Human-Centered AI: Reliable, Safe & Trustworthy

Part 1

Ben Shneiderman  @benbendc

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

Member, National Academy of Engineering
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
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,
The Goal of Visualization is Insight, Not Pictures
The Goal of Visualization is Insight, Not Pictures

Information Visualization Mantra

- Overview first, zoom & filter, then details-on-demand
- Overview first, zoom & filter, then details-on-demand
- Overview first, zoom & filter, then details-on-demand
- Overview first, zoom & filter, then details-on-demand
- Overview first, zoom & filter, then details-on-demand
- Overview first, zoom & filter, then details-on-demand
- Overview first, zoom & filter, then details-on-demand
- Overview first, zoom & filter, then details-on-demand

(IEEE, VLC, 1996)
“so much of what we read about AI strikes us as pure fantasy, predicated on a confidence in AI’s imagined strengths that bears no relation to current technological capabilities”
What is Human-Centered AI?
What is Human-Centered AI?

Amplify, Augment, Empower & Enhance People
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 Values
Rights, Justice & Dignity

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

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

Stakeholders
Users
Policy Makers
Business Leaders
Developers
Researchers
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
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

- **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

- **Frameworks**:
  - HCAI Framework
  - Design Metaphors
  - Governance Structures

Oxford University Press (Early 2022)  https://hcil.umd.edu/human-centered-ai/
UN Sustainable Development Goals

https://sdgs.un.org/goals
People are not computers;
Computers are not people
HCAI Attributes that Are Candidates for Assessment

General virtues of the system itself

- **Trustworthy**: Can users trust the system to perform correctly?
- **Responsible/Humane**: Has the system been designed, developed, and tested in a responsible way?
- **Ethical Design**: Were stakeholders involved in the design?
- **Ethical Data**: Was the data collected in an ethical manner?
- **Ethical Use**: Will the system’s outcome be used in an ethical manner?
- **Well-being/Benevolence**: Does the system support human health, comfort, and values?
- **Secure**: How vulnerable is the system to attack?
- **Private**: Does the system protect a person’s identity and data?

Perform well in practice

- **Robust/Agile**: Does the system perform well when inputs change?
- **Reliable/Dependable**: Does the system do the right thing?
- **Available**: Is the system running when needed?
- **Resilient/Adaptive**: Can the system recover from disruptions?
- **Testable/Verifiable/Validatable/Certifiable**: Can be tested to verify adherence to requirements?
- **Safe**: Does the system have a history of safe use?
Clarity to stakeholders
- **Accurate**: Does the system deliver correct results on test cases and real world cases?
- **Fair/Unbiased**: Are the system’s biases understood and reported?
- **Accountable/Liable**: Who or what is responsible for the system’s outcome?
- **Transparent**: Is it clear to an external observer how the system’s outcome was produced?
- **Interpretable/Explainable/Intelligible/Explicable**: Can the system explain the outcome?
- **Usable**: Can a human use it easily?

Enables independent oversight
- **Auditable**: Can the system be audited by others for retrospective forensic analysis of failures?
- **Trackable**: Does the system display status and next steps so human intervention is possible?
- **Traceable**: Is the system designed to allow tracing back from an outcome to the root cause?
- **Redressable**: Is there a process for those harmed to request review and compensation?
- **Insurable**: Does the design permit insurance companies to offer policies?
- **Recorded**: Does the system record activity for retrospective forensic review?
- **Open**: Is code and data publicly available for others to review?
- **Certifiable**: Can it be certified and approved for use?

Complies with accepted practices
- **Compliant with standards**: Does the system comply with relevant standards, e.g. IEEE P7000 series?
- **Compliant with accepted software engineering workflows**: Was a trusted process used?
HCAI Framework
Designing the User Interface

Balancing automation & human control

First Edition: 1986
Designing the User Interface

Balancing automation & human control

Human control

Computer automation

First Edition: 1986
LEVELS OF DRIVING AUTOMATION

0: NO AUTOMATION
Manual control. The human performs all driving tasks (steering, acceleration, braking, etc.).

1: DRIVER ASSISTANCE
The vehicle features a single automated system (e.g. it monitors speed through cruise control).

2: PARTIAL AUTOMATION
ADAS. The vehicle can perform steering and acceleration. The human still monitors all tasks and can take control at any time.

3: CONDITIONAL AUTOMATION
Environmental detection capabilities. The vehicle can perform most driving tasks, but human override is still required.

4: HIGH AUTOMATION
The vehicle performs all driving tasks under specific circumstances. Geofencing is required. Human override is still an option.

