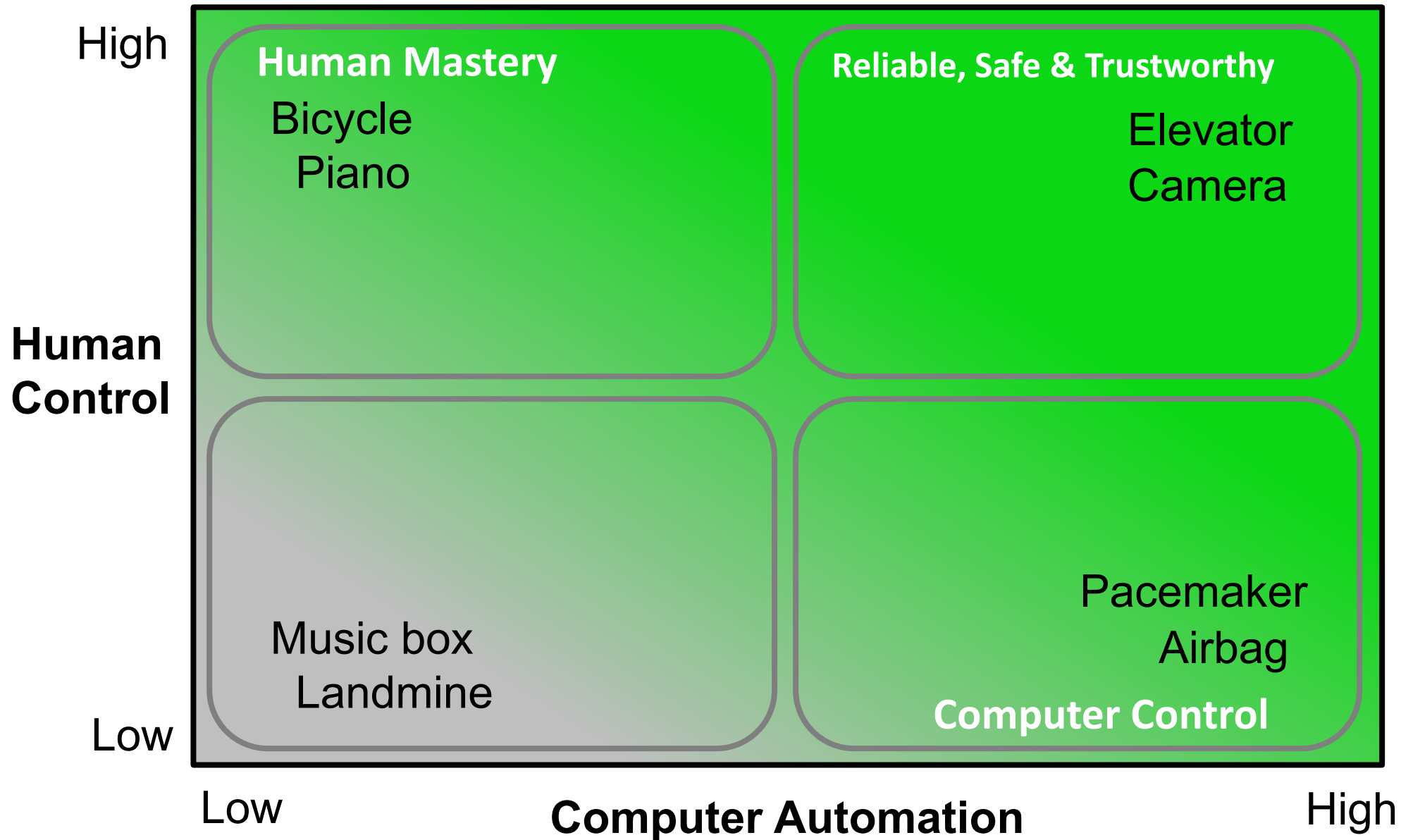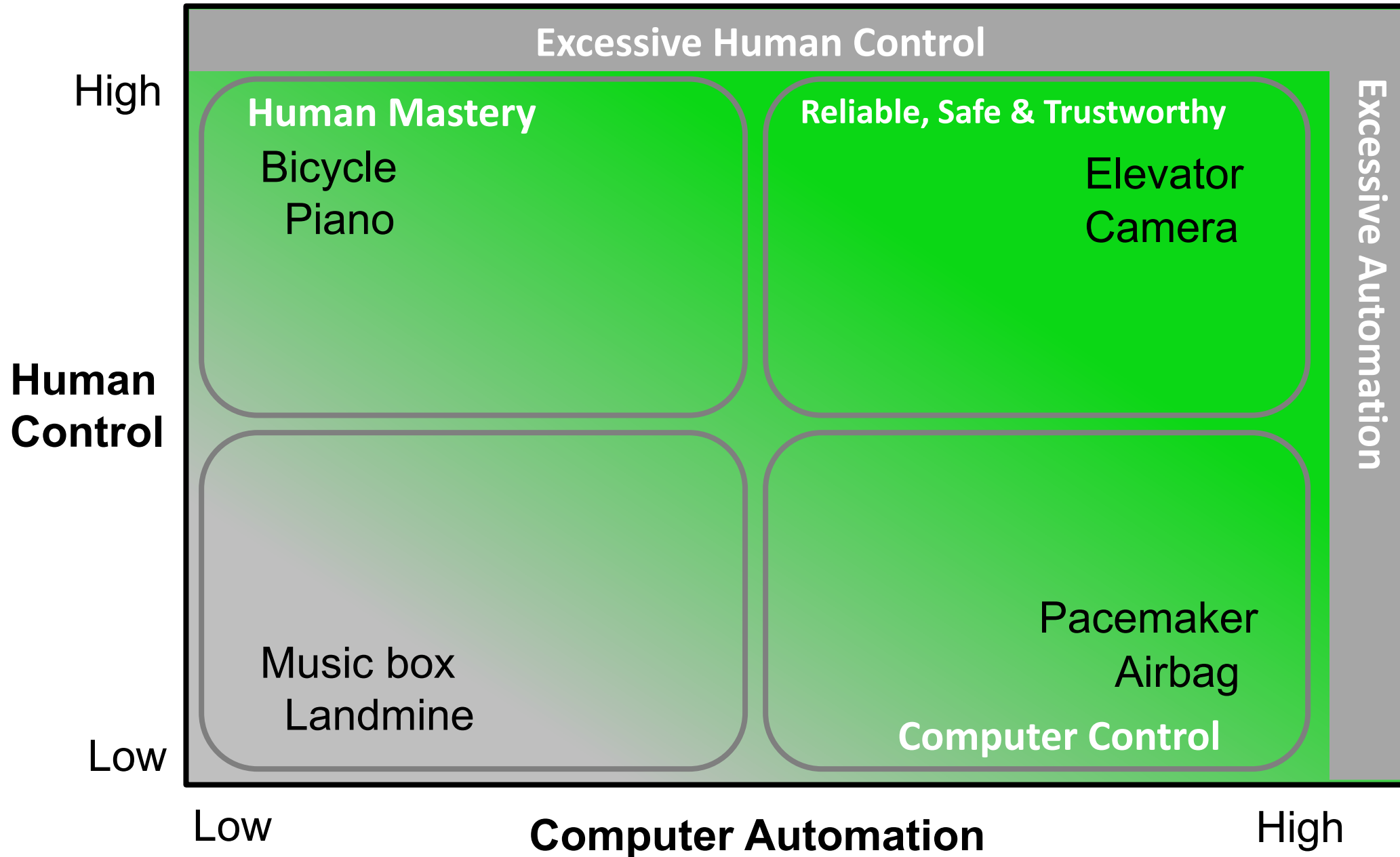
Human-Centered AI

