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

BayCHI, June 8, 2021
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

Sixth Edition: 2016
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

HCAI Framework

Threats
- Malicious Actors
- Bias
- Flawed Software

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

Malicious Actors
Bias
Flawed Software
Malicious Actors
Bias
Flawed Software

Threats
Malicious Actors
Bias
Flawed Software

 Researchers
Developers
Business Leaders
Policy Makers
Users

Stakeholders

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

Design Metaphors

HCAI Framework

Design Metaphors

HCAI Framework
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


Sixth Edition: 2016
Human-Centered AI

- Reliable, Safe & Trustworthy
  - Elevator Camera

Human Control

Computer Automation

- Low
  - Low
- High
  - High
Human-Centered AI

- High Human Control
  - Human Mastery (Bicycle, Piano)
  - Music box (Landmine)

- Low Human Control
  - Reliable, Safe & Trustworthy (Elevator, Camera)
  - Pacemaker, Airbag
  - Computer Control
Pain Control Designs

- Excessive Human Control
  - Human Mastery
  - Reliable, Safe & Trustworthy
- Computer Control
  - Morphine drip bag
  - Computer Control

Human Control: High, Low
Computer Automation: Low, High
Pain Control Designs

Excessive Human Control

High

Human Mastery

Reliable, Safe & Trustworthy

Morphine drip bag

Automatic dispenser

Computer Control

Low

Low

Computer Automation

High
Pain Control Designs

- **Excessive Human Control**
  - High Human Control
  - Patient-guided dispenser
  - Reliable, Safe & Trustworthy
- **Excessive Automation**
  - Low Human Control
  - Morphine drip bag
  - Computer Control

- **Computer Automation**
  - Low Computer Automation
  - Automatic dispenser
  - Human Mastery
Pain Control Designs

- **Excessive Human Control**
  - Human Mastery
    - Patient-guided dispenser
  - Reliable, Safe & Trustworthy
    - Patient-guided & clinician-monitored system

- **Excessive Automation**
  - Morphine drip bag
  - Automatic dispenser

- **Low Human Control, Low Computer Automation**
  - Computer Control

- **High Human Control, Low Computer Automation**
  - Human Mastery

- **Low Human Control, High Computer Automation**
  - Reliable, Safe & Trustworthy

- **High Human Control, High Computer Automation**
  - Excessive Human Control
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

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

Cuisinart Grind & Brew Coffee Maker

Panasonic Rice Cooker

Nutri Ninja Blender

Miele Dishwasher

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.”

https://www.davincisurgery.com/
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

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

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

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**

<table>
<thead>
<tr>
<th>Enter amounts to request mortgage:</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mortgage amount requested</td>
<td>375000</td>
</tr>
<tr>
<td>Household monthly income</td>
<td>7000</td>
</tr>
<tr>
<td>Liquid assets</td>
<td>48000</td>
</tr>
</tbody>
</table>

Submit
Mortgage Loan Explanations

Post-hoc Report

Enter amounts to request mortgage:

<table>
<thead>
<tr>
<th>Description</th>
<th>Amount</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mortgage amount requested</td>
<td>375000</td>
</tr>
<tr>
<td>Household monthly income</td>
<td>7000</td>
</tr>
<tr>
<td>Liquid assets</td>
<td>48000</td>
</tr>
</tbody>
</table>

Submit

Enter amounts to request mortgage:

<table>
<thead>
<tr>
<th>Description</th>
<th>Amount</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mortgage amount requested</td>
<td>375000</td>
</tr>
<tr>
<td>Household monthly income</td>
<td>7000</td>
</tr>
<tr>
<td>Liquid assets</td>
<td>48000</td>
</tr>
</tbody>
</table>

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
Post-hoc Report

Enter amounts to request mortgage:

- Mortgage amount requested: 375000
- Household monthly income: 7000
- Liquid assets: 48000

Submit

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

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
Recommender Control Panels

Modify Attributes
- Acousticness: 40
- Instrumentalness: 60
- Danceability: 80
- Valence: 60
- Energy: 40

Recommended Songs
- Calculate Recommendations
- Task: Make a playlist of songs to listen to during your personal maintenance.
- Click '👍' or '👎' to keep or dismiss a song. More info: 🌟
- oblivion
- Grimes
- My December
- Linkin Park
- I’ve got that tune
- Chinese Man
- Got the Life
- Korn
- Good Riddance (Time of Y... Green Day
- Burn It To The Ground
- Nickelback

Create Your Better Life Index
Rate the topics according to their importance to you:

- Housing
- Income
- Jobs
- Community
- Education
- Environment
- Civic Engagement
- Health
- Life Satisfaction
- Safety
- Work-Life Balance
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/
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


**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