The Role of Human-Centered AI in User Modeling, Adaptation, and Personalization—Models, Frameworks, and Paradigms

Helma Torkamaan, Mohammad Tahaei, Stefan Buijsman, Ziang Xiao, Daricia Wilkinson, and Bart P. Knijnenburg

Abstract This chapter explores the principles and frameworks of human-centered Artificial Intelligence (AI), specifically focusing on user modeling, adaptation, and personalization. It introduces a four-dimensional framework comprising paradigms, actors, values, and levels of realization that should be considered in the design of human-centered AI systems. This framework highlights a perspective-taking approach with four lenses of technology-centric, user-centric, human-centric, and future-centric perspectives. Ethical considerations, transparency, fairness, and accountability, among other aspects, are highlighted as values when developing and deploying AI systems. The chapter further discusses the corresponding human values for each of these concepts. Opportunities and challenges in human-centered AI are examined, including the need for interdisciplinary collaboration and the complexity of addressing diverse perspectives. Human-centered AI provides valuable insights

The original version of the chapter has been revised: A correction to this chapter can be found at https://doi.org/10.1007/978-3-031-55109-3_11

H. Torkamaan (✉) · S. Buijsman
Technical university of Delft, Delft, The Netherlands
e-mail: h.torkamaan@acm.org

S. Buijsman
e-mail: s.n.r.buijsman@tudelft.nl

M. Tahaei
UC Berkeley's International Computer Science Institute, Berkeley, CA, USA
e-mail: mtahaei@icsi.berkeley.edu

Z. Xiao
Johns Hopkins University, Baltimore, MD, USA
e-mail: ziang.xiao@jhu.edu

D. Wilkinson
Microsoft Research, Baltimore, MD, USA
e-mail: dwilkinson@microsoft.com

B. P. Knijnenburg
Clemson University, Clemson, SC, USA
e-mail: bartk@clemson.edu

B. Ferwerda et al. (eds.), *A Human-Centered Perspective of Intelligent Personalized Environments and Systems*, Human–Computer Interaction Series,
https://doi.org/10.1007/978-3-031-55109-3_2

for designing AI systems that prioritize human needs, values, and experiences while considering ethical and societal implications.

1 Introduction

In the rapidly evolving landscape of Artificial Intelligence (AI), the concept of *human-centered AI* plays an essential role in designing and developing AI systems. This chapter delves into this multifaceted concept, exploring diverse interpretations and perspectives that shape this rapidly evolving field. By examining the intersections of AI, human values, interaction design, and societal considerations, our aim is to provide a comprehensive understanding of the subject and guide future research and development endeavors in user modeling, adaptation, and personalization.

As AI systems increasingly permeate our daily lives, both online and in the physical world, the AI research community has faced challenges and mounting criticism regarding its failure and need to prioritize a human-centered perspective, e.g., [16, 28, 82, 112, 129, 131, 155]. In light of this, AI researchers can draw valuable lessons from the field of user modeling and personalization research, which has long recognized the significance of various human factors, as discussed in chapter "Human Factors in User Modeling for Intelligent Systems", as the core principles of human-centered intelligent solutions. Although some of these discussions and considerations may seem novel to AI researchers, they have been an integral part of the practices followed by user modeling and personalization researchers for many years.

Traditionally, the focus in user modeling and personalization has largely been on users actively engaging with systems. This perspective, however, has often overlooked a more comprehensive human-centered approach, which encompasses a boarder spectrum of individuals, including non-users and aspects of users' lives beyond their interaction with systems. In this chapter, we describe the emergence of human-centered AI, which presents new opportunities for considering and understanding humans in a more comprehensive manner, while also promoting the development of future-proof and responsible AI systems. By adopting human-centered AI frameworks, AI researchers, as well as researchers in the field of user modeling and personalization, can gain insights into human behavior, values, and well-being, paving the way for the creation of AI systems that better align with human needs. This shift in focus is essential for ensuring responsible integration and maximizing the usefulness and effectiveness of AI systems. Furthermore, it underscores the need for ongoing research, collaboration, and interdisciplinary approaches to address the complexities inherent in human-centered AI.

The concept of human-centered AI encompasses a broad spectrum of interpretations, reflecting different levels of human involvement in the design and integration of intelligent systems. Emerging interpretations in the literature range from a technology-centric perspective, primarily focused on improving prediction accuracy through algorithmic enhancements, to more human-centric perspectives that

prioritize human values, human factors in interaction design, augmenting human capabilities, and user-perceived perspectives. In addition, legal, moral, or sociotechnical considerations, the roles and responsibilities of governing bodies, and the long-term impacts of technical and functional decisions, as well as notions such as responsible innovation, social justice, empathetic behavior, sustainability, autonomy, causing no harm, and assurance and accountability, are all discussed as interpretations—or sometimes components—of human-centered AI. The complexity of human-related phenomena and the diverse needs and expectations of multiple actors contribute to the complexity surrounding human-centered AI.

Human-centered AI, in principle, centers around designing and evaluating intelligent systems based on human needs, experiences, expectations, values, and well-being. However, operationalizing this concept remains intricate. To address these challenges, this chapter proposes a comprehensive framework (Sects. 2 and 3) that provides a structured approach to understanding and implementing human-centered AI as well as reflecting on research and practices of developing AI systems. At the core of this framework are four paradigmatic approaches to AI systems (described in Sect. 2): the technology-centric, user-centric, human-centric, and future-centric perspectives, each offering a unique lens through which AI systems can be designed, viewed, and refined. This framework also incorporates additional critical components as discussed in Sect. 3: Actors, levels of realization and evaluation, and values that must be included in the design and evaluation of AI systems, where ethical and legal considerations point us to values such as social justice, accountability, privacy, empathy, understanding humans, and sustainable AI, among others.

However, the interplay between these values, along with the challenges of addressing conflicting perspectives and accommodating diverse stakeholders, presents significant opportunities and challenges for researchers and practitioners in the field (discussed in Sect. 5). Before delving into these opportunities and challenges, Sect. 4 provides an overview of the models, techniques, and frameworks described in the literature that facilitate the human-centered design of AI systems.

By embracing the principles and practices of human-centered AI, researchers and practitioners can develop AI systems that better align with human needs, values, and well-being. This alignment ensures the responsible integration of AI systems while maximizing their usefulness and effectiveness. However, achieving this alignment and successfully navigating the associated challenges require ongoing research, collaboration, and interdisciplinary approaches. By addressing the complexities inherent in human-centered AI, we can advance the field and pave the way for a future where AI systems genuinely serve and prioritize humanity's interests.

2 Empowering Humanity: Exploring the Paradigms of Human-Centered AI

As outlined in the introduction, we note that scholars differ on the exact definition of human-centered AI, and as such, there is no single framework that encompasses all aspects of this concept. A literature review by Capel and Brereton [21] emphasizes the ambiguity regarding how to frame, design, and evaluate human-centered AI. They divide the existing literature into four main categories: ethical AI, humans teaming with AI, explainable and interpretable AI, and human-centered approaches to the design and evaluation of AI, pointing to the need for more guidance and tools in this area. In this chapter, we attempt to provide an inclusive framework of AI research paradigms that portrays four paradigms and demonstrates how the human-centered AI research paradigm is rooted in the *user*-centric AI research paradigm—which in itself evolved from the traditional technology-centric paradigm—and how it is at the brink of evolving into a more comprehensive *future-centric* AI research paradigm.