Human-Centered AI

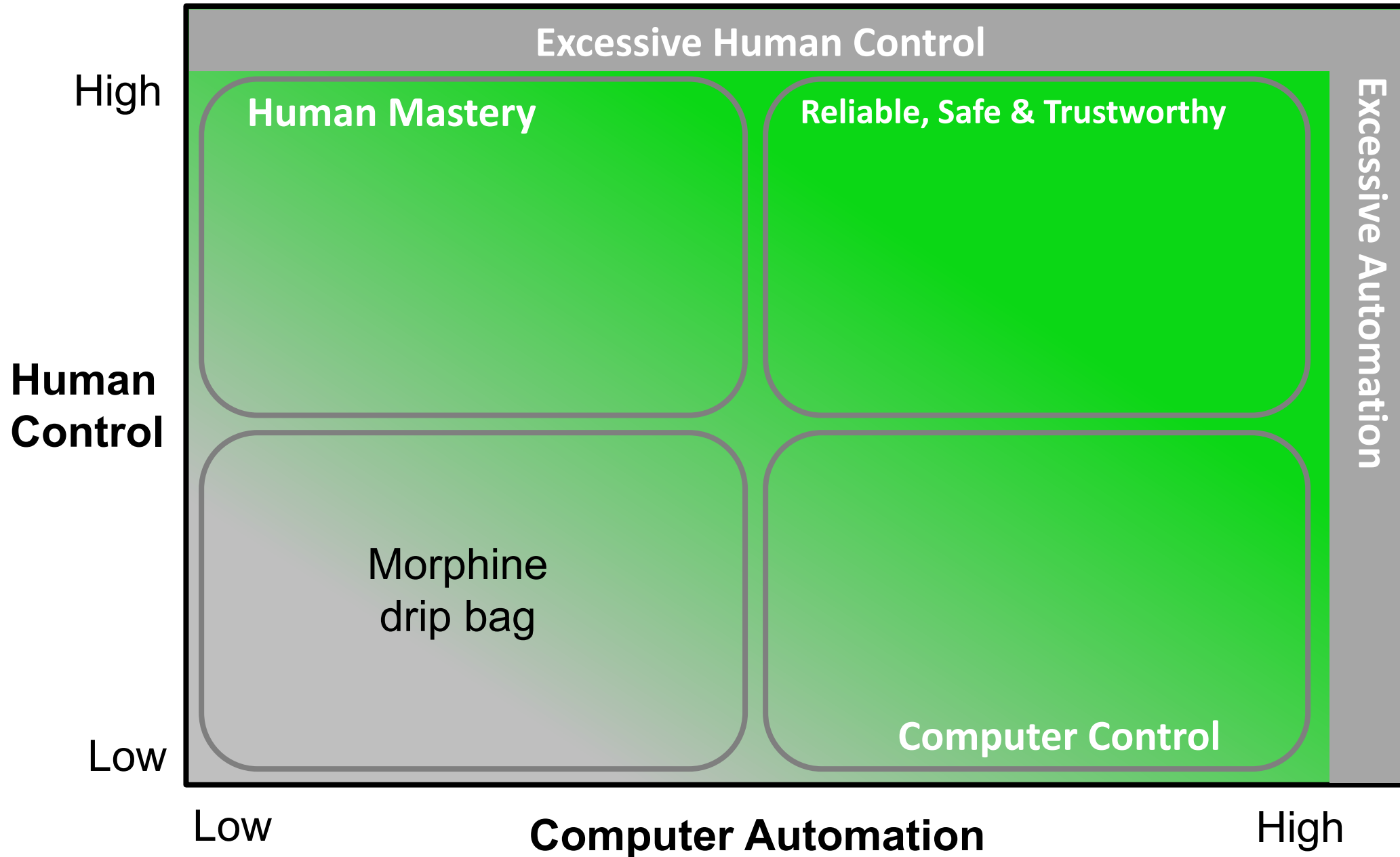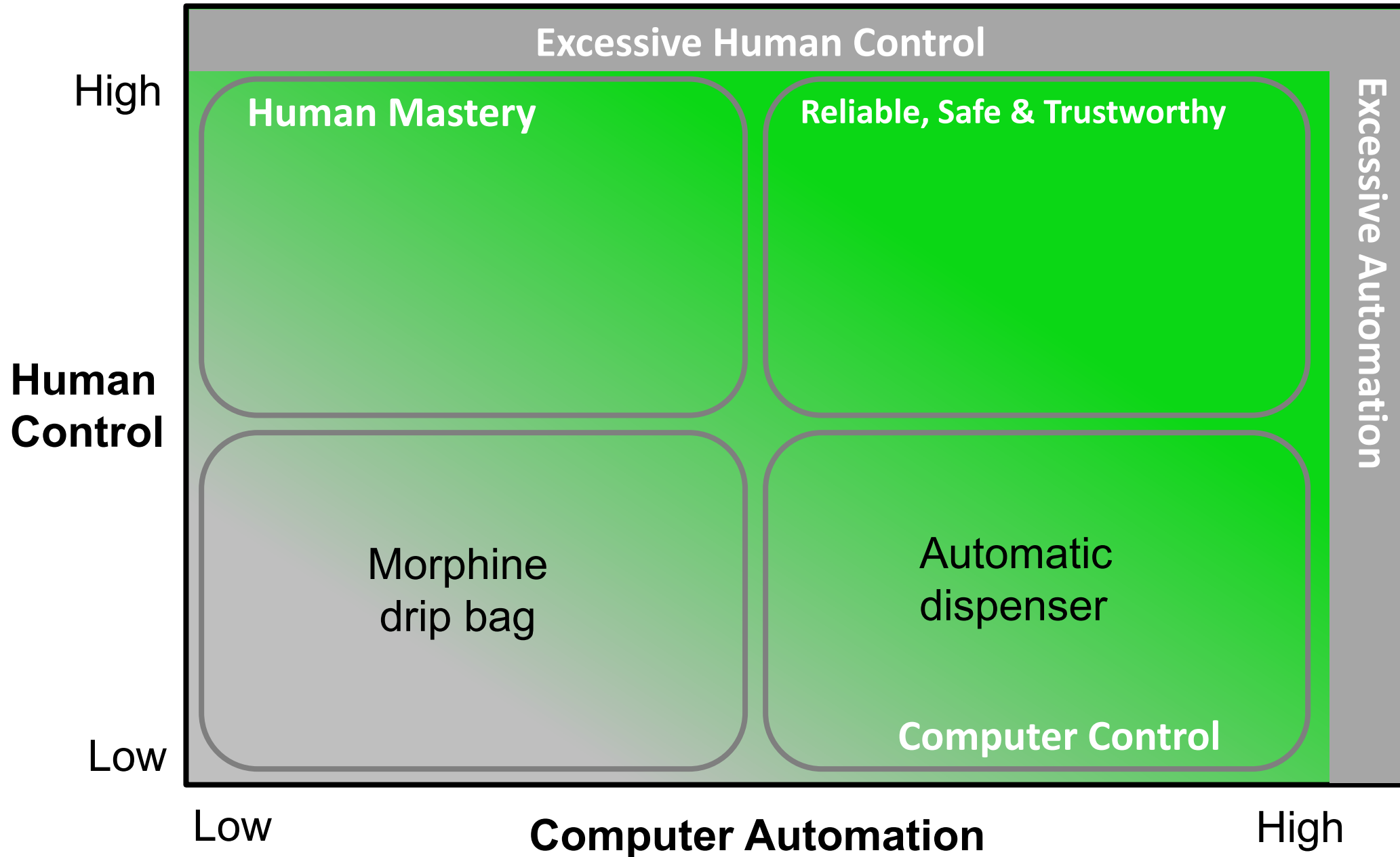