Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-centered AI Systems

BEN SHNEIDERMAN, University of Maryland, College Park, MD

This article attempts to bridge the gap between widely discussed ethical principles of Human-centered AI (HCAI) and practical steps for effective governance. Since HCAI systems are developed and implemented in multiple organizational structures, I propose 15 recommendations at three levels of governance: team, organization, and industry. The recommendations are intended to increase the reliability, safety, and trustworthiness of HCAI systems: (1) reliable systems based on sound software engineering practices, (2) safety culture through business management strategies, and (3) trustworthy certification by independent oversight. Software engineering practices within teams include audit trails to enable analysis of failures, software engineering workflows, verification and validation testing, bias testing to enhance fairness, and explainable user interfaces. The safety culture within organizations comes from management strategies that include leadership commitment to safety, hiring and training oriented to safety, extensive reporting of failures and near misses, internal review boards for problems and future plans, and alignment with industry standard practices. The trustworthiness certification comes from industry-wide efforts that include government interventions and regulation, accounting firms conducting external audits, insurance companies compensating for failures, non-governmental and civil society organizations advancing design principles, and professional organizations and research institutes developing standards, policies, and novel ideas. The larger goal of effective governance is to limit the dangers and increase the benefits of HCAI to individuals, organizations, and society.

CCS Concepts: • **Computing methodologies → Philosophical/theoretical foundations of artificial intelligence**; • **Human-centered computing → HCI theory, concepts and models**; • **Software and its engineering → Software development process management**;

Additional Key Words and Phrases: Human-centered AI, Human-Computer Interaction, Artificial Intelligence, reliable, safe, trustworthy, software engineering practices, management strategies, independent oversight, design

ACM Reference format:
Ben Shneiderman. 2020. Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-centered AI Systems. *ACM Trans. Interact. Intell. Syst.* 10, 4, Article 26 (October 2020), 31 pages.
https://doi.org/10.1145/3419764

Authors' address: B. Shneiderman, Department of Computer Science, University of Maryland, College Park, MD 20742; email: bshneide@umd.edu.

2160-6455/2020/10-ART26 $15.00
https://doi.org/10.1145/3419764

1 INTRODUCTION

The widespread application of artificial intelligence (AI) comes with high expectations of benefits for many domains, including healthcare, education, cybersecurity, and environmental protection. However, there are equally dire predictions of out-of-control robots, biased decision making, unfair treatment of minority groups, privacy violations, adversarial attacks, and challenges to human rights. While the AI research community is shifting to emphasize Human-centered Artificial Intelligence (HCAI), there is also resistance to change [101].

While both AI and HCAI have multiple definitions [124], traditional AI science research focused on emulating (some would say simulating) human behavior, while AI engineering emphasized replacing human performance. Typical technologies and applications include pattern recognition (images, speech, facial, signal, etc.), natural language processing and translation, bipedal robots, emotionally responsive human faces, and game playing (checkers, chess, go). In contrast, HCAI focuses on amplifying, augmenting, and enhancing human performance [103] in ways that make systems reliable, safe, and trustworthy [102]. These systems also support human self-efficacy, encourage creativity, clarify responsibility, and facilitate social participation.

HCAI systems emerge when designers, software engineers, and managers adopt user-centered participatory design methods by engaging with diverse stakeholders. Then usability testing will help ensure that these systems support human goals, activities, and values. The shift from measuring only algorithm performance to evaluating human performance and satisfaction is a strong signal of the shift to HCAI.

Variant HCAI definitions come from prominent institutions such as Stanford University (http://hai.stanford.edu), which seeks "to serve the collective needs of humanity" by understanding "human language, feelings, intentions and behaviors." There is a shared belief that machine learning is a frequent component for HCAI systems. However, modern data-driven machine learning methods, such as deep learning, make it more difficult to know where the failure points may be.

HCAI represents a second Copernican revolution. In the past, researchers and developers focused on building AI algorithms and systems, stressing the autonomy of machines rather than human control through user interfaces. In contrast, HCAI puts the human users at the center of design thinking, emphasizing user experience design. Researchers and developers for HCAI systems focus on measuring human performance and satisfaction, valuing customer and consumer needs, and ensuring meaningful human control [98].

This second Copernican revolution will take decades until it is widely accepted, as it represents a fundamental shift in thinking from machine-centered to human-centered outlooks. This article offers 15 practical recommendations to encourage and accelerate this shift. However, at least two sources of HCAI system complexity make it difficult to implement all 15 recommendations described in this article. First, individual components can be carefully tested, reviewed, and monitored, but complete HCAI systems, such as self-driving cars, aircraft autopilots, or electronic healthcare systems need higher levels of independent oversight and reviews of failures and near misses. Second, complete HCAI systems are woven together from many products and services, including chips, software development tools, training data suppliers, web-based services, and equipment maintenance providers, which may change, sometimes on a daily basis. These are grand challenges for software engineers, managers, and policy makers, so the recommendations in this article are meant to launch much needed discussions that can lead to constructive changes.

There are more than 300 reports describing aspirational HCAI principles from companies, professional societies, governments, consumer groups, and non-government organizations [95]. The Berkman Klein Center discusses the upsurge of policy activity, followed by a thoughtful summary of 36 of the leading and most comprehensive reports. They identify eight HCAI themes for deeper

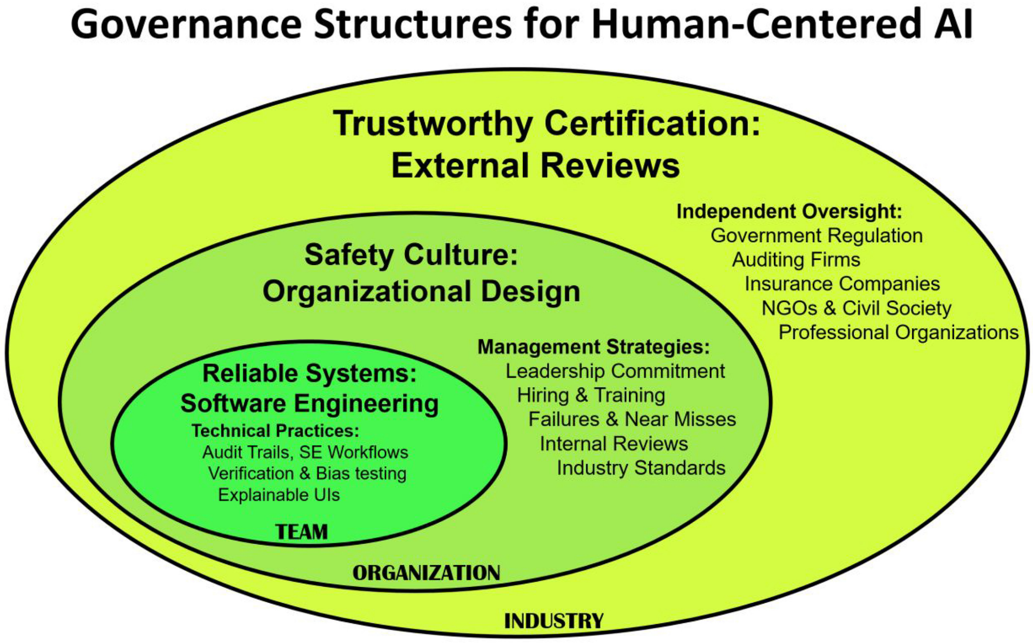


Fig. 1. Governance structures for Human-centered AI with three levels: reliable systems based on software engineering (SE) practices, a well-developed safety culture based on sound management strategies, and trustworthy certification by external review.

commentary and detailed principles: privacy, accountability, safety & security, transparency & explainability, fairness & non-discrimination, human control of technology, professional responsibility, and promotion of human values [36].

Other reports stress ethical principles, such as IEEE's extensive, "Ethically Aligned Design," which emerged from a 3-year effort involving more than 200 people. The report offered clear statements about eight general principles: human rights, well-being, data agency, effectiveness, transparency, accountability, awareness of misuse, and competence. It went further with strong encouragement to ensure that advanced systems "shall be created and operated to respect, promote, and protect internationally recognized human rights" (https://ethicsinaction.ieee.org/) [56]. These and other ethical principles are an important foundation for clear thinking, but as Winfield and Jirotka [122] note: "The gap between principles and practice is an important theme."

To help bridge this gap, this article offers a three-layer governance structure for HCAI systems: (1) reliable systems based on sound software engineering practices, (2) safety culture through proven business management strategies, and (3) trustworthy certification by independent oversight (Figure 1). The inner oval covers the many software engineering teams, which apply technical practices relevant to each project. These teams are part of a larger organization where safety culture management strategies influence each project team. In the largest oval, independent oversight boards review many organizations in the same industry, giving them a deeper understanding, while spreading successful practices.

Reliability, safety, and trustworthiness are vital concepts for everyone involved in technology development, whether driven by AI or other methods. These concepts and others, including privacy, security, environmental protection, social justice, and human rights are also strong concerns at every level: software engineering, business management, and independent oversight.

While governments and corporations often make strong positive statements about their commitment to serving stakeholder needs, when leaders have to make difficult decisions about power and money, they may favor their personal needs, political pressures, and stockholder expectations [58]. Current movements for Human Rights and Corporate Social Responsibility are helpful in building public support, but technology decisions by software engineers, managers, and review boards have to be guided by clear principles and actionable recommendations. This is especially true in emerging technologies, such as HCAI, where government leaders and corporate managers may need to be informed about the range of options available to them.

These proposed governance structures are small practical steps based on existing practices, which have to be adapted to fit new HCAI technologies. They are meant to clarify who takes action and who is responsible. Each idea requires research and testing to validate effectiveness, but they are oriented to designing HCAI systems that are reliable, safe, and trustworthy, which in turn bring benefits to individuals, organizations, and society [68, 119]. These governance structures are a starting point. Newer approaches will be needed as technologies advance or when market forces and public opinion shape the products and services that become successful. For example, such forces dramatically shifted business practices over facial recognition technologies in 2020, when leading developers withdrew from selling to police departments because of potential misuse and abuse.

In the rest of the article, Section 2 describes five technical practices of software engineering teams that enable reliable HCAI systems: audit trails, workflows, verification and validation testing, bias testing, and explainable user interfaces. Section 3 covers the ways that the organizations that manage software engineering projects can develop a safety culture through leadership commitment, hiring and training, reporting failures and near misses, internal reviews, and industry standards. Section 4 shows how independent oversight methods by external review organizations can lead to trustworthy certification and independent audits of products and services. These independent oversight methods create a trusted infrastructure to investigate failures, continuously improve systems, and gain public confidence. Independent oversight methods include government, auditing firms, insurance companies, NGOs and civil society, and professional organizations. Section 5 concludes with concerns and optimism that well-designed HCAI systems will bring meaningful benefits to individuals, organizations, and society.

The shift to human-centered thinking will be difficult for those who have long seen AI algorithms and systems as the goal. They will question the validity of this second Copernican revolution, but human-centered thinking and practices offer a hope-filled vision of future technologies that support human self-efficacy, creativity, responsibility, and social connectedness among people.

2 RELIABLE SYSTEMS BASED ON SOUND SOFTWARE ENGINEERING PRACTICES

Reliable HCAI systems are produced by applying sound technical practices to software engineering teams [75, 3]. These technical practices clarify human responsibility, such as audit trails for accurate records of who did what and when, and histories of who conducted design, coding, testing, and revisions [17].

2.1 Audit Trails and Analysis Tools

The success of Flight Data Recorders (FDR) in making civil aviation remarkably safe provides a clear guide for the design of any product or service that has consequential or life-critical impacts. This history of FDRs and the Cockpit Voice Recorders (CVR) demonstrates the value of using these tools to understand aviation crashes [42, 16, 59], which have contributed strongly to safe civil aviation. Beyond accident investigations, FDRs have proven to be valuable in showing what was done right to avoid accidents to improve training and equipment design. FDR data is now being used to detect changes in equipment behavior over time to guide preventive maintenance.

FDRs provide important lessons for HCAI designers of audit trails (also called product logs) to record the actions of robots [89, 121, 106, 74]. These robot versions of aviation flight data recorders have been called software, smart, ethical, or robot black boxes, but the consistent intention of designers is to collect relevant evidence for retrospective analyses of failures [105, 31]. These retrospective analyses are often to assign liability in legal decision-making and to provide guidance for continuous improvement of these systems. They also clarify responsibility, which exonerates those who have performed properly.

Similar proposals have been made for highly automated (also called self-driving or driverless) cars [86, 125], which extend current work on electronic logging devices, which are installed on many cars to support better maintenance. Secondary uses of vehicle logging devices are to improve driver training, monitor environmentally beneficial driving styles, and verify truck driver compliance with work and traffic rules. In some cases, these logs have provided valuable data in analyzing the causes of accidents, but controversy continues about who owns this data and what rights manufacturers, operators, insurance companies, journalists, and police have to gain access.