Our proposed framework represents four paradigmatic approaches to AI systems (Fig. 1), with each additional layer representing a more inclusive and comprehensive focus than the previous one. It is important to note that each paradigm may give rise to a distinct set of requirements and specifications that can be implemented at different levels of design, development, and deployment of the AI system, such as algorithms, data preparation, datasets, and interfaces. In this section, we provide a general description of each of these four paradigms, with subsequent sections delving into further details of the framework and its components and highlight specific AI research topics (e.g., privacy, fairness, explainability) using these four paradigms as theoretical lenses for further investigation.

2.1 Technology-Centric Perspective

The roots of AI research can be traced back to cognitive science [99] and computer-supported collaborative work [76]. Initially, the focus in AI, particularly in fields like user modeling, adaptation, and personalization, was predominantly technology (or system)-centric, emphasizing the enhancement of predictive accuracy of intelligent or adaptive systems through algorithmic improvements. This technology-centric perspective, a hallmark of early AI efforts, prioritized the assessment of algorithms isolatedly using benchmark datasets [153]. While such studies are easy to conduct, early research showed that superior prediction accuracy in offline studies does not always translate to better performance in real-world settings [86, 145].

To address this disparity, researchers have turned to online A/B tests involving real users [75]. Nevertheless, the majority of work within this paradigm still heavily relies on offline studies using benchmark datasets. Offline studies have limitations as they lean on historical data as ground truth and are bound by the quality of the datasets used to build the systems. In addition, they also factor adaptations that

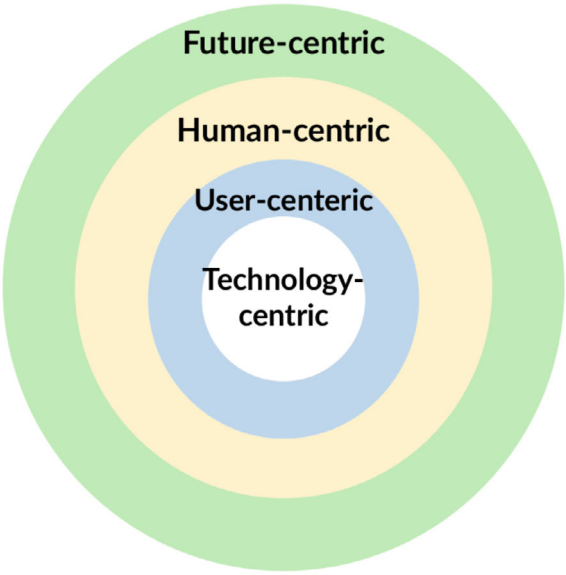

Fig. 1 Our proposed framework consists of four paradigmatic approaches to AI systems. Each paradigm surpasses the previous one in terms of inclusivity and comprehensiveness, encompassing and building upon the foundations of the previous paradigm

align with the more "predictable" aspects of users' preferences or behaviors [72]. Such studies assume that the desirability of adaptations is given (when in fact, users may not always want a system to be adaptive [37]) and that adaptations have no effect on users' behaviors (when in fact, adaptations tend to be persuasive [1]—i.e., users are disproportionally more likely to follow a system's suggestions [67]). While online A/B tests may account for these discrepancies in human behavior, they do not explain *why* these discrepancies occur [74]. Furthermore, research within this paradigm has predominantly focused on objectively measurable properties of algorithms, overlooking other critical factors and subjective outcomes.

2.2 User-Centric Perspective

In response to the shortcomings of the technology-centric perspective, researchers have gradually adopted a more *user-centric* approach to evaluating AI systems [71]—although even to date, this is by no means a dominant paradigm in AI research [76]. The shift toward a user-centric perspective stems from the recognition that prediction accuracy alone is insufficient for assessing the quality of an AI algorithm [87] and that the success of AI systems depends on factors beyond the algorithm itself [88].

Research in the user-centric paradigm often covers the interface between the user and the AI system—this includes the part of the system that gathers the input (e.g., user preferences) on which adaptations are based [111] and the mechanism by which the adaptations are implemented [98]. It evaluates these aspects not only in terms of

their impact on user behavior but also from a subjective standpoint, considering factors like usability and user experience [74]. It also examines how the user's personal characteristics may influence the effect of these input and output mechanisms on the user experience [70, 74]. One extensively studied aspect within modern AI systems under the user-centric paradigm is the transparency of the adaptation mechanism, leading to the emergence of Explainable AI (XAI) as a field of research [95, 141]. Furthermore, the user-centric paradigm has become more prominent with the advent of conversational AI systems [63], such as Amazon Alexa and ChatGPT.

In general, the user's needs, preferences, expectations, and experience of interacting with the system serve as the basis for the evaluation and interpretation of metrics in this paradigm. Common research methods employed in the user-centric paradigm include user research [17, 23], which investigates the needs and desires of existing or potential AI system users through interviews and observations. User-centered design [33, 52] is subsequently used to translate user needs into system designs. And once (prototypes of) those system designs are implemented, user-centric evaluation [73] can be used to assess them using qualitative studies, controlled user experiments, or online field trials. Evaluating the system from the user's perspective often involves considering the user's perception of different criteria, such as privacy, autonomy, transparency, explainability, fairness, and bias. Additionally, it is crucial to consider user behavior and understanding of the system.

While the user-centric paradigm represents a significant departure from the traditional technology-centric perspective by focusing on the interaction between the user and the AI system, it does have limitations. Notably, it often overlooks individuals who are not the users of the AI system but are nevertheless affected by the system. This omission precludes considerations and evaluations of aspects that concern all system actors and/or societal values, such as fairness and algorithmic bias. Another limitation of the user-centric perspective is its tendency to view users as actors who only interact with the systems for a limited duration, which overlooks the full scope of their experience. In contrast, a more expansive view, as explored in the previous chapter, considers human factors in user modeling. This approach acknowledges that users are complex individuals whose lives, personalities, and emotions extend beyond system usage. These elements are crucial for a comprehensive understanding of their needs and interactions with intelligent systems. Consequently, they merit consideration even during periods of non-use, ensuring a more holistic and human-centric evaluation of the system's influence and reach.

2.3 Human-Centric Perspective

By broadening the scope of AI research beyond the user-centric paradigm, researchers have started adopting a more multi-stakeholder [62] perspective that we refer to as the *human-centric paradigm*. This paradigm enables exploration of the broader impacts of AI systems in terms of societal impact, equitable outcomes, ethical and

legal requirements, and more. The human-centered paradigm builds on top of the user-centric paradigm and is more inclusive and comprehensive.

Notable research in this paradigm includes studies on the equitable distribution of content producers in recommendation outcomes [31], potential biases in prediction outcomes that disadvantage underrepresented communities [14], the amplification of historical biases [11], the privacy concerns of people whose data are used for training AI algorithms [42], gender fairness in recommender systems [35], and more.