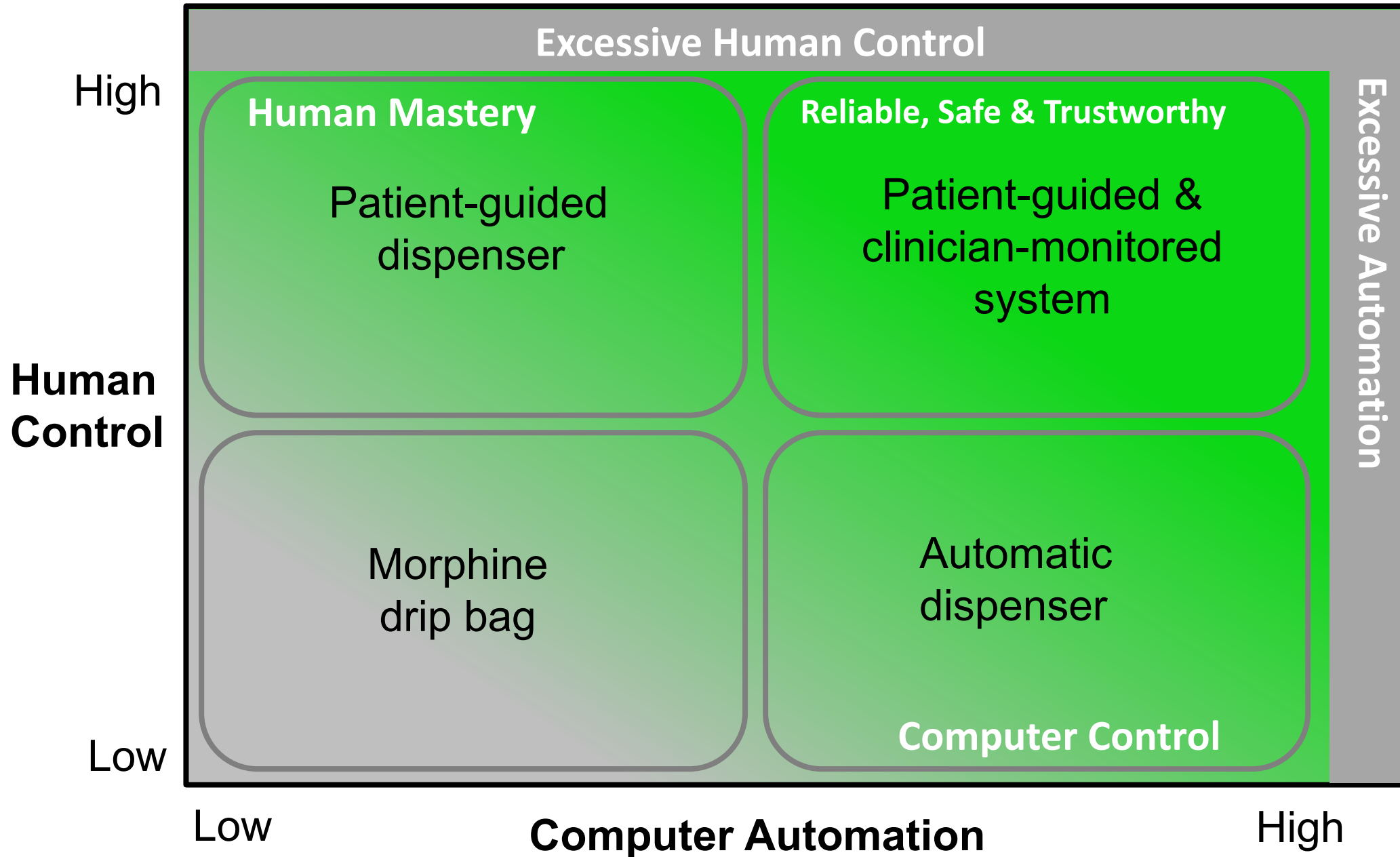**Human Control** (vertical axis: Low to High)

