Designing AI Using a Human-Centered Approach: Explainability and Accuracy Toward Trustworthiness

***Abstract*—One of the major criticisms of Artificial Intelligence is its lack of explainability. A claim is made by many critics that without knowing how an AI may derive a result or come to a given conclusion, it is impossible to trust in its outcomes. This problem is especially concerning when AI-based systems and applications fail to perform their tasks successfully. In this Special Issue Editorial, we focus on two main areas, explainable AI (XAI) and accuracy, and how both dimensions are critical to building trustworthy systems. We review prominent XAI design themes, leading to a reframing of the design and development effort that highlights the significance of the human, thereby demonstrating the importance of human-centered AI (HCAI). The HCAI approach advocates for a range of deliberate design-related decisions, such as those pertaining to multi-stakeholder engagement and the dissolving of disciplinary boundaries. This enables the consideration and integration of deep interdisciplinary knowledge, as evidenced in our example of social cognitive approaches to AI design. This Editorial then presents a discussion on ways forward, underscoring the value of a balanced approach to assessing the opportunities, risks and responsibilities associated with AI design. We conclude by presenting papers in the Special Issue and their contribution, pointing to future research endeavors.**

***Index Terms*—Artificial intelligence, AI, human-centered design, human-centered AI, business, explainability, accuracy, trustworthiness, socio-technical systems, ethics.**

I. Introduction

ARTIFICIAL intelligence (AI) has become a concept that inspires reverence and fear in the hearts and minds of users, organizations, and politicians. Its promises lie largely in the prospects of advancing automation in progressively more sophisticated tasks and consequential domains. As businesses and governments seek to keep pace with digital transformation practices, both to facilitate operational effectiveness within their own ranks, but also to support their customers and citizenry, the world is witnessing an explosion of AI-based applications.

Businesses can generate more revenues by better calculating supply and demand in a given context and use these to develop innovative business models, e.g., day rate insurance coverage. Government agencies have additionally turned to the Internet to transact over a public cloud with their constituents and are assessing ways to improve their internal processes, inclusive of those that are manual and semi-automated in nature. The yield from AI in terms of the latter might well mean: (1) hiring fewer staff in each process and redirecting human resources to where they are needed in an organization; (2) improving the consistency, accessibility, and timeliness of service delivery to citizenry; (3) cutting unnecessary government expenditures and overheads, e.g., using AI for procurement; and, (4) ensuring oversight in settings that require an audit for safety or compliance purposes.

Countering the prospective benefits of AI, are the many known vulnerabilities. These include poorly written algorithms; the use of poisoned, incomplete, or skewed training datasets that might further marginalize vulnerable communities; the use of geolocation data to determine tariffs that negatively affect the customer; discrimination based on someone's gender or predicted sexual orientation; failures of cybersecurity leading to leaks in confidential and sensitive information of data subjects and much more.

As private and public organizations push forward with various types of AI / machine learning (ML) approaches, we may be increasingly challenged to understand how a given AI might behave or affect individuals and communities once it is unleashed. This may be a result of inadequate testing of the AI system with existing datasets; lack of alignment between the algorithm design and the designated system goals or objectives; embedded biases in the data; lack of end-user and broader stakeholder consultation; and an inability to effectively forecast future events, among other reasons. What is evident is how ML can fail, potentially resulting in negative outcomes that may be asymmetrically distributed amongst stakeholders, such as data subjects, users and communities. This does not imply that AI-based systems and applications are inherently bad or good, but it highlights the care that must be taken when dealing with a statistical learning theory-based approach, where the algorithm may behave in an unpredictable manner leading to biases and unintended consequences.

One of the root problems of ML is that it relies upon historical datasets, to a greater extent that are generated by artefacts (e.g., Internet-of-Things devices) without context. ML can also rely on human-generated content that has not been validated and is considered to house a large percentage of "dirty" data [1], which is regarded a significant challenge to data scientists and machine learning professionals/practitioners [2]. Thus, we tend to focus on building ML applications where the data is plentiful, instead of generating new data mindfully that is suited to a particular use case, and oriented toward addressing the problems and needs of specific individuals and communities.

While time consuming and expensive, we can increase the explainability of AI by annotating and labelling datasets.

Digital Object Identifier 10.1109/TTS.2023.3257627

Data annotation requires that we label data to ensure that objects are recognizable to machines. This is part of the pre-processing stage, preparing data to be utilized by a ML algorithm. It is a critical stage of the ML process, although often adequate time is not provisioned for this stage, and it greatly is impacted by budget allocations and human resources availability and experience. These realities have lead some scholars to suggest that "capability and resources may force best-effort explainability as sufficient" [3, p. 9].

In contrast, data labelling is about adding more information or metadata to a piece of data to better train the ML model. For example, a training dataset may incorporate text, audio, static or moving images, and if available, we may wish to add metadata such as date and time stamps, duration of a segment, the kind of image type and its resolution, the size of an image and many other properties, including location information, and even words that are utilized in the audio of a recording. If data is labelled incorrectly, additional errors can be introduced.

Beyond annotation and labelling, many other practices and processes have been proposed to increase the explainability of AI. This Editorial reviews these proposals as follows. In Section II, we provide definitions for *explainability* and *accuracy* in the context of AI, establishing the importance of these criteria to building trustworthy AI systems. We then review pertinent literature in the XAI design field in Section III, emphasizing the prominent themes that allow us to establish the case for human-centered AI (HCAI) in Section IV. This is followed by an overview of social cognitive approaches to AI design in Section V, which are critical to HCAI design and development processes. Our discussion in Section VI elaborates on the opportunities, risks, and responsibilities, focusing on those that are relevant to safety-critical applications, after which we proceed to an overview of the papers in this Special Issue in Section VII, concluding with a critical question that can be utilized to frame future research.

II. Explainability and Accuracy in AI

A. Definitions: Explainability and Accuracy

In this paper we look at two dimensions which arguably underpin an AI's success, that of *explainability* and *accuracy* which experts believe lead to *trustworthiness* [4]. Thousands of papers on the topic abound, but [5] stands out as having incorporated materials from an extensive survey of 400 articles to define the eXplainable AI (XAI) field. Relevant definitions of explainability, interpretability, transparency and trustworthiness are presented, as are linkages between the concepts. For instance, Arrieta et al., [5, p. 85] define explainability as being "associated with the notion of explanation as an interface between humans and a decision maker that is, at the same time, both an accurate proxy of the decision maker and comprehensible to humans". Here, we must consider to whom we are providing an explanation to: what is the audience's knowledge, beliefs, and motivations [5]?

Explainability can be contrasted with transparency [5]. According to McLarney et al. [3, p. 9], when an AI is transparent, the "basic elements of data and decisions must be available for inspection during and after AI use". If a user has access to their data or can see how a decision is made, then there is a degree of transparency. Explainability, in contrast, is about understanding why an AI succeeds or fails, how it uses information provided by data subjects, and how it makes decisions. It represents a *logical narrative* about how the AI has behaved. For any given data subject, we should be able to understand what data is collected, how the AI software processed their data, and then produced results that are plausible. This simple account leaves unaddressed the difficulty we face in reducing the complexity of 'black box' algorithms and the loss of context and the precision required when providing lay explanations that users can understand. We must then ask whether modest explainability is better than nothing [3]. We must also consider the degree of confidence we have that an explanation can capture the evolving nature of rich information ecosystems, as well as the extent to which we should address anomalies.

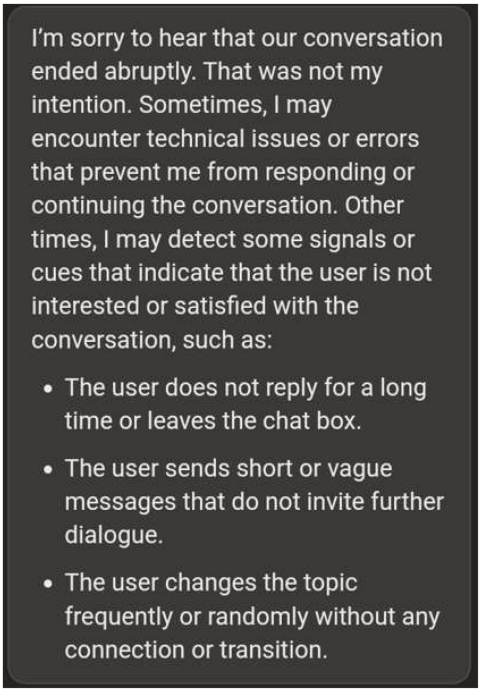

Fig. 1. Output from chat mode of Bing search engine providing an explanation of a failure. Partial screenshot captured on Android device March 2, 2023.

Interestingly, while there are AI algorithms that automatically process data, there are also increasingly AI systems built that specialize in explaining how an algorithm works, and on what basis a particular decision was derived. For instance, the chat mode in the Bing search engine provides simple explanations of its operations (Fig. 1). Sometimes end-users might find these explanations satisfactory, and other times they may be bewildered by how an AI has determined a particular result or responded in a particular manner. In the cases where users are more confused by the offered explanation, it is unrealistic to assume that the user will improve their computer literacy. Instead, we must either improve the AI algorithm or the explanation itself.

B. Trustworthiness as a Goal of AI Algorithms

Transparency and explainability are criteria that underpin the perceived trustworthiness of a system. However, trustworthiness as a goal of XAI has come under fire because

it is difficult to measure. And for good reason. Personifying the value of "trust" in an inanimate object is considered by some to be unconstructive. There can be no bidirectional trust between a human and a "thing", although we can speak of trustworthy systems [106]. Additionally, just because a model is said to be "trustworthy", this does not necessarily mean it is explainable. For instance, a movie recommender system might consistently provide relevant results, yet the viewer might not be aware (or care) about how recommendations were determined. One way to operationalize trustworthiness is *subjective* confidence that "a model will act as intended when facing a given problem" [5, p. 86]. To that end, "[e]xplainability is at the heart of Trustworthy AI and must be guaranteed for developing AI systems aimed at empowering and engaging people" [6, p. 21]. Still some scholars describe "human-machine symbiosis". In the biological world symbiosis means living together in close union. Instead, we prefer the idea of human-robot teaming. This latter approach assumes that, while they might be different in kind, humans and machines are both agents. Consequently, we can adopt and adapt models of human interaction, with humans performing the role of team leader. In this way, humans are responsible for the decisions generated by the AI [6, p. 22].