Industrial robots are another application area for audit trails, to promote safety and reduce deaths in manufacturing applications. Industry groups such as the Robotic Industries Association (www.robotics.org) has promoted voluntary safety standards and some forms of auditing since 1986.

Audit trails for stock market trading algorithms are now widely used to log trades so that managers, customers, and the U.S. Securities and Exchange Commission can study errors, detect fraud, or recover from flash crash events [99, 112]. Other audit trails from healthcare, cybersecurity, and environmental monitoring enrich the examples from which HCAI audit trails can be designed.

Challenging HCAI research questions remain, such as what data is needed for effective retrospective forensic analysis and how to efficiently capture and store high volume video, sound, or LIDAR data, with proper encryption to prevent falsification. Logs should also include machine learning algorithms used, the code version, as well as the associated training data at the time of an incident. Then research questions remain about how to analyze the large volume of data in these logs? Issues of privacy and security complicate the design, as do legal issues such as who owns the data and what rights manufacturers, operators, insurance companies, journalists, and police have to access, benefit from, or publish these data sets. Effective user interfaces, visualizations, statistical methods, and secondary AI systems enable investigators to explore the audit trails to make sense of the voluminous data. Audit trails require up-front effort, but by reducing failures they reduce injuries, damage, and costs.

2.2 Software Engineering Workflows

As AI technologies and machine learning algorithms are integrated into HCAI applications, software engineering workflows are being updated. The new challenges include new forms of benchmark testing for verifying and validating algorithms and data (Section 2.3), improved bias testing to enhance fairness (Section 2.4), and agile programming team methods [3]. All these practices have to be tuned to the different domains of usage, such as healthcare, education, environmental protection, and defense. To support users and legal requirements, software engineering workflows have to support explainable user interfaces (Section 2.5).

Zhang et al. [126] describe five typical tasks for machine learning, which may need distinctive workflows:

"(1) **Classification**: to assign a category to each data instance, e.g., image classification, handwriting recognition. (2) **Regression**: to predict a value for each data instance, e.g., temperature/

age/income prediction. (3) **Clustering**: to partition instances into homogeneous regions; e.g., pattern recognition, market/image segmentation. (4) **Dimension reduction:** to reduce the training complexity, e.g., dataset representation, data pre-processing. (5) **Control**: to control actions to maximize rewards, e.g., game playing."

These are helpful, but there are other tasks, such as anomaly detection and recommender systems. Workflows for all these tasks require expanded efforts with user requirements gathering, data collection, cleaning, and labeling, with use of visualization and data analytics to understand abnormal distributions, errors and missing data, clusters, gaps, and anomalies. Then model training and evaluation become a multi-step process that starts with early in-house testing, proceeds to deployment, and remains in place for vital work on continuous monitoring. User experience design promotes improved interfaces, which guide users to ensure that they understand how decisions are made and have recourse when they wish to challenge the decision. These traditional human-computer interaction methods and guidelines [104, 100] are being updated by leading corporations and researchers to meet the needs of HCAI [102, 103].

2.3 Verification and Validation Testing

For AI and machine learning algorithms embedded in HCAI systems, novel processes for algorithm verification and validation are needed, as well as usability testing with typical direct users and indirect stakeholders. The goal is to strengthen the possibility that the HCAI system does what users expected, while reducing the possibility that there will be unexpected harmful outcomes. Civil aviation provides good models for approval of new designs, careful verification and validation testing during early and continuing use, and certification testing for pilots.

Collecting a large set of test cases opens the minds of designers to consider extreme situations and possible failures. Formal statements about requirements are difficult to make for machine learning applications, but specific examples of desired outcomes for given inputs are possible. Contexts of use, such as self-driving cars, image recognition, planning algorithms, or machine translation, also need specialized testing. In situations that involve users, such as mortgage, parole, or job interview requests, usability testing with users is needed. As developers make software changes, the test cases can be rerun to verify safety and effectiveness. A key part of verification is to develop test scenarios to detect adversarial attacks, which would prevent malicious use by criminals, hate groups, and terrorists. As new requirements are added or the context of use changes, new test cases can be added [30].

The U.S. National Security Commission on AI [109] stresses that "AI applications require iterative testing, evaluation, verification, and validation that incorporates user feedback." This broad encouragement is refined by Zhang et al. [126], who make important distinctions in testing for three forms of machine learning:

"**Supervised learning:** a type of machine learning that learns from training data with labels as learning targets… **Unsupervised learning:** a learning methodology that learns from training data without labels and relies on understanding the data itself. **Reinforcement learning:** a type of machine learning where the data are in the form of sequences of actions, observations, and rewards."

However, since machine learning is highly dependent on the training data, different datasets need to be collected for each context to increase accuracy and reduce biases. In validating a pneumonia detection AI system for chest x-rays, the results varied greatly across hospitals depending on what x-ray machine was used, unrelated patient characteristics, and even the angle of the machine [23]. Documenting these multiple datasets, which may be updated regularly, or even continuously, is a vital next step, but it presents substantial challenges that go well beyond what programming code repositories currently accomplish. Data curation concepts such as provenance tracking with blockchain [70], ensuring that the data is still representative in the face of change, and connecting with people responsible for datasets are promising possibilities.

Lessons from database systems [46] and information visualization [91] are useful. The history of testing should be recorded to enable reconstruction and document how and by whom problems were repaired [11, 126].

For mobile robotic devices, which could inadvertently harm nearby human workers, deadly weapons, and medical device, special care is needed during testing. Metrics for "safe operation, task completion, time to complete the task, quality, and quantity of tasks completed" will guide development [10]. Mature application areas such as aviation, medical devices, and automobiles, with a long history of benchmark tests for product certification, provide good models for newer products and services.

2.4 Bias testing to Enhance Fairness

As AI and machine learning algorithms were applied to consequential applications such as parole granting, mortgage loan approval, and job interviewing, many critics [81] raised questions about their:

- opacity (difficulty in challenging algorithmic decisions),
- scale (use by large companies and governments for major applications), and
- harm (unfair treatment that impacted people's lives).

A growing research community responded with influential conferences, such as the one on Fairness, Accountability, and Transparency in Machine Learning (https://facctconference.org/), which studied gender, racial, age, and other forms of bias. Commercial practices began to shift when serious problems emerged [51] from biased decisions that influenced parole granting, when hate-filled chatbots learned from malicious social media postings, and when job hiring biases were exposed [87].

An early review by Friedman and Nissenbaum [39] described three kind of bias, which remain as a helpful guide: preexisting biases based on social practices and attitudes, technical bias based on design constraints in hardware and software, and emergent bias that arises from changing the use context. Baeza-Yates [5] described additional forms of bias, such as geography, language, and culture, which were embedded in web-based algorithms, databases, and user interfaces. Questions of bias are closely tied to the IEEE's Ethically Aligned Design report that seeks to build a strong ethical foundation for all AI projects.

Morris [77] raised concerns about how AI systems often make life more difficult for users with disabilities, but wrote "AI technologies offer great promise for people with disabilities by removing access barriers and enhancing users' capabilities." Morris suggests that speech-based virtual agents and other HCAI applications could be improved by using training datasets that included users with physical and cognitive disabilities. However, since AI systems could also detect and prey on vulnerable populations such as those with cognitive disabilities or dementia, research is needed on how to limit such attacks.

Converting ethical principles and bias awareness into action begins with in-depth testing of training datasets to verify that the data is current and has a fair distribution of records.

Beyond detecting biases, researchers offer fairness-enhancing interventions [38] and companies have developed commercial grade toolkits for detecting and mitigating algorithmic bias [7]. These examples are a good start, but every development team should include a bias testing leader for the training datasets and the programs themselves. A library of test cases with expected results could then be used to verify that the HCAI system did not show obvious biases. Continuing monitoring of usage with reports returned to the bias testing leader will help to enhance fairness. However, since development teams may be resistant to recognize biases in their HCAI systems, someone outside the team, will also need to monitor performance and reports over time (see Section 3 on safety culture management practices).

These constructive steps are a positive sign, but the persistence of bias remains a problem as applications such as facial recognition become more widely used for police work and commercial applications [82, 84]. Simple bias tests for gender, race, age, and so on, were helpful in building more accurate face databases, but problems remained when the databases were studied for intersections such as gender and race [14]. Presenting these results in refereed publications and in widely seen media can pressure the HCAI systems builders to make changes that demonstrate clear improvements. Joy Buolamwini, who founded the Algorithmic Justice League (see Appendix A), was able to show gender and racial bias in facial recognition systems from Microsoft, IBM, and Amazon, which she presented in compelling ways through her public talks, op-eds, and videos [14]. Her efforts led to improvements and then corporate withdrawal of facial recognition products from police departments when evidence of excessive use of force became widespread in spring 2020. Effective bias testing for machine learning training data is one contribution to changing this long history of systemic bias in treatment of minorities.

2.5 Explainable User Interfaces

Designers of HCAI systems have come to realize that consequential decisions that influence people's lives, such as rejections for mortgages, parole, or job interviews, often raise questions from those who are affected. Therefore, these systems must come with explanations that enable people to understand the decisions, so they know what they need to change or whether they should challenge the decision. Furthermore, explanations have become a legal requirement in many countries based on the European Union's General Data Protection Regulation (GDPR) requirement of a "right to explanation" [41, 115].

This controversial GDPR requirement is vague and difficult to satisfy in general, but international research efforts at developing Explainable AI (XAI) have blossomed [9, 47, 48, 106, 116, 75, 71]. A useful and practical resource are the three reports on "Explaining decisions made with AI" from the U.K. Information Commissioner's Office and the Alan Turing Institute [57]. The three reports cover: (1) The basics of explaining AI, (2) Explaining AI in practice, and (3) What explaining AI means for your organization. These reports would inspire more confidence if sample screen designs were shown and user testing results were presented.

Weld and Bansal [118] make a strong case for explainability (sometimes called interpretability or transparency) that goes beyond satisfying users' desire to understand and legal requirements to provide explanations. They argue that explainability helps designers enhance correctness, identify improvements in training data, account for changing realities, support users in taking control, and increase user acceptance. An interview study with 22 machine learning professionals documented the value of explainability for developers, testers, managers, and users [52]. However, explainability methods are only slowly finding their way into widely used applications and possibly in ways that are different from the research.

As the AI research community learns more of the centuries of social science research, thoughtfully described by Miller [73], they can apply the relevant theories and rigorous evaluation methods. As AI developers learn more of HCI design methods and usability testing, they will increase the chances of user acceptance of their systems.

Du et al. [28] make useful distinctions between intrinsically understandable machine learning models, such as decision trees or rule-based models [67, 26, 97], and the more common approach of *post hoc* explanations. *Post hoc* (or retrospective) explanations are generated for users who are surprised by results and want an explanation of why a certain decision was made or not made, e.g., why a mortgage application was rejected. These *post hoc* explanations seem to be the preferred approach especially when deep-learning neural nets and other black box methods are used.

Several forms of *post hoc* explanations were tried and abandoned in favor of more effective approaches three decades ago in knowledge-based expert systems, intelligent tutoring systems, and user interface help systems.

In Knowledge-based Expert Systems, many projects struggled with providing *post hoc* explanations, including the long series of projects that began with the medically oriented diagnostic system, MYCIN, but expanded to domain independent systems [13]. Clancey [22] describes their pursuit of explainability by way of step-by-step processes, support for "how" and "why" questions tied to causality, counterfactual query support, and graphical overviews. In successful business rule-based systems, designers shifted from dependence on *post hoc* explanations to prospective designs that give users a better understanding of how the application works, so they can prevent mistakes and limit confusion [54, 95].

In user interface help systems, the designers found that *post hoc* explanations and error messages were difficult to design, leading them to shift to alternative methods [104]:

1. preventing errors to reduce the need for explanations: e.g., by replacing typing MMDDYYYY with selecting from a calendar. Replacing typing with selecting eliminates the need for extensive error detection and numerous explanatory messages. and
2. using progressive step-by-step processes in which each question leads to a new set of questions. The progressive step-by-step processes guide users incrementally toward their goals, simplifying each step while giving them the chance to go back and change earlier decisions. Effective examples are in TurboTax for income tax preparation or more simply in the Amazon four-step e-commerce checkout process.

