BEN SHNEIDERMAN

HUMAN-CENTERED AI

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford, OX2 6DP,
United Kingdom

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide. Oxford is a registered trade mark of
Oxford University Press in the UK and in certain other countries

Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America

British Library Cataloguing in Publication Data
Data available

Library of Congress Control Number: 2021940444

ISBN 978–0–19–284529–0

DOI: 10.1093/oso/9780192845290.001.0001

Printed in Great Britain by
Bell & Bain Ltd., Glasgow

CONTENTS

PART I

What Is Human-Centered Artificial Intelligence?

Researchers, developers, business leaders, policy-makers, and others are expanding the technology-centered scope of artificial intelligence (AI) to include human-centered AI (HCAI) ways of thinking. This expansion from an algorithm-focused view to embrace a human-centered perspective can shape the future of technology so as to better serve human needs. Educators, designers, software engineers, product managers, evaluators, and government agency staffers can build on AI-driven technologies to design products and services that make life better for the users, enabling people to care for each other. Humans have always been tool builders, and now they are super-tool builders, whose inventions can improve our health, family life, education, business, the environment, and much more. The remarkable progress in algorithms for machine and deep learning during the past decade has opened the doors to new opportunities, and some dark possibilities. However, a bright future awaits AI researchers, developers, business leaders, policy-makers, and others who build on AI algorithms by including HCAI strategies of design and testing. This enlarged vision can shape the future of technology so as to better serve human values and needs. As many technology companies and thought leaders have said, the goal is not to replace people but to empower them by making design choices that give humans control over technology.

James Watts' steam engine, Samuel Morse's telegraph, and Thomas Edison's electric light were technology breakthroughs that were put to work to open up new possibilities for transportation, communications, business, and families. They all moved beyond the existing and familiar technologies to demonstrate new products and services that enhanced life while suggesting ever more potent possibilities. Each positive step is also embraced by malicious actors such as criminals, hate groups, terrorists, and oppressive rulers, so careful attention to how technologies are used can reduce these threats. The human capacity for frontier thinking, to push beyond current examples, is amply visible in the Wright brothers' airplane, Tim Berners-Lee's World Wide Web, and Jennifer Doudna and Emmanuelle Charpentier's genome editing. Now, as new technologies blossom into ever more potent breakthroughs we have a choice to make about how these technologies will be applied.

The high expectations and impressive results from AI, such as the AlphaGo program winning at the game of Go, have triggered intense worldwide activity by researchers, developers, business leaders, and policy-makers. The promise of startling advances from machine learning and other algorithms energizes discussions while eliciting huge investments in medical, manufacturing, and military innovations.

The AI community's impact is likely to grow even larger by embracing a human-centered future, filled with supertools that amplify human abilities, empowering people in remarkable ways. This compelling prospect of HCAI builds on AI methods, enabling people to see, think, create, and act with extraordinary clarity. HCAI technologies bring superhuman capabilities, augmenting human creativity, while raising human performance and self-efficacy. These capabilities are apparent in familiar HCAI applications such as digital cameras that have high levels of human control but many AI supports in setting aperture, adjusting focus, and reducing jitter from hand movements. Similarly, HCAI navigation systems give walkers, bikers, drivers, and public transport users control over the many choices that are derived from AI programs which use real-time data to predict travel times.

Extending the power of AI-driven algorithms, *Human-Centered AI* shows how to make successful technologies that amplify, augment, empower, and enhance human performance. This expanded mindset should please readers as it describes a safer, more understandable, and more manageable future. A human-centered approach will reduce the out-of-control technologies, calm fears of robot-led unemployment, and give users the rewarding sense of mastery and accomplishment. Beyond individual experiences, HCAI will enable better control of privacy/security to limit misinformation and counter malicious actors. The dangers from AI and HCAI systems are legitimate concerns—any technology that empowers people to do good also empowers those who would do evil. Carefully designed controls, audit trails, and supervised autonomy are some of the strategies that stakeholders can adopt to achieve, reliable, safe, and trustworthy systems.

This book makes a coherent presentation of the fresh HCAI methods with numerous examples to guide researchers, developers, business leaders, and policy-makers. It offers an HCAI framework to guide innovation, design metaphors to combine disparate views, and governance structures to advance a human-centered approach. Benefitting people

becomes the driving force for making ever more potent supertools, tele-bots, active appliances, and control centers that empower users with extraordinary capabilities.

Reframing established beliefs with a fresh vision is among the most powerful tools for change. It can liberate researchers and designers, building on the past while allowing them to adopt new beliefs. The vast number of people embracing AI technologies are beginning to align with HCAI themes with an openness to human-centered thinking. I hope that the traditional AI technology-centered community, who have made so many important breakthroughs, will take in the human-centered perspectives, which offer a different vision of human destiny. A human-centered strategy will bring AI wider acceptance and higher impact by providing products and services that serve human needs. By encouraging a passionate devotion to empower people, enrich communities, and inspire hope, *Human-Centered AI* offers a vision of future technologies that values human rights, justice, and dignity.

CHAPTER 1

Introduction: High Expectations

The Analytical Engine has no pretensions whatever to originate anything. It can do whatever we know how to order it to perform.

Ada Lovelace *(1843)*

This book proposes a new synthesis in which AI-based intelligent algorithms are combined with human-centered thinking to make HCAI. This approach will increase the chance that technology will empower rather than replace people. In the past, researchers and developers focused on building AI algorithms and systems, stressing machine autonomy and measuring algorithm performance. The new synthesis gives equal attention to human users and other stakeholders by raising the value of user experience design and by measuring human performance. Researchers and developers for HCAI systems value meaningful human control, putting people first by serving human values such as rights, justice, and dignity, and supporting goals such as self-efficacy, creativity, responsibility, and social connections.[1]

This new synthesis reflects the growing movement to expand from technology-centered thinking to include human-centered aspirations that highlight societal benefit. The interest in HCAI has grown stronger since the 2017 Montreal Declaration for Responsible Development of AI. That declaration called for devotion to human well-being, autonomy, privacy, and creation of a just and equitable society. Enthusiasm for these human-centered goals is also part of the AI4GOOD movement,[2] DataKind,[3] and the IBM Watson AI XPRIZE Foundation,[4] which seek to apply AI methods, such as machine learning, "to solve some of society's biggest challenges." This admirable devotion to societal needs is aligned with the HCAI approach that applies rigorous design and evaluation methods to produce high-impact research. AI4GOOD sets appropriate goals which can be pursued with HCAI methods that guide

researchers and developers to determine how to effectively address genuine human needs, including meaningful problems in government, and vital challenges for businesses, schools, and healthcare systems. However, every opportunity for doing good is balanced by the dangers from the increased power of AI and HCAI systems, which can equally be used by malicious actors such as criminals, hate groups, terrorists, and oppressive politicians.