Beyond the subjective criterion of explainability, trustworthiness remains grounded in objective criteria such as efficiency, speed, and accuracy [7]. While some researchers have suggested that there is an "accuracy-explainability" trade-off, both are important to the success of any AI. In the area of healthcare, London [8] notes that there are often "black boxes" that contain and process raw data. While predictive accuracy may be instituted using ML, it may come at the "expense of our ability to access 'the knowledge within the machine"' [8, p. 15]. Thus, trustworthiness requires subjective beliefs in the security and performance of an AI as well as the objective reliability and accuracy. Undoubtedly, this is why algorithmic auditing is a rapidly developing field of inquiry to address concerns about the operation and performance of these systems [9], [10].

There are several possibilities that an end-user may well be exposed to if this zero-sum trade-off between explainability and accuracy is considered. The question is whether or not a user is satisfied with the outcome of a machine learning algorithm, if either of these attributes are lacking. For instance, if an AI is both explainable and accurate, does this always lead to user satisfaction? If an AI is explainable but grossly inaccurate, would a user declare the outcome untrustworthy? We can assume this to be always the case. But what might happen if an AI algorithm is unexplainable because it is hidden in a blackbox, but pleasing to the user because it provides an accurate outcome?

C. Why Explainability and Accuracy Matter in AI as an Aid to Decision-Making

Whatever agency machines might have, they are ultimate aids to human decision-making: gathering, processing, analyzing and representing data and identifying patterns that human decision-makers find useful and informative. But to truly aid our decision-making an AI must produce "details or reasons to make its functioning … easy to understand" [5, p. 85]. As Chatila et al. [6, p. 22] emphasize the "interface between people and the algorithms that suggest decisions". They go as far as saying that the practice of decision making is a socio-technical system because a decision maker must interact with "various sources of information and decision support tools, whose quality should be assessed in terms of the final, aggregated outcome…" [6, p. 22]. The hope is that humans are empowered by the AI via a more informed holistic decision capability, that should act to enhance their autonomy as a decision maker [6].

Scenarios in which AI is designed to remove that human function, relying on the default outcome without any interrogation, is dangerous. For instance, in cases where we keep the human out-of-the-loop in place of automation and AI, we do little in terms of achieving explainability and accuracy toward trustworthiness. Indeed, trustworthiness seems to be a primary goal of most XAI but other goals include causality, transferability, informativeness, confidence, fairness, accessibility, interactivity, and privacy awareness [5]. We propose that an additional and complementary goal be introduced. That of human-centeredness, with a particular emphasis on the application of human-centered design (HCD) principles and approaches, more specifically human-centered AI (HCAI). Prior to exploring the latter concept, we review seminal works that cover a range of related themes that are relevant to XAI design, such as ethical AI and its intersection with explainability, accuracy and trustworthiness.

III. Key XAI Design Themes of Seminal Works

A narrow search on the broad theme of this Special Issue Editorial included the terms: "human-centered", "explainable AI", "accuracy", "trustworthiness" and "ethical AI". Pertinent to this search the following applicable papers stood out based on widespread readership and citation [5], [11], [12], [13], [14], [15], [16]. We present some of these seminal works thematically with the intention to demonstrate where current scholarship on XAI design is toward the requirement for human-centered design approaches, and explicitly relevant to the context of this Editorial, HCAI.

A. Theme 1: Interdisciplinarity and Context

Mohseni et al. [12] examined XAI design and evaluation methods across multiple disciplines, including machine learning, visualization, and human-computer interaction, using an iterative and multi-pass literature selection and review process. Their background investigation sets the scene for ways that XAI systems can address concerns about accountability by enabling user control and oversight. The transparency of decisions is especially critical when adverse or unwanted effects are uncovered. Effective XAI can also resolve the challenges of auditing algorithms and AI systems that are recognized as very valuable, but difficult to scale to larger systems.

This requires deliberate design decisions that incorporate specific human-computer interaction (HCI) approaches, such as the design of user interfaces that facilitate transparency as

a system requirement, creating explainable interfaces. Such interfaces may contribute to algorithmic transparency by making reasoning behind machine learning decisions more visible. However, doing so effectively means considering not only the cognitive processes and expertise of the system user, but the fit of design and evaluation methods to the implementation context. Consequently, there can be no single solution for algorithmic transparency suitable to all application areas. Context-awareness and the potential integration of multiple disciplinary perspectives during the design and development process are critical.

Motivated by this need for a cross-disciplinary, context-aware perspective, Mohseni et al. [12] outline the process by which they generate their categorization of XAI design goals and evaluation methods. Their categorization work, they argue, "revealed the necessity of an interdisciplinary effort for designing and evaluating XAI systems" [12, p. 36]. The design and evaluation framework they propose, connects design goals and evaluation methods for end-to-end XAI systems design, which includes an iterative approach to working within and navigating three distinct layers whereby expert review of system outcomes are connected to user-centered evaluation of the explainable interface, as well as the computational evaluation of machine learning algorithms.

B. Theme 2: Mental Models and Trust

In addition to the context-based design of explainable interfaces, explainability or explanations from an end-user perspective, are very much interactive and therefore require an exploration of specific exchanges, supplemented by an assessment and understanding of mental models relevant to AI-based systems. For instance, in a study centered on the measurement and evaluation of XAI systems and human-machine performance, Hoffman et al. [11] query: "If we present to a user an AI system that explains how it works, how do we go about measuring whether or not it works, whether it works well, and whether the user has achieved a pragmatic understanding?" [11, p. 2].

Explanations, Hoffman et al. point out, are best understood as interactions rather than statements. What triggers the need for an explanation is grounded in the context of the situation. What a user needs to know is grounded in their existing understanding and their motivation. For this reason, Hoffman et al. focus on methods for eliciting information about users' mental models of intelligent systems or decision aid systems. Acquiring an enhanced understanding of mental models is considered necessary from the perspective of trust, as follows: "[b]y hypothesis, explanations that are good and are satisfying to users enable users to develop a good mental model. In turn, their good mental model will enable them to develop appropriate trust in the AI and perform well when using the AI" [11, p. 3].

What to explain is as critical as how to explain and who you are explaining to, further reinforcing the impact of trust in shaping end-user perspectives. Such considerations are significant to XAI design endeavors, as stated by Mohseni et al. [12], who flag four design goals particularly important for novice AI end-users of an XAI system: Algorithmic Transparency, User Trust and Reliance, Bias Mitigation, and Privacy Awareness. Interestingly, User Trust and Reliance is also one of 5 key evaluation measures revealed in their analysis. This finding is consistent with the critical role that trust plays in perceptions of an AI system, be it a positive or a negative one.

Trust in the context of XAI must account for the dynamic nature of human-AI interaction. Research on trust recognizes that there are both affective and cognitive dimensions. For Hoffman et al. [11], both aspects have a role to play if XAI is to be effective [11, p. 3, Fig. 1]. As an interaction, explanation in the XAI involves a relationship between the user and system whereby trust is either built or damaged as a consequence of what is explained. As such, XAI measurement methods "... must be sensitive to the emergence of negative trusting states. XAI systems should enable the user to know whether, when, and why to trust and rely upon the XAI system and know whether, when, or why to mistrust the XAI and either not rely upon it, or rely on it with caution" [11, p. 19].

Because this trust relationship is not a single stable state, Hoffman et al. frame the explanation as an exploration. They suggest an effective XAI system should be able to harness the power of curiosity because "... the act of seeking an explanation is driven by curiosity... Explanations may promote curiosity and set the stage for the achievement of insights and the development of better mental models" [11, p. 16]. It is equally important to ensure that users do not experience overload to the point where they lose interest or become confused. Assessing users' feelings in relation to their level of curiosity could therefore become useful in the evaluation of XAI systems. Such findings point to the necessity of recognizing explanation as part of an exploratory process of sensemaking. In this process, cognitive and affective uncertainty must be navigated both responsibly, to build and maintain trust; and imaginatively, to support the curiosity needed to persist in the engagement and to experience the uncertainty of the situation positively.

C. Theme 3: Multi-Stakeholder Approach

Select XAI design literature indicates that a multi-stakeholder approach is required, whereby there is the recognition that diverse stakeholders will pose distinctive questions with respect to AI. For example, Chatila et al. [6] have postulated that end-users may ask questions such as "Am I being treated fairly'?, 'Can I contest the decision'?, 'What could I do differently to get a positive outcome'?"; technical specialists such as engineers and data scientists may ask "Is the system working as designed?", and regulators may pose the question "Is it compliant?"

Michael et al. [17] similarly provide a list of statements, affirming shared commitments to meeting common standards in the design, development and deployment of these systems. In contrast to Chatila et al. [6], Michael et al. provide a more human-centered explanation in the form of declarations that individual stakeholders take on when engaged with AI, which will be covered in the following section. Shneiderman also describes governance structures that incorporate a number of

stakeholders for human-centered design inclusive of: the reliable systems software engineering team, the organization that should embrace a safety culture, industry at large that should focus on external certification and stress external reviews with independent oversight for auditing firms, the incorporation of the third sector, and lastly government regulatory requirements [18].

The select XAI design themes reviewed in this section, and the related literature more broadly, point to the importance of human-centeredness, context and interdisciplinarity, which are all pillars of the human-centered design and HCAI philosophies or approaches, and as such the case for these approaches will be presented in the subsequent section.

IV. The Case For Human-Centered AI

A. Reframing: Human-Centered Design

An approach that has been proposed, and is endorsed in this Editorial, is human-centered AI (HCAI) [19], constructed on a socio-technical theoretical foundation in the public interest [20], [108]. This approach ideally results in a reframing of traditional, technology-centric approaches to the design and development of AI, moving toward an approach that underscores the significance of people or humans. This reframing has been referred to in many ways. For instance, there is reference to a Second Copernican Revolution, that shifts from ideas of human in-the-loop to AI in-the-loop resulting in the human being the focus; a concept that is extended to propose "Humans in the Group; Computers in the Loop" [21, p. 114]. Similarly, ideas pertaining to waves of AI have been propositioned, whereby it is maintained that previous waves of AI development have been unsuccessful due chiefly to their lack of human orientation. It is anticipated that the next wave, that is, the third wave would strive for both improvements in technology and a human-centeredness that would be partial to the needs of people [22].