A key distinction between the user-centric and human-centric perspectives lies in the context of individuals for whom a product or service is intended and designed. In the human-centric paradigm, a human may not necessarily be the primary (or even secondary) user of an AI system, but they can still be influenced by its existence. The human-centric perspective adopts a more comprehensive approach, considering the entire ecosystem of people involved in or influenced by a technical solution. As such, its scope is often inclusive of the individuals and entities involved throughout the system's lifecycle, i.e., *stakeholders* who are affected by the system, interested in its outcome or a part of it, though they may not interact with the system directly as well as the *actors* who are directly involved in the system or its function, and interact with it. Given the data-driven nature of AI systems and their reliance on extensive human data, it becomes increasingly important to consider these actors and stakeholders, even if they do not use the system. Furthermore, these systems can have far-reaching impacts and may function on a large scale. AI systems are increasingly being employed not just for individual use but also for high-level decision-making processes that can impact large groups of people across different layers of society. Consequently, prioritizing human-centeredness in the design, development, and deployment lifecycles of intelligent systems becomes paramount.

To study these aspects, novel research methods on top of those used in previous paradigms can be employed. For instance, while a significant body of research on bias and fairness in the technology-centric paradigm often utilizes benchmark datasets to examine how the outcomes of various algorithmic predictions are distributed among different groups of users, the human-centric paradigm additionally incorporates multi-stakeholder research methods. These methods include social computing research methods [65, 80, 135], stakeholder analysis [108], research on the diversity and inclusiveness of research [126], implementation studies [36], and participatory design [10, 159].

Responsible innovation is a concept that intersects both the human-centric and the future-centric paradigms in AI research. Within the human-centric paradigm, responsible innovation highlights the importance of comprehending the societal and ethical implications associated with the design, development, and deployment of AI systems. It recognizes the broader impacts of AI on individuals and society and advocates for changes in design practices to make them more inclusive and equitable. Responsible innovation aims to align AI systems with diverse stakeholders' needs and values, and for this to happen, we need to be well aware of the actual impact of AI on diverse stakeholders and actors, as well as the needs of these stakeholders and actors.

While the human-centric paradigm plays a vital role in promoting inclusive AI practices, it focuses primarily on analyzing current practices and their real-world impacts. Although this analysis may lead to recommendations for future practices or adjusting technology-centric tools [154], its emphasis is not on long-term impacts. Furthermore, the human-centric paradigm focuses on immediate and existing societal concerns and often involves strategies for designing and evaluating AI systems, with an emphasis on current societal norms, legal frameworks, and ethical guidelines. As a result, while it creates an environment where self-actualization and agency may be more attainable, these aspects are not explicitly targeted or integrated into the core objectives of AI system design under this paradigm. The human-centric approach, focusing on the present and immediate future, often lacks the forward-looking, proactive strategies necessary to explicitly nurture and support long-term human development, growth, and empowerment.

2.4 Future-Centric Perspective

Expanding beyond the confines of the human-centered paradigm, we find the *future-centric paradigm* of AI research. This paradigm strives to transcend the existing state of AI research by focusing on designing systems that embody the values of diverse stakeholders. For instance, it seeks to develop sustainable AI systems (i.e., moving beyond an adversarial perspective toward AI systems that have a more symbiotic relationship with their environment) and support human agency (i.e., moving beyond equity toward self-directed learning and growth). This future-centric paradigm develops along two intertwined perspectives on AI research.

The first perspective is responsible innovation, which plays a pivotal role in envisioning and shaping the future impact of AI systems. While the basic principles of responsible innovation are rooted in the human-centered paradigm, within the future-centric paradigm, this perspective takes a *proactive approach* to anticipate and address the long-term societal implications of AI. By encouraging AI system owners and developers to acknowledge their accountability for the long-term consequences of their creations, responsible innovation encourages fostering open discussions regarding ethical principles and the desired future outcomes of the evolution of AI technology. It also helps steer AI research and development toward a future that aligns with human values and aspirations by integrating a societal lens and promoting future thinking and co-creation with users.

The key distinction between the human-centric paradigm and the future-centric paradigm lies in their respective considerations of impacts. While the human-centric paradigm primarily focuses on the immediate consequences, the future-centric paradigm takes a long-term perspective into account. In addition, the future-centric perspective looks at how technology *should* be developed. To effectively address this forward-looking approach and specify the implications of factors, such as social justice and explainability, a conceptual dimension is required. One valuable tool in the realm of responsible innovation is conceptual engineering [148], which has been

integrated to fulfill this purpose. The future-centric perspective, therefore, starts with a critical understanding of the current state of AI technology as well as a (normatively) good portrayal of the future society it aims to achieve. It then aims to help steer technological innovation to realize this future portrayal.

The second perspective is *self-actualization*, which aims to support users in discovering, developing, and pursuing their longer term goals [32, 72]. This perspective recognizes that adapting a system to users' preferences is bound to fail if these preferences are underdeveloped. Moreover, it acknowledges that numerous existing systems concentrate primarily on users' short-term interests rather than longer term goals. Focusing on self-actualization requires the development of systems that do not replace but rather support users' preference development and decision-making practices.

Future-centric research methods are characterized by a *reflexive* approach that meaningfully involves end-users' perspectives in the practice of developing and evolving AI technology through methods like critical theory [48, 89], value-sensitive design [43], and Design for Values [6, 53]. Within this perspective, there is also increasing attention to the fact that our values change over time (e.g., sustainability is considered to be much more important now than it was 40 years ago), prompting the need for methods to design for value change [107]. Furthermore, researchers in this paradigm conduct longitudinal studies to study the longer term personal and societal outcomes of AI system implementation [36]. These approaches entail prioritizing the ethical and social implications of AI systems and ensuring that they align with the values and needs of the wider community in the long run.

3 A Framework for Scoping Human-Centered AI

In the previous section, we presented the paradigmatic perspectives on AI systems. This section dives deeper into a comprehensive framework, offering an overview and a structured approach to understanding and implementing human-centered AI. This simplified framework consists of four dimensions: (i) Paradigm, (ii) Actors, (iii) Values, and (iv) Level of realization. These dimensions encompass several other components, which are described in this section. This framework suggests a four-dimensional space in which every point can be characterized by each of the four dimensions.

3.1 Paradigm

The paradigm forms the foundation of our framework and plays a vital role in comprehensively analyzing the AI system and its impact. Our framework adopts four distinct paradigms: the technology-centric perspective, the user-centric perspective, the human-centric perspective, and the future-centric perspective. As detailed in the

preceding section (Sect. 2), each subsequent perspective offers an increasingly comprehensive viewpoint on the system and its impact on the involved actors. Moreover, the paradigms contribute to integrating and prioritizing the corresponding expectations and requirements.