Computer Automation (horizontal axis: Low to High)

- **Human Mastery** (High Human Control, Low Computer Automation)
  - Bicycle
  - Piano

- **Reliable, Safe & Trustworthy** (High Human Control, High Computer Automation)
  - Elevator
  - Camera

- (Low Human Control, Low Computer Automation)
  - Music box
  - Landmine

- **Computer Control** (Low Human Control, High Computer Automation)
  - Pacemaker
  - Airbag

Human-Centered AI

Excessive Human Control

Human Mastery

Bicycle
Piano

Reliable, Safe & Trustworthy

Elevator
Camera

Music box
Landmine

Pacemaker
Airbag

Computer Control

Excessive Automation

High

Human Control

Low

Low **Computer Automation** High

Pain Control Designs

Excessive Human Control

High — **Human Mastery**   **Reliable, Safe & Trustworthy**

Human Control

Morphine drip bag

Computer Control

Low

Low   **Computer Automation**   **High**

Excessive Automation

Pain Control Designs

Pain Control Designs

Excessive Human Control

High — Human Mastery — Patient-guided dispenser

Reliable, Safe & Trustworthy

Human Control

Morphine drip bag

Automatic dispenser

Computer Control

Low

Excessive Automation

Low — High

Computer Automation
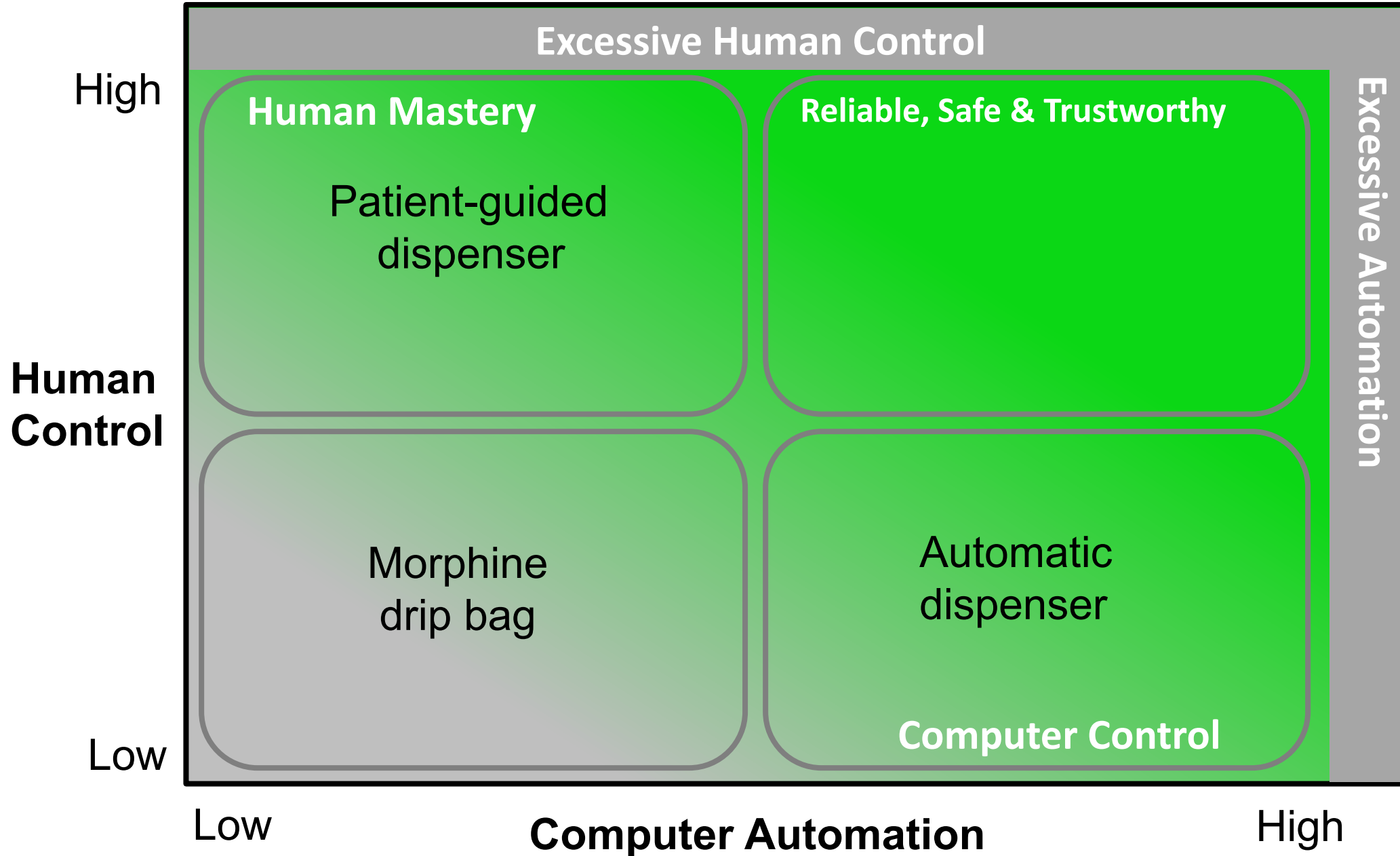
Pain Control Designs

Johns Hopkins University Hospital Control Center
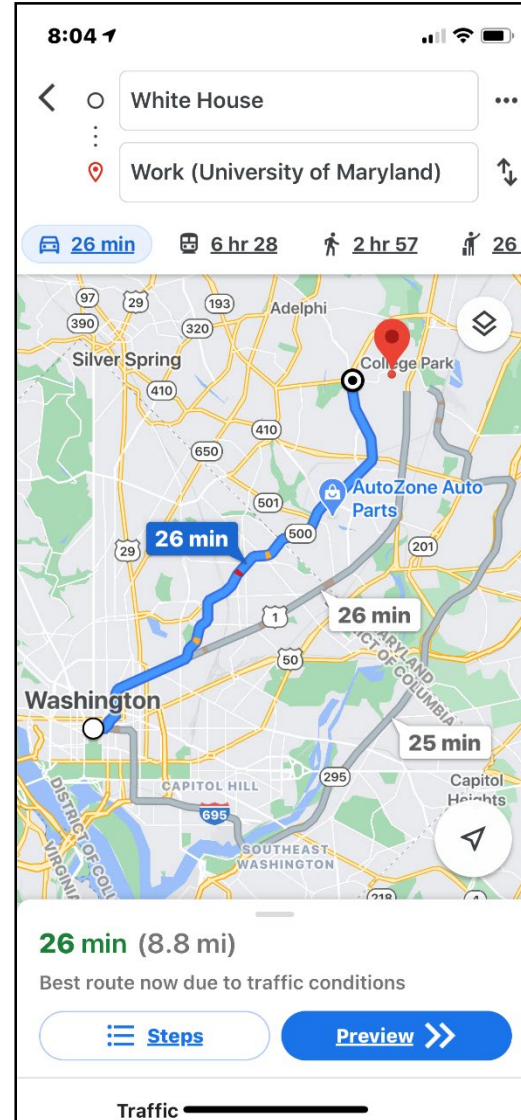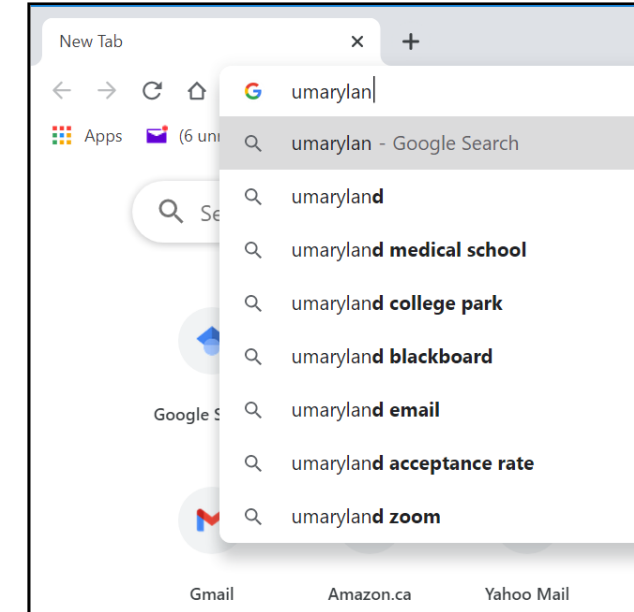
Design Metaphors

Design Metaphors

Intelligent Agent
Thinking Machine, Cognitive Actor, Artificial Intelligence, Knowledgeable

Humanoid Robot
Anthropomorphic, Android, Bionic, Bio-inspired

Simulated Teammate
Co-active Collaborator, Colleague, Helpful Partner, Smart Co-worker

Autonomous System
Independent, Self-directed, Goal-setting, Self-monitored

Combined Designs

SuperTools
Extend Abilities, Empower Users, Enhance Human Performance

Active Appliances
Steerable Equipment, Expendable, Increase Human Flexibility & Mobility

Tele-Bots
Dextrous Instrument, Powerful Prosthetic, Boost Human Perceptual & Motor Skills

Control Centers
Human Control & Oversight, Situation Awareness, Predictable Actions