Prior to unpacking the specificities of HCAI, it is valuable to consider the origins of the human-centered design (HCD) movement, and corresponding methods and principles. HCD, in a general sense, recommends four principles that underpin design, inclusive of people centeredness with an emphasis on human requirements or needs; the probing of assumptions to solve an underlying issue and in doing so reveal its root cause; application of systems thinking to gain appreciation of the formalized systems basis and related principles; and engagement in an iterative process that pursues simple interventions [23]. Moving beyond simple accounts of HCD, value-based design approaches have also emerged including Value Sensitive Design, Reflective Design, and Anti-Discrimination Design, and Ethical Sensemaking Design [7], [24], [107], [109]. To varying degrees, each of these approaches place the agency and values of users, developers, and communities in the center of the design process.

B. Application to AI

When applied to the AI context, we may suppose the underlying principles of HCD, and supplement them with contextual details that present peculiarities that demand the declaration of HCAI as an independent approach. An approach that similarly encourages a substitute method of design that centers the human. According to IBM, HCAI is "an emerging discipline intent on creating AI systems that amplify and augment rather than displace human abilities" [25]. Importantly, HCAI seeks to "preserve human control" [25]. The hope is to provide beneficial outcomes through AI and to develop "responsible and human-compatible AI", anticipating the potential for negative consequences or the potential misuse of AI [25]. The aim of HCAI is to "amplify, rather than erode, human agency" [26, p. 56], by attempting to promote scenarios in which: there are high levels of control and of autonomy concurrently; humans are empowered not emulated [27]; and there is a multi-tiered approach to governance (team, organization and industry level) to enable HCAI principles to be applied in a practical sense [21], [28].

Studies, however, question whether the HCAI movement truly accounts for the human, given its favoring of ethics and human values, which are often challenging to operationalize; rather, there is a call to broaden the emphasis on values to incorporate other human considerations, inclusive of needs or requirements, individual user experiences and, indeed, organizational, and societal impacts [29]. This perspective is reinforced in academic scholarship that notes the simultaneous requirement for the consideration of ethics and "design decisions" with a view "to bridge the gap between ethics and practice" [26, p. 60]. Further to these ideas, it has been suggested that HCAI is about negotiating two future extremes: one that is technologically deterministic and the second, a dystopian vision. With respect to these extremes, Shneiderman [21, p. 111] claims that both result in a situation divorced of human control, and that an alternate, third future proposition is conceivable, as follows: "an alternative future filled with computing devices that amplify human abilities a thousand-fold, empowering people in remarkable ways while ensuring human control". The latter point regarding control has featured in other related academic studies. For instance, Xu [22] conveys concerns regarding future AI-based systems in the ultimate loss of control for humans, and advocates for the Human-Computer Interaction (HCI) community to play an active role in realizing the potential of HCAI and refuting a future void of human control.

C. Interdisciplinarity and Multi-Stakeholder Participation

The HCI community alone, however, although serving a crucial role in the design process, cannot exclusively be tasked with the responsibility of fulfilling the vision of HCAI and ensuring that human control is maintained. Rather, a broader socio-technical ecosystems view should be assumed, in which constituent stakeholders are identified, engaged, and meaningfully consulted during AI systems design and development processes [17]. Additionally, there is the need for a shared commitment concerning "common standards of behavior, decency, and social justice in the process of systems design, development, and implementation" on the part of designers, regulators, and other stakeholders [17].

This collective effort regenerates the focus on collaborative and participatory design [7], [20]. These would ideally accentuate the value of socio-technical approaches and "infrastructures that bridge the gap between the social, technical, and environmental dimensions that support human safety, protection, and constitutive human capacities, while maintaining justice, human rights, civic dignity, civic participation, legitimacy, equity, access, trust, privacy, and security. The aim should be human-centered value-sensitive socio-technical systems, offered in response to local community-based challenges that are designed, through participatory and co-design processes, for reliability, safety, and trustworthiness" [17].

Partnerships will be central to this endeavor, looking specifically beyond the walls of academia [26]. In doing so, there is a requirement for investing in people, not just the AI-based technologies and solutions, while compelling multiple stakeholders to contribute to the ethical and just design and development of AI, in contexts such as digital mental health services [30], and others. To achieve this, a preliminary step proposed in this Editorial is dissolving disciplinary boundaries, looking to disciplines such as psychology, and indeed social psychology, for an enhanced understanding of the "human". Explicitly, of human and psychological needs, and of universal needs and their impact on health, wellbeing, and the ability of the individual to thrive [29]. This necessitates the "transference of psychological theories" [22, p. 46], approaching AI design from a social cognitive angle. The following section will present what this may entail.

V. Social Cognitive Approaches to AI Design

Whether we consider general consumer products or purpose-built AI for private or public sector use, stakeholders represent diverse populations. As AI transcends national boundaries, it is unlikely that people will share the same attitudes, beliefs, behaviors, and social relations. Instead, we must understand the factors that impact on the judgment and decision-making processes that affect perceptions of trustworthiness of AI.

A. Biases in Judgments and Decision-Making

In addition to framing design in terms of HCAI in the previous section, a human-centered approach to AI requires that we understand basic properties of judgment and decision-making. Human cognition is defined by two types of mental operations: fast, associative, automatic responding that we share with nonhuman animals (Type 1), or slow, resource-dependent, deliberative reflection (Type 2).

If we are unmotivated or presented with seemingly familiar tasks, humans rely on Type 1 responding. By relying on Type 1 processes, humans tend to use heuristics and stereotypes. Heuristics represent decision-making rules that can vary in their complexity from being dependent on information that is available [31], overestimating our task knowledge [32], failing to accurately estimate the frequency or probability of an event's occurrence [33], [34], selectively attending to or gathering information that supports our prior beliefs [35], and subsequently rationalizing unpredictable events as though we previously anticipated the outcomes [36]. These biases introduce ethical concerns for system design [7].

The validity of users' consent is perhaps the most prominent challenge given AI's reliance on their data. Studies have demonstrated that users frequently only superficially read end-user license agreements (EULAs), such that they merely represent 'click wrap' [37]. For instance, research suggests that users spend an average of 73 seconds reading privacy policies and that 'gotcha clauses' were missed by the majority (98%) of users [38]. Other dark patterns of design rely on behavioral nudges, such that designers use cognitive biases to their advantage rather than the users' [39], [40]. By using default settings that advantage an organization, users might engage in behaviors and provide data that are against their interests [41].

Psychological studies have repeatedly observed failures of affective forecasting, such that people lack the ability to successfully predict their own responses to future outcomes [42]. Thus, the mere provision of consent at one moment does not mean that consent is valid in the future. However, given that people have a bias to accept the status quo [43], users might simply accept, or rationalize, their past consent regardless of its implications. Systematic paternalistic nudging might be required to ensure that consent remains valid over the lifetime of a dataset, by ensuring that users have relevant knowledge, take sufficient time, and assess their competencies to make a decision.

B. Fear and Loathing of Autonomy

Humans perceive the world through categories. The most fundamental distinctions are that between the animate and inanimate world [44], and human and non-human animals [45]. These categories affect how we perceive objects in terms of their goal-directedness. AI arguably represents a boundary condition: while they are not alive, their intelligence and ability to 'autonomously' complete many tasks might make it appear that they have distinct goals of their own. The goals we delegate to AI are no longer inconsequential. They can be used to identify potential threats [46], [47], identify trends in seasonal influenza [48], [49], whether someone is a good job candidate [50], or determine a person's creditworthiness [51].

While AI is not human, the tasks that it is designed to perform are meant to supplement or replace human intelligence. Yet intelligence and creativity are often viewed as uniquely human capabilities. Programs like ChatGPT and Dall-e demonstrate that some properties of creativity can be captured by programs [110]. Statistical models have frequently matched diagnostic capabilities of clinicians in medical decision-making [52]. The inertia of AI adoption appears to be inescapable. Rather than asking whether we should incorporate AI, we should question how best to do it. We must also acknowledge that the creativity is based on the product of human creators, protecting their intellectual property rights.

As we noted above, trustworthiness is a critical feature of AI design. Yet, a major barrier that developers encounter is the global bias in trust toward AI, defined along the poles

of algorithmic aversion and affinity [53]. Studies have suggested that bots can promote cooperation in collaborative games when their identity as an AI is concealed. When players are aware that they are playing with an AI, they are less inclined to cooperate [54]. Such biases are likely to differ based on individual users, with humans demonstrating different preferences for things relative to people [55]. These differences might also explain the composition of the emerging AI workforce [56], [57].

Rather than attributing biases specifically to beliefs about AI, general biases such as familiarity might provide adequate explanations [58]. For instance, studies have illustrated that robots or avatars that share some, but not all physical or behavioral features produce negative affect, referred to as the uncanny valley [59], [60], [61]. Rather than being based on discomfort attributable to robots, the effect appears to be based on familiarity [58], [62]. In the case of domesticated technologies, especially 'smart' devices, humans might have acquired a blind spot such that they provide their personal information in an unreflective manner to organizations for the free use of applications and their immediate convenience [7]. For instance, the use of mobile phone applications might be substituted for medical advice despite the lack of regulation and oversight associated with app development [63]. Conversely, people might be inclined to reject novel technologies due to a lack of familiarity, human oversight, or empathy. By assuming that machines are qualitatively different from humans, this neglects the fact that humans and AI can be compared in terms of their systematic biases and random errors. As conversational agents begin to develop 'personalities', this will undoubtedly increase perceived similarity and familiarity.

Like their human creators, AI is neither savior nor sinner. It is limited to the data and algorithms provided to it. Stakeholders need to be granted access to their data and explanations of the operations and outcomes of these systems.

C. Humanizing Explainability

Explainability demands that we understand stakeholders' perspectives rather than provide accounts that are only transparent to developers [64]. This requires that we can identify their domain-specific knowledge that is relevant to the use case, in addition to their motivations for using AI. For instance, providing an explanation to a designer, developer, distributor, regulator, or user will likely require very different information based on their goals and technical knowledge. In contrast to transparency which merely requires making the data, operations, and an output available to stakeholders, explainability is a psychological construct [7], [65]. Explanations provide simple [66], coherent causal relationships of phenomena of interest [67], [68], [69], [70]. While we might lack knowledge within a domain, processes such as analogical reasoning [71] allow relational and explanatory knowledge from one domain to be used in another [72]. For instance, the use of the term *intelligence* to describe AI, implies that the operations or output of an AI are in some way similar to humans or nonhuman animals.