This movement towards setting societal goals for AI is aligned with the United Nations AI for Good Global Summit, an annual gathering of ardent research leaders, serious business executives, and conscientious policy-makers since 2017. The conference organizers state that their "goal is to identify practical applications of AI and scale those solutions for global impact."[5] The efforts of United Nations agencies and member countries are guided by the seventeen United Nations' Sustainable Development Goals (SDGs), which were established in 2015 to set aspirations for 2030 (Figure 1.1).[6] These goals include elimination of poverty, zero hunger, quality education, and reduced inequalities. Other ambitions address environmental issues such as climate action, life on land, life below water, and sustainable cities and communities. While many social, political, and psychological changes are needed, technology will play a role in finding solutions, including AI and HCAI.

Fig 1.1 The seventeen United Nation's Sustainable Development Goals (SDGs)

Source: https://sdgs.un.org/goals

Policy-makers assess progress toward these goals by 169 target indicators for each country, such as proportion of population living in households with access to basic services, maternal mortality ratio, and proportion of population using safely managed drinking water services.[7]

A related set of goals is captured in the notion of human *well-being,* which is the basis for a recent IEEE P7010 standard whose authors hope will encourage HCAI researchers and developers to "assess, manage, mitigate, and improve the well-being impacts on human and societal well-being, extending from individual users to the public."[8] Successful HCAI methods and applications could do much to advance these efforts. Human-centered methods and design thinking for all technologies will be helpful, but HCAI could be a potent combination that proves to be especially valuable for these grand challenges.

A key question is what do we mean by HCAI and what makes it different from AI? There are many definitions, but there are two key aspects:

1) Process: HCAI builds on user experience design methods of user observation, stakeholder engagement, usability testing, iterative refinement, and continuing evaluation of human performance in use of systems that employ AI and machine learning.
2) Product: HCAI systems are designed to be supertools which amplify, augment, empower, and enhance human performance. They emphasize human control, while embedding high levels of automation by way of AI and machine learning. Examples include digital cameras and navigation systems, which give humans control yet have many automated features.

The goal is to increase human self-efficacy, creativity, responsibility, and social connections while reducing the impact of malicious actors, biased data, and flawed software.

This book has three fresh ideas for changing technology design so as to bring about a new synthesis with its human-centered orientation.

HCAI framework that guides creative designers to ensure human-centric thinking about highly automated systems. The examples include familiar devices, such as thermostats, elevators, self-cleaning ovens, and cellphone cameras, as well as life critical applications, such as highly automated cars and patient-controlled pain relief devices. The new aspiration is to have *high levels of human control AND high levels of automation.*

Design metaphors suggest how the two central goals of AI research, science and innovation, are both valuable, but researchers, developers, business leaders, and policy-makers will need to be creative in finding effective ways of combining them to benefit the users. There are four design metaphors that can be used to combine the two goals of AI research:

1) intelligent agents and supertools;
2) teammates and tele-bots;
3) assured autonomy and control centers; and
4) social robots and active appliances.

Journalists, headline writers, graphic designers, and Hollywood producers are entranced by the possibilities of robots and AI, so it will take a generation to change attitudes and expectations towards a human-centered view. With fresh thinking, researchers, developers, business leaders, and policy-makers can find combined designs that will accelerate HCAI thinking. A greater emphasis on HCAI will reduce unfounded fears of AI's existential threats and raise people's belief that they will be able to use technology for their daily needs and creative explorations. It will increase benefits for users and society in business, education, healthcare, environmental preservation, and community safety.

Governance structures bridge the gap between widely discussed ethical principles and the practical steps needed to realize them. Software team leaders, business managers, and organization leaders will have to adapt proven technical practices, management strategies, and independent oversight methods, so they can achieve the desired goals of:

1) **Reliable** systems based on proven software engineering practices;
2) **Safety** culture through business management strategies; and
3) **Trustworthy** certification by independent oversight and government regulation.

Technical practices for designers, software engineers, programmers, team leaders, and product managers include audit trails to enable analysis of failures, just like the flight data recorders (aircraft black boxes, which are really orange boxes) that have made civil aviation such a success story. Part 4 suggests how sound existing practices can be applied to software engineering workflows,

verification and validation testing, bias testing to enhance fairness, and explainable user interfaces.

Management strategies for creating a safety culture begin with leadership commitment to safety that leads to better hiring practices and training oriented to safety. Other management strategies are extensive reporting of failures and near misses, which are collected internally from employee reports and gathered externally from users who make incident reports, internal review boards, and alignment with industry standard practices.

Trustworthy certification and clarity about liability comes from accounting firms that conduct independent audits and insurance companies that compensate for failures. Then there are non-governmental and civil society organizations that advance design principles, and professional organizations that develop voluntary standards and prudent policies. Further support for trustworthiness will come from government legislation and regulation, but advocates of certification and independent oversight will have to cope with resistance to regulation and "revolving door" movements in which corporate leaders make jobs in oversight organizations.