5: FULL AUTOMATION
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
Designing the User Interface

Ensuring human control while increasing automation

Low  High

Human Control

Low  High

Computer Automation

Human-Centered AI

Human Control

Computer Automation

High
Low

Reliable, Safe & Trustworthy
Human-Centered AI

- High
- Low

Human Control

Computer Automation

High
Low

Elevator Camera

Reliable, Safe & Trustworthy
Human-Centered AI

- **High Human Control, Low Computer Automation**: Human Mastery (Bicycle, Piano)
- **High Human Control, High Computer Automation**: Reliable, Safe & Trustworthy (Elevator, Camera)
- **Low Human Control, Low Computer Automation**: Music box (Landmine)
- **Low Human Control, High Computer Automation**: Pacemaker (Airbag, Computer Control)
Human-Centered AI

Human Mastery
- Bicycle
- Piano

Reliable, Safe & Trustworthy
- Elevator
- Camera

Human-Centered AI

Excessive Human Control
- Music box
- Landmine

Computer Control
- Pacemaker
- Airbag

Human Control
- Low
- High

Computer Automation
- Low
- High
Pain Control Designs

- **Human Mastery** (High Human Control, Excessive Human Control)
- **Reliable, Safe & Trustworthy** (High Computer Control, Excessive Automation)
- **Morphine drip bag** (Low Human Control, Low Computer Automation)
- **Computer Control** (Low Human Control, High Computer Control)
Pain Control Designs

- High Human Control
  - Human Mastery
  - Reliable, Safe & Trustworthy
- Low Human Control
  - Morphine drip bag
  - Automatic dispenser
- High Computer Automation
  - Computer Control
- Low Computer Automation
Pain Control Designs

- **Human Mastery**: Patient-guided dispenser
- **Reliable, Safe & Trustworthy**: 
- **Morphine drip bag**: 
- **Automatic dispenser**: 
- **Computer Control**
Pain Control Designs

- **Excessive Human Control**
  - **High Human Control**
    - **Human Mastery**
      - Patient-guided dispenser
  - **Low Human Control**
    - **Morphine drip bag**

- **Excessive Automation**
  - **High Computer Automation**
    - **Reliable, Safe & Trustworthy**
      - Patient-guided & clinician-monitored system
  - **Low Computer Automation**
    - **Automatic dispenser**

- **Computer Control**

- **Patient-guided & clinician-monitored system**
Johns Hopkins University Hospital Control Center
Excessive Human Control

High

Human Mastery

Reliable, Safe & Trustworthy

Low

Human Control

Computer Automation

1940 Car

Computer Control

High
Car Control Designs

<table>
<thead>
<tr>
<th>Human Control</th>
<th>Computer Automation</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>High</td>
</tr>
<tr>
<td>Human Mastery</td>
<td>Reliable, Safe &amp; Trustworthy</td>
</tr>
<tr>
<td>1980 Car</td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>1940 Car</td>
<td>Computer Control</td>
</tr>
<tr>
<td>Low</td>
<td></td>
</tr>
</tbody>
</table>

Excessive Human Control

Excessive Automation
Car Control Designs

Human Mastery
1980 Car

Reliable, Safe & Trustworthy

1940 Car

Computer Control
2020 Car

High Human Control

Excessive Human Control

Excessive Automation

Low Human Control

Computer Automation

Low
Wheelchair Designs

Human Mastery
- Hand-powered & user guided

Reliable, Safe & Trustworthy
- Motorized, joystick controlled, tele-operated & programmable

Push chair
- (requires caretaker)

Robotic
- (navigates to destination)

Computer Control
Application Categories

• Recommenders
• Consequential
• Life-critical
Application Categories

• Recommenders
• Consequential
• Life-critical

• Rapid performance
• Long duration
• Remote locations
Micro-Structure of Design

• Automate where
  + Reliable performance is possible
  - But allow overrides
Micro-Structure of Design

• Automate where
  + Reliable performance is possible
    - But allow overrides

• Give human control where
  + Desired for creative flexibility
  + Automation is uncertain
    - But prevent human errors
Micro-Structure of Design

• Automate where
  + Reliable performance is possible
  - But allow overrides

• Give human control where
  + Desired for creative flexibility
  + Automation is uncertain
  - But prevent human errors

• Design supervisory control
  + Teleoperate remotely
  + Collect aggregate data
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

HCAI Framework

Design Metaphors

Governance Structures

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

Threats
- Malicious Actors
- Bias
- Flawed Software

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

## Science Goal

<table>
<thead>
<tr>
<th>Intelligent Agents</th>
<th>Teammates</th>
<th>Assured Autonomy</th>
<th>Social Robots</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thinking Machine, Cognitive Actor, Artificial Intelligence, Knowledgeable</td>
<td>Co-active Collaborator, Colleague, Helpful Partner, Smart Co-worker</td>
<td>Independent, Self-directed, Goal-setting, Self-monitored</td>
<td>Anthropomorphic, Humanoid, Android, Bionic, Bio-inspired</td>
</tr>
</tbody>
</table>

## Innovation Goal

<table>
<thead>
<tr>
<th>Supertools</th>
<th>Tele-operated Devices</th>
<th>Supervised Autonomy</th>
<th>Active Appliances</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extend Abilities, Empower Users, Enhance Human Performance</td>
<td>Steerable Instrument, Powerful Prosthetic, Boost Human Perceptual &amp; Motor Skills</td>
<td>Human Control &amp; Oversight, Situation Awareness, Predictable Actions</td>
<td>Consumer-oriented, Wide Use, Low Cost Comprehensible Control Panels</td>
</tr>
</tbody>
</table>
Teammate & Supertool

• Social Teammate:
  Since many people respond socially to robots,
  -> design robots to be human-like social teammates.
Teammate & Supertool

• **Social Teammate:**
  Since many people respond socially to robots,
  -> design robots to be human-like social teammates.