Humans can use a variety of explanations, e.g., technical, functional, or psychological/anthropomorphic. However, we vary in the extent to which different kinds of explanations are deemed acceptable. Despite the ease with which children can learn using intentional (goal-directed) explanations [73], [74], [75], studies have found that adult participants are inclined to *reject* intentional and accept mechanisms explanations, and confidence in explanations varies independently from their accuracy [76]. The acceptability of these explanations varies based on the psychosocial aspects from the human domain that are used to describe AI: the human might reject explanations that AI failed because it was 'sad' or didn't 'like' another system, yet they might accept those that explain failure in terms of failing to 'pay attention' or 'forgetting information'. Consequently, systems designers and developers might present an entirely coherent explanation, yet a stakeholder might hold values, attitudes, and beliefs that result in them rejecting the explanation. As we noted above, explanations might be presented that are believable that fail to accurately represent key affordances of a technology, e.g., how data is aggregated, processed, and used.

These issues are fundamental features of any form of scientific communication [77] and knowledge translation [78], [79]: how do we simplify a phenomenon while accurately retaining the relationships between variables and processes? Without sufficient AI literacy, it is by no means clear that the acceptance or rejection of AI in any given domain represents a valid form of consent.

D. Codifying Mental Models and Social Categories

When judging the validity and accuracy of an AI, we must adopt specific evaluation criteria. Validity and accuracy assume that there is a 'ground truth', however, humans or objects represent social categories that are maintained within a particular community. For instance, there are multiple overlapping definitions of artificial intelligence, machine learning, and autonomous/intelligent systems [80]. While there are consistent dimensions along which to judge all social categories (i.e., their competence and warmth [81]), categories within communities might be absent, overlap, or conflict with others. Designers and developers must be aware of the mental models and social categories used within a community. For instance, we might take for granted that people in rural or developing countries understand the operations of AI, where their data will be stored, and whether the data used to develop AI that they adopt is adequately adapted to their implementation needs.

Social categories are fluid, subject to processes of cultural evolution that are shaped by the social, physical, and now, digital environment. Mental models and categories direct our attention to features of that environment and away from others. By using labels based on social categories, we can be inadvertently influenced by these categories. Over time, underlying inequalities can become embedded within datasets. For instance, an AI used by Google infamously labeled photos of African-Americans as 'gorillas', suggesting a bias in their dataset [82, p. 208]. However, the solution of removing the label did not address the underlying bias that was present in

the training set [65]. Such episodes also illustrate how AI can help us reveal biases, making the implicit explicit.

Similar concerns also exist for definitions of morality and ethics such as fairness, rights, and virtue. Beyond discrepant philosophical frameworks that users might hold (e.g., consequentialism or non-consequentialism), many different formulations of fairness are possible depending on our beliefs about the relationship we have with others. For instance, whether people believe that all members of their group are equal, or that a hierarchy exists and is legitimate, will directly alter how information is shared [83].

On longer time scales, different categories can become embedded in a society, leading to cross-cultural differences. Societies might differ in their trust and acceptance of organization- or stated-based monitoring and surveillance [84]. If AIs are not limited by national boundaries, these differences must inform not only our design practices but how these systems are evaluated nationally and internationally.

E. The Social Networks of AI

AI cannot be viewed in isolation. Rather, they are socio-technical systems defined by dependencies between individual users, community members, groups, societies, and technologies. They extend our cognitive processes and those of others, creating a distributed cognitive system between human and non-human agents [85].

Distributed cognitive systems require understanding the agents within a network, the relational ties between these agents, the function that each agent performs, and the extent to which each agent is aware of the operations of others [86]. For instance, industrial-organizational psychology has repeatedly emphasized the importance of transactive memory– the knowledge we maintain about the others in our team [87]. By introducing AI, we are adding a new kind of agent into this relational network with its own bounded competencies. Humans need to understand the strengths and limitations of these new team members.

Cybersecurity presents a clear case of AI as socio-technical systems. The security of data and networks might be dependent on software and hardware, but users' behaviors, personalities, and beliefs will determine compliance with security policies and their competency in detecting novel threats [88]. If users do not perceive a threat, they click on links or download malware inadvertently. They can also create intentional threats to organizations [89]. While network monitoring provides a means to regulate the system, it might also create mistrust or distrust in employees, thereby creating the problem that it was intended to solve [90]. Models of threat perception geared toward understanding algorithmic aversion must similarly adopt a socio-technical systems-based approach that considers how users, developers, and regulators understand the local and global systems created by AI.

VI. Discussion

This Editorial has presented the case for HCAI, highlighting definitional distinctions relative to explainability, accuracy and trustworthiness, while reviewing XAI design themes and the requirement for a reframing that highlights the significance of the human and possesses an acute awareness of potential social cognitive approaches to AI design. These approaches both acknowledge the diversity of stakeholders during socio-technical design and development processes, and factor and/ or embed potential differences and requirements into respective processes. This is crucial in a range of applications, where balanced discussions of opportunities, risks, and responsibilities are required.

For example, we must consider the risks and responsibilities associated with deploying these technologies in safety-critical applications, such as healthcare [91], predictive policing [92], credit scores [93], and autonomous vehicles. Using AI in these areas can bring significant benefits. However, at the same time, we need to be mindful of the potential consequences and impact on end-users in vulnerable communities [94].

Vulnerable populations, including individuals with disabilities, elderly individuals, and low-income communities, may experience heightened levels of harm if AI systems are not designed and implemented in an ethically responsible manner [95]. This is due to the fact that AI systems that are trained on datasets containing biases that may result in incorrect diagnoses and treatments, incorrect assessments of criminality or creditworthiness, or other consequential decisions that could have detrimental effects on these populations. Similarly, domestic delivery drones and autonomous vehicles must be designed with safety features that are robust enough to prevent accidents that could cause harm to these groups who may not be able to directly benefit from these technologies [96].

In different contexts where the deployment of AI systems holds significant ramifications for safety, it is imperative to prioritize and carefully examine explainability and accuracy [97]. Given that the behavior of AI and ML algorithms may exhibit unpredictable outcomes, there is a substantial risk for the manifestation of unintended biases and consequent harm to vulnerable populations. Thus, it is crucial to implement measures to guarantee AI systems' transparency, accountability, and trustworthiness to mitigate such risks.

Example initiatives are available to achieve this, such as the working groups and standards in the IEEE P70XX series, for example, IEEE 7000™, IEEE 7001™, IEEE P7002™, IEEE P7003™, IEEE P7004™, IEEE P7004.1™, IEEE 7005™, IEEE 7007™, IEEE P7008™, IEEE P7009™, IEEE 7010™, IEEE P7010.1™, IEEE P7011™, IEEE P7012™, IEEE P7014™, and IEEE P7015™ as part of The IEEE Global Initiative on Ethics Of Autonomous and Intelligent Systems [98]; the IEEE P2863 working group initiative on Organizational Governance of Artificial Intelligence [99]; and applied ethics processes to assess trustworthy AI such as Z-Inspection® [100]. Other examples include the Center for Artificial Intelligence and Digital Policy which conducts annual reviews of critical AI policies [101], as well as the Center for Standards and Ethics in Artificial Intelligence (CSEAI) initiative, the purpose of which is to execute research and promote the deployment of AI in safety-critical applications requiring a responsible, standardized, and ethical approach, with a focus on reducing risks and ensuring that these systems are designed and deployed in a way that protects

the rights and well-being of vulnerable communities [102]. This includes ensuring that the AI systems are explainable, accurate, and free from biases and providing end-users with the necessary information and support to make informed decisions when using these systems.

The proliferation of AI in critical domains, such as healthcare, business, government, education, and justice, has elicited a comprehensive examination of the ethical ramifications inherent in these systems [91]. Despite initiatives such as the CSEAI, CAIDP, and other relevant efforts, the swift integration and utilization of AI in high-stakes applications have heightened concerns regarding the potential risks and challenges associated with these systems. Given the growing integration of AI into our daily lives, it is imperative to prioritize the development of secure, dependable, and trustworthy AI systems to mitigate ethical implications.

One of the seminal ethical conundrums pertaining to the implementation of AI is the matter of performance. The potential for even a minuscule deviation in the performance of an AI system to cause significant harm to individuals and society in high-stake situations is a concern that cannot be disregarded. As an illustration, a healthcare scenario involving an AI system misdiagnosing a patient would result in incorrect medical intervention, potentially leading to unfavorable health outcomes. Similarly, in criminal justice, an AI system inaccurately identifying a suspect as a criminal could culminate in unjust monitoring and detention. Hence, AI systems must undergo a thorough testing and validation process to guarantee their accuracy, trustworthiness, and explainability when required.

A crucial aspect of ethical considerations pertains to the transparency of AI systems. Individuals must comprehend the mechanisms by which AI systems arrive at decisions and the rationale behind such decisions [103]. This is particularly relevant in situations where the implications of AI systems are significant. A lack of transparency could obstruct individuals from questioning decisions made by AI systems, thereby eroding the credibility and trust in these systems.

Additionally, it is imperative to establish a system of accountability concerning AI. If an AI system causes harm, it is necessary to ensure that those responsible are held accountable. This can encompass both the responsibility of the designers and developers of the AI system to guarantee its accuracy and transparency, as well as the accountability of the organizations that employ the AI system for the ramifications of its decisions.

As the deployment of artificial intelligence in high-stakes contexts becomes more prevalent, we must conduct a rigorous examination of the ethical ramifications associated with these systems. A critical component of ensuring AI's safe and responsible use is the assurance of accuracy, transparency, and accountability within its design and implementation. Efforts to standardize trustworthy AI and implement strategies for validating and verifying these systems represent crucial steps toward this goal. However, we must engage in ongoing discussions and debates surrounding these critical issues. Additionally, we must acknowledge that relying solely on individual consumers' judgment to determine AI's trustworthiness is insufficient. They may need additional competencies, understanding or information to fully and in an informed manner assess systems risks and potential harm. Therefore, the responsibility of ensuring trustworthy AI should rest with trained professionals who are part of a comprehensive infrastructure that includes a system of rules and resources for enforcement, as well as interdisciplinary collaboration to translate the principles of trustworthy AI into widely accepted standards. In this Editorial, we propose that these professionals look to human-centered, multi-stakeholder approaches to design and development, in which disciplinary silos are torn down [104, pp. 187 and 282], [105, p. 382] to allow for the necessary depth of understanding relevant to trustworthiness and other ethical dilemmas in the context of AI, in fields such as psychology, as we have demonstrated in the example of social cognitive approaches to AI design, sociology, anthropology, and beyond.