The three fresh ideas are covered in Parts 2, 3, and 4 of this book. They are the foundation for achieving the aspirations, goals, and human values shown in Figure 1.2, which is a compact overview of this book The stakeholders participate in every aspect, while the threats from malicious actors, bias, and flawed

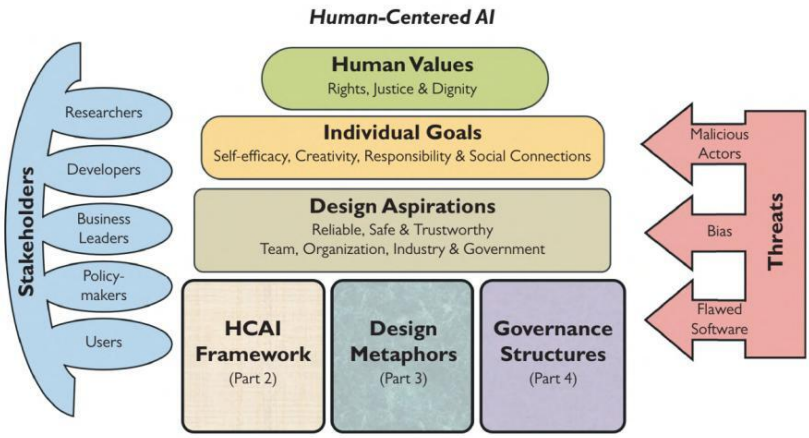

Fig 1.2 The three ideas of this book support the aspirations, goals, and human values, while recognizing the needs of stakeholders and the dangers of threats.

software remain prominent in stakeholder minds. The three fresh ideas are the HCAI framework, design metaphors, and governance structures.

Successful automation is all around us. Navigation applications give drivers control by showing times for alternative routes. E-commerce websites offer shoppers relevant options, customer reviews, and clear pricing so they can find and order the goods they need. Elevators, clothes-washing machines, and airline check-in kiosks, too, have meaningful controls that enable users to get what they need done quickly and reliably. When modern cameras assist photographers in taking properly focused and exposed photos, users have a sense of mastery and accomplishment for composing the image, even as they get assistance with optimizing technical details. These and millions of other mobile device applications and cloud-based web services enable users to accomplish their tasks with self-confidence and sometimes even pride.

In a flourishing automation-enhanced world, clear, convenient interfaces could let humans control automation to make the most of people's initiative, creativity and responsibility. The most successful machines could be powerful supertools that let users carry out ever-richer tasks with confidence, such as helping architects find innovative ways to design energy-efficient buildings and giving journalists tools to dig deeper into data to uncover fraud and corruption. Other HCAI supertools could enable clinicians to detect emerging medical conditions, industry watchdogs to spot unfair hiring decisions, and auditors to identify bias in mortgage loan approvals.

Designers of AI algorithms and HCAI user interfaces must work diligently to ensure that their work brings more benefits than harms. Charting a path between utopian visions of happy users, thriving businesses, and smart cities and the dystopian scenarios of frustrated users, surveillance capitalism, and political manipulations of social media is the real challenge we face. Training researchers, developers, business leaders, and policy-makers to consider downside risks will do much to limit harm. A good start is the growing database of more than a thousand AI incident and accident reports[9] that provide disturbing examples of what can go wrong.[10]

Humans are accomplished at building tools that expand their creativity—and then at using those tools in even more innovative ways than their designers intended. It's time to let more people be more creative more of the time. Technology designers who appreciate and amplify the key aspects of humanity are most likely to invent the next generation of what I call supertools, tele-bots, and active appliances. These designers will shift from trying to replace human

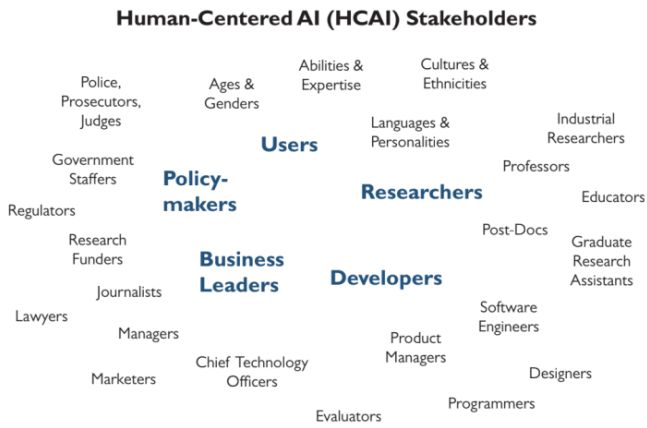

Fig 1.3 HCAI stakeholders with core professionals who are researchers, developers, business leaders, and policy-makers.

behavior in machines to building the wildly successful applications that people love to use.

This book is intended for diverse readers who play a role in shaping technology and its uses. I refer to researchers, developers, business leaders, and policy-makers who shape HCAI systems and the users who benefit from them. Figure 1.3 names some of the larger set of diverse users and professionals who are all stakeholders with a role to play.

If AI technology developers increase their use of information visualization, their own algorithmic work will improve and they will help many stakeholders to better understand how to use these new technologies. The traditional AI research community favors statistical machine learning and neural net-inspired deep learning algorithms that do tasks automatically or autonomously. However, that attitude is changing as information visualization has proven its value in understanding deep learning methods, improving algorithms, and reducing errors. Visual user interfaces have become appreciated for providing developers, users, and other stakeholders with a better understanding of and more control over how algorithmic decisions are made for parole requests, hiring, mortgages, and other consequential applications.

My education about how AI systems could be evaluated came when serving on a National Academy of Sciences panel during 2006–2008, whose task was to prepare a report on *Protecting Individual Privacy in the Struggle Against Terrorists: A Framework for Program Assessment*.[11] This twenty-one person panel was filled with a diverse collection of impressive people and co-chaired by two

remarkable individuals: William J. Perry, former US Secretary of Defense, and Charles M. Vest, President of the National Academy of Engineering and former President of MIT. Our job was to recommend evaluation methods for troubling technologies such as data mining, machine learning, and behavioral surveillance, so they could be used safely. The challenges were to protect individual privacy and limit inappropriate use by rigorously assessing the enthusiastic claims for these emerging technologies. One of my roles was to study independent oversight methods to clarify how they have been used and how they could be applied for these emerging technologies. Our statistical testing process, described as "a framework for evaluating information-based programs to fight terrorism or serve other important national goals," became a model for government agencies and others. The panel's recommendations included: "Any information-based counterterrorism program of the U.S. government should be subjected to robust, independent oversight . . . All such programs should provide meaningful redress to any individuals inappropriately harmed by their operation." Our takeaway message was that careful evaluations coupled with independent oversight were strong partners in advancing safe use of technology.