To illustrate the significance of these perspectives, let us consider the concept of *bias*. When adopting a technology-centric perspective, the focus is on understanding how biases can be addressed throughout the data collection and preparation stages. The emphasis may be on the utilization of bias-aware techniques and mitigation strategies. Furthermore, the outcomes are evaluated quantitatively using the bias metrics defined within the system. A user-centric perspective may suggest capturing user perception of bias through interviews or surveys conducted after the users interact with the system in a live setting. From a human-centric perspective, the definition of bias would be influenced by moral and legal guidelines. Moreover, it may recommend conducting both quantitative and qualitative studies involving diverse populations and actors, regardless of their interaction with the AI system. This inclusive approach can ensure that biases are identified and addressed for all individuals, promoting fairness and equity. Finally, a future-centric perspective would focus on approaches that strive to bring us closer to an ideal future, e.g., for the betterment of humanity. This perspective requires a substantive theory of social (likely to consist of distributive and procedural) justice that is linked to long-term thinking. It takes a prescriptive approach and considers the potential impact of bias mitigation techniques and policies on society as a whole. From the conceptual work on justice, it would then connect to actions on an algorithmic and sociotechnical level to ensure a more just society in the future.

3.2 Actors

The second dimension of our framework focuses on the involved actors. Within this context, "actors" refer to a broad range of individuals or entities, as well as stakeholders, as well as other individuals or any parties influenced by either the AI system, its use or users, or the impacts arising from system-user interactions. This expansive view encompasses all those affected by, or impacting on, the AI system's deployment and operation, either directly or indirectly.

The AI landscape often includes two main types of actors: *individuals* (e.g., end-users, data subjects, and AI developers) and *organizations* (e.g., private technology companies, UNICEF, and legal bodies) [27, 151]. Since the actors involved in an AI system often have divergent or conflicting values and goals, the integration of AI systems can lead to tensions and conflicts in terms of the design and prioritization of requirements stemming from these divergent values.

For example, consider the scenario of a new AI system that aims to enhance services and products or provide a personalized experience by collecting more user data: this data collection benefits the business (i.e., it increases profits) but may face objections from other actors, such as regulatory bodies. Individuals and business

parties often have different goals (e.g., see [62]). Individual values can also clash with business goals. For instance, a company might opt to maximize profits by using an AI system that monitors employees' working hours or surveils their behavior. And yet, such a decision could be perceived as unethical from the standpoint of some employees, potentially violating their rights and leading them to voice their concerns, quit their jobs, or experience negative effects on their mental well-being and trust. Even among the individuals involved in an AI, there could be different views on what AI should do. For example, the general public (non-expert AI users) may want an AI system to be safer and protect their privacy, while an AI researcher may heavily focus on creating fairer AI systems [61]. Finally, even a single individual may have conflicting opinions about an AI application. For instance, a web user may dislike the advent of widespread online activity monitoring but appreciate its contribution to national security.

Legal and governing bodies have the difficult task of reconciling these conflicting opinions—a task that is further hampered by an invariable lack of a coherent and consistent approach to AI-related regulations and guidelines, as they struggle to keep pace with technological advancements and the widespread use of AI. Furthermore, the existence of conflicting regulatory approaches makes developing AI systems that may be used cross-border and considering their social impacts particularly challenging. Data protection regulations, for instance, have only been established in some countries since the early 2010s, while in many other regions, they are still under discussion. Similarly, the global discourse surrounding AI-related regulations is relatively recent. Many countries are still in the process of assessing how to manage and handle AI (mis)use cases, highlighting the weight of establishing responsible AI development frameworks.

Given the widespread use of AI systems, a category of actors known as malicious actors can emerge, which refers to individuals or organizations that intentionally exploit or misuse AI systems for nefarious purposes. For example, consider hackers, cybercriminals, and state-sponsored entities. Malicious actors may intentionally abuse or use the system for harmful purposes, such as criminal behavior, spreading misinformation, conducting attacks, and exerting control. When designing AI systems, it is crucial to consider misuse scenarios and assess associated risks, ensuring the systems' responsible and ethical development while mitigating potential harms.

The dynamics and relationships of various actors can add further complexities. Interactions between different individuals, personas, and organizations may involve bidirectional impacts, collaborations, competitions, negotiations, and conflicts and even result in a change in values and priorities for actors. Therefore, it is essential to foster responsible, effective, useful, and human-centered AI systems to consider the dynamics and relationships of all actors on top of each actor's characteristics and needs and to address conflicts to achieve the desired societal outcomes.

3.3 Values

The third dimension of our framework evolves around *values*. In the literature, a number of values are employed to specify or discuss human-centered AI. Distinguishing between these values can be challenging due to their overlapping components and the frequent ambiguity surrounding their discussions. We argue that these values are not absolute, and additional values may be relevant depending on the specific context. Nevertheless, we aim to offer an overview of the most frequently addressed values in the (human-centered) AI community.

We start this discussion by acknowledging that each relevant value in the context of human-centered AI systems corresponds with one or more basic human values. We build upon a foundation laid by the work of Frankena [41], who presented a comprehensive list of intrinsic human values, and proposed a list of human values (listed in Fig. 2) that can be utilized for a need-based design of AI systems.

The values described in this section are often referenced in both legal settings (e.g., in the EU AI Act [25], UNESCO Recommendation on the Ethics of AI for protecting and promoting human rights [146]) and in surveys of AI ethics (e.g., [55, 64, 94]). In order to align with and promote human values, AI systems should adhere to these values, which includes—but crucially goes beyond—complying with these ethical and legal requirements. Additional context-specific values may also be relevant in practice. From a normative design perspective, understanding and operationalizing various social norms and values in the AI system is key to ensuring that the system objective aligns with human needs. Therefore, these values are design-consequential and are linked to concrete (socio)technical requirements and tools. Context-specific values can be elicited and accounted for using methodological approaches such as value-sensitive design [43] and Design for Values [6, 53]. These approaches typically start with a step to identify relevant (normative) values for the specific system and a specific group of people. The relevant values are then translated into concrete design requirements, realized in prototype systems, and evaluated for their effectiveness.

Engaging with values in design is complicated by the fact that they often conflict in practice. For example, improving the fairness of an AI system may require increased access to sensitive information, which can, in turn, negatively affect the value of preserving privacy. Providing a fair distribution of benefits (or harms) can also lower the overall accuracy of the algorithm, thereby negatively affecting the values related to functionality and user experience. Resolving such conflicts between values can be approached in different ways [106]: (1) maximizing the score among alternative solutions to the conflict (for example, assigning explicit weights to the values of privacy and fairness, then optimizing for the sum score); (2) satisficing among alternatives (for example, defining a lower bound of fairness and privacy, and finding a solution that provides a sufficient level of both); (3) re-specifying design requirements to ones that still fit the relevant norms but no longer conflict (for example, defining fairness in a way that no longer requires additional data-sharing); (4) innovating on existing approaches to overcome the conflict between values (for example, by using synthetic data in fairness measures to circumvent the privacy/fairness conflict), among other