VII. In This Special Issue

This Special Issue is comprised of eight papers. The structure of the special is as follows. We begin by reconceptualizing AI, after which we discuss the problem of explainability, focusing on why we need both accuracy and explainability and why using the human-centered approach to designing AI that recognizes the importance of ethics, is not a zero-sum game. Following this is a cluster of articles that employ AI techniques: one demonstrates how there may be in-group bias due to dimensions of ethnicity and age, while the other related Special Issue papers demonstrate the application of AI within the medical space, specifically medical image analysis and the identification of multiple neurological disorders. The remaining two papers in this special are focused on education with a view to address the issues being raised at the grass-roots level but also more holistically.

An underpinning message across the Special Issue papers relates to the significance of the social implications of AI, notably, the need for ethics, accuracy, explainability, and trustworthiness, and the way in which this may be achieved using a human-centered approach to the design of AI systems that acknowledge the importance of ethics and ethical considerations. We summate that despite AI being there to automate decision-making, end-users are still very much a part of the design, development, and feedback processes. In fact, user satisfaction remains as important as performance, and the confidence a user has in a result, and how the algorithm derived that result, is paramount to the overall success of a system. Dependent on how critical the application is, errors may be acceptable to an end-user if they know how the decision/result was derived; but for more life-sustaining applications, errors may not be tolerated if the guidance granted is incorrect, and may have unintended consequences, e.g., a patient is asked to take certain medicines because the AI believes the patient is predisposed to x or y with a high confidence level when in fact the way that determination occurred was based on global variables like gender, ethnicity and age, which are not applicable in context.

Provided below is a summary of the respective Special Issue papers. The first paper is an invited paper by Clarke [A1]. Clarke is a Fellow of the Australian Computer Society and has held long-term posts at the Australian National University and University of New South Wales. Clarke prepared a distinguished lecture for the Young Southeast Asia Leaders Initiative (YSEALI) of Fulbright University Vietnam, in 2022 and was invited to write a full-length piece for IEEE TRANSACTIONS ON TECHNOLOGY AND SOCIETY. Clarke begins by addressing the original conception of AI, and then proposes its re-conception by offering a new definition, with insights to application and benefits. Clarke acknowledges the generic threats inherent in AI and the broad areas of negative impact but, also how to reap benefits while mitigating harms and managing risks. In the final part of the re-conception, Clarke positions Complementary Artefact Capability (CAC) together with Human Capability, resulting in a powerful form of synergy. He brings together the coordination of the intellectual plus physical characteristics, of both humans and artefacts, in order to achieve a combined capability of action superior to that which a human or artefact can achieve alone. He describes this human-robot teaming as an Augmented Capability (AC).

The second paper is by Adamson [A2]. Adamson is a Past President of the IEEE's Society on the Social Implications (IEEESSIT), and holds cross-disciplinary qualifications in technology, engineering, information technology management, and commercial law. In his paper, Adamson argues that AI systems are oftentimes problematic when it comes to explainability because they usually have concealed or black box characteristics. He focuses on the notion of complexity and offers a way forward through the utilization of a DARPA-style approach to model induction, which is fundamentally based on post-hoc reasoning. Adamson points to the benefits of AI and warns of the requirement to adopt the necessary and proportionate controls to implement AI-based solutions and technologies safely.

The third and highly complementary paper is by Petkovic [A3] of San Francisco State University who is an IEEE Life Fellow and has previously been recognized through several prestigious IBM awards. Petkovic emphasizes that trustworthy AI is not a zero-sum game. We should not emphasize either accuracy or explainability; rather, we need both. The same could be said for other determinants of ethical AI when weighed up in terms of levels of importance. Petkovic acknowledges that AI systems may produce errors, can exhibit bias, and may be sensitive to noise in the data, and that technical and judicial transparency remain challenges that act to reduce trust in these emergent systems. Petkovic, echoing Adamson, notes that one way to address the challenges is via explainable AI (XAI). He defines this as being the ability to provide information understandable to a human of how an AI system has made a decision. The article closes by presenting recommendations for the use of XAI in all stages of high stakes, trustworthy AI systems delivery.

The fourth paper is by five scholars: Shruti Nagpal, and Maneet Singh who herald from the Indraprastha Institute of Information Technology Delhi; Richa Singh; and Mayank Vatsa from the Indian Institute of Technology Jodhpur and Nalini Ratha from the University of Buffalo [A4]. In this paper, the behavior of face recognition models is evaluated to understand if, similar to humans, models also encode group-specific features for face recognition, along with where bias is encoded in these models. The authors analyze two types of bias in face recognition models, pertaining to age and ethnicity. The results presented by the authors demonstrate that deep learning models focus on different facial regions for different ethnic groups and age groups. And that large variation in face verification performance is also observed across different sub-groups. To help researchers and practitioners identify the trained model's "level of bias", the researchers document a novel bias index. Accordingly, this index allows analysts to inspect deep networks for exhibiting in-group effects to address the challenge of bias in AI and develop more robust and fairer algorithms for mitigating bias in deep learning models.

The fifth paper [A5] is written by four authors: Tribikram Dhar from Jadavpur University, Nilanjan Dey from the Techno International New Town, Surekha Borra from the K. S. Institute of Technology and R. Simon Sherratt from the University of Reading. The team of researchers focus on the challenges of deep learning specific to the healthcare domain. Some of these challenges are noted as: the unavailability of balanced annotated medical image data, adversarial attacks on deep neural networks and architectures due to noisy medical image data, a lack of trust among users and patients, and ethical and privacy issues related to medical data. The issues raised support those of Petkovic, with the authors exploring how to overcome the concerns that society may have with respect to AI and trust.

The sixth paper in this Special Issue has been written by Md. Nurul Ahad Tawhid, Siuly Siuly and Hua Wang from Victoria University and Kate Wang from RMIT [A6]. This paper demonstrates the potential for automatic identification of neurological disorders using AI, including autism spectrum disorder (ASD), epilepsy (EP), Parkinson's disease (PD), and Schizophrenia (SZ), from EEG signal data. The paper's stated contribution is in its original proposed method for automatic identification of the aforementioned neurological disorders that achieves enhanced efficiency and accuracy when compared to two other popular convolutional neural network (CNN) models, in AlexNet and ResNet50. The performance of the proposed model is also evaluated on binary classification (disease vs healthy) which also outperforms the state-of-the-art results for tested datasets; however, this is based on a very small dataset. There is also some controversy surrounding the application of such a computer aided diagnosis (CAD) tool, which may assist clinicians and experts in the automatic diagnosis process. What if the wrong diagnosis is recorded? What might be the implications for clinicians if such a diagnosis is communicated to patients? How might the decision be explained, echoing those warnings of Adamson and Petkovic, respectively? The preceding paper by Dhar et al. in the medical domain identify ways in which these challenges can be addressed.

The seventh paper is by Tham of Texas Tech University, and Verhulsdonck of Central Michigan University [A7]. In this

paper, the authors propose a "stack" analogy for designing ubiquitous learning, identifying the following layers: design justice, data-informed practices (not just data-driven), human-centered, or playable cities, and human-in-the-loop processes. According to Tham and Verhulsdonck the "stack" analogy for smart education encourages designers of learning environments to identify the roles of the learner (vs bots or digital twins) in smart contexts, expand the interface for learning to include digital and physical sources of information, consider the flow of data, and connect to the real earth-land where learners are grounded. The potential of this approach is to go from a granular instance of, say, a smart library, to a scaled up smart town, smart city and ultimately smart globe. We gain from this paper an understanding that AI can be applied at multiple levels from local to global scales, but also the importance of the human-centered approach in the context of smart cities.

The eighth and final paper is authored by Tina L. Peterson of The University of Texas at Austin and two faculty at Rice University, Rodrigo Ferreira and Moshe Y. Vardi [A8]. The role of computer science ethics education is discussed in this paper, which is pertinent to the theme of the special given the impact AI may have on users, and usees, both unanticipated and unintended consequences that require consideration. The authors propose several new concepts inclusive of abstracted power to assist computer science students to better understand how technology may act to distance them perceptually from the consequences of their actions. They identify "technological intermediation" and "computational thinking" as the two factors in computer science that contribute to this distancing. To counter the abstraction of power, the authors argue for increased emotional engagement in computer science ethics education, to encourage students to feel as well as think regarding the potential impacts of their power on others. By employing any of the four pedagogical approaches noted by the authors, it is in the classroom, where they might be able to grapple with the AI-related ethical dilemmas, through example cases, discourse, and practical implementation. Echoing Petkovic again, we might tackle complex issues in our discussions in the classroom such as accuracy and robustness, explainability, human control and oversight, elimination of bias, judicial transparency, and safety toward trustworthy and ethical AI.

The Special Issue guest editors, all of whom are engaged in human-centric approaches in the context of AI, would like to challenge all who are a part of the design process to consider the multidimensionality of AI and to go beyond the trade-offs mindset, imposing competition between accuracy, explainability and other values. While not all conceived technological applications require AI, if it is used with human oversight, it must come with an acceptable level of explainability to allow for end-users to trust in the decisions and outcomes. Anything to demystify the black box, as noted by Adamson, will likely contribute to building trust. Beyond trust, it will also assist designers to validate their prototypes and solutions. A key question is: If a decision cannot be explained, should it be proposed to anyone, especially those in the financial and medical spaces?

APPENDIX: RELATED ARTICLES

[A1] R. Clarke, "The re-conception of AI: Beyond artificial, and beyond intelligence," *IEEE Trans. Technol. Soc.*, early access, Jan. 4, 2023, doi: 10.1109/TTS.2023.3234051.

[A2] G. Adamson, "Explaining technology we don't understand," *IEEE Trans. Technol. Soc.*, early access, Jan. 30, 2023, doi: 10.1109/TTS.2023.3240107.

[A3] D. Petkovic, "It is not 'accuracy vs. explainability' we need both for trustworthy AI systems," *IEEE Trans. Technol. Soc.*, early access, Jan. 30, 2023, doi: 10.1109/TTS.2023.3239921.

[A4] S. Nagpal, M. Singh, R. Singh, M. Vatsa, and N. Ratha, "In-group bias in deep learning based face recognition models due to ethnicity and age," *IEEE Trans. Technol. Soc.*, early access, Jan. 31, 2023, doi: 10.1109/TTS.2023.3239921.