In the time since that report, AI's success with machine and deep learning has catapulted it to the top of agendas for business leaders and government policy-makers. Bestselling books, such as Nick Bostrom's *Superintelligence: Paths, Dangers, Strategies* and Stuart Russell and Peter Norvig's textbook on *Artificial Intelligence: A Modern Approach*, celebrated the accomplishments, suggested continuing opportunities, and raised fears of what could go wrong.[12] Their work and many others led to intense interest from technology corporations, which quickly shifted to being AI corporations, and government commitments internationally of billions of dollars for AI applications in business, medical, transportation, military, and other applications.

On the other hand, cautionary voices sounded alarms. Cathy O'Neil's groundbreaking 2016 book *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* laid out the case of how numerous widely used algorithms were opaque and harmful.[13] As a Harvard-trained Wall Street analyst she is greatly respected, and her writing is powerful and clear. Her book and the European Union's General Data Protection and Regulation (GDPR) accelerated efforts to develop explainable AI (XAI) so that mortgage applicants, parole requestors, or job seekers who were rejected could get a meaningful explanation. Such explanations would help them adjust their

requests or challenge an unfair decision. Information visualization methods became an increasing part of the designs in the growing XAI community.

The interest in HCAI blossomed with 500+ reports from public interest groups, professional societies, and governments encouraging responsible, ethical, and humane approaches. These efforts accelerated as the need to have better human control over computers was highlighted by stock market flash crashes, deadly failures of excessively autonomous Patriot missile systems in the 2003 Iraq War, and fatal accidents involving self-driving cars. While the two Boeing 737 MAX crashes in 2018 and a few months later in 2019 were not directly related to AI systems, the belief in autonomous systems misled designers and regulators. They believed that embedded algorithms could perform perfectly so that pilots were not even informed about their presence. When the angle of attack sensor failed, the embedded algorithms forced the plane to turn nose down, resisting the repeated attempts of the confused pilots to turn the nose up. The often mentioned ironies, dilemmas, conundrums, paradoxes, and myths of autonomy turned into a deadly tragedy of autonomy.

This book is meant as a guidebook to hope and a roadmap to realistic policies. To succeed, the HCAI community will have to change the language, metaphors, and images of technology that suggest human-like robots to collaboration among people who are using computers. The clichéd images of a human hand touching a robot hand or a humanoid robot walking with children already seem archaic and misguided. While control panels for washing machines or clothes dryers are a modest starting point, their successors are likely to become the next commercial successes. Tele-operated drones, remotely activated home controls, and precise surgical devices will spread. The ambitious control rooms for NASA's Mars Rovers, transportation management centers, patient-monitoring displays, and financial trading rooms are compelling prototypes for many applications. Medical monitors and implanted devices will be operated by smartphone apps, giving control to users and supervisory control to clinicians and product managers who can monitor thousands of these devices so as to improve their designs.

The future is human-centered—filled with supertools, tele-bots, and active appliances that amplify human abilities, empowering people in remarkable ways while ensuring human control. This compelling HCAI prospect enables people to see, think, create, and act in extraordinary ways by combining engaging user experiences with embedded AI algorithms to support services that users want.

However, I am well aware that my vision for the future is still a minority position, so there is much work to be done to steer researchers, developers, managers, and policy-makers to a human-centered agenda.

Underlying belief systems have long been at the heart of technology discussions. Three of those foundational issues are covered in the next three chapters:

- Chapter 2: How does rationalism or empiricism provide sound foundations?
- Chapter 3: Are people and computers in the same category?
- Chapter 4: Will automation, AI, and robots lead to widespread unemployment?

Then Chapter 5 summarizes this part and reminds readers of why they might be skeptical about my approach.

CHAPTER 2

How Do Rationalism and Empiricism Provide Sound Foundations?

The contrast between AI and HCAI is a continuation of the 2000-year-old clash between Aristotle's rationalism, based on logical analyses, and Leonardo da Vinci's empiricism, based on sensory exploration of the world. Both are valuable and worthy of understanding.

The differences came through when using a Roomba robot vacuum cleaner from iRobot. I was eager to buy this device, which has been refined for thirty years and has sold 30 million of these active appliances. The online customer reviews were 70% positive ("I love it," "impressive") with only a few percent having a bad experience ("absolutely despise this product," "sending it back"). Roombas are a good model to follow, but there is room for improvement. The design goal was for it to vacuum your home or apartment on its own, so there are only three buttons with a few colored lights. The sparse "Owner's Guide" (instead of a "User's Guide") has only a few paragraphs on how to use the three buttons and what the lights mean.

In short, Roomba was designed to do the job on its own, which is what many users want, so maybe the stripped-down user interface was a good decision. An alternate design could give users meaningful controls so they know what it will do next and control the order it cleans rooms. The smartphone app shows a floor map that Roomba detects, but design improvements could give users more insight into what is happening, such as where it will go next and how long it will take. Roomba was designed by rationalist thinking to do the job on its own, rather than by empiricist thinking that would give users greater

sense of control. By contrast, hugely successful, much loved, digital camera apps emerged from empiricist thinking, which puts the users first by giving them a simple point and shoot device, and also has easy-to-use controls so they can chose from many modes of operation, including selfies and portrait lighting, and then preview the image that they will get. Users can explore alternatives like videos or panoramic views, take dozens of photos, edit them or add annotations, and then immediately share them with friends and family. A few decades ago, only professional photographers could reliably take high-quality images and they might take days or weeks to print so they could mail copies out.