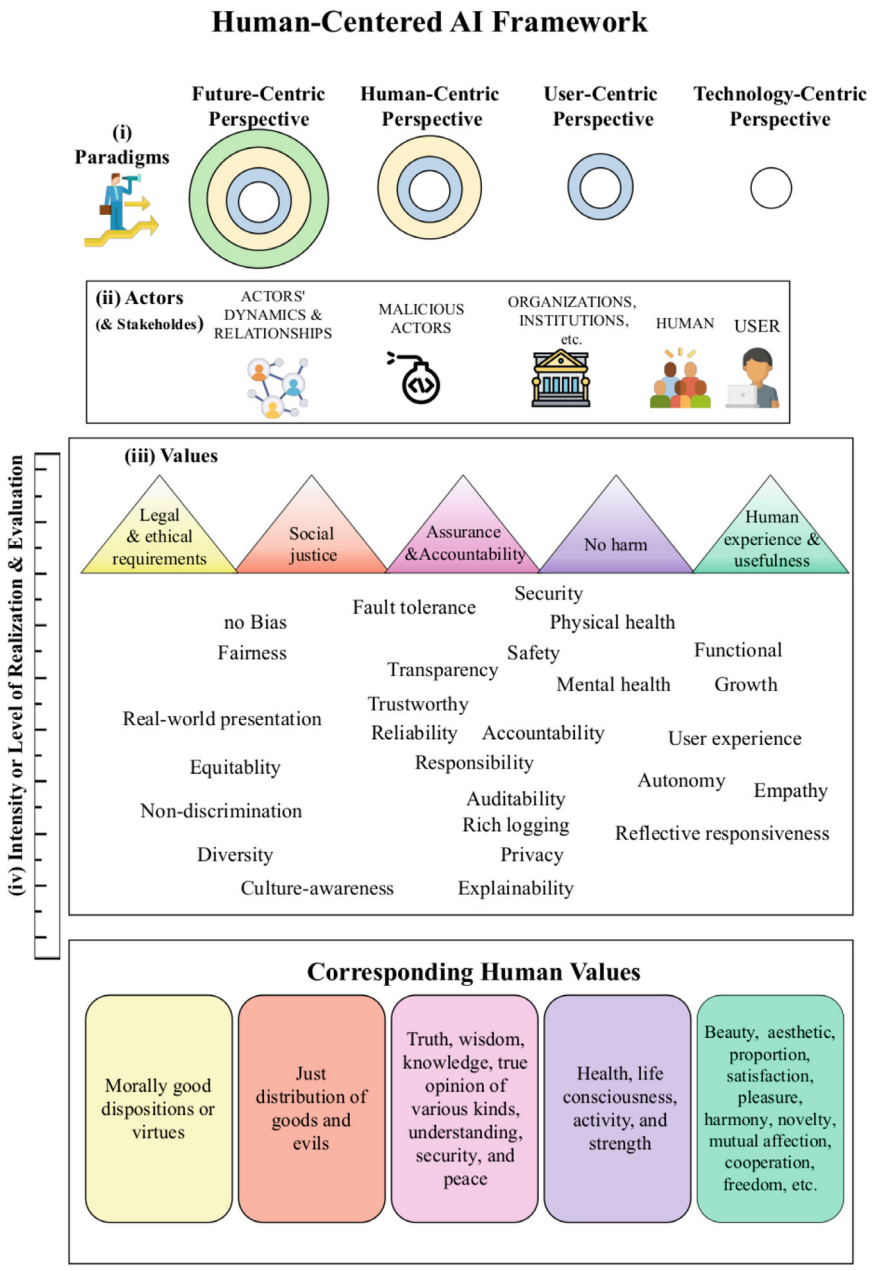

Fig. 2 This diagram illustrates a simplified framework for human-centered AI comprising four dimensions: (I) Paradigm, figure has changed. Similar font to the text is now used in the updated version point within the space represents a specific combination of these dimensions. The box on the bottom displays an overview of human values corresponding to the framework's values

ways. In all these cases, choices have to be made from the four different paradigms in order to find a responsible way to handle conflicts. To further show how these choices are made and particularly how the perspective from each paradigm influences the treatment of different values, the rest of this section highlights a number of important values for human-centered AI systems. This list is not meant to be exhaustive but does give a sense of the many values that are explored in this area.

3.3.1 Preserving Security, Privacy, and Safety

AI enables massive data analysis, encouraging more data collection and raising privacy concerns. Privacy has been a hot topic among academics and regulators for a while, e.g., [7, 13, 50, 57, 83, 139]. Smartphones, apps, e.g., social media, big tech, and advertising companies were and are among the foremost data collectors. These data can make users feel uncomfortable while also causing severe harm [138]. For example, the cases of intimate partner abuse of a lack of usability and affordances in security and privacy controls of smartphones are still an issue [84]. Moving toward more automated decision-making systems with higher impacts, which require even more user data collection to provide decisions, exacerbates the situation. There are already several cases of controversies, e.g., related to the use of facial recognition or predictive policing, concerning the violation of fundamental rights [46].

Preserving privacy, security, and safety values are linked to several other human-centered AI values, such as the trustworthiness of the system and its reliability. These values correspond with several human values, such as peace, security, power, freedom, honor, esteem, and good reputation. Violating privacy, security, and safety values often leads to catastrophic consequences, contradicting the principle of avoiding harm. The approach to building a resilient and safe system depends on various dimensions of the presented framework. The specific measures required depend on the actors involved, the paradigms employed, and the level of value realization.

For example, given a malicious actor and taking a technology-centric perspective, various poisoning or adversarial attacks, e.g., [140], could be investigated and considered, which may motivate the desire to opt out of various data collection practices. On the other hand, a user-centric perspective may focus on design guidelines that provide users with a sense of safety and control, minimizing errors and harmful actions (e.g., [122, 152]). A human-centric perspective would consider the data sources underlying the models and ensure the system's safety for society [56, 69]. Meanwhile, a future-centric perspective would consider long-term impacts and propose additional preventive measures that are realizable on different paradigms, such as conducting long-term privacy impact assessments or prioritizing privacy from the outset rather than an afterthought [22, 109].[1] In addition, the extent to which these

[1] Note that the examples provided in this section are merely a few illustrations, and a comprehensive discussion of the various approaches, methods, evaluations, and resulting guidelines extends beyond the scope of this chapter. We have only presented a concise overview of one example per paradigm for brevity (for more details on privacy, see [68]).

values are realized and for which actors the security, safety, and privacy measures are designed are crucial factors that determine the appropriate provisions.

3.3.2 Equity, Fairness, and Controlled Bias

Ensuring equity and fairness and removing (or mitigating) bias are crucial values in human-centered AI, interconnected with other values, such as trustworthiness, transparency, non-discrimination, social justice, legal and ethical compliance, and avoiding harm. These values align with *human values* of "morally good dispositions or virtues" and, more specifically, the "just distribution of goods and evils".

Despite their significance, AI systems, particularly those that involve human data, have faced challenges and controversies in achieving equity and fairness (examples highlighted in [54, 85, 102]). At first glance, it may appear contradictory that AI applications, which aim to offer optimization and efficiency through objectivity, struggle with issues of equity and fairness. However, the pursuit of objectivity alone cannot prevent biases in society from permeating into the recommendations. An illustrative example of this phenomenon is the bias reinforcement effect observed in recommender systems, where biases in the training data are reinforced algorithmically (cf. [29]).