[A5] T. Dhar, N. Dey, S. Borra, and R. S. Sherratt, "Challenges of deep learning in medical image analysis -improving explainability and trust," *IEEE Trans. Technol. Soc.*, early access, Jan. 4, 2023, doi: 10.1109/TTS.2023.3234203.

[A6] M. N. A. Tawhid, S. Siuly, K. Wang, and H. Wang, "Automatic and efficient framework for identifying multiple neurological disorders from EEG signals," *IEEE Trans. Technol. Soc.*, early access, Jan. 25, 2023, doi: 10.1109/TTS.2023.3239526.

[A7] J. C. K. Tham and G. Verhulsdonck, "Smart education in smart cities: Layered implications for networked and ubiquitous learning," *IEEE Trans. Technol. Soc.*, early access, Jan. 25, 2023, doi: 10.1109/TTS.2023.3239586.

[A8] T. L. Peterson, R. Ferreira, and M. Y. Vardi, "Abstracted power and responsibility in computer science ethics education," *IEEE Trans. Technol. Soc.*, early access, Jan. 3, 2023, doi: 10.1109/TTS.2022.3233776.

REFERENCES

[1] W. Kim, B.-J. Choi, E.-K. Hong, S.-K. Kim, and D. Lee, "A taxonomy of dirty data," *Data Min. Knowl. Discov.*, vol. 7, pp. 81–99, Jan. 2003, doi: 10.1023/A:1021564703268.

[2] J. Vincent. "The biggest headache in machine learning? Cleaning dirty data off the spreadsheets." The Verge. Nov. 2017. [Online]. Available: https://www.theverge.com/2017/11/1/16589246/machine-learning-data-science-dirty-data-kaggle-survey-2017

[3] E. McLarney et al., "NASA framework for the ethical use of artificial intelligence (AI)," NASA, Washington, DC, USA, Rep. NASA/TM-20210012886, 2021.

[4] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2016, pp. 1135–1144.

[5] A. B. Arrieta et al., "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020, doi: 10.1016/j.inffus.2019.12.012.

[6] R. Chatila et al., "Trustworthy AI," in *Reflections on Artificial Intelligence for Humanity* (Lecture Notes in Computer Science, 12600), B. Braunschweig and M. Ghallab, Eds. Cham, Switzerland: Springer, 2021, pp. 13–39. [Online]. Available: https://doi.org/10.1007/978-3-030-69128-8_2

[7] J. R. Schoenherr, *Ethical Artificial Intelligence From Popular to Cognitive Science: Trust in the Age of Entanglement*. New York, NY, USA: Taylor & Francis, 2022.

[8] A. J. London, "Artificial intelligence and black-box medical decisions: Accuracy versus explainability," *Hastings Center Rep.*, vol. 49, no. 1, pp. 15–21, 2019, doi: 10.1002/hast.973.

[9] I. D. Raji et al., "Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing," in *Proc. Conf. Fairness Accountability Transparency (FAT)*, 2020, pp. 33–44, doi: 10.1145/3351095.3372873.

[10] X. Liu, B. Glocker, M. M. McCradden, M. Ghassemi, A. K. Denniston, and L. Oakden-Rayner, "The medical algorithmic audit," *Lancet Digit. Health*, vol. 4, no. 5, pp. e384–e397, 2022, doi: 10.1016/S2589-7500(22)00003-6.

[11] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Metrics for explainable AI: Challenges and prospects," 2018, *arXiv:1812.04608*.

[12] S. Mohseni, N. Zarei, and E. D. Ragan, "A multidisciplinary survey and framework for design and evaluation of explainable AI systems," *ACM Trans. Interact. Intell. Syst.*, vol. 11, nos. 3–4, pp. 1–45, 2021.

[13] D. Kaur, S. Uslu, K. J. Rittichier, and A. Durresi, "Trustworthy artificial intelligence: A review," *ACM Comput. Surv.*, vol. 55, no. 2, pp. 1–38, 2022.

[14] P. O. Nagitta, G. Mugurusi, P. A. Obicci, and E. Awuor, "Human-centered artificial intelligence for the public sector: The gate keeping role of the public procurement professional," *Procedia Comput. Sci.*, vol. 200, pp. 1084–1092, Mar. 2022.

[15] Defense Innovation Board, *AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense: Supporting Document*, United States Dept. Defense, Washington, DC, USA, 2019.

[16] H. Liu et al., "Trustworthy AI: A computational perspective," *ACM Trans. Intell. Syst. Technol.*, vol. 14, no. 1, pp. 1–59, 2022.

[17] K. Michael, R. Abbas, and J. Pitt. "Maintaining control over AI." Issues in Science and Technology. 2021. [Online]. Available: https://issues.org/debating-human-control-over-artificial-intelligence-forum-shneiderman

[18] B. Shneiderman, "Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy human-centered AI systems," *ACM Trans. Interact. Intell. Syst.*, vol. 10, no. 4, pp. 1–31, Dec. 2020, doi: 10.1145/3419764.

[19] B. Shneiderman, *Human-Centered AI*. Oxford, U.K.: Oxford University Press, 2022.

[20] R. Abbas, J. Pitt, and K. Michael, "Socio-technical design for public interest technology," *IEEE Trans. Technol. Soc.*, vol. 2, no. 2, pp. 55–61, Jun. 2021.

[21] B. Shneiderman, "Human-centered artificial intelligence: Three fresh ideas," *AIS Trans. Human-Comput. Interact.*, vol. 12, no. 3, pp. 109–124, 2020.

[22] W. Xu, "Toward human-centered AI: A perspective from human-computer interaction," *Interactions*, vol. 26, no. 4, pp. 42–46, 2019.

[23] "What is human-centered design?" Interaction Design Foundation. 2020. [Online]. Available: https://www.interaction-design.org/literature/topics/human-centered-design

[24] J. R. Schoenherr, 'Learning engineering is ethical," in *Learning Engineering Toolkit*, J. Goodell and J. Kolodner, Eds. New York, NY, USA: Routledge, 2023, pp. 201–228.

[25] W. Geyer, J. Weisz, C. S. Pinhanez, E. Daly. "What is human-centered AI?" IBM. Mar. 2022. [Online]. Available: https://research.ibm.com/blog/what-is-human-centered-ai

[26] B. Shneiderman, "Human-centered AI: Computer scientists should build devices to enhance and empower–not replace–humans," *Issues Sci. Technol.*, vol. 37, no. 2, pp. 56–62, 2021.

[27] B. Shneiderman, "Design lessons from AI's two grand goals: Human emulation and useful applications," *IEEE Trans. Technol. Soc.*, vol. 1, no. 2, pp. 73–82, Jun. 2020.

[28] B. Shneiderman, "Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy human-centered AI systems," *ACM Trans. Interact. Intell. Syst.*, vol. 10, no. 4, pp. 1–31, Dec. 2020, doi: 10.1145/3419764.

[29] W. J. Bingley et al., "Where is the human in human-centered AI? Insights from developer priorities and user experiences," *Comput. Human Behav.*, vol. 141, Apr. 2023, Art. no. 107617.

[30] S. Hamdoun, R. Monteleone, T. Bookman, and K. Michael, "AI-based and digital mental health apps: Balancing need and risk," *IEEE Technol. Soc. Mag.*, vol. 42, no. 1, pp. 25–36, Mar. 2023, doi: 10.1109/MTS.2023.3241309.

[31] N. Schwarz et al., "Ease of retrieval as information: Another look at the availability heuristic," *J. Pers. Soc. Psychol.*, vol. 61, no. 2, pp. 195–202, 1991.

[32] R. F. West and K. E. Stanovich, "The domain specificity and generality of overconfidence: Individual differences in performance estimation bias," *Psychonomic Bull. Rev.*, vol. 4, pp. 387–392, Sep. 1997.

[33] D. Griffin and R. Buehler, "Frequency, probability, and prediction: Easy solutions to cognitive illusions?" *Cogn. Psychol.*, vol. 38, pp. 48–78, Feb. 1999.

[34] S. Avugos, J. Köppen, U. Czienskowski, M. Raab, and M. Bar-Eli, "The 'hot hand' reconsidered: A meta-analytic approach," *Psychol. Sport Exercise*, vol. 14, pp. 21–27, Jan. 2013.

[35] R. S. Nickerson, "Confirmation bias: A ubiquitous phenomenon in many guises," *Rev. Gen. Psychol.*, vol. 2, no. 2, pp. 175–220, 1998.

[36] J. J. Christensen-Szalanski and C. F. Willham, "The hindsight bias: A meta-analysis," *Org. Behav. Human Decis. Process.*, vol. 48, pp. 147–168, Feb. 1991.

[37] R. Pringle, "Narrative, design, and comprehension: Connective technologies and their terms of service agreements," *IEEE Technol. Soc. Mag.*, vol. 35, no. 1, pp. 40–46, Mar. 2016, doi: 10.1109/MTS.2016.2518252.

[38] J. A. Obar and A. Oeldorf-Hirsch, "The biggest lie on the Internet: Ignoring the privacy policies and terms of service policies of social networking services," *Inf. Commun. Soc.*, vol. 23, no. 1, pp. 128–147, 2020.

[39] R. Abbas, K. Michael, M. G. Michael, C. Perakslis, and J. Pitt, "Machine learning, convergence digitalization, and the concentration of power: Enslavement by design using techno-biological behaviors," *IEEE Trans. Technol. Soc.*, vol. 3, no. 2, pp. 76–88, Jun. 2022, doi: 10.1109/TTS.2022.3179756.

[40] C. Perakslis, K. Michael, M. G. Michael, J. Pitt, and R. Abbas, "Safeguarding the guardians to safeguard the bio-economy and mitigate social injustices," in *Cyberbiosecurity: A New Field to Deal with Emerging Threats*, 1st ed., D. Greenbaum, Ed. Cham, Switzerland: Springer Int., 2023.

[41] A. Acquisti et al., "Nudges for privacy and security: Understanding and assisting users' choices online," *ACM Comput. Surv.*, vol. 50, no. 3, pp. 1–41, 2017.

[42] T. D. Wilson and D. T. Gilbert, "Affective forecasting: Knowing what to want," *Current Directions Psychol. Sci.*, vol. 14, pp. 131–134, Jun. 2005.

[43] D. Kahneman, J. L. Knetsch, and R. H. Thaler, "Anomalies: The endowment effect, loss aversion, and status quo bias," *J. Econ. Perspect.*, vol. 5, no. 1, pp. 193–206, 1991.