There are many nuanced discussions of rationalism and empiricism, but here's how I understand that debate. Rationalists believe in logical thinking, which can be accomplished in the comfort and familiarity of their office desk or research lab. They have confidence in the perfectability of rules and the strength of formal methods of logic and mathematical proofs. They assume the constancy of well-defined boundaries—like hot and cold, wet and dry. Aristotle recognized important distinctions, such as the differences between vertebrates and invertebrates, or the four categories of matter: earth, water, air, and fire. These categories are useful, but can become limiting in seeing other options, middle grounds, and newer patterns.

Aristotle's devotion to rational reflection rather than empirical observation, sometimes led him astray, as in his belief that women had only twenty-eight teeth, when a simple examination would have corrected his error. Followers of rationalism have included René Descartes, Baruch Spinoza, Immanuel Kant, and in the twentieth century, the famed statistician Ronald Fisher. His overly strong commitment to statistics led him to reject early data on smoking, so he continued his smoking habit, eventually dying of lung cancer. Rationalism, especially as embodied in logical mathematical thinking, is the basis for much of AI science research, in which logical thinking leads to algorithms that are treasured for their elegance and measured by their efficiency.

Rational thinking leads to medical information systems that require clinicians to enter reports about human health with a limited set of categories or codes. This formalization has benefits in forcing agreement about the categories, but has limitations because human health deserves more than a set of checkboxes or numeric ratings, which is why free text reports by clinicians are valued. Similarly, rules-based or decision tree models have their benefits and limitations. AI innovation researchers, who seek to make commercial products and services, realize that a rational approach may be a good start,

but they know there are benefits from adding a human-centered empirical approach.

Empiricists believe that researchers must get out of their offices and labs to sense the real world in all its contextual complexity, diversity, and uncertainty. They understand that beliefs have to be continuously refined to respond to changing realities and new contexts. Leonardo da Vinci developed fluid dynamics principles by using his keen eyesight to study bird flight and generalized by watching water flowing around obstacles. Galileo Galilei followed da Vinci by noticing the rhythmic swaying of a chandelier in church, leading him to the formula for pendulum swing times. In the 1830s, Charles Darwin traveled to distant destinations, including the Galapagos, to observe the rich diversity of nature, which enabled him to develop the theory of evolution through natural selection.

Other empiricists were John Locke, David Hume, and, in the twentieth century, the statistician John Tukey, who believed in looking at data graphically. Empiricism and empathic observation of people are the basis for much of the user-experience design community, which assesses human performance so as to improve it. Empiricists question simple dichotomies and complex ontologies, because these may limit thinking, undermining analysts' capacity to see importance nuances and non-hierarchical relationships.

The rationalist viewpoint is a strong pillar of the AI community, leading researchers and developers to emphasize data-driven programmed solutions based on logic. Fortunately, an increasing component of AI research bends to the empirical approach, such as in affective computing, healthcare, and the conference on Empirical Methods in Natural Language Processing. The interest in empirical thinking grows when the goal is to build widely used consumer devices.

Rationalism also favors the belief that statistical methods and machine learning algorithms are sufficient to achieve AI's promise of matching or exceeding human intelligence on well-defined tasks. The strong belief in data-driven statistical methods is in contrast to deeply engaging with domain experts who understand causal relationships among variables. AI advocates have gone so far as to say that theories about causal relationships are no longer needed and that machine learning replaces expertise.[1] Others, such as Turing Award winner Judea Pearl, believe that the next step for AI will be to deal with causality.[2]

The troubling belief is that predictions no longer require causal explanations, suggesting that statistical correlations are sufficient to guide decisions.

Yes, machine learning can reveal patterns in the data used to "train" algorithms, but it needs to be extended to deal with surprising extreme cases, such as when the Tesla self-driving car that failed to distinguish a white truck from the sky and smashed into it, killing the driver. Furthermore, machine learning needs supplements that recognize hidden biases and the absence of expected patterns. Improvements to machine learning techniques could make them less brittle in novel situations, which humans cope with by human common sense and higher cognition.[3] Human curiosity and desire to understand the world means that humans are devoted to causal explanations, even when there is a complex set of distant and proximate causes for events.

Both philosophies, rationalism and empiricism, offer valuable insights, so I apply rational thinking for its strengths, but I know that balancing it with an empirical outlook helps me see other possibilities and use observational strategies. Watching users of technology has always led me to fresh insights, so I am drawn to usability studies, interviews, naturalistic observations, and repeated weeks-long case studies with users doing their work to complement the rationalist approach of controlled experiments in laboratory settings.

I think a design philosophy that begins with empathy for users and pushes forward with humility about the limits of machines and people will help build more reliable, safe, and trustworthy systems. Empathy enables designers to be sensitive to the confusion and frustration that users might have and the dangers to people when AI systems fail, especially in consequential and life-critical applications. Humility leads designers to recognize the need for audit trails that can be retrospectively analyzed when the inevitable failures occur. Rationalists tend to expect the best and design for optimal performance; empiricists are always on the lookout for what could go wrong and what could be made better. They thrive on feedback from users.

Implications for Design

Future technology designs are closely tied to beliefs in rationalism or empiricism. The sympathy for rationalism leads some researchers to favor autonomous designs in which computers operate reliably without human oversight. While critics have pointed out the ironies, paradoxes, conundrums, deadly myths, and dangers of imperfect autonomous devices, this approach is still favored by many people. The discussion about autonomy becomes especially fierce when lethal autonomous weapons systems (LAWS)[4] are

debated by military thinkers who see them as an important option and those who fear the dangers of misuse.

Autonomous vehicles or self-driving cars are vigorous technology directions which could have adequate levels of safety if designers took an empiricist's outlook to enable meaningful human control, even as the levels of automation increase.[5] Shifting from *self-driving* to *safety-first* cars might lead to more rapid improvements of proven methods such as collision avoidance, lane following, and parking assist. The shift to using terms like advanced driver assistance systems (ADAS) is an indication of awareness that improving driver performance is a more constructive goal than pushing for self-driving cars. Then further improvements will come from vehicle-to-vehicle communication, improved highway construction, and advanced highway management control centers that build on the strategies of air traffic control centers.