A large body of work within this space has focused on what is considered allocative harm—that is, adverse impact from interactions with an AI system that affects people from a particular social group where those people are unfairly deprived of access to opportunities, resources, or impact resulting in economic loss [12]. Allocative harms move well beyond mundane content filtering tasks as AI systems are being leveraged to help determine which neighborhoods to police [19], who gets employment benefits [9, 51], who gets picked out from a crowd for a criminal investigation [147], and which unhoused person should get housing [144]. These harms could result in what Mimi Onuhoha describes as algorithmic violence as algorithmic systems withhold information, opportunities, or resources that affect people's well-being further "preventing people from meeting their basic needs" [103]. Another type of harm caused by unfair systems is representative harm—that is, AI systems may replicate the subordination of social groups regarding aspects of their identity (e.g., [127, 128]).

Often, addressing these issues has been an afterthought, only brought to attention through independent studies or by tracing the harmful effects they have caused, e.g., [26, 85]. The harm resulting from a lack of equity extends far beyond immediate consequences and can have long-lasting and far-reaching societal impacts.

The definition and criteria for fairness have evolved over time [58]. Different definitions and criteria of fairness and types of biases manifest across various paradigms. For example, biases related to data or algorithms may be addressed through a technology-centric lens [31], while behavioral biases or user-perceived biases are better approached from a user-centric perspective [134]. However, these categorizations are not absolute, and achieving fairness driven by a human-centric lens may

necessitate modifications in algorithms or data. A survey by Mehrabi et al. [90] categorizes biases into three groups: data-to-algorithm biases (e.g., measurement bias, representation bias, sampling bias), algorithm-to-user biases (e.g., algorithmic bias, presentation bias, popularity bias), and user-to-data biases (e.g., historical bias, population bias, behavioral bias). A feedback loop between the user, algorithm, and data represents the interplay between biases and their definitions.

When viewed through a human-centric lens, the public's perception of fairness may impact its technical realization, or perhaps, the notion of group fairness gains more attention than individual fairness. Equity, fairness, and bias can also be viewed through a future-centric lens, for instance, looking into the long-term fairness goals [124]. Additionally, the definition and criteria for fairness and bias can vary when considering the actors involved. For instance, fairness can be categorized based on the question of "Fair for Who" [150, p. 52:10].

3.3.3 Transparency and Explainability

Transparency and its associated value *explainability* afford opportunities for stakeholders to better understand how a model or system works (i.e., the process used to generate output given a certain input). Transparency serves as an opportunity to communicate the strengths and limitations of the system, which provides not only informational merit, but also significantly improves humans' *informational agency, and relational and systemic practices* [34]. Thus, both transparency and explainability work in tandem with other connected human-centered AI values, including safety, auditability, and trustworthiness that directly map with a core need for understandability in human values.

From a technology-centric perspective, transparency and explainability have traditionally been of relatively low priority. Recently, though, developers of AI algorithms have realized the importance of demystifying the increasingly complex models used in intelligent systems, which have caused some to revert to simpler, more inherently explainable algorithmic solutions.

Taking a user-centric perspective moves the focus to individual users of the system whose information is being processed. A recent surge of research in the explainable AI (XAI) field has been fundamental in the production of interpretable explanations that either (a) directly expose the inner workings of a system, or (b) provide post-hoc explanations for black-box models that cannot be accessed [92, 95]. In determining the appropriate content of explanations, scholars have put forward the following questions as grounding starting points: "(1)*What did the system do?*, (2) *Why did the system do it?*, (3) *Why did the system not do X?*, (4) *What would the system do if Y happens?*, (5) *How can I get the system to do Z, given the current context?* and (6) *What information does the system contain?*" [45, 81]. Attempting to answer these questions, user-centric frameworks of explainability have posed pathways for the design and development of tailored explanations for different users to avoid potential cognitive biases through presentation styles with appropriate levels of data abstraction and design values that are relevant to the users [116].

Meanwhile, human-centric approaches have been largely interdisciplinary to effectively operationalize values of transparency and explainability to alleviate ethical concerns and meet regulatory compliance [117]. Working closely to harmonize values of transparency with a human-centered lens, various parties have invested in examining the reasons why people seek explanations in AI systems. This foundational work has led to the development of various taxonomies that demonstrate a connection with human values related to knowledge acquisition, trust development, satisfaction, and user-perceived factors such as scrutability (the ability to tell the system it is wrong) that facilitate control [101]. This paradigm also acknowledges that transparency and explainability needs may vary significantly for various actors. For instance, beyond transparency for end-users, specialized transparency mechanisms and tools could also be tailored toward developers and practitioners—these actors, who likely have a heightened level of responsibility, seek deeper engagement with their models to shed light on system behavior, algorithmic parameters, openness with datasets, and data processing techniques to set appropriate benchmarks and identify areas for further refinement [3].

To further support the varying need to understand systems, frameworks such as Transparency by Design (TbD) have been proposed to offer guidelines to achieve transparency throughout the development cycle [34]. However, gaps still persist in understanding how, as values, transparency and explanations could be further developed within future-centric discourse to propel responsible innovation. Importantly, future-centric research should critically consider not only *what* information should be relayed in an explanation, but also *how* that information is relayed, and *who* is receiving that information. In answering these questions, transparency mechanisms can take a longer term perspective by providing space to justify system behavior, offer room for improvement, push moments of discovery, and impart pathways for control [3].

3.3.4 Reliability, Accountability, and Auditability

In order to allow for effective oversight and control over the development and use of AI systems, it is important to have strong accountability and quality assurance. Audits can give us both: they lead to assurances about the functioning of AI systems, and they enable us to identify mistakes and undesirable effects and then make improvements. Concerning the corresponding human values, these human-centered AI values contribute to the need for peace and security, truth, and ensuring the realization of other human-centered AI values, such as privacy, security, safety, transparency, equity, and health.

The very idea of accountability, namely, is that there are clearly identifiable people responsible for explaining mistakes and responsible for fixing them. In certain cases, it may involve apportioning blame, for example, for not conducting sufficient safety tests or ensuring proper reliability of systems. However, the primary focus of accountability is to prevent undesirable outcomes (through assurances) and to rectify

any issues that arise (through accountability). As with the other values, these values filter through the different dimensions: actors, paradigms, and levels of realization.

The emphasis on accountability and transparency is reflected in new legislation, such as the EU AI Act [25], which aims to establish a Europe-wide ecosystem for AI auditing based on specific standards, including non-discrimination and robustness, among others. While the legislation itself employs broad, human- and future-centric language, it will be accompanied by detailed technical standards from technology- and user-centric perspectives. Although these technical standards are still being developed, various frameworks for conducting audits have been proposed, e.g., [79, 114] in anticipation of the legislation, e.g., [40].