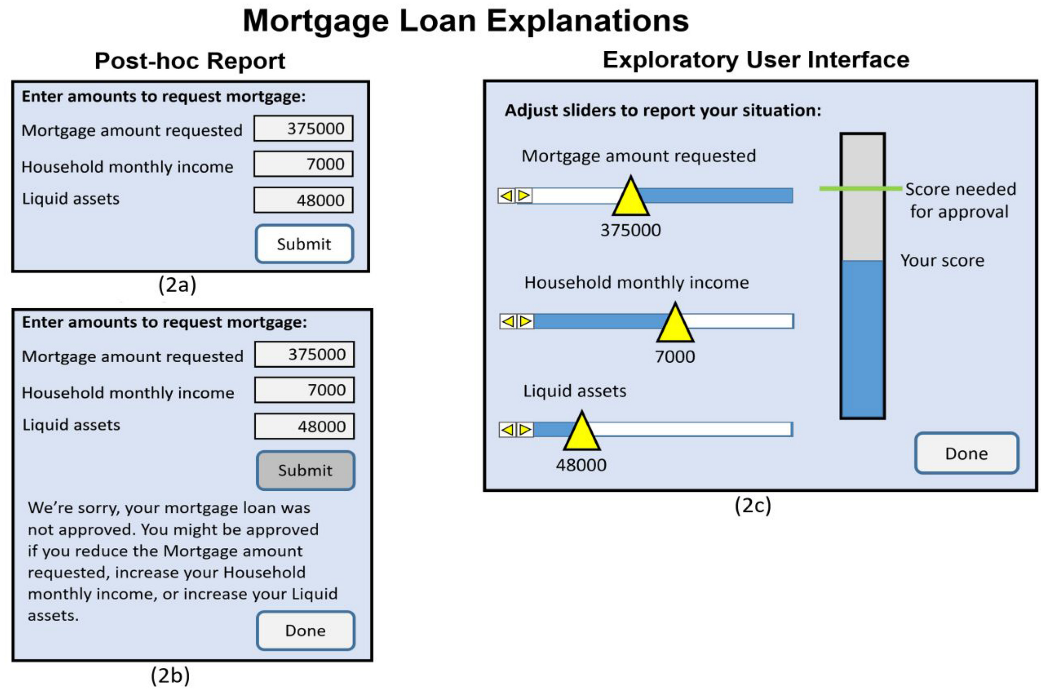
These explanatory or prospective methods may be better described as exploratory user interfaces to signal the capacity of users to probe the algorithm boundaries with different inputs. These exploratory user interfaces will have to be adapted to the five machine learning tasks Zhang et al. describe [126], but the following example for regression decision-making for mortgages suggests what is possible in this simplified case. Figures 2(a) and 2(b) show a *post hoc* explanation for a mortgage rejection, while Figure 2(c) shows an exploratory user interface that enables users to investigate how their input choices affect the outcome, thereby reducing the need for explanations.

In general, exploratory user interfaces are welcomed by users who spend more time developing an understanding of the sensitivity of variables and digging more deeply into aspects that interest them, leading to greater satisfaction and compliance with recommendations [80, 27, 21]. Further gains come from enabling adaptable user interfaces to fit different needs and personalities [127, 119]. Satisfied users increase product acceptance, thereby increasing sales.

For complex decisions, more elaborate user interfaces and data analytics should clarify which of the many features are the most relevant [49], even in deep learning for image understanding [50].

Interactive HCAI approaches are endorsed by Weld and Bansal [118], who recommend that designers should "make the explanation system interactive so users can drill down until they are satisfied with their understanding." Exploration works best when the user inputs are actionable, that is, users have control and can change the inputs. Alternatives are needed when users do not have control over the input values or when the input is from sensors such as in image and face recognition applications. For the greatest benefit, explanatory and exploratory user interfaces should be designed to support accessibility by users with visual, hearing, motor, or cognitive disabilities.

In addition to the distinctions between intrinsically understandable models, *post hoc*, and prospective explanations, Du et al. [28] follow other researchers in distinguishing between global explanations that give an overview of what the algorithm does and local explanations that deal


Mortgage Loan Explanations
Post-hoc Report
Exploratory User Interface
Submit
(2a)
Submit
Done
(2b)
Score needed for approval
Your score
Done
(2c)

Enter amounts to request mortgage:
Mortgage amount requested 375000
Household monthly income 7000
Liquid assets 48000

Enter amounts to request mortgage:
Mortgage amount requested 375000
Household monthly income 7000
Liquid assets 48000

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.

Adjust sliders to report your situation:
Mortgage amount requested 375000
Household monthly income 7000
Liquid assets 48000

Fig. 2. Panels (2a) and (2b) show a typical *post hoc* explanation, while panel (2c) shows an exploratory user interface, which allows users to move the triangular sliders to see how their changes update their score on the right. This interface facilitates investigation that leads to better understanding for this simple mortgage loan application.

with specific outcomes, such as why a prisoner is denied parole or a patient receives a certain treatment recommendation. Local explanations support user comprehension and future actions, such as a prisoner who is told that they could be paroled after four months of good behavior or the patient who is told that if they lost ten pounds they would be eligible for a non-surgical treatment.

Heer [45] shows how to implement exploratory user interfaces that increase user control of AI processes in data cleaning, exploratory data visualization, and natural language translation. Increased user control is also appearing in recommender systems that are moving toward exploratory user interfaces that offer more transparent approaches, especially for consequential medical or career choices [27].

To achieve even greater user satisfaction, Chen et al. propose designing explanations that consider user sentiments when explaining system decisions [21].

3 SAFETY CULTURE THROUGH BUSINESS MANAGEMENT STRATEGIES

Charles Perrow's influential book, *Normal Accidents* [88], made a strong case for organizational responsibility for safety, rather than criticism of specific designs or operator error. His analysis, which emerged from political science and sociology, emphasized the dangers of organizational complexity and overly tight coupling of units with too much centralized control and insufficient redundancy to cope with disruptions. Perrow's work was extended by Klein [60] and criticized by others who pointed to the lack of metrics for the two dangers [53]. Other critics found Perrow's belief that in complex organizations with tight coupling, normal accidents were inevitable was unreasonably pessimistic.

More positive approaches were offered in descriptions of High Reliability Organizations [61, 12, 25], which emerged from organizational design and business administration, and Resilience Engineering [123, 44], which grew out of cognitive science and human factors engineering. A fourth theme described how organizations cultivate safety cultures [43, 8], especially in healthcare [32] by open management strategies. Leveson [68] develops a systems engineering approach to safety engineering that includes design, hazard analysis, and failure investigations. She thoughtfully distinguishes between safety and reliability, pointing out that they are separable issues, demanding different responses.

My distillation of the HCAI-related issues includes these five themes: (1) Leadership commitment to safety, (2) hiring and training oriented to safety, (3) extensive reporting of failures and near misses, (4) internal review boards for problems and future plans, and (5) alignment with industry standards and accepted best practices.

3.1 Leadership Commitment to Safety

Top organizational leaders can make their commitment to safety clear with explicit statements about values, vision, and mission. They have a "preoccupation with failure" [117] with positive statements about building a safety culture, which includes values, beliefs, policies, and norms, and a safety climate, which includes atmosphere, context, and attitudes. The culture is more durable, while the climate may change over time because of internal or external factors.

Leadership commitment is made visible to employees by frequent restatements of that commitment, positive efforts in hiring, repeated training, and dealing openly with failures and near misses. Reviews of incidents, such as monthly hospital review board meetings can bring much increased patient safety. Berry et al. [8] report that, "Improved safety and teamwork climate… are associated with decreased patient harm and severity-adjusted mortality."

Safety-focused leaders stress internal review boards for discussion of plans and problems, as well as adherence to industry standards and practices, such as the software engineering capability maturity model (Section 3.5).

Safety cultures require effort and budget to ensure that there are sufficient and diverse staff involved with sufficient time and resources to do their work. This may imply redundancy to ensure knowledgeable people are available when problems emerge. Safety, reliability, and resilience raise continuing costs, but the reduction of costly destructive events is the payoff. In addition, safety efforts often result in increased productivity, reduced expenses for employee injuries, and savings in operations and maintenance costs. Even so, leadership statements raise commitment from stakeholders, which is necessary when critics question spending to prepare for rare and unpredictable events.

While some literature on safety culture focuses on employee and facility safety, for HCAI systems, the focus must be on those whose lives are impacted by these systems. Therefore, a safety culture for HCAI systems will be built by strong connections with users, such as patients, physicians, or caretakers in a healthcare domain. Outreach to affected communities means two-way communications to inform stakeholders, continuous data collection on usage, and easy reporting of adverse events.

Skeptics fear that corporate safety culture pronouncements are merely public relations attempts to deal with unacceptable risks in many industries such as nuclear power, chemical production, or social media platforms. They also point to cases in which failures were blamed on operator error rather than improper organizational preparation and inadequate operator training. One approach to ensuring safety is to appoint an internal ombudsperson to privately hear staff and stakeholder concerns, while enabling fair treatment of whistleblowers who report serious safety threats.

Implementations of safety cultures in HCAI-based organizations is emerging with initial efforts to support AI governance in medical care [19, 93].

3.2 Hiring and Training Oriented to Safety

When safety is included in job hiring position statements, that commitment becomes visible to current employees and potential new hires. Diversity in hiring also demonstrates commitment to safety by including senior staff that representative the diversity of employees and skills. Safety cultures may need experienced safety professionals from health, human resources, organizational design, ethnography, and forensics.

Safety-first organizations conduct training exercises regularly, such as industrial control room operators carrying out emergency plans, pilots flying simulators, and hospital workers running multiple day exercises for mass casualties or pandemics. When routine practices are similar to emergency plans, employees are more likely to succeed during the emergency. Thoughtful planning includes ranking of emergencies, based on past frequency of occurrence or severity, with an analysis of how many internal responders are needed in each situation, plus planning for how to engage external services when needed. Well-designed checklists can reduce errors in normal operations and remind operators what to do in emergencies.

The training needed for computer software and hardware designers has become easier due to the guidelines documents from leading technology companies such as Apple's Human Interface Guidelines (https://developer.apple.com/design/human-interface-guidelines/ios/overview/themes/) and Google's design guidebook (https://pair.withgoogle.com/intro/), which both contain useful example screen designs. In addition, Microsoft's 18 Guidelines for AI-Human Interaction (https://www.microsoft.com/en-us/research/project/guidelines-for-human-ai-interaction/) and IBM's Design for AI website (https://www.ibm.com/design/ai/fundamentals/) rely on thoughtful general principles, which will need refinement.

These guidelines build on a long history of user interface design [104, 100] and newer research on designing interfaces for HCAI systems [29, 4, 103]. However, guidelines have to be taught to user interface designers, programmers, AI engineers, product managers, and policy makers, whose practices gain strength if there are organizational mechanisms for ensuring enforcement, granting exemptions, and making enhancements.

As HCAI systems are introduced, the training needs for consumers with self-driving cars, clinicians struggling with electronic healthcare systems, and operators of industrial control rooms become more complex. These users need to understand what aspects of the HCAI systems they control and how machine learning works, including its potential failures and associated outcomes. Daugherty and Wilson [24] underline the importance of training: "Companies must provide the employee training and retraining required so that people will be prepared and ready to assume any new roles... investing in people must be the core part of any company's AI strategy."

3.3 Extensive Reporting of Failures and Near Misses

Safety-oriented organizations regularly report on their failures (sometimes referred to as adverse events) and near misses (sometimes referred to as "close calls") [32]. Near misses can be small mistakes that are handled easily or dangerous practices that can be avoided, thereby limiting serious failures. These include near misses, such as an occasional water leak, forced equipment restart, operator error, or electrical outage. If near-miss errors of omission or commission are reported and logged, then patterns become clear to equipment and facility managers, so they can focus attention on avoiding more serious failures. Since near misses typically occur much more often than failures, they provide richer data to guide maintenance, training, or redesign.

The U.S. National Safety Council (http://www.nsc.org) makes the surprising recommendation to avoid rewarding managers whose units have few failures, but rather to reward those managers whose units have high rates of near-miss reports. By making near-miss reporting a common and virtuous practice, staff attention is more focused on safety and ready to make near-miss reports, rather than cover them up.

Civil aviation has a much-deserved reputation for safety. This stems, in part, from a rich culture of near-miss reporting such as through the U.S. Federal Aviation Administration Hotline (https://hotline.faa.gov/), which invites passengers, air traffic controllers, pilots, and the general public to report incidents, anonymously if they wish. The U.S. National Transportation Safety Board, whose public reports are trusted and influential in promoting improvements, thoroughly investigates crashes with injuries or loss of life. In addition, the Aviation Safety Reporting System (https://asrs.arc.nasa.gov/) is a voluntary use website that "captures confidential reports, analyzes the resulting aviation safety data, and disseminates vital information to the aviation community." These public reporting systems are good models for HCAI systems.

In software engineering, code development environments, such as Github, record the author of every line of code and document who made changes. Then, when systems are in operation, bug reporting tools, such as Bugzilla (https://www.bugzilla.org/about/) guide project teams to frequent and serious bugs, with a tracking system to record resolution and testing. Fixing early users' annoyance with problems they find, could prevent other users, possibly with less capacity to recognize problems and find work arounds, from making more serious mistakes.

Bug reporting is easier for interactive systems with clear errors in displays than for highly automated systems without displays, such as elevators, manufacturing equipment, self-driving cars, or drones. This shows the utility of having status displays for equipment to make problems more visible by end users or maintenance personnel. Web-based public reporting systems such as the U.S. Food and Drug Administration's Adverse Event Reporting System (https://www.fda.gov/safety/medwatch-fda-safety-information-and-adverse-event-reporting-program) provide a model for public reporting of problems with HCAI systems.