• **Human-Centered Conjecture:**
  Since only humans can be responsible &
  computers have distinct capabilities (speed, storage, display...)
  -> design computers to be supertools
Teammate & Supertool

• Social Teammate:
  Since many people respond socially to robots,
  -> design robots to be human-like social teammates.

• Human-Centered Conjecture:
  Since only humans can be responsible &
  computers have distinct capabilities (speed, storage, display...)
  -> design computers to be supertools
  -> invite users to fix, personalize & extend the design
Active Appliances

Coffee maker, Rice cooker, Blender

Dishwasher, Clothes Washer/Dryer
Implanted Cardiac Pacemakers
NASA Mars Rovers are Tele-Operated
Da Vinci Tele-Operated Surgery

“Robots don’t perform surgery. Your surgeon performs surgery with da Vinci by using instruments that he or she guides via a console.”

https://www.davincisurgery.com/
Bloomberg Terminal
Counter Terrorism Center
Design Guidelines

<table>
<thead>
<tr>
<th>Eight Golden Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Strive for consistency</td>
</tr>
<tr>
<td>2. Seek universal usability</td>
</tr>
<tr>
<td>3. Offer informative feedback</td>
</tr>
<tr>
<td>4. Design dialogs to yield closure</td>
</tr>
<tr>
<td>5. Prevent errors</td>
</tr>
<tr>
<td>6. Permit easy reversal of actions</td>
</tr>
<tr>
<td>7. Keep users in control</td>
</tr>
<tr>
<td>8. Reduce short-term memory load</td>
</tr>
</tbody>
</table>

https://www.cs.umd.edu/~ben/goldenrules.html
### Design Guidelines

<table>
<thead>
<tr>
<th>Eight Golden Rules</th>
<th>Eight Silver Slogans for HCAI Systems</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Strive for consistency</td>
<td>1. Store rich data from powerful sensors</td>
</tr>
<tr>
<td>2. Seek universal usability</td>
<td>2. Design information abundant displays</td>
</tr>
<tr>
<td>3. Offer informative feedback</td>
<td>3. Provide interactive information visualization</td>
</tr>
<tr>
<td>4. Design dialogs to yield closure</td>
<td>4. Make predictive models visual</td>
</tr>
<tr>
<td>5. Prevent errors</td>
<td>5. Smooth human-to-human communication</td>
</tr>
<tr>
<td>6. Permit easy reversal of actions</td>
<td>6. Create clear control panels</td>
</tr>
<tr>
<td>7. Keep users in control</td>
<td>7. Implement audit trails</td>
</tr>
<tr>
<td>8. Reduce short-term memory load</td>
<td>8. Develop incident reporting websites</td>
</tr>
</tbody>
</table>

[https://www.cs.umd.edu/~ben/goldenrules.html](https://www.cs.umd.edu/~ben/goldenrules.html)
Human-Centered AI

HCAI Framework
Design Metaphors
Governance Structures

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

Oxford University Press (Early 2022)  https://hcil.umd.edu/human-centered-ai/
Human-Centered AI Framework

Excessive Human Control

High Human Control
- Human Mastery
  - Bicycle
  - Piano
- Reliable, Safe & Trustworthy
  - Elevator
  - Camera

Low Computer Automation
- Music box
- Landmine

Low Human Control

High Computer Automation
- Pacemaker
  - Airbag
- Computer Control
Design Metaphors

Science Goal

Intelligent Agents
- Thinking Machine, Cognitive Actor, Artificial Intelligence, Knowledgeable

Teammates
- Co-active Collaborator, Colleague, Helpful Partner, Smart Co-worker

Assured Autonomy
- Independent, Self-directed, Goal-setting, Self-monitored

Social Robots
- Anthropomorphic, Humanoid, Android, Bionic, Bio-inspired

Innovation Goal

Supertoools
- Extend Abilities, Empower Users, Enhance Human Performance

Tele-operated Devices
- Steerable Instrument, Powerful Prosthetic, Boost Human Perceptual & Motor Skills

Supervised Autonomy
- Human Control & Oversight, Situation Awareness, Predictable Actions

Active Appliances
- Consumer-oriented, Wide Use, Low Cost Comprehensible Control Panels
A Case for Cooperation Between Machines and Humans

A computer scientist argues that the quest for fully automated robots is misguided, perhaps even dangerous. His decades of warnings are gaining more attention.

By John Markoff

May 21, 2020   Updated 3:09 p.m. ET


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/](https://aisel.aisnet.org/thci/vol12/iss3/1/)


**Summary & resources:** [https://hcil.umd.edu/human-centered-ai/](https://hcil.umd.edu/human-centered-ai/)
The Future is Human-Centered