Supertools

Digital Camera Controls   **Navigation Choices**   **Texting Autocompletion**

Spelling correction
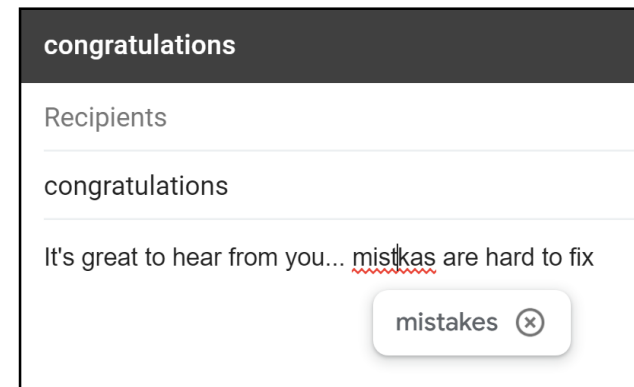
Supertool: Bloomberg Terminal

Active Appliances

Coffee maker, Rice cooker, Blender
Dishwasher, Clothes Washer/Dryer

Panasonic Rice Cooker

Miele Dishwasher

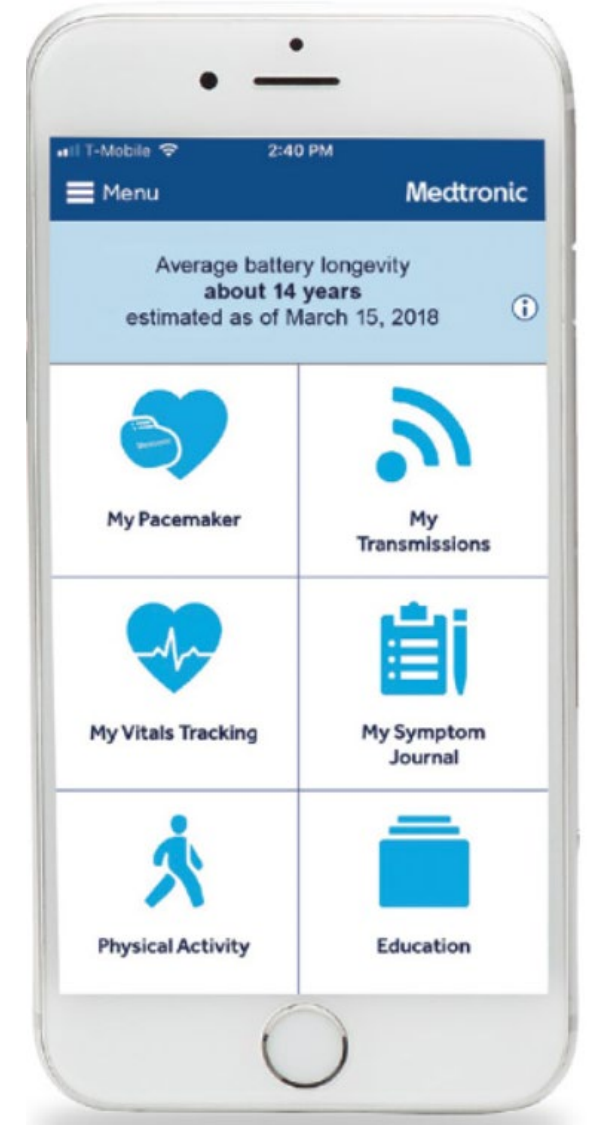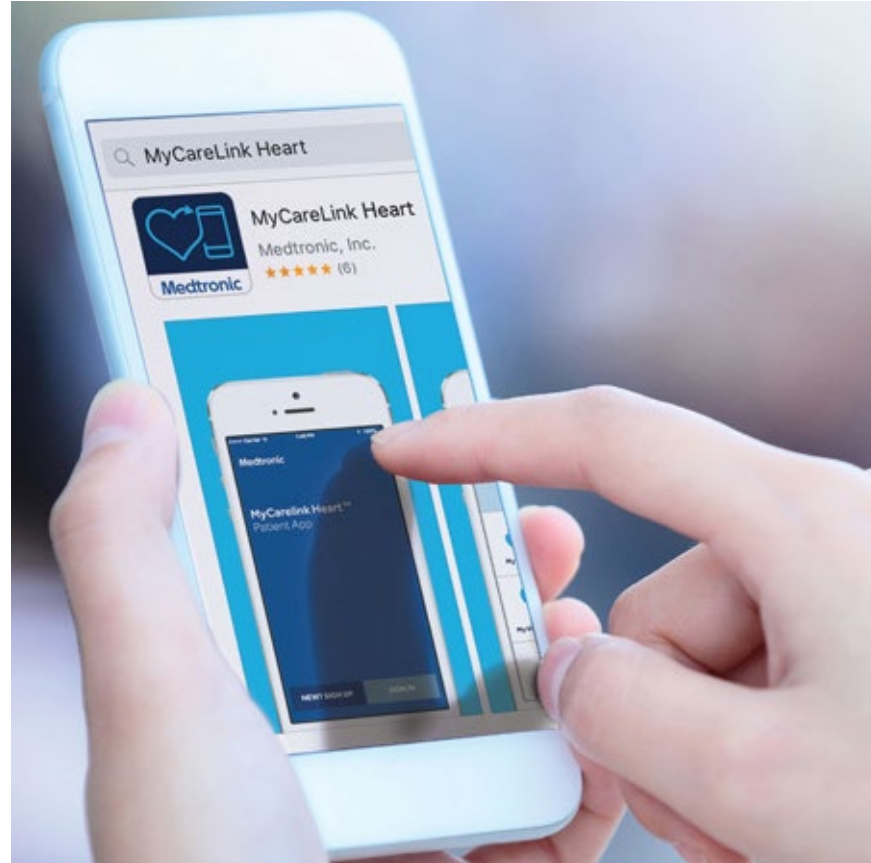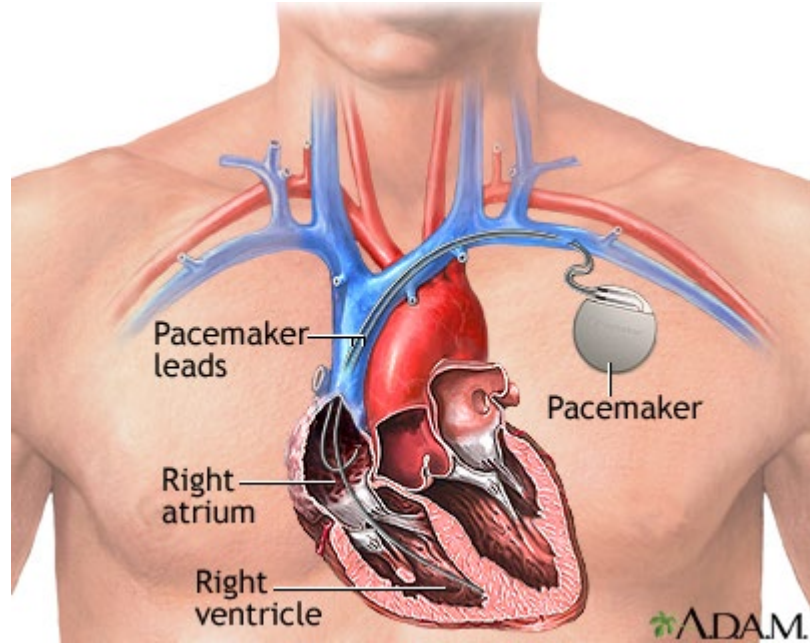Cuisinart Grind & Brew Coffee Maker
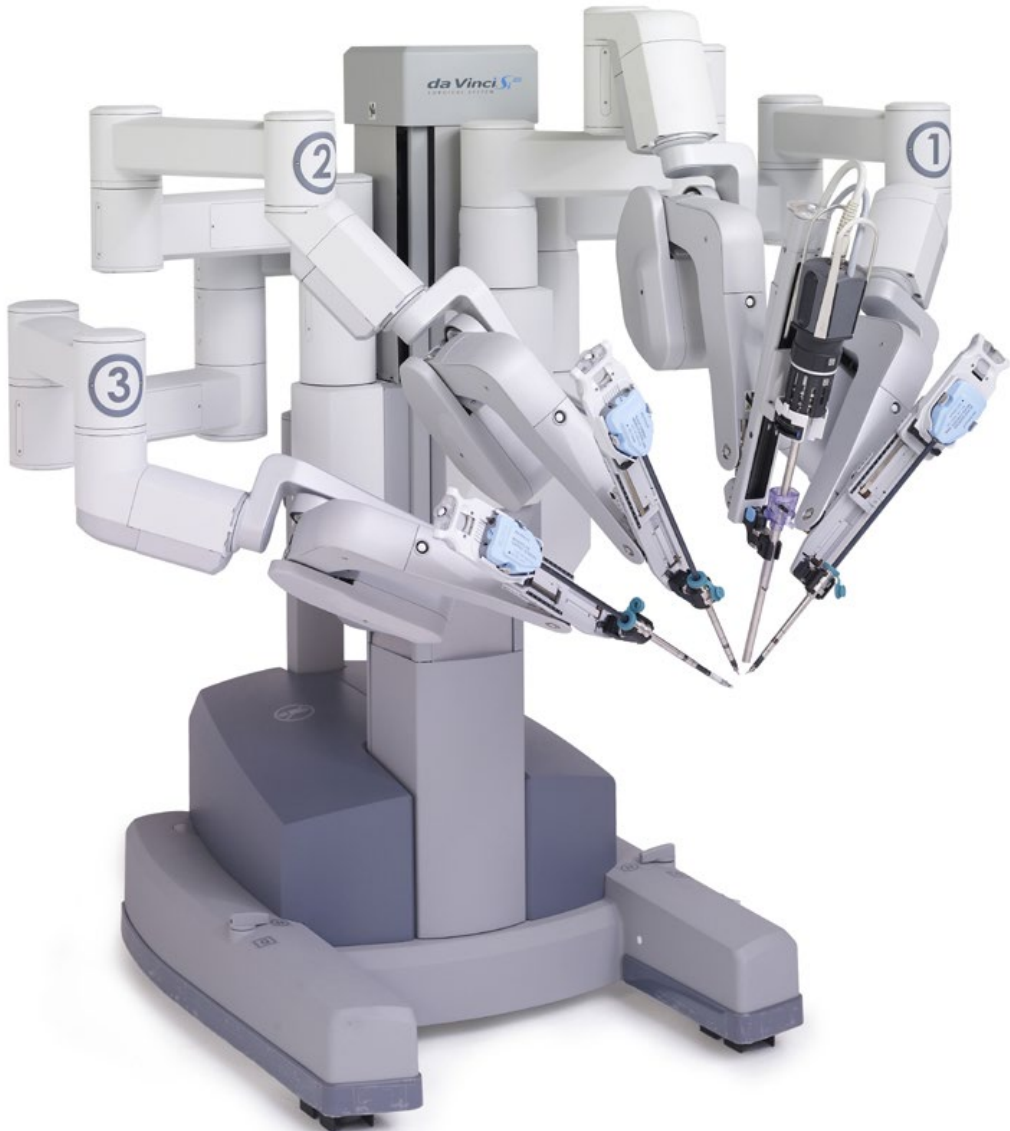
Nutri Ninja Blender