U.S. Army methods of After-Action Reviews have also been used in healthcare, transportation, industrial process control, environmental monitoring, and firefighting, so they might be useful for studying HCAI failures and near misses [72]. Investigators try to understand what was supposed to happen, what actually happened, and what could be done better in the future. A complete report that describes what went well and what could be improved will help encourage acceptance of recommendations. As After-Action Review participants gain familiarity with the process, their analyses are likely to improve and so will the acceptance of their recommendations.

3.4 Internal Review Boards for Problems and Future Plans

Commitment to a safety culture is shown by regularly scheduled monthly meetings to discuss failures and near misses, as well as to celebrate resilient efforts in the face of serious challenges. Standardized statistical reporting of events allows managers and staff to understand what metrics are important and to suggest new ones. Internal and sometimes public summaries emphasize the importance of a safety culture. Review boards may include managers, staff, and others, who offer diverse perspectives on how to promote continuous improvement. In some industries, such as aviation, monthly reports of on-time performance or lost bag rates drive healthy competition, which serves the public interest. Similarly, hospitals may report patient care results for various conditions or surgeries, enabling the public to choose hospitals, in part, by their performance.

A surprising approach to failures is emerging in many hospitals, which are adopting Disclosure, Apology and Offer programs [6]. Medical professionals usually provide excellent care for their patients, but when problems arise, there has been a tendency to do the best they can for the patient. However, fear of malpractice lawsuits limit physician willingness to report problems to patients, their families, or hospital managers. The emerging approach of Disclosure, Apology, and Offer programs shifts to full disclosure to patients and their families with a clear apology and an offer of treatments to remedy the problem and/or financial compensation. While some physicians and managers feared that this would increase malpractice lawsuits, the results were dramatically different. Patients and their families appreciated the honest disclosure and especially the clear apology. As a result, lawsuits were often cut in half, while the number of medical errors decreased substantially because of physicians' awareness of these programs. Professional and organizational pride also increased [6].

Internal review and auditing teams can also improve HCAI practices to limit failures and near misses. Google's five-stage internal algorithmic auditing framework, which is designed "to close the AI accountability gap," provides a good model for others to follow and build on [92]:

1. Scoping: identify the scope of the project and the audit, raise questions of risk
2. Mapping: create stakeholder map and collaborator contact list, conduct interviews and select metrics
3. Artifact Collection: document design process, datasets, and machine learning models
4. Testing: conduct adversarial testing to probe edge cases and failure possibilities
5. Reflection: consider risk analysis, failure remediation, and record design history

The authors include a Post Audit Review for self-assessment summary report and mechanisms to track implementation. However, they are well aware that "internal audits are only one important aspect of a broader system of required quality checks and balances."

Initial corporate efforts, include Facebook's oversight board, set up in mid-2020 for content monitoring and governance on their platform (https://about.fb.com/news/2020/05/welcoming-the-oversight-board/). Microsoft's AI and Ethics in Engineering and Research (AETHER) Committee advises their leadership on responsible AI issues, technology, processes, and best practices that "warrant people's trust" (https://www.microsoft.com/en-us/ai/our-approach). Microsoft's Office of Responsible AI implements company-wide rules for governance, team readiness, and dealing with sensitive use cases. They also help shape new HCAI-related "laws, norms, and standards…for the benefit of society at large."

3.5 Alignment with Industry Standard Practices

In many consequential and life-critical industries there are established industry standards, often promulgated by professional associations, such as the Robotics Industry Association (RIA, www.robotics.org). The RIA, founded in 1974, works with the American National Standards Institute (www.ansi.org) to drive "innovation, growth, and safety" by developing voluntary consensus standards for use by its members. Their work on robotics is a model for other forms of HCAI.

The International Standards Organization (ISO) has a Technical Committee on Robotics (https://committee.iso.org/home/tc299) whose goal, since 1983, is "to develop high quality standards for the safety of industrial robots and service robots…by providing clear best practices on how to ensure proper safe installations, as well as providing standardized interfaces and performance criteria." The IEEE P7000 series of standards is directly tied to HCAI issues such as transparency, bias, safety, and trustworthiness (https://ethicsstandards.org/p7000/).

The World Wide Web Consortium (W3C) supports website designers who pursue universal or inclusive design goals with the Web's Content Accessibility Guidelines (https://www.w3.org/TR/WCAG21/). Another source is the U.S. Access Board, whose Section 508 standards guide designers in government agencies to "give disabled employees and members of the public access to information that is comparable to the access available to others" (https://digital.gov/resources/

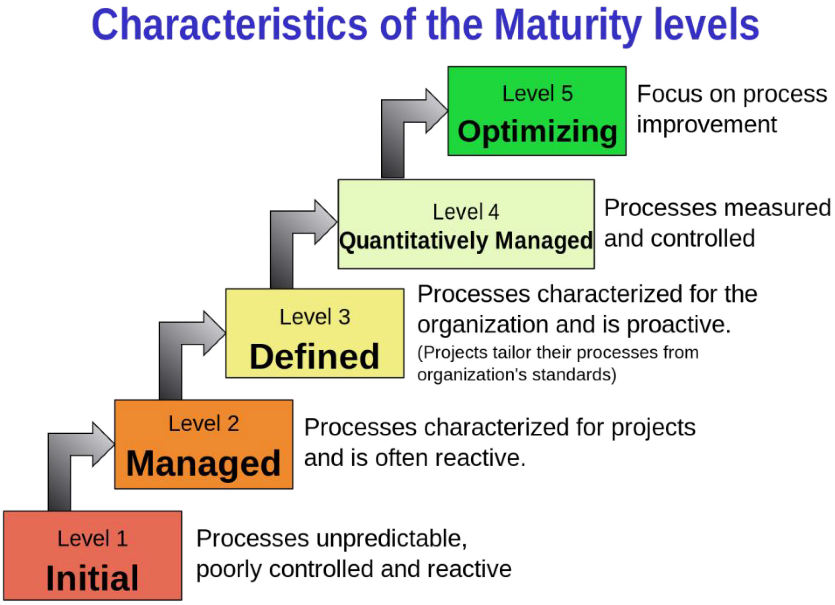

Characteristics of the Maturity levels
Level 5
Optimizing
Focus on process improvement
Level 4
Quantitatively Managed
Processes measured and controlled
Level 3
Defined
Processes characterized for the organization and is proactive.
(Projects tailor their processes from organization's standards)
Level 2
Managed
Processes characterized for projects and is often reactive.
Level 1
Initial
Processes unpredictable, poorly controlled and reactive

Fig. 3. Characteristics of the Capability Maturity Model in five levels (Wikimedia Commons, CMU).

introduction-accessibility/). These accessibility guidelines will be needed to ensure universal usability of HCAI systems, and they provide helpful models of how to deploy other guidelines for HCAI systems.

By working with organizations that develop these guidelines, companies can contribute to future guidelines and standards, learn about best practices, and provide education for staff members. Customers and the public may see participation and adherence to standards as indication of a safety-oriented company. However, skeptics are concerned that corporate participation in developing voluntary standards leads to weak standards whose main goal is to prevent more rigorous government regulation or other interventions.

Another approach to improve software quality is the Capability Maturity Model, developed by the Software Engineering Institute in the late 1980s [55], which is regularly updated [85, 113]. The 2018 Capability Maturity Model Integration version comes with the claim that it helps "integrate traditionally separate organizational functions, set process improvement goals and priorities, provide guidance for quality processes, and provide a point of reference for appraising current processes" [37, 62].

The Capability Maturity Model Integration is a guide to software engineering organizational design with five levels of maturity from Level 1 in which processes are unpredictable and vary across groups. Higher levels define orderly software development processes with detailed metrics for management control and organization-wide discussions of how to optimize performance (Figure 3). Training for staff and management help ensure that the required practices are understood and followed. Many U.S. Government software development contracts, especially from defense agencies, stipulate which maturity level is required, using a formal appraisal process.

Skeptics question whether the Capability Maturity Models lead to top-heavy management structures, which may slow the popular agile and lean development methods. Still, proposals for HCAI Capability Maturity Models are emerging for medical devices, transportation, and other industries [2] with advocates pointing to higher quality software.

HCAI Capability Maturity Models might describe Level 1 initial use of HCAI that is guided by individual preferences and knowledge, making it haphazard and error prone. Level 2 use might call for uniform staff training in tools and processes, while Level 3 might describe repeated use of tools and processes that are reviewed for their efficacy and refined to meet the application domain needs and organization style. Assessments would cover testing for biased data, validation and verification of HCAI systems performance, and reviews of customer complaints. Level 4 might require measurement of HCAI systems and developer performance, with reporting of tracking data to understand how failures and near misses occurred. Level 5 might have repeated measures across many groups and over time to support continuous improvement and quality control.

4 TRUSTWORTHY CERTIFICATION BY INDEPENDENT OVERSIGHT

The third governance layer is independent oversight by external review organizations. Even established large companies, government agencies, and other organizations that build consequential HCAI systems are venturing into new territory, so they will face new problems. Therefore, thoughtful independent oversight reviews will be valuable in achieving trustworthy systems that receive wide public acceptance. However, designing successful independent oversight structures is still a challenge, as shown by reports on more than forty variations that have been used in government, business, universities, non-governmental organizations, and civic society [107, 101].

The key to independent oversight is to support the legal, moral, and ethical principles of human or organizational responsibility and liability for their products and services. Responsibility is a complex topic, with nuanced variations such as legal liability, professional accountability, moral responsibility, and ethical bias [56, 36]. A deeper philosophical discussion of responsibility is useful, but for this article, I assume that humans and organizations are legally liable (responsible) for the products and services that they create, operate directly, maintain, or use indirectly [83].

Professional engineers, physicians, lawyers, aviation specialists, business leaders, and so on, are aware of their personal responsibility for their actions, but the software field has largely avoided certification and professional status for designers, programmers, and managers. In addition, contracts often contain “hold harmless” clauses that stipulate developers are not liable for damages, since software development is often described as a new and emerging activity, even after 50 years of experience. The HCAI movement has raised these issues again with frequent calls for Algorithmic Accountability and Transparency [40], Ethically Aligned Design [56], and Professional Responsibility [36]. Professional organizations, such as the AAAI, ACM, and IEEE, have ethical codes of conduct for their members, but penalties for unethical conduct are rare.

When damages occur, the allocation of liability is a complex legal issue. Many legal scholars however believe that existing laws are sufficient to deal with HCAI systems, although novel precedents will help clarify the issues [15].

Independent oversight is widely used [107] by businesses, government agencies, universities, non-governmental organizations, and civic society to stimulate discussions, review plans, monitor on-going processes, and analyze failures. The goal of independent oversight is to promote continuous improvement to ensure reliable, safe, and trustworthy products and services.

The individuals who serve on independent oversight boards need to be respected leaders whose specialized knowledge makes them informed enough about the organizations they review to be knowledgeable, but far enough away that they are independent. Conflicts of interest, such as previous relationships with the organization that is being reviewed, are likely to exist, so they must be disclosed and assessed. Diverse membership representing different disciplines, age, gender, ethnicity, and other factors help build robust oversight boards.

Their capacity to investigate may include the right to examine private data, compel interviews, and even subpoena power to legally require certain evidence. Their reports may be reviewed by the

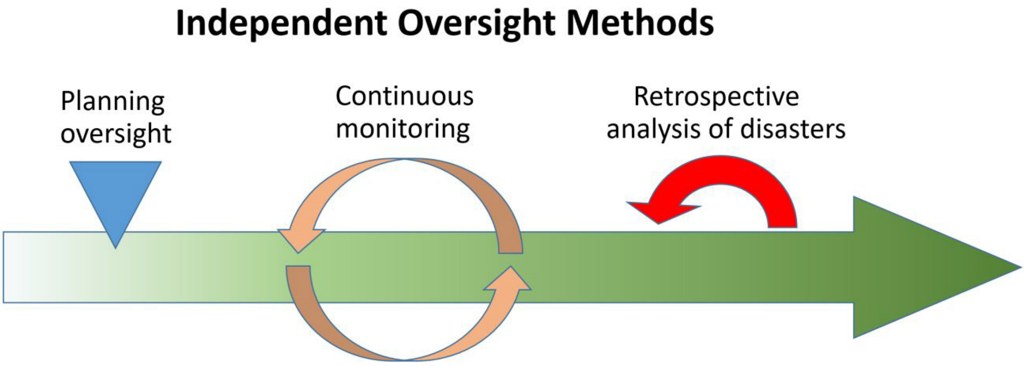

Independent Oversight Methods
Planning oversight
Continuous monitoring
Retrospective analysis of disasters

Fig. 4. Independent Oversight Methods: Planning oversight, continuous monitoring, and retrospective analysis of disasters.

parties in advance, but the decision to publish should reside in the board. Their recommendations will have impact if there is a requirement to respond and make changes within a specified time period, usually measured in weeks or months.