These frameworks cover technology-centric information, such as the maintaining logs, Datasheets [44], and Model Cards [93], which provide insights into the technical aspects of AI systems. They also consider user-centric perspectives [78] to ensure that auditing involves end-users and to ensure the reliability and accountability of AI models. Additionally, more human- and future-centric social impact assessments have been proposed, such as the Dutch "Fundamental Rights and Algorithm Impact Assessment" [157] and ethics-based auditing methods [97]. This assessment framework focuses on the potential impact of AI systems on human rights.

3.3.5 Health

The essence of responsible design, in the first place, is to *do no harm* and, even further, to *do good*. While all values discussed in human-centered AI, address this essence to some degree, a specific case of doing no harm is explicitly concerned with individuals' mental and physical well-being. A human value relevant to this principle is preserving and promoting health and strength. AI systems can impact human health in various ways.

AI systems can be purposefully designed for health applications, serving various purposes, such as medical image interpretation [115], e.g., detecting skin melanoma [5, 105], providing decision support for healthcare professionals [100], or providing health recommendations [125]. These applications have the potential to significantly influence the health outcomes of the users they serve.

From a technology-centric perspective, realizing the idea of "do no harm and do good" can be achieved through algorithm design and evaluation. An illustrative example, of doing good involves personalizing behavior change recommendations according to a user's ability level, resulting in promoting mental well-being [143]. There is also increasing research investigating various algorithms and techniques in different diseases striving for better prediction and detection accuracy [77]. It is crucial to note that a false negative outcome in many health applications can have life-threatening consequences.

From a user-centric perspective, this value can be achieved through user-centered design practices and evaluation, prioritizing the well-being and safety of individual users. For example, a prescriptive or descriptive presentation of biased recommenda-

tions impacts human judgment in a mental health emergency decision-making [4]. A human-centric perspective takes into account other actors, including the caretakers of the patient, during the design and evaluation process. For example, it is essential to consider how AI systems can impact public health and whether their use can extend to national health systems. Furthermore, human-centric design ensures the incorporation of ethical and legal principles and the consideration of sociotechnical implications.

A future-centric perspective necessitates considering how the design and use of an AI system today can impact, over time, the physical and mental health of multiple actors and stakeholders, e.g., patients, healthcare workers, caregivers, their social circles, and communities encompassing present and future generations. This involves addressing ethical considerations on what it means to do no harm and what obligations we have toward others. Part of adopting a future-centric perspective involves investigating how to approach the ideal situation in which the 'no-harm' values are met as effectively as possible. This investigation has to be connected to the actual impacts on actors in the context of the human-centered AI framework and consider technical possibilities to reduce negative consequences in the future.

In addition to AI health systems, general AI systems should also adhere to this value. For example, consider a flawed self-driving car that can cause stress or physical harm. Especially when systems are not necessarily built for the health application domain, it is possible for their impact on physical and mental health to be overlooked. For example, highly accurate recommendations may have potential negative emotional consequences on users [142].

Last but not the least, it is our responsibility to acknowledge that when a system is released to the public and integrated into everyday applications, users may employ them for unintended purposes. For instance, users may use ChatGPT to seek safety-related advice that can cause harm [104]. Therefore, it is essential to remain vigilant in our design and evaluation processes to mitigate potential negative impacts on well-being and promote responsible, human-centered, and ethical AI use.

3.3.6 Human Growth and Reflective Responsiveness

These values emphasize the importance of thoughtful and introspective engagement with users in the design and operation of AI systems, considering their goals and, accordingly, helping them grow. They also go beyond immediate responses and encourage AI systems to consider user needs, goals, satisfaction, preferences, feedback, and experiences in a reflective manner. These values may address human values related to harmony, proportion in objects contemplated, pleasure and satisfaction of all or certain kinds, happiness, beatitude, contentment, harmony and proportion in one's own life, mutual affection, love, friendship, cooperation, power and experiences of achievement, self-expression, freedom, and esteem.

The value of growth and reflective responsiveness recognizes the complexity of human interactions and may involve empathy, understanding, personalization, and continuously learning and adapting to ensure meaningful and responsive interactions.

For example, empathy focuses on affect-aware and affective system design, considering the thoughts and experiences of individuals and communities. This ensures that the system understands and reacts appropriately to fulfill human-centered objectives. It also effectively enhances or diminishes certain affect, thoughts, or experiences in alignment with desired outcomes. Furthermore, empathy should be taken into account when actively involving users in developing AI systems, conducting user research, and need assessment. This entails being compassionate and understanding and demonstrating empathy toward the user's context and values, and providing personalized assistance to individuals based on their unique needs and circumstances.

3.3.7 User Experience

User experience is another critical value to consider in human-AI systems. It not only directly impacts the user adoption and overall success of the technology, but it also fosters trust, aids interaction outcomes, and ensures value alignment. This value corresponds with human values of satisfaction of all or certain kinds, adventure and novelty, and aesthetic experience. By considering user experience, developers can drive continuous improvement and enhance the overall effectiveness of AI technologies, ultimately benefiting both individuals and society. For example, in recommender systems, user experience influences the user's decision to continue using the service [74] and long-term user adoption enables more robust user modeling [130]. While user experience has not received substantial attention in most technology (or system)-centric AI research, technology aspects are certainly essential determinants of the user experience of human-AI systems [74]. Particularly, the user experience of AI systems is influenced by factors such as the system's performance, reliability, and adaptability [24]. An AI system should consistently deliver accurate results, be responsive, and minimize errors or failures to meet user expectations. Additionally, the system should be adaptable, adjusting to changes in user needs and preferences over time.

A user-centric perspective takes a more subjective approach to user experience, and emphasizes the importance of usability (including factors, such as perceived novelty and aesthetic experience, among others) and user experience in achieving *user satisfaction*. A well-designed AI system should be easy to use, with intuitive interfaces, clear instructions, and helpful feedback mechanisms. Several frameworks exist that can guide researchers in measuring user satisfaction and its impact on user's behavioral intentions [74, 110]. Personalization plays a significant role in enhancing user satisfaction, as AI systems that cater to individual preferences and needs provide personalized content, recommendations, and assistance based on user data and behavior patterns. At the same time, AI researchers must acknowledge the limitations of personalization and limit its potential negative effects, such as the filter bubble [91], choice overload [18], and negative emotional impact [142].

From a human-centric perspective, user experience encompasses the satisfaction of all individuals, considering users as humans not only during their interaction with the system but also beyond that time. Additionally, it takes into account the expe-

riences of non-users and uses participatory design and co-creation where possible. The human perception of values can be influenced by their individual characteristics and interaction experiences. For instance, fairness perception is impacted by the presented explanations and their style as well as individual characteristics, such as personality and demographics [134]. Designers must therefore consider individuals with different characteristics, namely cultural backgrounds, abilities, and expectations when creating AI systems [15].

A future-centric perspective on user experience emphasizes the long-term impact of AI systems on society. This perspective acknowledges the importance of designing AI systems that not only meet the needs of current users but also consider the potential consequences and implications for future generations. Important in this regard is to avoid the tendency of AI systems to tailor to users' immediate needs rather than their longer term goals and desires [32, 72].