General Electric Washer

General Electric Dryer

Active Appliance: Implanted Cardiac Pacemakers

Tele-Bot: NASA Mars Rovers

Tele-Bot: Da Vinci Surgical System

"Robots don't perform surgery. Your surgeon performs surgery with da Vinci by using instruments that he or she guides via a console."
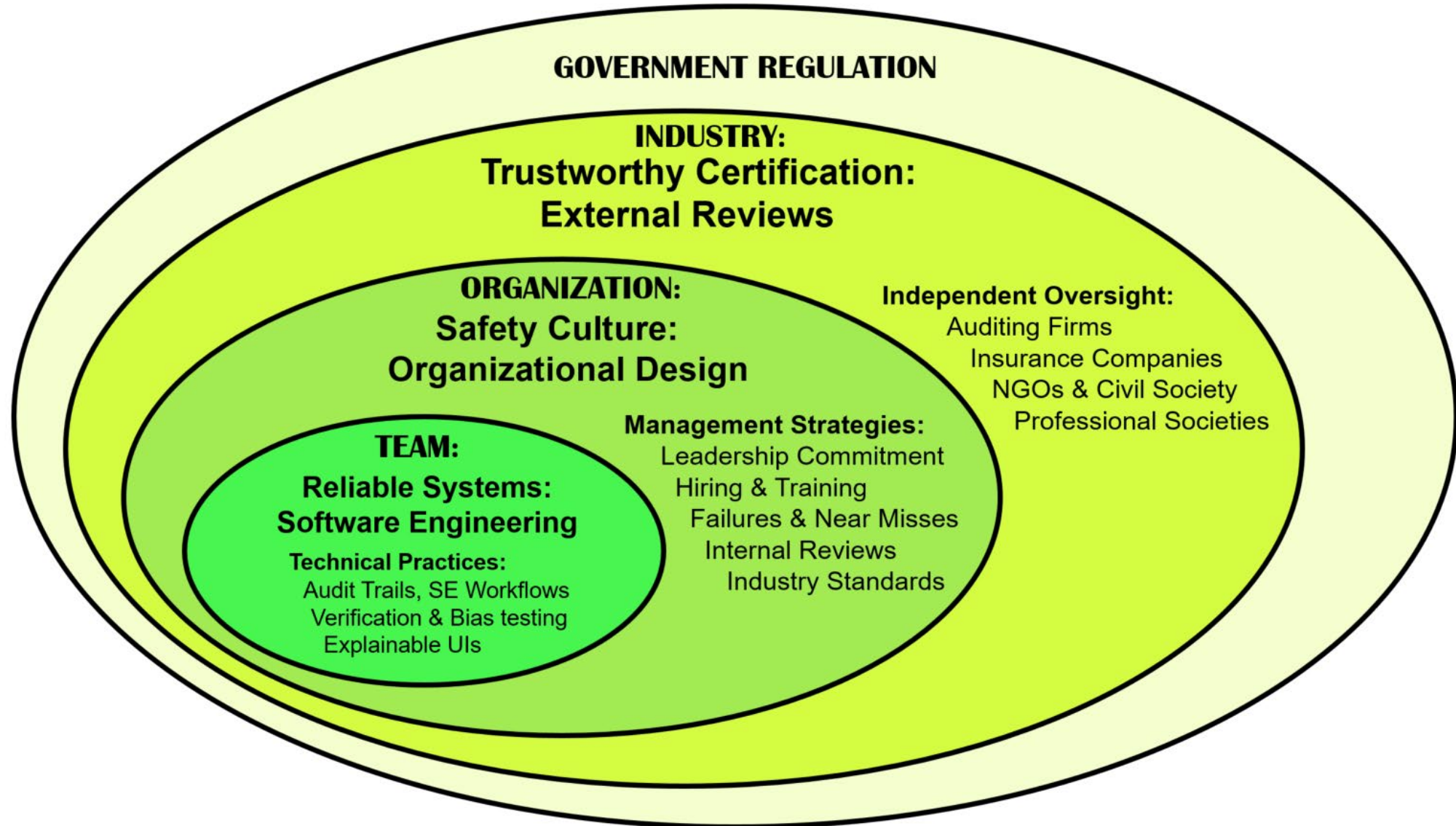
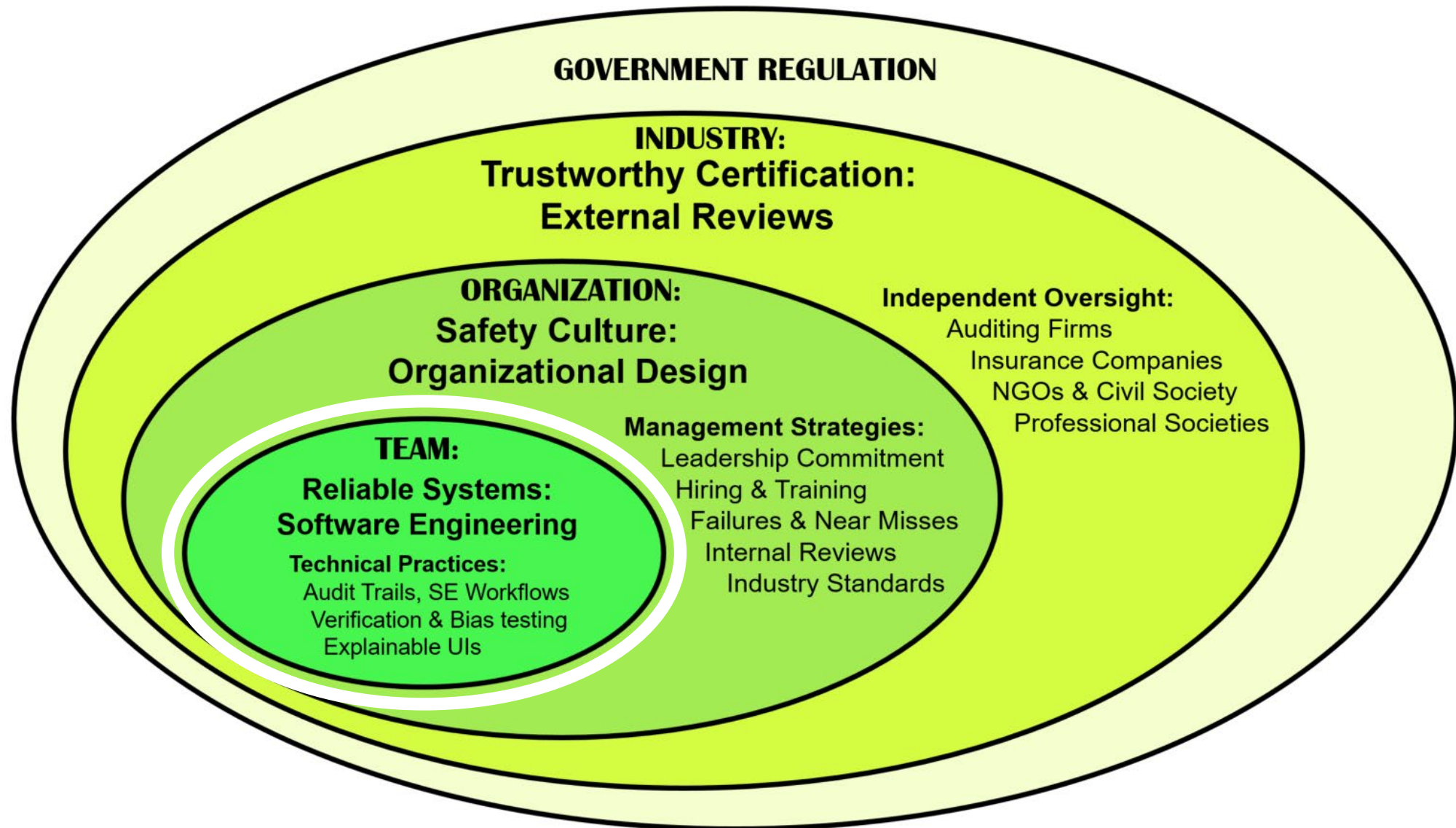# Control Center: Hospital

Control Center: Counter Terrorism

Governance Structures

Governance Structures for Human-Centered AI

GOVERNMENT REGULATION

INDUSTRY:
Trustworthy Certification:
External Reviews

ORGANIZATION:
Safety Culture:
Organizational Design

Independent Oversight:
Auditing Firms
Insurance Companies
NGOs & Civil Society
Professional Societies

Management Strategies:
Leadership Commitment
Hiring & Training
Failures & Near Misses
Internal Reviews
Industry Standards

TEAM:
Reliable Systems:
Software Engineering

Technical Practices:
Audit Trails, SE Workflows
Verification & Bias testing
Explainable UIs

Governance Structures for Human-Centered AI

GOVERNMENT REGULATION

INDUSTRY:
Trustworthy Certification:
External Reviews

ORGANIZATION:
Safety Culture:
Organizational Design

Independent Oversight:
Auditing Firms
Insurance Companies
NGOs & Civil Society
Professional Societies

TEAM:
Reliable Systems:
Software Engineering

Technical Practices:
Audit Trails, SE Workflows
Verification & Bias testing
Explainable UIs

Management Strategies:
Leadership Commitment
Hiring & Training
Failures & Near Misses
Internal Reviews
Industry Standards

ACM TIIS (Oct 2020) `https://dl.acm.org/doi/10.1145/3419764`

Reliable systems based on software engineering practices