Three independent oversight methods are common (Figure 4) [101]:

1. **Planning oversight:** proposals for new HCAI systems or major upgrades are presented for review in advance so that feedback and discussion can influence the plans. Planning oversight is similar to zoning boards, which review proposals for new buildings that are to adhere to building codes. A variation is the idea of algorithmic impact assessments, which are similar to environmental impact statements that enable stakeholders to discuss plans before implementation [94].
2. **Continuous monitoring:** this is an expensive approach, but the U.S. Food and Drug Administration has inspectors who work continuously at pharmaceutical and meat packing plants, while the U.S. Federal Reserve Board continuously monitors practices at large banks. One form of continuous monitoring is periodic inspections, such as quarterly inspections for elevators or annual financial audits for publicly traded companies. Continuous monitoring of mortgage or parole granting HCAI systems would reveal problems as the profile of applicants' changes or the context shifts, such as has happened during the COVID-19 crisis.
3. **Retrospective analysis of disasters:** the U.S. National Transportation Safety Board conducts widely respected thorough reviews with detailed reports about aircraft, train, or ship crashes. Similarly, the U.S. Federal Communications Commission is moving to review AI systems in social media and web services, especially disability access and fake news attacks. Other agencies in the U.S. and around the world are developing principles and policies to enable study and limitation of HCAI failures. A central effort is to develop voluntary industry guidelines for audit trails and analysis for diverse applications.

Skeptics point to failures of independent oversight methods, sometimes tied to lack of sufficient independence, but the value of these methods is widely appreciated.

In summary, clarifying responsibility for designers, engineers, managers, maintainers, and users of advanced technology will improve safety and effectiveness, since these stakeholders will be aware of their liability for negligent behavior. The five technical practices for software engineering teams (Section 2) are first steps to develop reliable systems. The five management strategies for organizations (Section 3) with many software engineering teams build on existing strategies to promote safety cultures. This section offers five paths to trustworthy certification within an industry by independent oversight reviews, in which knowledgeable industry experts bring successful practices from one organization to another. Skeptics fear that companies will seek to subvert independent oversight, so the business case will have to be developed to supplement public and professional pressure.

4.1 Government Interventions and Regulation

Government agencies already play key roles in improving automated systems. The U.S. National Transportation Safety Board (NTSB) has a long history as a trusted investigator of aviation, ship, train, and other disasters, in part, because the U.S. Congress funds it as an independent agency outside the usual executive branch departments. Their skilled teams arrive at accident scenes to collect data, which become the basis for thoughtful reports with recommendations for improvement.

For example, its report on the deadly May 2016 Tesla crash criticized the manufacturer's and operator's "overreliance on the automation and a lack of understanding of system limitations." The report recommended that future designs include "a standardized set of retrievable data...to enable independent assessment of automated vehicle safety and to foster automation system improvements" [110]. Furthermore, the report cautioned "this crash is an example of what can happen when automation is introduced 'because we can' without adequate consideration of the human element."

The idea of a National Algorithms Safety Board [101] provoked discussions, but adding HCAI expertise to existing oversight boards is a more practical approach. This is just what several U.S. Government agencies have done in their current efforts to review HCAI projects. Good examples are the work of the National Institutes of Health (NIH), Department of Transportation (DOT), Federal Aviation Administration (FAA), and Food and Drug Administration (FDA).

A U.S. White House report emphasizes the need to "Ensure the safety and security of AI systems. Advance knowledge of how to design AI systems that are reliable, dependable, safe, and trustworthy" (U.S. National Science and Technology Council, June 2019). The report stresses research to achieve these goals "but does not describe or recommend policy or regulatory actions related to the governance or deployment of AI."

A broad review of policymaking pointed to AI's distinctive features of intentionality, intelligence, and adaptability, which necessitated new initiatives from government and other bodies [120]. This includes policy decisions, regulatory actions, legal liability, corporate self-policing, and public opinion that demands reasonable safeguards.

Many current AI industry leaders and government policy makers fear that government regulation would limit innovation, but when done carefully, regulation can accelerate innovation as it did with automobile safety and fuel efficiency. A U.S. government memorandum for Executive Branch Departments and Agencies offered ten principles for "stewardship of AI applications" [114]. The memorandum suggests that: "The private sector and other stakeholders may develop voluntary consensus standards that concern AI applications, which provide nonregulatory approaches to manage risks associated with AI applications that are potentially more adaptable to the demands of a rapidly evolving technology."

A later White House report [111] seeks to "ensure that regulations guiding the development and use of AI are supportive of innovation and not overly burdensome." The principles were to "ensure public engagement, limit regulatory overreach, and promote trustworthy AI." However, there is a need to protect the public from biased systems and regulatory capture, in which industry advocates set weak regulations [66].

These cautions leave U.S. industry leaders unsure about how to promote voluntary standards, but other countries, especially in Europe [33], are developing regulatory approaches and specific checklists for developers and managers [34]. The European effort stresses seven principles:

1. human agency and oversight,
2. technical robustness and safety,
3. privacy and data governance,
4. transparency,
5. diversity, non-discrimination and fairness,
6. environmental and societal well-being, and
7. accountability.

Given the many concerns about governmental regulation, non-governmental approaches that are discussed in the following subsections, provide alternative paths to the desired benefits of reliable, safe, and trustworthy HCAI systems.

4.2 Accounting Firms Conduct External Audits

The U.S. Securities and Exchange Commission (SEC) requires publicly traded businesses to have annual internal and external audits, with results posted on the SEC website and published in corporate annual reports. This SEC mandate, which required use of the Generally Accepted Accounting Principles (GAAP), is generally considered to have limited fraud and offered investors more accurate information. However, there were massive failures such as the Enron and MCI WorldCom problems, which led to the Sarbanes–Oxley Act of 2002, known as the Corporate and Auditing Accountability, Responsibility, and Transparency Act, but remember that no system will ever completely prevent malfeasance and fraud. New mandates about reporting on HCAI projects, such as descriptions of fairness and usability test results, could standardize and strengthen reporting methods, to increase investor trust by allowing comparisons across corporations.

Independent financial audit firms, which analyze corporate financial statements to certify that they are accurate, truthful, and complete, could develop reviewing strategies for corporate HCAI projects to provide guidance to investors. They would also make recommendations to their client companies about what improvements to make. These firms often develop close relationships with internal auditing committees, so that there is a good chance that recommendations will be implemented.

Leading independent auditing firms could be encouraged by public pressure or SEC mandate to increase their commitment to support HCAI projects. The big four firms, PricewaterhouseCoopers (https://www.pwc.com/gx/en/issues/data-and-analytics/artificial-intelligence.html), Deloitte (https://www2.deloitte.com/us/en/pages/deloitte-analytics/solutions/analytics-ai.html), Ernst & Young (https://www.ey.com/en_us/ai), and KPMG (https://advisory.kpmg.us/services/data-analytics/artificial-intelligence.html), all claim expertise in AI. The Deloitte website makes a promising statement that "AI tools typically yield little direct outcome until paired with human-centered design," which leans in the directions recommended by this article. Accounting firms have two potential roles, consulting and independent audit, but these roles must be kept strictly separated, as stipulated by the Sarbanes-Oxley Act.

A compelling example of independent oversight of corporate projects is contact tracing for COVID-19. Apple and Google partnered to produce mobile device apps that would alert users if someone they came in contact with developed COVID-19. However, the privacy threats immediately raised concerns, leading to calls for independent oversight boards and policies. One thoughtful proposal offers over 200 items for an independent oversight board of governance to assess and adjudicate during an audit (https://forhumanity.center/contact-tracing-audit). For controversial projects that involve privacy, security, industry competition, or potential bias, independent oversight panels could play a role in increasing public trust.

If the big four auditing firms stepped forward, then their credibility with corporations and general public trust could lead to independent HCAI audits that had substance and impact. A model to build on is the Committee of Sponsoring Organizations (www.coso.org), which brought together five leading accounting organizations to improve enterprise risk management, internal controls, and fraud deterrence. This form of auditing for HCAI could reduce pressures for government regulation and improve business practices. Early efforts could attract attention and enlist trusted public figures and organizations to join such review boards.

4.3 Insurance Companies Compensate for AI Failures

The insurance industry is a potential guarantor of trustworthiness, as it is in the building, manufacturing, and medical domains. Insurance companies could specify requirements for insurability of HCAI systems in manufacturing, medical, transportation, industrial, and other domains. They have long played a key role in ensuring building safety by requiring adherence to building codes for structural strength, fire safety, flood protection, and many other features.

Building codes could be a model for software engineers, as described in Landwehr's proposal for "a building code for building code" [64, 65]. He extends historical analogies to plumbing, fire, or electrical standards by reviewing software engineering for avionics, medical devices, and cybersecurity, but the extension to HCAI systems seems natural.

Builders must satisfy the building codes to gain the inspector's approval, which allows them to gain liability insurance. Software engineers could contribute to detailed software design, testing, and certification standards, which could be used to enable insurance company to conduct risk assessment and develop insurance pricing. Requirements for audit trails of performance and monthly or quarterly reports about failures and near misses would give insurance companies data they need. Actuaries would become skillful in developing risk profiles for different applications and industries, with guidelines for compensation when damage occurs. Liability law from related technologies would have to be interpreted to HCAI systems [15, 83].

A natural next step would be for insurance companies to gather data from multiple companies in each industry they serve, which would accelerate their development of risk metrics and underwriting evaluation methods. This would also support the refinement of building codes for each industry to educate developers and publicly record expected practices. The development of building codes also guides to companies about how to improve their HCAI products and services.

In some industries such as healthcare, travel, and car or home ownership, consumers purchase insurance, which provides no-fault protection to cover damages for any reason. But in some cases, providers purchase insurance to cover the costs of medical malpractice suits, transportation accidents, and building damage from fire, floods, or storms. For many HCAI systems, it seems reasonable that the providers would be the ones to purchase insurance to provide protection for the large numbers of consumers who might be harmed. This will drive up the costs of products and services, but as in many industries, consumers are ready to pay these costs. Insurance companies will have to develop risk assessments for HCAI systems, but as the number of applications grow, sufficient data on failures and near misses will emerge to guide refinements.

Skeptics fear that the insurance companies are more concerned with profits than with protecting public safety and they worry about the difficulty of pursuing a claim when injured by a self-driving car, mistaken medical recommendation, or biased treatment for job hiring, mortgage, or parole. However, as the history of insurance shows, having insurance will benefit many people in their difficult moments of loss. Developing realistic insurance from the damages caused by HCAI systems is a worthy goal.

4.4 Non-governmental and Civil Society Organizations

In addition to government efforts, auditing by accounting firms, and warranties from insurance companies, the U.S. and many other countries have a rich set of non-governmental and civil society organizations that have already been active in promoting reliable, safe, and trustworthy HCAI systems (Appendix A). These examples have various levels of support, but collectively they are likely to do much to promote improved systems and public acceptance.

These non-governmental organizations (NGOs) are often funded by wealthy donors or corporations who believe that an independent organization has greater freedom to explore novel ideas and lead public discussions in rapidly growing fields such as AI. Some of the NGOs were started by individuals who have the passion necessary to draw others in and find sponsors, but the more mature ones may have dozens or hundreds of paid staff members who share their enthusiasm. Some of these NGOs disappear after initial funding and enthusiasm dissipate, while others develop beneficial services or training courses on new technology policy issues that bring in funding and further expand their networks of contacts.

An inspiring example is how the Algorithmic Justice League was able to get large technology companies to improve their facial recognition products to reduce gender and racial bias within a two-year period. Their pressure also was likely to have been influential in the Spring 2020 decisions of leading companies to halt their sales to police agencies in the wake of the intense movement to limit police racial bias.

NGOs have proven to be early leaders in developing new ideas about HCAI principles and ethics, but now they will need to increase their attention to developing new ideas about implementing software engineering practices and business management strategies. They will also have to expand their relationships with government policy makers, liability lawyers, insurance companies, and auditing firms so they can influence the external oversight mechanisms that have long been part of other industries.

However, NGOs have limited authority to intervene. Their role is to point at problems, raise possible solutions, stimulate public discussion, support investigative journalism, and change public attitudes. Then governmental agencies respond with policy guidance to staff, auditing companies change their processes to accommodate HCAI, and insurance companies update their risk assessment as they underwrite new technologies. NGOs could also be influential by conducting independent oversight studies to analyze widely used HCAI systems. Their reports could provide fresh insights and assessment processes tuned to the needs of diverse industries.

4.5 Professional Organizations and Research Institutes

Professional organizations have proven effective in developing voluntary guidelines and standards. Established and new organizations (Appendix B) have been vigorously engaged in international discussions on ethical and practical design principles for responsible AI. They are already influential in producing positive outcomes. However, skeptics caution that industry leaders often dominate professional organizations, so they may push for less restrictive guidelines and standards.