Many AI thinkers continue to imagine a future where social robots will become our teammates, partners, and collaborators. But making machines that pretend to have emotions seems counterproductive and focuses AI designers on a questionable goal. Computers don't have emotions; people do. Today, human-like social robots remain novelties, mostly confined to entertainment.

The AI community's sympathy for rationalism continues to lead developers to favor autonomous designs in which computers operate reliably without human oversight. While there is a growing choir who chant the chorus of a "human-in-the-loop," this phrase often implies a grudging acceptance of human control panels. Those who seek a complete and perfect system are resistant to the idea that there needs to be human intervention, oversight, and control.

A more compelling chorus for me would recognize that humans are happily woven into social networks and that computers should play a supportive role. Humans thrive in social structures of supervisors, peers, and staff whom they want to please, inspire, and respect. They also want feedback, appreciation for their accomplishments, and supportive guidance about how to do better. They use computers to amplify their ability to work in competent, safe, or extraordinary ways. This attitude fits nicely into a bumper sticker "Humans in the group; computers in the loop" (Figure 2.1).

Progress in technology design is likely to accelerate as recognition spreads that humans must have meaningful control of technology and are clearly responsible for the outcomes of their actions. This human-centered, empiricist-driven strategy would seem to be appropriate in military applications where responsibility within a chain of command is a core value.

Fig 2.1 The bumper sticker "Humans in the Group; Computers in the Loop" reminds us that people are social and that they can use computers to support their performance.

Automation is invoked by humans, but they must be able to anticipate what happens, because they are responsible. One effective way to enable users to anticipate what happens is with direct manipulation designs—the objects and actions are represented on the screen; humans choose which actions to carry out; the actions and objects are all visible. Users drop a file into the trash can, accompanied by a clanging sound to signal that it has arrived. Touch screen pinches, taps, and swipes left and right become natural quickly. Visual interfaces provide an overview first, then allow users to zoom in on what they want and filter out what they don't want, and then get details on demand. Where possible, humans are in control and computers are predictable.

Humans want feedback to know that their intent is being carried out by the computer. They want to know what the computer will do next, in enough time to stop or change the action. That's why dialog boxes have a "Cancel" button, so there is a way to stop performance of undesirable actions and go back to a previous state.

Devotees of autonomous design often assume machines will do the right thing, with little interest in giving adequate feedback and even less interest in logging activity to support retrospective review of failures. A better strategy would be to follow civil aviation by installing a "flight data recorder in every robot." Adding audit trails, also called activity or product logs, would signal appropriate humility in addressing consequential and life-critical applications, thereby enabling retrospective analyses of failures and near misses and review of aggregate patterns of usage.

As flaws in AI-driven systems emerged to shatter the belief in their perfectibility, AI researchers have been forced to address issues such as biased evaluations for mortgage applications or parole requests. They began to take on fairness, accountability, transparency, explainability, and other design features

that gave human developers, managers, users, and lawyers a better understanding of what was happening than in the previously closed black boxes. The good news is that a growing community of AI and HCAI researchers are shifting to empirical thinking as they study how to detect bias, what kinds of explanations are successful, and what redress methods for grievances work well.

The HCAI community's belief in empiricism leads participants to design systems with users at the center of attention. HCAI designers start by observing users in their homes and workplaces, interviewing users to get their feedback, and testing hypotheses with empirical studies. Designers conduct user-experience tests to guide repeated design revisions, and follow through with continuous monitoring to gain user feedback during use. HCAI thinking suggests incident reporting and suggestion box schemes, such as the FDA's Adverse Event Reporting System (AERS)[6] and the FAA's Aviation Safety Reporting System.[7]

While AI projects are often focused on replacing humans, HCAI designers favor developing information-rich visualizations and explanations built in, rather than added on. Today, the vast majority of apps are giving users more control—by showing highway navigation routes on maps, exercise histories in bar charts, and financial portfolios in line graphs. These information-abundant displays give users a clear understanding of what is happening and what they can do. Visual displays are now frequently complemented by audio interfaces based on speech recognition and generation, opening up new possibilities for diverse users to accomplish their tasks.

Those who share the rationalists' belief that computers are on the way to replacing people assume that future computers will be as intelligent as people, and also share human emotions. In short, they see no separation between people and what computers can become, so let me explain why I think people are in a different category from computers.

CHAPTER 3

Are People and Computers in the Same Category?

A second contrast between AI and HCAI advocates is the issue of whether people are in the same category as computers or if they are distinct. The Stanford University AI-100 report states that "the difference between an arithmetic calculator and a human brain is not one of kind, but of scale, speed, degree of autonomy, and generality,"[1] which suggests that humans and computers are in the same category. In contrast, many HCAI sympathizers believe that there is a vast difference: "People are not computers. Computers are not people."

It's not that humans have a soul, a spirit, or are a mythical spark of life; it's just that the extraordinary human capabilities, formed by lengthy physical and cultural evolution, deserve appreciation. Human life can only be seen in the context of the remarkable tools people have refined over the generations, such as language, music, art, and mathematics, and technologies, such as clothing, housing, planes, and computers. Human creativity is also apparent in astonishing successes, such as agriculture, healthcare, cities, and legal systems. I believe that our historical role is to add to these technologies, tools for thinking, and cultural systems. Making a robot that simulates what a human does has value, but I'm more attracted to making supertools that dramatically amplify human abilities by a hundred- or thousand-fold. Past accomplishments have produced these kinds of astonishing technological advances, as do computers, the World Wide Web, email/texts, and mobile devices.

Maybe I should be more open to speculative discussions of what is possible in the long run. Maybe I should allow imaginative science fiction stories to

open my mind to new possibilities of sentient computers, conscious machines, and superintelligent AI beings. I am driven by the assumption that people are uniquely creative and that my efforts are best directed towards making well-designed supertools that boost human performance.