1) Audit trails and analysis tools

2) Software engineering workflows

3) Verification & validation testing

4) Bias testing to improve fairness

5) Explainable user interfaces

Reliable systems based on software engineering practices

1) Audit trails and analysis tools

2) Software engineering workflows

3) Verification & validation testing

4) Bias testing to improve fairness

5) Explainable user interfaces

Reliable Systems

Software engineering practices for a TEAM

1) Audit trails and analysis tools

"Flight Data Recorder for Every Robot"

- Retrospective analysis of failures
- Understanding near misses
- Analysis to support preventive maintenance

Reliable Systems

Software engineering practices for a TEAM

5) Explainable user interfaces

- Retrospective explanations (local & global)

New Goal: **Prevent** confusion and surprise
- Prospective user interfaces
- Interactive, visual, exploratory

Mortgage Loan Explanations

Post-hoc Report

Enter amounts to request mortgage:

Mortgage amount requested   375000

Household monthly income   7000

Liquid assets   48000
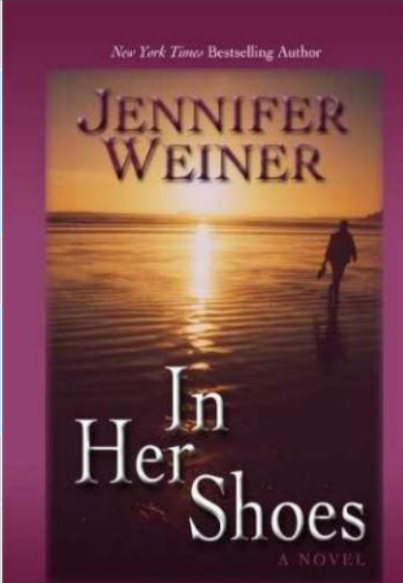
Submit

Mortgage Loan Explanations

Post-hoc Report

Enter amounts to request mortgage:

Mortgage amount requested   375000

Household monthly income   7000

Liquid assets   48000

Submit

Enter amounts to request mortgage:

Mortgage amount requested   375000

Household monthly income   7000

Liquid assets   48000

Submit

We're sorry, your mortgage loan was not approved. You might be approved if you reduce the Mortgage amount requested, increase your Household monthly income, or increase your Liquid assets.