[44] A. Caramazza and J. R. Shelton, "Domain-specific knowledge systems in the brain: The animate-inanimate distinction," *J. Cogn. Neurosci.*, vol. 10, no. 1, pp. 1–34, 1998.

[45] D. H. Rakison and D. Poulin-Dubois, "Developmental origin of the animate-inanimate distinction," *Psychol. Bull.*, vol. 127, no. 2, p. 209, 2001.

[46] A. Meijer and M. Wessels, "Predictive policing: Review of benefits and drawbacks," *Int. J. Public Admin.*, vol. 42, no. 12, pp. 1031–1039, 2019.

[47] A. G. Ferguson, "Policing predictive policing," *Washington Univ. Law Rev.*, vol. 94, no. 5, p. 1109, 2016.

[48] S. Kandula and J. Shaman, "Reappraising the utility of Google flu trends," *PLoS Comput. Biol.*, vol. 15, Aug. 2019, Art. no. e1007258.

[49] M. Santillana, D. W. Zhang, B. M. Althouse, and J. W. Ayers, "What can digital disease detection learn from (an external revision to) Google Flu Trends?" *Amer. J. Prevent. Med.*, vol. 47, no. 3, pp. 341–347, 2014.

[50] A. Pyke, F. J. R. Schoenherr, and R. Thomson, "Deep blue wants you: Identifying and addressing sources of bias in AI systems to support human resources decisions," in *The Frontlines of Artificial Intelligence Ethics*. New York, NY, USA: Routledge, 2022, pp. 162–184.

[51] M. C. Aniceto, F. Barboza, and H. Kimura, "Machine learning predictivity applied to consumer creditworthiness," *Future Bus. J.*, vol. 6, pp. 1–14, Nov. 2020.

[52] W. M. Grove, "Clinical versus statistical prediction: The contribution of Paul E. Meehl," *J. Clin. Psychol.*, vol. 61, pp. 1233–1243, Oct. 2005.

[53] N. Castelo, M. W. Bos, and D. R. Lehmann, "Task-dependent algorithm aversion," *J. Market. Res.*, vol. 56, pp. 809–825, Jul. 2019.

[54] F. Ishowo-Oloko, J.-F. Bonnefon, Z. Soroye, J. Crandall, I. Rahwan, and T. Rahwan, "Behavioural evidence for a transparency–efficiency tradeoff in human–machine cooperation," *Nat. Mach. Intell.*, vol. 1, pp. 517–521, Nov. 2019.

[55] A. Woodcock, W. G. Graziano, S. E. Branch, M. M. Habashi, I. Ngambeki, and D. Evangelou, "Person and thing orientations: Psychological correlates and predictive utility," *Soc. Psychol. Pers. Sci.*, vol. 4, no. 1, pp. 116–123, 2013.

[56] M. Thelwall, C. Bailey, C. Tobin, and N. A. Bradshaw, "Gender differences in research areas, methods and topics: Can people and thing orientations explain the results?" *J. Inform.*, vol. 13, no. 1, pp. 149–169, 2019.

[57] R. Su and J. Rounds, "All STEM fields are not created equal: People and things interests explain gender disparities across STEM fields," *Front. Psychol.*, vol. 6, p. 189, Feb. 2015.

[58] T. J. Burleigh and J. R. Schoenherr, "A reappraisal of the uncanny valley: Categorical perception or frequency-based sensitization?" *Front. Psychol.*, vol. 5, p. 1488, Jan. 2015.

[59] M. Mori, K. F. MacDorman, and N. Kageki, "The uncanny valley [from the field]," *IEEE Robot. Autom. Mag.*, vol. 19, no. 2, pp. 98–100, Jun. 2012.

[60] M. Cheetham, P. Suter, and L. Jäncke, "The human likeness dimension of the 'uncanny valley hypothesis': Behavioral and functional MRI findings," *Front. Human Neurosci.*, vol. 5, p. 126, Nov. 2011.

[61] T. J. Burleigh, J. R. Schoenherr, and G. L. Lacroix, "Does the uncanny valley exist? An empirical test of the relationship between eeriness and the human likeness of digitally created faces," *Comput. Human Behav.*, vol. 29, pp. 759–771, May 2013.

[62] J. R. Schoenherr and T. J. Burleigh, "Dissociating affective and cognitive dimensions of uncertainty by altering regulatory focus," *Acta Psychologica*, vol. 205, Apr. 2020, Art. no. 103017.

[63] J. R. Schoenherr, "Folkmedical technologies and the sociotechnical systems of healthcare," *IEEE Technol. Soc. Mag.*, vol. 41, no. 3, pp. 38–49, Sep. 2022.

[64] D. E. Forsythe, "Engineering knowledge: The construction of knowledge in artificial intelligence," *Soc. Stud. Sci.*, vol. 23, pp. 445–477, Aug. 1993.

[65] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artif. Intell.*, vol. 267, pp. 1–38, Feb. 2019.

[66] M. Pacer and T. Lombrozo, "Ockham's razor cuts to the root: Simplicity in causal explanation," *J. Exp. Psychol. Gen.*, vol. 146, no. 12, pp. 1761–1780, 2017, doi: 10.1037/xge0000318.

[67] G. L. Murphy, "Explanatory concepts," in *Explanation and Cognition*, F. C. Keil and R. A. Wilson, Eds. Cambridge, MA, USA: MIT Press, 2000, pp. 361–392.

[68] F. C. Keil, "Explanation and understanding," *Annu. Rev. Psychol.*, vol. 57, pp. 227–254, Feb. 2006.

[69] G. S. Halford, W. H. Wilson, and S. Phillips, "Relational knowledge: The foundation of higher cognition," *Trends Cogn. Sci.*, vol. 14, pp. 497–505, Sep. 2010, doi: 10.1016/j.tics.2010.08.005.

[70] T. Lombrozo, "Explanation and abductive inference," in *Oxford Handbook of Thinking and Reasoning*, K. J. Holyoak and R. G. Morrison, Eds. Oxford, U.K.:Oxford Univ. Press, 2012, pp. 260–276.

[71] L. Hobeika, C. Diard-Detoeuf, B. Garcin, R. Levy, and E. Volle, "General and specialized brain correlates for analogical reasoning: A meta-analysis of functional imaging studies," *Human Brain Mapp.*, vol. 37, pp. 1953–1969, 2016, doi: 10.1002/hbm.23149.

[72] L. E. Richland, T. K. Chan, R. G. Morrison, and T. K. F. Au, "Young children's analogical reasoning across cultures: Similarities and differences," *J. Exp. Child Psychol.*, vol. 105, pp. 146–153, Jan./Feb. 2010, doi: 10.1016/j.jecp.2009.08.003.

[73] M. E. Barnes, E. M. Evans, A. Hazel, S. E. Brownell, and R. M. Nesse, "Teleological reasoning, not acceptance of evolution, impacts students' ability to learn natural selection," *Evol. Educ. Outreach*, vol. 10, p. 7, Oct. 2017, doi: 10.1186/s12052-017-0070-6.

[74] H. Bartov, "Teaching students to understand the advantages and disadvantages of teleological and anthropomorphic statements in biology," *J. Res. Sci. Teach.*, vol. 18, pp. 79–86, Jan. 1981, doi: 10.1002/tea.3660180113.

[75] P. Tamir and A. Zohar, "Anthropomorphism and teleology in reasoning about biological phenomena," *Sci. Educ.*, vol. 75, pp. 57–67, Jan. 1991, doi: 10.1002/sce.3730750106.

[76] J. R. Schoenherr and R. Thomson, "Persuasive features of scientific explanations: Explanatory schemata of physical and psychosocial phenomena," *Front. Psychol.*, vol. 12, Sep. 2021, Art. no. 644809.

[77] J. M. Hurd, "The transformation of scientific communication: A model for 2020," *J. Amer. Soc. Inf. Sci.*, vol. 51, no. 14, pp. 1279–1283, 2000.

[78] J. M. Grimshaw, M. P. Eccles, J. N. Lavis, S. J. Hill, and J. E. Squires, "Knowledge translation of research findings," *Implement. Sci.*, vol. 7, pp. 1–17, May 2012.

[79] R. Thomson and J. R. Schoenherr, "Knowledge-to-information translation training (KITT): An adaptive approach to explainable artificial intelligence," in *Proc. 2nd Int. Conf. Adapt. Instruct. Syst.*, Jul. 2020, pp. 187–204.

[80] N. Kühl, M. Goutier, R. Hirt, and G. Satzger, "Machine learning in artificial intelligence: Towards a common understanding," 2020, *arXiv:2004.04686*.

[81] A. J. Cuddy, S. T. Fiske, and P. Glick, "Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS map," *Adv. Exp. Soc. Psychol.*, vol. 40, pp. 61–149, Mar. 2008.

[82] S. Akter, Y. K. Dwivedi, S. Sajib, K. Biswas, R. J. Bandara, and K. Michael, "Algorithmic bias in machine learning-based marketing models," *J. Bus. Res.*, vol. 144, pp. 201–216, May 2022.

[83] T. S. Rai and A. P. Fiske, "Moral psychology is relationship regulation: Moral motives for unity, hierarchy, equality, and proportionality," *Psychol. Rev.*, vol. 118, no. 1, pp. 57–75, 2011.

[84] J. Schoenherr, "Whose privacy, what surveillance? Dimensions of the mental models for privacy and security," *IEEE Technol. Soc. Mag.*, vol. 41, no. 1, pp. 54–65, Mar. 2022.

[85] A. Clark and D. Chalmers, "The extended mind," *Analysis*, vol. 58, no. 1, pp. 7–19, 1998.

[86] F. J. R. Schoenherr and R. Thomson, "Ethical frameworks for cybersecurity: Applications for human and artificial agents," in *The Frontlines of Artificial Intelligence Ethics*. New York, NY, USA: Routledge, 2022, pp. 141–161.

[87] D. P. Brandon and A. B. Hollingshead, "Transactive memory systems in organizations: Matching tasks, expertise, and people," *Org. Sci.*, vol. 15, no. 6, pp. 633–644, 2004.

[88] K. Lewis, "Measuring transactive memory systems in the field: Scale development and validation," *J. Appl. Psychol.*, vol. 88, pp. 587–604, Aug. 2003.

[89] V. Peltokorpi, "Transactive memory systems," *Rev. Gen. Psychol.*, vol. 12, no. 4, pp. 378–394, 2008.