Blurring the boundaries between people and computers diminishes appreciation of the richness, diversity, and creativity of people. I prefer to celebrate human accomplishments and treasure the remarkable cultural achievements in language, art, music, architecture, and much more, including machine learning and related algorithms, which are among the great human accomplishments. Clarifying the distinctions between people and computers increases respect for human responsibility and guides people in the appropriate ways to use computer power.[2]

Humans have bodies. Having a body makes you human. It puts us in touch with pain and pleasure, with sadness and joy. Crying and laughing, dancing and eating, love-making and thinking are all parts of being human. Emotions and passions are worth celebrating and fearing. Human emotions go far beyond the seven basic emotions that Paul Ekman described as universal: anger, contempt, disgust, enjoyment, fear, sadness, and surprise.[3] His work, which has been used in the AI community, oversimplifies the complexity of human emotions and their facial expressions. One direction for richer views of emotion is to do what many sentiment-analysis programs do, which is to assume that there are many more emotions (Figure 3.1).

Fig 3.1 Emotions, including the seven from Paul Ekman (blue), the negative emotions (red), the positive emotions (green), and some others (gray).

Source: Adapted from Susannah Paletz, Emotions Annotation Guide for Social Media, Version 3.32, January 21, 2020

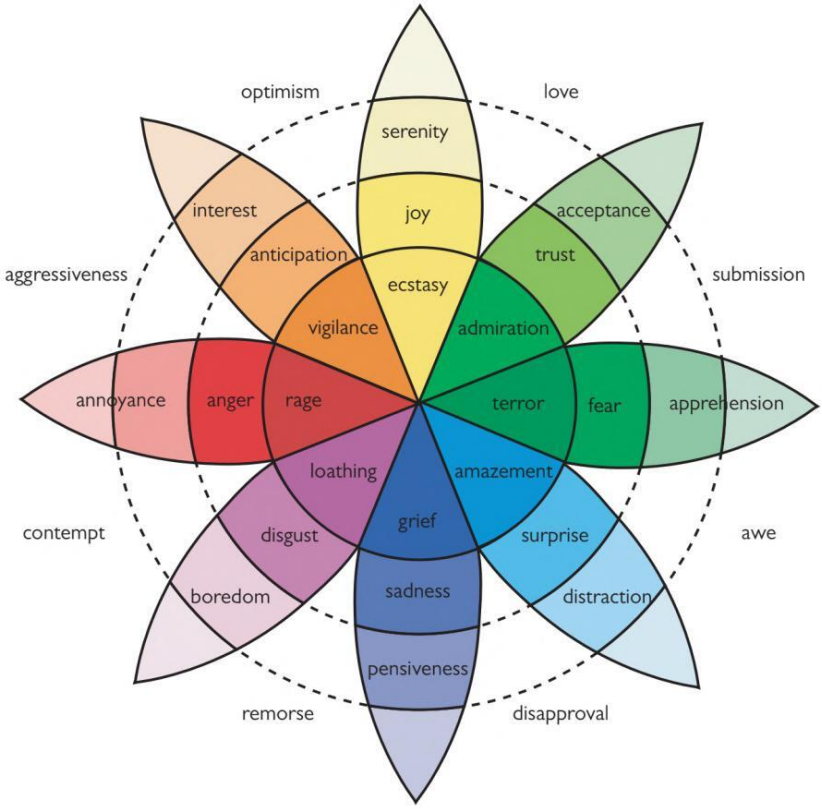

Fig 3.2 Wheel of human emotions.

Source: Robert Plutchik, 1980. Wiki Commons.
https://commons.wikimedia.org/wiki/Category:Plutchik%27s_Wheel_of_Emotions

For those who prefer a visual representation that also suggests stronger and weaker variations and the potential for emotions that fall between two categories, Robert Plutchik's wheel-like diagram with thirty-two emotions may be more appealing (Figure 3.2).

Another line of research sharply criticizes Paul Ekman's model, which might be called the classical view of emotions. In this mechanistic model of human behavior, internal emotional states of mind trigger accurate automatic expressions of emotion, which are the same for all humans. However, recent research, such as the work of Northeastern University psychologist Lisa Feldman-Barrett, favors a theory of *constructed emotions*, which are generated from sensory perceptions, cultural norms, and personal experiences, so expressions of emotions will be very different across individuals.[4] She describes emotional reactions as being constructed based on many factors, rather than automatically triggered

and beyond human control. Therefore, facial expressions and body language are weak indicators of criminal intent, happiness, or fear. Human behavior is more complex than developers of simple algorithms for facial recognition programs assume. Measurable human actions, such as interrupting a speaker, moving closer to or farther away from someone, or making eye contact, can provide valuable feedback to help people change their behaviors.

Human emotions are extraordinarily complex and defy simple processes for recognizing them. Wikipedia's summary on emotions says: "There is currently no scientific consensus on a definition. Emotions are often intertwined with mood, temperament, personality, disposition, creativity, and motivation." Efforts to use facial recognition to determine personality, criminal intent, or political orientation can be as dangerous as the discredited ideas of phrenology, which suggested head and skull structures indicated mental abilities, personality, or criminal tendencies.[5]

While there is a large body of work on how computers can detect human emotional states and then respond to them, many researchers now question how accurate these can be. Even if it were possible, the idea of enabling social robots to express emotions in facial features, body language, and spoken language is troubling. Deceptive practices, whether banal or intended, can undermine the very trust that designers seek to build.[6] Emotional reactions by computers may be useful in entertainment or game applications, which may be enough to justify the research, but for most applications users want to get their tasks done with minimal distraction. Some users may be annoyed by or distrust computers that pretend to express emotion.

A more promising and reliable strategy is sentiment analysis, which analyzes text in social media posts, product reviews, or newspaper headlines. These aggregate data analyses, not attempt to identify the current emotions of an individual, can show differences in language usage by men and women, Democrats and Republicans, ethnic groups, or socioeconomic clusters. Sentiment analysis can also show changes over time, for example, to show that newspaper headlines have become increasingly negative.

Mimicking or imitating a human by computer is an enjoyable pursuit for some people, but a technology designer's imagination could be liberated by using other inspirations. More ambitious goals lead to valued innovations such as the World Wide Web, information visualization, assistive technology, Wikipedia, and augmented reality. These innovations extend human abilities to enable more people to be more creative more often.