Done

Mortgage Loan Explanations

Post-hoc Report

Enter amounts to request mortgage:

Mortgage amount requested | 375000

Household monthly income | 7000

Liquid assets | 48000

Submit

Enter amounts to request mortgage:

Mortgage amount requested | 375000

Household monthly income | 7000

Liquid assets | 48000

Submit

We're sorry, your mortgage loan was not approved. You might be approved if you reduce the Mortgage amount requested, increase your Household monthly income, or increase your Liquid assets.

Done

Prospective User Interface

Adjust sliders to report your situation:

Mortgage amount requested

◁▷ 375000

Household monthly income

◁▷ 7000

Liquid assets

◁▷ 48000
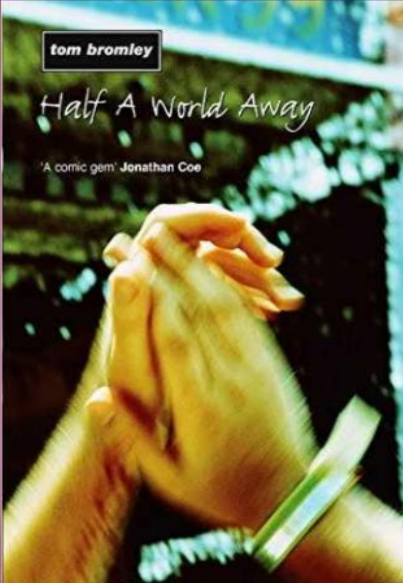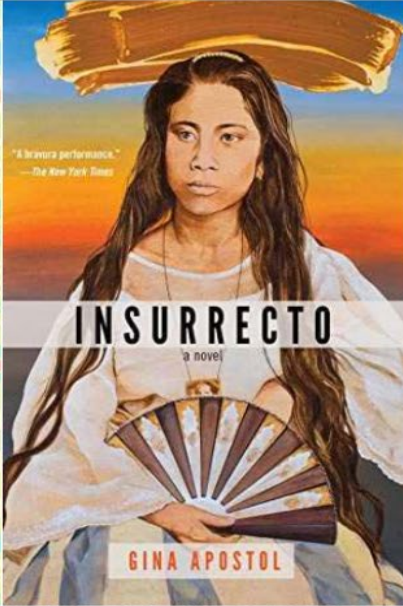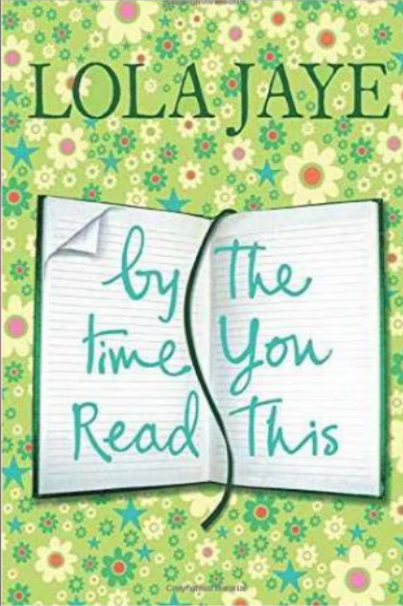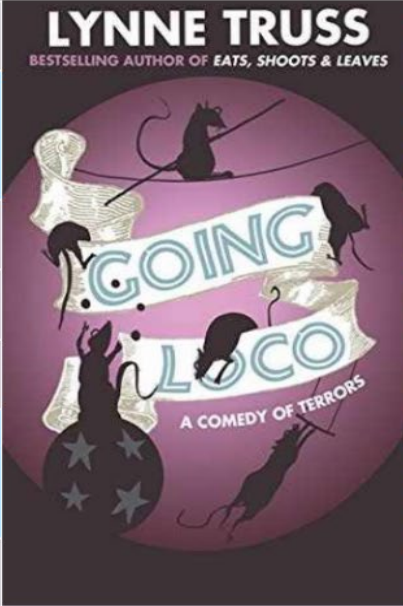
Score needed for approval

Your score

Done

Recommenders: Whichbook.net

Recommender Control Panels

Human-Centered AI

Human Values
Rights, Justice & Dignity

Individual Goals
Self-efficacy, Creativity, Responsibility & Social Connections

Design Aspirations
Reliable, Safe & Trustworthy
Team, Organization, Industry & Government

Stakeholders
- Researchers
- Developers
- Business Leaders
- Policy Makers
- Users

Threats
- Malicious Actors
- Bias
- Flawed Software

HCAI Framework

Design Metaphors

Governance Structures

Governance Structures for Human-Centered AI

GOVERNMENT REGULATION

INDUSTRY:
Trustworthy Certification:
External Reviews

ORGANIZATION:
Safety Culture:
Organizational Design

Independent Oversight:
Auditing Firms
Insurance Companies
NGOs & Civil Society
Professional Societies

Management Strategies:
Leadership Commitment
Hiring & Training
Failures & Near Misses
Internal Reviews
Industry Standards

TEAM:
Reliable Systems:
Software Engineering

Technical Practices:
Audit Trails, SE Workflows
Verification & Bias testing
Explainable UIs

Human-Centered Artificial Intelligence: Reliable, safe & trustworthy, *International Journal of Human-Computer Interaction 36,* 6 (March 2020). https://doi.org/10.1080/10447318.2020.1741118

Design lessons from AI's two grand goals: Human emulation and useful applications, *IEEE Transactions on Technology & Society 1,* 2 (June 2020). https://ieeexplore.ieee.org/document/9088114

Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy Human-Centered AI systems, *ACM Trans. on Interactive Intelligent Systems 10,* 4 (Oct 2020). https://dl.acm.org/doi/10.1145/3419764

Human-Centered Artificial Intelligence: Three fresh ideas,
AIS Trans. on Human-Computer Interaction 12, 3 (Oct 2020). https://aisel.aisnet.org/thci/vol12/iss3/1/

Human-Centered AI, *NAS ISSUES 37, 2* (Winter 2021). https://issues.org/human-centered-ai/

Summary & resources: https://hcil.umd.edu/human-centered-ai/

The Future is Human-Centered

Google Group

`https://groups.google.com/g/human-centered-ai`

Twitter Account

`@HumanCenteredAI`

Website

`https://hcai.site`

The Future is Human-Centered