Professional societies, such as the IEEE, have long been effective in supporting international standards, with current efforts on the P7000 series addressing topics such as transparency of autonomous systems, algorithmic bias considerations, fail-safe design for autonomous and semi-autonomous systems, and rating the trustworthiness of news sources (https://ethicsstandards.org/p7000/). The ACM's U.S. Technology Policy Committee has subgroups that address accessibility, AI/Algorithmic accountability, digital governance, and privacy. The challenge for professional societies is to increase the low rates of participation of their members in these efforts.

Academic institutions have long conducted research on AI, but they have now formed large centers to do research and promote interest in ethical, design, and research themes around HCAI. Early efforts have begun to add ethical concerns and policymaking strategies to education, but much more remains to be done so that graduates are more aware of the impact of their work. Example institutions include:

- Stanford University (Human-centered AI (HAI) Institute),
- Harvard University (Berkman Klein Center for Internet and Society),
- University of Oxford (Internet Institute, Future of Humanity Institute),
- University of Cambridge (Leverhulme Centre for the Future of Intelligence),
- Columbia University (Data Science Institute), and
- University of California-Berkeley (Center for Human-compatible AI).

There are also many others research labs and educational programs devoted to understanding the long-term impact of AI and exploring ways to ensure it is beneficial for humanity. The challenge for these organizations is to build on their strength in research by bridging to practice, to promote better software engineering processes, organizational management strategies, and independent oversight methods. University-industry-government partnerships could be a strong pathway for influential actions.

Responsible industry leaders have repeatedly expressed their desire to use artificial intelligence and advanced technologies in safe and effective ways. Microsoft's CEO Satya Nadella [78] proposed six principles for responsible use of advanced technologies. He wrote that artificially intelligent systems must:

- Assist humanity and safeguard human workers.
- Be transparent.... Ethics and design go hand in hand.
- Maximize efficiencies without destroying the dignity of people.
- Be designed for intelligent privacy.
- Have algorithmic accountability so that humans can undo unintended harm.
- Guard against bias.... So that the wrong heuristics cannot be used to discriminate.

Similarly, Google's CEO Sundar Pichai [90] offered seven objectives for artificial intelligence applications:

- Be socially beneficial.
- Avoid creating or reinforcing unfair bias.
- Be built and tested for safety.
- Be accountable to people.
- Incorporate privacy design principles.
- Uphold high standards of scientific excellence.
- Be made available for uses that accord with these principles.

Skeptics will see these statements as self-serving corporate whitewashing, designed to generate positive public responses. However, they can produce important internal efforts such as Google's internal review and algorithmic auditing framework [92] (see Section 3.4). Corporate statements can help raise public expectations, but the diligence of internal commitments should not be a reason to limit external independent oversight. Since support for corporate social responsibilities may be countered by pressures for a profitable bottom line, corporations and the public benefit from comments and questions from knowledgeable journalists and external review boards.

5 LIMITATIONS, CONCLUSIONS, AND FUTURE DIRECTIONS

HCAI systems represent a second Copernican revolution that puts human performance and human experience at the center of design thinking. This article offers 15 recommendations to create reliable, safe, and trustworthy HCAI by enabling designers to translate widely discussed ethical principles into professional practices in large organizations with clear schedules. This article describes three levels of organizational structures: (1) reliable systems based on sound software engineering practices in development teams, (2) safety culture through business management strategies, and (3) trustworthy certification by external reviews from independent oversight organizations (Figure 1).

These diverse concerns mean that drawing researchers and practitioners from diverse disciplines is more likely to lead to success [73, 35]. These HCAI systems will be well received if they go beyond statements about fairness, transparency, accountability, security, and privacy to support specific practices that raise human self-efficacy, encourage creativity, clarify responsibility, and facilitate social participation.

The proposed governance structures face many challenges. No industry will implement all fifteen recommendations in the three levels. Each recommendation requires research and testing to validate effectiveness, and refinement based on the realities of each implementation. Another complexity is that real HCAI systems have many components from diverse providers, which means that some recommendations, such as software, data, and usability testing can be accomplished for a component, but may be more difficult for a complete system. Formal methods and thorough testing may be possible for an airbag deployment algorithm, but independent oversight reviews may be more relevant for a self-driving car system.

Just as each home evolves over time, HCAI systems will be adapted to meet changing desires and demands. Adaptations will integrate new HCAI technologies, the needs of different application domains, and the changing expectations of all stakeholders.

Global interest in HCAI systems is demonstrated by the activities of the United Nations International Telecommunications Union and its 35 UN partner agencies. They seek to apply AI to the 17 influential UN Sustainable Development Goals (https://sustainabledevelopment.un.org/), all of which combine technology developments with behavioral changes, to improve healthcare, wellness, environmental protection, and human rights.

The massive interest in ethical, social, economic, human rights, social justice, and responsible design is a positive sign for those who wish to see HCAI applied for social good. Skeptics fear that poor design will lead to failures, bias, privacy violations, and uncontrolled systems, while malicious actors will misuse AI's powers to spread misinformation, threaten security, and disrupt civil society. These are legitimate concerns, but the intense effort by well-intentioned researchers, business leaders, government policy makers, and civil society organizations suggests more positive outcomes are possible. A second Copernican revolution will take decades to be widely adopted, but putting people at the center can shift thinking to build future societies of which we can all be proud.

APPENDICES

APPENDIX A: NON-GOVERNMENTAL AND CIVIL SOCIETY ORGANIZATIONS WORKING ON HCAI

There are hundreds of organizations in this category, so this brief listing only samples some of the prominent ones.

Underwriters Laboratories, established in 1894, has been "working for a safer world" by "empowering trust". They began with testing and certifying electrical devices and then branched out worldwide to evaluate and develop voluntary industry standards. Their vast international network has been successful in producing better products and services, so it seems natural for them to address HCAI. https://www.ul.com/about/mission.

Brookings Institution, founded in 1916, is a Washington, DC non-profit public policy organization, which is home to an Artificial Intelligence and Emergy Technology Initiative (AIET). They focus on governance issues by publishing reports and books, bringing together policy makers and researchers at conferences, and "seek to bridge the growing divide between industry, civil society, and policymagers." https://www.brookings.edu/project/artificial-intelligence-and-emerging-technology-initiative/.

Electronic Privacy Information Center (EPIC), founded in 1994, is a Washington, DC-based public interest research that that focuses "public attention on emerging privacy and civil liberties issues and to protect privacy, freedom of expression, and democratic values in the information age." They run conferences, offer public education, file amicus briefs, pursue litigation, and testify before Congress and governmental organizations. Their recent work has emphasized AI issues such as surveillance and algorithmic transparency. http://epic.org.

Algorithmic Justice League, which stems from IT and Emory University, seeks to lead "a cultural movement toward equitable and accountable AI". They combine "art and research to illuminate the social implications and harms of AI". With funding from large foundations and individuals that have done influential work on demonstrating bias, especially for face recognition systems. There work productively led to algorithmic and training data improvements in leading corporate systems. https://www.ajlunited.org.

AI Now Institute at New York University "is an interdisciplinary research center dedicated to understanding the social implications of artificial intelligence." This institute emphasizes "four core domains: Rights & Liberties, Labor & Automation, Bias & Inclusion, Safety & Critical Infrastructure." It supports research, symposia, and workshops to educate and examine "the social implications of AI". https://ainowinstitute.org.

Data & Society, an independent New York-based non-profit that "studies the social implications of data-centric technologies & automation.... We produce original research on topics including AI and automation, the impact of technology on labor and health, and online disinformation." https://datasociety.net.

Foundation for Responsible Robotics is a Netherlands-based group whose tag line is: "Accountable innovation for the humans behind the robots". They say their mission is "to shape a future of responsible (AI based) robotics design, development, use, regulation, and implementation. We do this by organizing and hosting events, publishing consultation documents, and through creating public-private collaborations." https://responsiblerobotics.org.

AI4ALL, an Oakland, CA-based nonprofit works "for a future where diverse backgrounds, perspectives, and voices unlock AI's potential to benefit humanity". They sponsor education projects such as summer institutes in the U.S. and Canada for diverse high school and university students, especially women and minorities to promote AI for social good. http://ai-4-all.org.

ForHumanity is a public charity, which examines and analyzes the downside risks associated with AI and automation, such as "their impact on jobs, society, our rights and our freedoms." They believe that independent audit of AI systems, covering trust, ethics, bias, privacy and cybersecurity at the corporate and public-policy levels, is a crucial path to building an infrastructure of trust. They believe that "if we make safe and responsible artificial intelligence & automation profitable whilst making dangerous and irresponsible AI & automation costly, then all of humanity wins" https://www.forhumanity.center/.

Future of Life Institute is a Boston-based charity working on AI, biotech, nuclear, and climate issues. They seek to "catalyze and support research and initiatives for safeguarding life and developing optimistic visions of the future, including positive ways for humanity to steer its own course considering new technologies and challenges." https://futureoflife.org.

APPENDIX B: PROFESSIONAL ORGANIZATIONS AND RESEARCH INSTITUTES WORKING ON HCAI

There are hundreds of organizations in this category, so this brief listing only samples some of the prominent ones. A partial listing is at https://en.wikipedia.org/wiki/Category:Artificial_intelligence_associations.

Institute for Electrical and Electronics Engineers (IEEE) launched a global initiative for ethical considerations in the design of AI and autonomous systems. It's an incubation space for new standards and solutions, certifications and codes of conduct, and consensus building for ethical implementation of intelligent technologies. https://standards.ieee.org/industry-connections/ec/autonomous-systems.html.

IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (2019) originates with the large professional engineering society, collected more than 200 people over 3 years to prepare an influential report: *Ethically Aligned Design: A Vision for Prioritizing Human Wellbeing with Autonomous and Intelligent Systems* (IEEE, 2019) https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html https://ethicsinaction.ieee.org/.

ACM, a professional society with 100,000 members working in the computing field have been active in developing principles and ethical frameworks for responsible computing. ACM's Technical Policy Committee delivered a report with seven principles for algorithmic accountability and transparency (Garfinkel et al., 2017). https://www.acm.org/.

Association for the Advancement of Artificial Intelligence (AAAI) is a "nonprofit scientific society devoted to advancing the scientific understanding of the mechanisms underlying thought and intelligent behavior and their embodiment in machines. AAAI aims to promote research in, and responsible use of, artificial intelligence." They run very successful conferences, symposia, and workshops, often in association with ACM, that bring researchers together to present new work and train newcomers to the field. https://www.aaai.org/.

Robotic Industries Association (RIA), founded in 1974, is a North American trade group that "drives innovation, growth, and safety in manufacturing and service industries through education, promotion, and advancement of robotics, related automation technologies, and companies delivering integrated solutions." https://www.robotics.org.

Machine Intelligence Research Institute (MIRI) is a research nonprofit studying the mathematical underpinnings of intelligent behavior. Their mission is to develop formal tools for the clean design and analysis of general-purpose AI systems, with the intent of making such systems safer and more reliable when they are developed. https://intelligence.org.

Open AI is a San Francisco-based research organization that "will attempt to directly build safe and beneficial Artificial General Intelligence (AGI)… that benefits all of humanity." Their research team is supported by corporate investors, foundations, and private donations. https://openai.com.

The Partnership on AI, established in 2016 by six of the largest technology companies, has more than 100 industry, academic, and other partners who "shape best practices, research, and public dialogue about AI's benefits for people and society." They funded the Partnership on AI, which "conducts research, organizes discussions, shares insights, provides thought leadership, consults with relevant third parties, responds to questions from the public and media, and creates educational material." https://www.partnershiponai.org.

ACKNOWLEDGMENTS

I greatly appreciate supportive comments on early versions of this article from Salvatore Andolina, Michael Bernstein, Dan Berry, Ryan Carrier, Thomas Diettrich, Mengnan Du, Greg Falco, Gerhard Fischer, Simson Garfinkel, Christopher Hart, Jason Hattrick-Simpers, Harry Hochheiser, Robert M. Hoffman, Eric Hughes, Gary Klein, Sean Koon, Meredith Ringel Morris, Lawrence Nadel, Sharon Oviatt, Catherine Plaisant, Philip Resnik, Arnon Rosenthal, Todd Shock, and Wei Xu. Their thoughtful guidance helped shape this article in important ways. Editors Michelle Zhou and Remco Chang were especially helpful in arranging for extensive reviews that guided further changes.

REFERENCES

[1] A. Abdul, J. Vermeulen, D. Wang, B. Y. Lim, and M. Kankanhalli. 2018. Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda. *Proceedings of the CHI Conference on Human Factors in Computing Systems.* ACM, 1–18.

[2] S. Alsheibani, C. Messom, and Y. Cheung. 2019. Towards an artificial intelligence maturity model: From science fiction to business facts. *Proceedings of the 23rd Pacific Asia Conference on Information Systems.* Association for Information Systems. Retrieved from http://www.pacis2019.org/wd/Submissions/PACIS2019_paper_146.pdf.