Another central goal for me is to support human-to-human communication and cooperation, which have spawned success stories around email/texting, video conferencing, document sharing, and social media. Web-accessed videos, music, and game-playing are huge successes as well, often energized by going social to share favorites with friends, reach out to communicate with large audiences, and promote businesses to broad markets. All these successes have downsides of reducing face-to-face contacts, allowing mischievous scams, and permitting malicious actors to carry out crimes, spread hatred, or recruit terrorists. Just as residents can limit who comes into their homes, users should have rich controls to limit what kinds of messages they receive from autonomous anonymous bots. Social media platforms have yet to do their job to restrict misuses by giving users better controls over what they see.

Another question: what value is there in building computers that look and act like people? As we'll see in Part 3, there is a large community of people who believe that human-like, also called anthropomorphic, humanoid, or android, computers are the way of the future. This community wants to make social robots with human faces, arms, legs, and speech capabilities that could move around in a human world, maybe as older adult caretakers or disaster-response robots. This notion has lead to a long history of failures. Advocates say that this time is different because computers are so much more powerful and designers are so much more knowledgeable.

Human–human communication and relationships are just one model, and sometimes a misleading one, for the design of user interfaces. Humans relate to humans; humans operate computers. Improved interfaces will enable more people to carry out more tasks more rapidly and effectively. Voice is effective for human–human interaction, but visual designs of interfaces will be the dominant strategy because they enable users to operate computers rapidly. Voice user interfaces, such as Alexa and Siri, have an important role, especially when hands are busy and mobility is required (Chapter 16), even though the ephemeral and slow nature of voice communication limits its utility. Furthermore, human generation of speech commands requires substantial cognitive effort and working memory resources, limiting the parallel effort possible when using hand gestures and controls.

Interface designs that are consistent, predictable, and controllable are comprehensible, thus enabling mastery, satisfaction, and responsibility. They will be more widely used than ones that are adaptive, autonomous, and anthropomorphic.

Amplifying human abilities is a worthy goal. Telescopes and microscopes are extensions of the human eye that amplify human abilities. Calculators, digital libraries, and email enable users to do things that no human could do unaided. We need more powerful augmentation and amplification tools that empower people. One approach is the development of creativity support tools that give artists, musicians, poets, playwrights, photographers, and videographers more flexibility to explore alternatives and creatively produce something novel, interesting, and meaningful. Cameras and musical instruments have extended the possibilities of what people can do, but the human is still the driving creative source. Newer devices are likely to carry forward that tradition.

However, some researchers claim that AI technologies do more than empower people; these new technologies are the creators themselves. This claim goes back to the early days of computer art, at least to the time when Jasia Reichardt curated the Cybernetic Serendipity exhibition in London in 1968. Soon after that, Harold Cohen began working on a program he called AARON, which generated images of plants, people, and more abstract forms that were widely appreciated because they resembled watercolor paintings that appeared to have been made by a human. However, Harold Cohen was clearly the creator and therefore the recipient of the 2014 Lifetime Achievement Award in Digital Art from the largest computer graphics professional society, ACM's SIGGRAPH.

Other forms of computer-generated art has more geometric patterns in them, often with algorithmically generated features, adding to the suggestion that the art pieces go beyond the artist's imagination. Leading contributors such as Paul Brown[7] and Ernest Edmonds[8] have exhibited around the world and their work has been collected by major art galleries and museums. Brown uses evolving generative patterns and seeks art "that makes itself," but his books and exhibits list him as the artist. Edmonds, who is also a respected computer scientist, pursues interactive art that changes depending on who is viewing the art. He uses computation "to extend and amplify my creative process not to replace it." Like Harold Cohen, Ernest Edmonds received the ACM SIGGRAPH Lifetime Achievement Award in Digital Art in 2017.

Current AI art producers see their work as a step forward in that they create ever-more ambitious images that are surprising even to the programmers. These artists like, Alexander Mordvintsev,[9] produce something magical in that their Generative Adversarial Networks (GANs) use machine learning algorithms, trained on a set of images, so that the program can act autonomously to make novel images. Mordvintsev's DeepDream program[10] produces engaging

and sometimes troubling images of distorted animals with multiple heads, eyes looking through furry limbs, and pets merging into their backgrounds in ways that challenge our view of reality.

While Harold Cohen considered AARON to be acting autonomously to generate images that were surprising to him, he told me that ultimately he was the creator of the artworks. While potent algorithms and technologies give artists new ways of producing art, the artists are still the source of the creative passion. AARON's work and recent AI art have gotten attention by being sold at auction, yet the proceeds and copyright still come to the artist-creator.

Computer-generated music also stimulates lively discussions of whether the music is produced by the human programmer or by the AI-driven computer program. Computer music algorithms have long been able to generate new music in the style of Bach or the Beatles, Mozart or Madonna, but critics disagree about whether the credit should go to the author or the algorithm. Some algorithms trained on databases of popular songs generate lyrics and music, giving a still richer sense of innovation. However, musicians, like jazz performer and musical therapist Daniel Sarid, suggest that these explorations are "an interesting exercise in understanding human cognition and esthetic organization as well as what constitutes musical language, but, has nothing to do with art." Sarid suggests that composers have a higher goal—they are on "a quest into the collective unconscious of the society/community within which he/she creates."[11]

Some enthusiasts would like to grant computer algorithms intellectual property rights for the images and music produced, but the US copyright office will only grant ownership to humans or organizations. Similar efforts have been made to have computer algorithms hold patents, which have yet to win legal approval. The debate continues, even though it is still unclear how algorithms would rise up to defend their intellectual property, pay damages, or serve jail time for violations.

As time passes, we will see more clearly that people are not computers and computers are not people. As people develop more ambitious embedded computerized applications, the computer as an object of attention will vanish, just as steel, plastics, and glass have become largely invisible parts of our surroundings. Even as computers become more powerful, the notion of computers being intelligent will be seen as naïve and quaint, just as alchemy and astrology are seen now.

That's the future, but let's examine the more immediate question that is on many people's minds: will automation, AI, and robots lead to widespread unemployment?