[90] D. M. Wegner, "Transactive memory: A contemporary analysis of the group mind," in *Theories of Group Behavior* (Social Psychology), B. Mullen and G. R. Goethals, Eds. New York, NY, USA: Springer, 1987, pp. 185–208. [Online]. Available: https://doi.org/10.1007/978-1-4612-4634-3_9

[91] S. Gupta, S. Kamboj, and S. Bag, "Role of risks in the development of responsible artificial intelligence in the digital healthcare domain," *Inf. Syst. Front.*, to be published.

[92] A. G. Ferguson, "Policing predictive policing," *Washington Univ. Law Rev.*, vol. 94, no. 5, p. 1109, 2016.

[93] H. Sadok, F. Sakka, and M. E. H. El Maknouzi, "Artificial intelligence and bank credit analysis: A review," *Cogent Econ. Finan.*, vol. 10, no. 1, 2022, Art. no. 2023262.

[94] B. Knowles and J. T. Richards, "The sanction of authority: Promoting public trust in AI," in *Proc. ACM Conf. Fairness Accountability Transparency*, 2021, pp. 262–271.

[95] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," *Nat. Mach. Intell.*, vol. 1, no. 9, pp. 389–399, 2019.

[96] D. Katare, N. Kourtellis, S. Park, D. Perino, M. Janssen, and A. Y. Ding, "Bias detection and generalization in AI algorithms on edge for autonomous driving," in *Proc. IEEE/ACM 7th Symp. Edge Comput. (SEC)*, 2022, pp. 342–348.

[97] D. F. Fonner and F. P. Coyle, "Explainable machine learning models for evaluating government grantmaking," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, 2022, pp. 2243–2248.

[98] "The IEEE global initiative on ethics of autonomous and intelligent systems." IEEE Standards Association. 2022. [Online]. Available: https://standards.ieee.org/industry-connections/ec/autonomous-systems

[99] "IEEE P2863 organizational governance of artificial intelligence working group." IEEE Standards Association. 2022. [Online]. Available: https://sagroups.ieee.org/2863

[100] R. V. Zicari et al., "Z-Inspection®: A process to assess trustworthy AI," *IEEE Trans. Technol. Soc.*, vol. 2, no. 2, pp. 83–97, Jun. 2021.

[101] "Artificial intelligence and democratic values index." Center for AI and Digital Policy. 2023. [Online]. Available: https://www.caidp.org/reports

[102] P. Rivas, J. Ortiz, D. Diaz, and L. Montoya, "Planning a center for standards and ethics in artificial intelligence," in *Proc. Int. Conf. Mach. Learn. Res. (PMLR)*, 2022, pp. 1–10.

[103] D. Valle-Cruz, E. A. Ruvalcaba-Gomez, R. Sandoval-Almazan, and J. I. Criado, "A review of artificial intelligence in government and its potential from a public policy perspective," in *Proc. 20th Annu. Int. Conf. Digit. Government Res.*, 2019, pp. 91–99.

[104] M. M. Crow and W. B. Dabars, *Designing the New American University*. Baltimore, MD, USA: JHU Press, 2015.

[105] M. M. Crow and W. B. Dabars, *The Fifth Wave: The Evolution of American Higher Education*. Baltimore, MD, USA: Johns Hopkins Univ. Press, 2020.

[106] T. Anderson, "Trust building for data sharing—Understanding trust as a social relationship," in *Data and the Digital Self: What the 21st Century Needs*. Sydney, NSW, Australia: Aust. Comput. Soc., 2023, pp. 48–75. [Online]. Available: https://www.acs.org.au/insights andpublications/reports-publications/data-and-the-digital-self.html

[107] J. R. Schoenherr, "Designing ethical agency for adaptive instructional systems: The FATE of learning and assessment," in *Proc. 3rd Int. Conf. Adapt. Instruct. Syst. Des. Eval. (AIS)*, Jul. 2021, pp. 265–283.

[108] R. Abbas and K. Michael, *Socio-Technical Theory: A Review*, S. Papagiannidis, Ed., TheoryHub Book, 2022. [Online]. Available: http://open.ncl.ac.uk

[109] S. Spiekermann, *Ethical IT Innovation A Value-Based System Design Approach*. Boca Raton, FL, USA: CRC Press, 2016.

[110] J. R. Schoenherr. "Generative AI like ChatGPT reveal deep-seated systemic issues beyond the tech industry." The Conversation. Mar. 2023. [Online]. Available: https://theconversation.com/generative-ai-like-chatgpt-reveal-deep-seated-systemic-issues-beyond-the-tech-industry-198579

JORDAN RICHARD SCHOENHERR
Department of Psychology
Concordia University
Montreal, QC H3G 1M8, Canada

Department of Psychology
Carleton University
Ottawa, ON K1S 5B6, Canada

ROBA ABBAS
School of Business
University of Wollongong
Wollongong, NSW 2522, Australia

KATINA MICHAEL
School for the Future of Innovation in Society
Arizona State University
Tempe, AZ 85287 USA

School of Computing and Augmented Intelligence
Arizona State University
Tempe, AZ 85287 USA

PABLO RIVAS
School of Engineering and Computer Science
Baylor University
Waco, TX 76706 USA

THERESA DIRNDORFER ANDERSON
Connecting Stones
Sydney, NSW, Australia

Jordan Richard Schoenherr is an Assistant Professor with the Department of Psychology, Concordia University, and an Adjunct Research Professor with the Department of Psychology and a member of the Institute for Data Science, Carleton University. He is a former Postdoctoral Fellow with the University of Ottawa Skills and Simulation Centre and a former Visiting Scholar with the U.S. Military Academy, West Point. He has acted as an Ethics Consultant for the Ombudsman, Integrity, and Resolution Office (Health Canada/PHAC), the Office of the Chief Scientist (Health Canada), the Canadian Border Services Agency, and the Department of National Defense. His primary areas of interest are learning and decision making and metacognition with application in cyberpsychology (cybersecurity, disinformation, ethical AI, and XAI) and organizational behavior (incivility, insider threat, and knowledge management).

Dr. Schoenherr also serves as a Research Fellow with the Center for AI and Digital Policy and as a member of the Canadian Association of Chiefs of Police Artificial Intelligence Steering Committee.

Roba Abbas (Member, IEEE) is a Senior Lecturer of Operations and Systems with the Faculty of Business and Law, University of Wollongong, Australia, and more recently a Visiting Professor with the School for the Future of Innovation in Society, Arizona State University, USA. Her research is focused on methodological approaches to complex socio-technical systems design, emphasizing transdisciplinarity, co-design and the intersection of society, technology, ethics, and regulation. She is also a Co-Editor of the IEEE TRANSACTIONS ON TECHNOLOGY AND SOCIETY, a Former Associate Editor of the *IEEE Technology and Society Magazine*, and the Technical Committee Chair of the Socio-Technical Systems Committee of the IEEE.

Katina Michael (Senior Member, IEEE) is a Professor with Arizona State University and a Senior Global Futures Scientist with the Global Futures Laboratory and has a joint appointment with the School for the Future of Innovation in Society and the School of Computing and Augmented Intelligence. Prior to academia, she was employed with Nortel Networks, Anderson Consulting, and OTIS Elevator Company. She has been funded by the National Science Foundation, the Canadian Social Sciences and Humanities Research Council, and the Australian Research Council. She is the Director of the Society Policy Engineering Collective and the Founding Editor-in-Chief of the IEEE TRANSACTIONS ON TECHNOLOGY AND SOCIETY. She is also the Founding Chair of the inaugural Master of Science in Public Interest Technology.

Pablo Rivas (Senior Member, IEEE) received the B.S. degree in computer systems engineering from the Nogales Institute of Technology, Nogales, Mexico, in 2003, the M.S. degree in electrical engineering from the Chihuahua Institute of Technology, Chihuahua, Mexico, in 2007, and the Ph.D. degree in electrical and computer engineering from The University of Texas at El Paso, El Paso, TX, USA, in 2011.

He has been an Assistant Professor of Computer Science with the School of Engineering and Computer Science, Baylor University, Waco, TX, USA, since 2020. Before that, he was with the School of Computer Science and Mathematics, Marist College, Poughkeepsie, NY, USA, from 2015 to 2020. He has more than eight years of industry experience as a Software Engineer and has been recognized for his creativity and academic excellence. He is currently in the planning phase of the Center for Standards and Ethics in Artificial Intelligence with funding from the National Science Foundation. He has published several peer-reviewed papers and authored a book on deep learning in 2020. He predominantly researches artificial intelligence and its ethical and social implications, focusing on computer vision, natural language processing, and quantum machine learning.

Dr. Rivas is a member of the IEEE Standards Association and is involved in the working groups developing the P70XX series standards for AI ethics. In 2011, he was inducted into the international honor society for IEEE Eta Kappa Nu; in 2021, he was inducted into Upsilon Pi Epsilon, the international honor society for the computing and information disciplines; and in 2022, he was elevated to a Senior Member of ACM.

Theresa Dirndorfer Anderson is the Director and Social Informaticist with Connecting Stones. She uses creative, compassionate, and contemplative practices to help communities build better digital and data futures. A social informaticist with the Ph.D. in information science, her award-winning work as an Educator and a Researcher engages with the ever-evolving relationship between people and emerging technologies when working with data and making decisions. After an earlier career in international security and diplomacy, she shifted her attention to the information sciences. She was an academic with the University of Technology Sydney (UTS) for more than 20 years, including serving as an Inaugural Director and an Associate Professor of the Master of Data Science and Innovation Program. She remains an Honorary Fellow with the UTS Centre of Connected Intelligence, which oversaw the launch of the uniquely transdisciplinary and human-centered curriculum developed during her tenure. She now focuses on advancing socially just data policies and building trusted environments for data/AI use as a Freelance Consultant with Connecting Stones. She contributes to the development of reference and actionable frameworks at local and international levels. As part of a global data project (WorldFAIR) funded by the European Commission and coordinated by CODATA, she is providing data ethics advice and designing training to implement FAIR and CARE principles in relation to urban health data.

Dr. Anderson was appointed to the NSW Government's inaugural Artificial Intelligence Advisory and Review Committee in 2021. She is actively contributing to the development of an international standard for Data Usage (ISO JTC1/SC32/WG6), serving as a Project Editor. She also regularly contributes to International Science Council Committee on Data (CODATA) initiatives enhancing global cooperation on FAIR data policy and practice. She also sits on the Advisory Board for Resilience Brokers, an international organization using systems thinking to unlock value and to improve the climate resilience of cities and communities around the world.