[3] S. Amershi, A. Begel, C. Bird, R. DeLine, H. Gall, E. Kamar, N. Nagappan, B. Nushi, and T. Zimmermann. 2019a. Software engineering for machine learning: A case study. In *Proceedings of the IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP'19).* IEEE, 291–300.

[4] S. Amershi, D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, and E. Horvitz. 2019b. Guidelines for human-AI interaction. In *Proceedings of the CHI Conference on Human Factors in Computing Systems.* ACM, 1–13.

[5] R. Baeza-Yates. 2018. Bias on the web. *Commun. ACM* 61, 6 (2018), 54–61.

[6] S. K. Bell, P. B. Smulowitz, A. C. Woodward, M. M. Mello, A. M. Duva, R. C. Boothman, and K. Sands. 2012. Disclosure, apology, and offer programs: Stakeholders' views of barriers to and strategies for broad implementation. *Milbank Quart.* 90, 4 (2012), 682–705.

[7] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, and S. Nagar. 2019. AI fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM J. Res. Dev.* 63, 4/5 (2019), 4–1.

[8] J. C. Berry, J. T. Davis, T. Bartman, C. C. Hafer, L. M. Lieb, N. Khan, and R. J. Brilli. 2016. Improved safety culture and teamwork climate are associated with decreases in patient harm and hospital mortality across a hospital system. *J. Patient Safety* (Jan. 2016). Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/26741790.

[9] O. Biran and C. Cotton. 2017. Explanation and justification in machine learning: A survey. In *Proceedings of the International Joint Conference on Artificial Inteeligence Workshop on Explainable AI (XAI'17).*

[10] R. Bostelman, T. Hong, and J. Marvel. 2016. Survey of research for performance measurement of mobile manipulators. *J. Res. Natl. Inst. Standards Technol.* 121, 3 (2016), 342–366.

[11] E. Breck, N. Polyzotis, S. Roy, S. E. Whang, and M. Zinkevich. 2019. Data validation for machine learning. In *Proceedings of the Conference on Systems and Machine Learning (SysML'19).* Retrieved from https://www.sysml.cc/doc/2019/167.pdf.

[12] H. Brown. 2018. Keeping the lights on: A comparison of normal accidents and high reliability organizations. *IEEE Technol. Soc. Mag.* 37, 2 (2018), 62–70.

[13] B. G. Buchanan and E. H. Shortliffe (Eds.). 1985. *Rule-based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project.* Addison-Wesley Publishing Company.

[14] J. Buolamwini and T. Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proc. Mach. Learn. Res.* 81, (2018), 77–91.

[15] Calo Ryan. 2016. Robots in American law. University of Washington School of Law Research Paper. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2737598.

[16] N. Campbell. 2007. The evolution of flight data analysis. *Proceedings of Australian Society of Air Safety Investigators.* Retrieved from https://asasi.org/papers/2007/The_Evolution_of_Flight_Data_Analysis_Neil_Campbell.pdf.

[17] Canadian Government. 2019. Responsible use of artificial intelligence (AI). Retrieved from https://www.canada.ca/en/government/system/digital-government/modern-emerging-technologies/responsible-use-ai.html.

[18] J. V. Carvalho, Á. Rocha, J. Vasconcelos, and A. Abreu. 2019. A health data analytics maturity model for hospitals information systems. *Int. J. Info. Manage.* 46 (2019), 278–285.

[19] R. Challen, J. Denny, M. Pitt, L. Gompels, T. Edwards, and K. Tsaneva-Atanasova. 2019. Artificial intelligence, bias and clinical safety. *BMJ Qual. Safety* 28, 3 (2019), 231–237.

[20] L. Chen, D. Yan, and F. Wang. 2019. User evaluations on sentiment-based recommendation explanations. *ACM Trans. Interact. Intell. Syst.* 9, 4 20 (2019), 38 pages. DOI: https://doi.org/10.1145/3282878

[21] H. F. Cheng, R. Wang, Z. Zhang, F. O'Connell, T. Gray, F. M. Harper, and H. Zhu. 2019. Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders. In *Proceedings of the CHI Conference on Human Factors in Computing Systems ACM*, 1–12.

[22] W. J. Clancey. 1986. From GUIDON to NEOMYCIN and HERACLES in twenty short lessons. *AI Mag.* 7, 3 (1986), 40–40.

[23] J. Couzin-Frankel. 2019. Medicine contends with how to use artificial intelligence. *Science* 354, 6446 (2019), 1119–1120.

[24] P. R. Daugherty and H. J. Wilson. 2018. *Human+ Machine: Reimagining Work in the Age of AI.* Harvard Business Press.

[25] T. G. Dietterich. 2019. Robust artificial intelligence and robust human organizations. *Front. Comput. Sci.* 13, 1–3 https://doi.org/10.1007/s11704-018-8900-4

[26] F. Doshi-Velez and B. Kim. 2017. Towards a rigorous science of interpretable machine learning. *Arxiv Preprint Arxiv*:1702.08608.

[27] F. Du, C. Plaisant, N. Spring, K. Crowley, and B. Shneiderman. 2019. EventAction: A visual analytics approach to explainable recommendation for event sequences. *ACM Trans. Interact. Intell. Syst.* 9, 4 (2019), 1–31.

[28] M. Du, N. Liu, and X. Hu. 2020. Techniques for interpretable machine learning. *Commun. ACM* 63, 1 (2020), 68–77.

[29] J. J. Dudley and P. O. Kristensson. 2018. A review of user interface design for interactive machine learning. *ACM Trans. Interact. Intell. Syst.* 8, 2 (2018), 8.

[30] C. Ebert and M. Weyrich. 2019. Validation of autonomous systems. *IEEE Softw.* 36, 5 (2019), 15–23.

[31] S. Elbaum and J. C. Munson. 2000. Software black box: An alternative mechanism for failure analysis. In *Proceedings of the 11th International Symposium on Software Reliability Engineering (ISSRE'00).* IEEE, 365–376. Retrieved from https://ieeexplore.ieee.org/abstract/document/885887.

[32] S. M. Erickson, J. Wolcott, J. M. Corrigan, and P. Aspden (Eds.). 2004. *Patient Safety: Achieving a New Standard for Care.* National Academies Press, Washington, DC.

[33] European Commission. 2020a. White paper on artificial intelligence—A European approach to excellence and trust, Brussels. Retrieved from https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf.

[34] European Commission. 2020b. The assessment list for trustworthy artificial intelligence (ALTAI) for self-assessment, independent high-level expert group on artificial intelligence, Brussels. Retrieved from https://ec.europa.eu/digital-single-market/en/news/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment.

[35] G. Falco, M. Eling, D. Jablanski, M. Weber, V. Miller, L. A. Gordon, S. S. Wang, J. Schmit, R. Thomas, M. Elvedi, T. Maillart, E. Donovan, S. Dejung, E. Durand, F. Nutter, U. Scheffer, G. Arazi, G. Ohana, and H. Lin. 2019. Cyber risk research impeded by disciplinary barriers. *Science, 366* (6469), 1066–1069.

[36] J. Fjeld, N. Achten, H. Hilligoss, A. Nagy, and M. Srikumar. 2020. Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. *Berkman Klein Center Research Publication.* Retrieved from https://cyber.harvard.edu/publication/2020/principled-ai.

[37] P. Fraser, J. Moultrie, and M. Gregory. 2002. The use of maturity models/grids as a tool in assessing product development capability. In *Proceedings of the IEEE International Engineering Management Conference.* IEEE, 244–249.

[38] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth. 2019. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, ACM, 329–338. https://doi.org/10.1145/3287560.3287589

[39] B. Friedman and H. Nissenbaum. 1996. Bias in computer systems. *ACM Trans. Info. Syst.* 14, 3 (1996), 330–347.

[40] S. Garfinkel, J. Matthews, S. S. Shapiro, and J. M. Smith. 2017. Toward algorithmic transparency and accountability. *Commun. ACM* 60, 9 (2017), 5–5.

[41] B. Goodman and S. Flaxman. 2017. European union regulations on algorithmic decision-making and a "right to explanation." *AI Mag.* 38, 3 (2017), 50–57.

[42] D. R. Grossi. 1999. *Aviation Recorder Overview.* In *Proceedings of the International Symposium on Transportation Recorders.* 153–164.

[43] F. W. Guldenmund. 2000. The nature of safety culture: a review of theory and research. *Safety Sci.* 34, 1-3 (1999), 215–257.

[44] T. K. Haavik, S. Antonsen, R. Rosness, and A. Hale. 2019. HRO and RE: A pragmatic perspective. *Safety Sci.* 117 (2019), 479–489.

[45] J. Heer. 2019. Agency plus automation: Designing artificial intelligence into interactive systems. *Proc. Natl. Acad. Sci. U.S.A.* Retrieved from https://www.pnas.org/content/116/6/1844.

[46] M. Herschel, R. Diestelkämper, and H. B. Lahmar. 2017. A survey on provenance: What for? what form? what from? *VLDB J.* 26, 6 (2017), 881–906.

[47] R. R. Hoffman and G. Klein. 2017. Explaining explanation, part 1: Theoretical foundations. *IEEE Intell. Syst.* 32, 3 (2017), 68–73.

[48] R. R. Hoffman, S. T. Mueller, and G. Klein. 2017. Explaining explanation, part 2: Empirical foundations. *IEEE Intell. Syst.* 32, 4 (2017), 78–86.

[49] F. Hohman, A. Head, R. Caruana, R. DeLine, and S. M. Drucker. 2019. Gamut: A design probe to understand how data scientists understand machine learning models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, 1–13.

[50] F. Hohman, H. Park, C. Robinson, and D. H. P. Chau. 2019. SUMMIT: Scaling deep learning interpretability by visualizing activation and attribution summarizations. *IEEE Trans. Visual. Comput. Graph.* 26, 1 (2019), 1096–1106.

[51] K. Holstein, J. Wortman Vaughan, H. Daumé III, M. Dudik, and H. Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, 1–16. https://doi.org/10.1145/3290605.3300830

[52] S. Hong, J. Hullman, and E. Bertini. 2020. Human factors in model interpretability: Industry practices, challenges, and needs. *Proc. ACM Hum.-Comput. Interact.* 4 (2020), 1–26.

[53] A. Hopkins. 1999. The limits of normal accident theory. *Safety Sci.* 32, 2 (1999), 93–102.

[54] R. Hull, B. Kumar, D. Lieuwen, P. F. Patel-Schneider, A. Sahuguet, S. Varadarajan, and A. Vyas. 2003. Everything personal, not just business: Improving user experience through rule-based service customization. In *Proceedings of the International Conference on Service-Oriented Computing*. Springer, Berlin, 149–164.

[55] W. S. Humphrey. 1988. Characterizing the software process: a maturity framework, *IEEE Softw.* 5, 2 (1988), 73–79. DOI: 10.1109/52.2014

[56] IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (2019). *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*, 1st ed. IEEE. Retrieved from https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html https://ethicsinaction.ieee.org/.

[57] Information Commissioner's Office and Alan Turing Institute (2019). Explaining decisions made with AI. Retrieved from https://ico.org.uk/about-the-ico/ico-and-stakeholder-consultations/ico-and-the-turing-consultation-on-explaining-ai-decisions-guidance/.

[58] P. Kalluri. 2020. Don't ask if artificial intelligence is good or fair, ask how it shifts power. *Nature* 583 (7815), 169.

[59] K. M. Kavi. 2010. Beyond the black box. *IEEE Spectrum* 47, 8 (2010), 46–51.

[60] G. A. Klein. 2017. *Sources of Power: How People Make Decisions*. MIT Press, Cambridge, MA.

[61] T. R. La Porte. 1996. High reliability organizations: Unlikely, demanding and at risk. *J. Conting. Crisis Manage.* 4, 2 (1996), 60–71.

[62] T. C. Lacerda and C. G. von Wangenheim. 2018. Systematic literature review of usability capability/maturity models. *Comput. Standards Interfaces* 55, 95–105.

[63] P. Landon, P. Weaver, and J. P. Fitch. 2016. Tracking minor and near-miss events and sharing lessons learned as a way to prevent accidents. *Appl. Biosafety* 21, 2 (2016), 61–65.

[64] C. E. Landwehr. 2013. A building code for building code: Putting what we know works to work. In *Proceedings of the 29th Annual Computer Security Applications Conference (ACSAC'13)*. 139—147. Retrieved from http://www.landwehr.org/2013-12-cl-acsac-essay-bc.pdf.

[65] C. Landwehr. 2015. We need a building code for building code. *Commun. ACM* 58, 2 (2015), 24–26.

[66] N. T. Lee, P. Resnick, and G. Barton. 2019. Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms. Center for Technology Innovation, Brookings. Retrieved from https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/.

[67] B. Letham, C. Rudin, T. H. McCormick, and D. Madigan. 2015. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *Ann. Appl. Stat.* 9, 3 (2015), 1350–1371.

[68] N. Leveson. 2011. *Engineering a Safer World: Systems Thinking Applied to Safety*. MIT Press, Cambridge, MA.

[69] F.-F. Li. 2018. How to make A.I. that's good for people. *The New York Times* (Mar. 7, 2018). Retrieved from https://www.nytimes.com/2018/03/07/opinion/artificial-intelligence-human.html.

[70] X. Liang, S. Shetty, D. Tosh, C. Kamhoua, K. Kwiat, and L. Njilla. 2017. Provchain: A blockchain-based data provenance architecture in cloud environment with enhanced privacy and availability. In *Proceedings of the 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID'17)*, 468–477.

[71] Q. V. Liao, D. Gruen, and S. Miller. 2020. Questioning the AI: Informing design practices for explainable ai user experiences. *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*. ACM, New York, 1–15.

[72] T. Mai, R. Khanna, J. Dodge, J. Irvine, K. H. Lam, Z. Lin, N. Kiddle, E. Newman, S. Raja, C. Matthews, C. Perdriau, M. Burnett, and A. Fern. 2020. Keeping it "organized and logical": After-action review for AI (AAR/AI). In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. ACM, 465–476.

[73] T. Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artific. Intell. 267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

[74] A. Mitrevski, S. Thoduka, A. O. Sáinz, M. Schöbel, P. Nagel, P. G. Plöger, and E. Prassler. 2018. Deploying robots in everyday environments: Toward dependable and practical robotic systems. In *Proceedings of the 29th International Workshop Principles of Diagnosis (DX'18).* Retrieved from http://www.ropod.org/downloads/dx18.pdf.

[75] B. Mittelstadt, C. Russell, and S. Wachter. 2019. Explaining explanations in AI. In *Proceedings of the Conference on Fairness, Accountability, and Transparency.* ACM, 279–288. https://doi.org/10.1145/3287560.3287574

[76] M. Modarres, M. P. Kaminskiy, and V. Krivtsov. 2016. *Reliability Engineering and Risk Analysis: A Practical Guide.* CRC Press.

[77] M. R. Morris. 2020. AI and accessibility: A discussion of ethical considerations. *Commun. ACM* 63, 6 (2016), ACM 35–37. 10.1145/3356727

[78] Nadella Satya (2016). The partnership of the future, *Slate.* Retrieved from https://slate.com/technology/2016/06/microsoft-ceo-satya-nadella-humans-and-a-i-can-work-together-to-solve-societys-challenges.html.

[79] J. Nicas, N. Kitroeff, D. Gelles, and J. Glanz. 2019. Boeing built deadly assumptions into 737 max, blind to a late design change. *The New York Times* (2019). Retrieved from https://www.nytimes.com/2019/06/01/business/boeing-737-max-crash.html.

[80] S. Nourashrafeddin, E. Sherkat, R. Minghim, and E. E. Milios. 2018. A visual approach for interactive keyterm-based clustering. *ACM Trans. Interact. Intell. Syst.* 8, 1 (2018), 1–35.

[81] C. O'Neil. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* Crown Publishers, New York.

[82] F. Pasquale. 2015. *The Black Box Society: The Secret Algorithms that Control Money and Information.* Harvard University Press, Cambridge, MA.

[83] F. Pasquale. 2017. Toward a fourth law of robotics: Preserving attribution, responsibility, and explainability in an algorithmic society. *Ohio State Law J.* 78 (2017), 1243–1255.

[84] F. Pasquale. 2018. When machine learning is facially invalid. *Commun. ACM* 61, 9 (2018), 25–27. DOI: 10.1145/3241367

[85] M. C. Paulk, B. Curtis, M. B. Chrissis, and C. V. Weber. 1993. Capability maturity model, version 1.1. *IEEE Softw.* 10, 4 (1993), 18–27.

[86] A. Pérez, M. I. García, M. Nieto, J. L. Pedraza, S. Rodríguez, and J. Zamorano. 2010. Argos: An advanced in-vehicle data recorder on a massively sensorized vehicle for car driver behavior experimentation. *IEEE Trans. Intell. Transport. Syst.* 11, 2 (2010), 463–473.

[87] C. C. Perez. 2019. *Invisible Women: Exposing Data Bias in a World Designed for Men.* Random House.

[88] C. Perrow. 2011. *Normal Accidents: Living with High Risk Technologies—Updated edition.* Princeton University Press.

[89] O. Pettersson. 2005. Execution monitoring in robotics: A survey. *Robot. Auton. Syst.* 53, 2 (2005), 73–88.

[90] Pichai Sundar (2018). AI at Google: Our principles. Retrieved from https://www.blog.google/technology/ai/ai-principles/.

[91] E. D. Ragan, A. Endert, J. Sanyal, and J. Chen. 2015. Characterizing provenance in visualization and data analysis: An organizational framework of provenance types and purposes. *IEEE Trans. Visual. Comput. Graph.* 22, 1 (2015), 31–40.

[92] I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, and P. Barnes. 2020. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT'20).* ACM, 33–44. https://doi.org/10.1145/3351095.3372873

[93] S. Reddy, S. Allan, S. Coghlan, and P. Cooper. 2020. A governance model for the application of AI in health care. *J. Amer. Med. Info. Assoc.* 27, 3 (2020), 491–497.

[94] D. Reisman, J. Schultz, K. Crawford, and M. Whittaker. 2018. Algorithmic impact assessments: A practical framework for public agency accountability. AI Now Institute, 1–22. Retrieved from https://ainowinstitute.org/aiareport2018.pdf.

[95] F. Rosenberg and S. Dustdar. 2005. Design and implementation of a service-oriented business rules broker. In *Proceedings of the 7th IEEE International Conference on E-Commerce Technology Workshops.* IEEE, 55–63.

[96] M. Rotenberg. 2020. *The AI Policy Sourcebook 2020.* Electronic Privacy Information Center, Washington, DC.

[97] C. Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Mach. Intell.* 1, 5 (2019), 206–215.

[98] F. Santoni de Sio and J. Van den Hoven. 2018. Meaningful human control over autonomous systems: A philosophical account. *Front. Robot. AI 5*, 15.

[99] J. J. Seddon and W. L. Currie. 2017. A model for unpacking big data analytics in high-frequency trading. *J. Bus. Res.* 70 (2017), 300–307.

[100] H. Sharp, J. Preece, and Y. Rogers. 2019. *Interaction Design: Beyond Human-Computer Interaction*, 5th ed. Wiley Publishers.

[101] B. Shneiderman. 2016. Opinion: The dangers of faulty, biased, or malicious algorithms requires independent oversight. *Proc. Natl. Acad. Sci. U.S.A.* 113, 48 (2016), 13538–13540. Retrieved from http://www.pnas.org/content/113/48/13538.full.

[102] Ben Shneiderman. 2020a. Human-centered artificial intelligence: Reliable, safe, & trustworthy. *Int. J. Hum.-Comput. Interact.* 36, 6 (2020a), 495–504. https://doi.org/10.1080/10447318.2020.1741118

[103] Ben Shneiderman. 2020b. Design lessons from ai's two grand goals: Human emulation and useful applications. *IEEE Trans. Technol. Soc. 1, 2 (Early Access).* Retrieved from https://ieeexplore.ieee.org/document/9088114.

[104] B. Shneiderman, C. Plaisant, M. Cohen, S. Jacobs, and N. Elmqvist. 2016. *Designing the User Interface: Strategies for Effective Human-Computer Interaction*, 6th ed. Pearson.

[105] G. Siegel. 2014. *Forensic Media: Reconstructing Accidents in Accelerated Modernity.* Duke University Press.

[106] A. Theodorou, R. H. Wortham, and J. J. Bryson. 2017. Designing and implementing transparency for real time inspection of autonomous robots. *Connect. Sci.* 29, 3 (2017), 230–241. Retrieved from https://doi.org/10.1080/09540091.2017.1310182.

[107] U. S. National Research Council. 2008. *Protecting Individual Privacy in the Struggle Against Terrorists: A Framework for Program Assessment.* National Academies Press, Washington, DC. Retrieved from http://www.nap.edu/catalog.php?record_id=12452.

[108] U. S. National Science and Technology Council (2019). The National Artificial Intelligence Research and Development Strategic Plan: 2019 Update. *Executive Office of the President.* Retrieved from https://www.nitrd.gov/pubs/National-AI-RD-Strategy-2019.pdf.

[109] U. S. National Security Commission on Artificial Intelligence(2019). Interim Report. Retrieved from https://epic.org/foia/epic-v-ai-commission/AI-Commission-Interim-Report-Nov-2019.pdf.

[110] U. S. National Transportation Safety Board (2017). Collision between a car operating with automated vehicle control systems and a tractor-semitrailer truck near Williston, Florida, May 7, 2016, Report HAR1702. Retrieved from https://dms.ntsb.gov/public/59500-59999/59989/609449.pdf.

[111] U. S. White House (2020). American artificial Intelligence Initiative: Year one annual report. Office of Science and Technology Policy. Retrieved from https://www.whitehouse.gov/wp-content/uploads/2020/02/American-AI-Initiative-One-Year-Annual-Report.pdf.

[112] G. R. Vishnia and G. W. Peters. 2020. AuditChain: A trading audit platform over blockchain. *Front. Blockchain* 3, 9.

[113] C. G. von Wangenheim, J. C. R. Hauck, A. Zoucas, C. F. Salviano, F. McCaffery, and F. Shull. 2010. Creating software process capability/maturity models. *IEEE Softw.* 27, 4 (2010), 92–94.

[114] R. T. Vought. 2019. Guidance for regulation of artificial intelligence applications. February 11, 2019, U.S. White House Announcement, Washington, DC. Retrieved from https://www.whitehouse.gov/wp-content/uploads/2020/01/Draft-OMB-Memo-on-Regulation-of-AI-1-7-19.pdf.

[115] S. Wachter, B. Mittelstadt, and C. Russell. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard J. Law Technol.* 31 (2017), 841–887.

[116] D. Wang, Q. Yang, A. Abdul, and B. Y. Lim. 2019. Designing theory-driven user-centric explainable AI. In *Proceedings of the CHI Conference on Human Factors in Computing Systems.* ACM, 1–15. https://doi.org/10.1145/3290605.3300831

[117] K. E. Weick, K. M. Sutcliffe, and D. Obstfeld. 1999. Organizing for high reliability: Processes of collective mindfulness. In *Research in Organizational Behavior*, Vol. 1, R.S. Sutton and B.M. Staw (Eds.). JAI Press, Stanford, Chap. 44, 81–123.

[118] D. S. Weld and G. Bansal. 2019. The challenge of crafting intelligible intelligence. *Commun. ACM* 62, 6 (2019), 70–79.

[119] J. Wenskovitch, M. X. Zhou, C. Collins, R. Chang, M. Dowling, A. Endert, and K. Xu. 2020. Putting the "I" in interaction: Interactive interfaces personalized to individuals. *IEEE Comput. Graph. Appl.* 40, 3 (2020), 73–82.

[120] D. M. West and J. R. Allen. 2020. *Turning Point: Policymaking in the Era of Artificial Intelligence.* Brookings Institution Press, Washington, DC.

[121] A. F. Winfield and M. Jirotka. 2017. The case for an ethical black box. In *Proceedings of the Annual Conference Towards Autonomous Robotic Systems*, Springer 262–273. Retrieved from https://link.springer.com/chapter/10.1007/978-3-319-64107-2_21.

[122] A. F. Winfield and M. Jirotka. 2018. Ethical governance is essential to building trust in robotics and artificial intelligence systems. *Philos. Trans. Roy. Soc. A: Math. Phys. Eng. Sci.* 376 (2018), 0085.

[123] D. D. Woods. 2017. Essential characteristics of resilience. In *Resilience Engineering: Concepts and Precepts*, E. Hollnagel, D. W. Woods, and N. Leveson (Eds.). Ashgate Publishing, 21–34.

[124] W. Xu. 2019. Toward human-centered AI: A perspective from human-computer interaction. *ACM Interact.* 26, 4 (2019), 42–46. doi.org/10.1145/3328485

[125] Y. Yao and E. Atkins. 2020. The smart black box: A value-driven high-bandwidth automotive event data recorder. *IEEE Trans. Intell. Transport. Syst.* Retrieved from https://ieeexplore.ieee.org/abstract/document/8995510/. DOI: 10.1109/TITS.2020.2971385

[126] J. M. Zhang, M. Harman, L. Ma, and Y. Liu. 2020. Machine learning testing: Survey, landscapes and horizons. *IEEE Trans. Softw. Eng.* Retrieved from https://ieeexplore.ieee.org/document/9000651. DOI: 10.1109/TSE.2019.2962027

[127] M. X. Zhou, G. Mark, J. Li, and H. Yang. 2019. Trusting virtual agents: The effect of personality. *ACM Trans. Interact. Intell. Syst.* 9 (2–3) (2019), 1–36.

Received May 2019; accepted August 2020