3.3.8 Autonomy

The value of autonomy is closely tied to human values such as power, freedom, self-expression, and experiences of achievement. It encompasses the ability to exercise free will and make independent decisions regarding one's objectives and actions [120]. Autonomy has been recognized as a fundamental value in human-centered AI across various frameworks, e.g., [20, 39, 132].

Similar to other human-centered AI values of our framework, autonomy is intertwined—and/or in conflict—with other values. For instance, the technology-centric perspective has long presumed that the epitome of AI functionality is a fully autonomous system—a goal that often directly contradicts the value of human autonomy [38]. Furthermore, when examining AI systems from a user-centric perspective, users may perceive a diminished sense of autonomy in situations where their privacy is at risk [119], or when a lack of user experience may make the system too difficult for novices to operate [70]. Depending on specific dimensions of our framework, such as the levels of autonomy, the actors involved, and the paradigmatic perspectives, different aspects and manifestations of autonomy may be realized. This may include user (or actor) control over the decision-making process, goal definition, pursuits, and realization method, e.g., choice for data collection, data utilization for modeling, and personalization options aligned with their goals. It can even extend to adaptive and controlled automation, providing users with a sense of empowerment and control as a result. Ensuring the preservation of autonomy in AI system design and deployment and striking a balance between automation and control are keys to building truly human-centered AI.

Finally, a future-centric perspective on autonomy may aim to move AI systems beyond consumerist perspectives, to a perspective that encourages discovery and personal growth [72]. Crucially, this may require systems to increase users' level of autonomy, so as to support rather than replace their decision-making practices.

3.3.9 Functionality

Basic system functionality serves as a core value in AI systems, providing the foundation for the realization of other values. Without functionality, a system cannot fulfill its intended purpose.

From a technology-centric perspective, functionality is often seen as synonymous with algorithmic accuracy: a system fulfills its function if and only if it provides accurate outputs (which are usually tested against some ground truth values, if available). An important question is *how accurate* a system must be to provide useful functionality. In some cases, any predictions that are better than random make a valuable improvement. In other cases, inaccurate predictions can be actively harmful—especially when the predictions are biased (see Sect. 3.3.2), inscrutable (see Sect. 3.3.4), or require a disproportionate amount of sensitive data to be produced (see Sect. 3.3.1).

Functionality can be viewed from other paradigmatic perspectives as well, offering valuable insights into its evaluation and realization. For example, a user-centric perspective may require looking beyond accuracy and algorithm [87]), a human-centric perspective may require adequate functionality concerning specific legal or moral requirements, and a future-centric perspective may critically reflect upon the purpose of AI systems, suggesting a shift toward a different set of evaluation criteria depending on the societal values that the system is supposed to adhere to [14, 66].

3.4 Intensity or Level of Realization and Evaluation

Values in our human-centered AI framework have different *intensities* or *levels of realization*, presenting a fourth dimension. For example, fairness can be realized to a greater or lesser extent, and responsibility can be more or less well-organized (e.g., a wider or narrower range of cases where a responsible party can be identified). This variation can affect compliance with legal frameworks and associated (legal and moral) risks. To ensure compliance and an appropriate level of realization of all relevant values, it is crucial to systematically incorporate values in design across all paradigms and openly communicate those values and their realization level. Previously in this chapter (Table 1), we elaborated on the methods, scopes, and success criteria associated with each paradigm, building upon the preceding one and outlining how each value should be assessed, and their success should be determined. However, it is important to note that the assessment and determination of success for and realization of each value, derived from the third dimension, are inherently reliant on the interplay with the other dimensions as well.

To illustrate, let's consider a specific actor, e.g., an individual user, for the second dimension and adopt a human-centric or future-centric perspective for the first dimension. A value-sensitive design approach proposes identifying the relevant value and subsequently incorporating its requirements into the system prototype, which can

Table 1 An overview of the four paradigms of human-centered AI

Paradigms	Technology (or system)-centric	User-centric	Human-centric	Future-centric
Scope	Data and algorithms	User-system interaction, input and output mechanisms, and the effects of user characteristics	Multi-actor perspective (content producers, data subjects, underprivileged), human factor, the impact of the system on all people and society as a whole, responsible innovation, and immediate consequences	Working beyond current perspectives, future-oriented responsible innovation, self-actualization
Methods	Offline evaluation, e.g., ML studies using benchmark dataset	User-centered design, user research, user-centric evaluation, e.g., interviews, surveys, observations, experiments, and online field trials	Human-centered design, social computing, stakeholder analysis, implementation studies, participatory design, co-creation	Reflexive methods (e.g., critical theory, design for values, value changes), longitudinal studies
Success criteria	Accurate predictions, algorithmic metrics; quantitative behavioral metrics (only in live studies)	Usability, user experience, user perception, e.g., of transparency and privacy	Societal good, equitable outcomes, etc.	Value alignment, e.g., for values such as sustainability and human agency
Limitations	• Historical data is used as ground truth • Adaptation is assumed to be desirable • Human behavior is assumed not to be influenced by the adaptations • Near-exclusive focus on objectively measurable properties of algorithms	• No concern for people who are not users • No focus on collective outcomes and societal values • Often limited to narrow and non-holistic user-centric approaches	• No future perspective • Limited focus on agency	• Difficult to anticipate future changes in values and society • Conceptual methods can remain too abstract, missing a link to concrete design of AI systems

then be evaluated in terms of its realization. Consequently, this approach enables the determination of the *intensity* or *levels of realization*, success criteria, and maturity of that value. Different intensities or levels of realization of a value, like *explainability*, can be identified as individual user, contextualization, and self-actualization [141], and accordingly, different metrics can be utilized to determine its success of realization. It is worth noting that these levels may be defined differently for industrial actors. Levels of realization of a value, like *autonomy*, can vary in terms of the degree of user involvement and having a say, depending on the actor, paradigm, or considered context within the respective paradigm.

The evaluation of human-centered AI realization varies depending on the particular perspective, actors involved, values considered, and defined level of realization. Depending on these criteria, the AI system can be evaluated, and its human-centeredness can be determined using a combination of quantitative and qualitative approaches. The defined intensities or levels of realization might have different steps that, upon evaluation, reveal the extent to which the AI system has achieved the envisioned maturity for a particular value.

In summary, although an ideal system definition may be proposed at some point, the fourth dimension suggests that human-centeredness can vary for each AI system and should be carefully determined by thoroughly considering other dimensions. For instance, the required intensity of legal compliance (including applicability, necessity, and realization) may differ among AI systems based on their functionality, context of use, actors, regional variations, and other relevant considerations. Such understanding calls for in-depth analysis across all dimensions for an accurate assessment.

The crucial understanding is that while the concept of human-centeredness may be subjective and vary from one AI system to another, it's essential to identify, document, and transparently communicate the characteristics of each AI system. This approach ensures reliability and transparency in evaluating the system, and its limitations and impact.

Concluding this section, our framework, informed by a range of existing literature, serves as a flexible and practical tool for researchers, AI teams, and companies. This framework is designed to help understand the scope and limitations of the solution under development, diverse actors considered or involved, their values, levels of realization, and the corresponding human values that are pertinent in this domain. At an organizational level, this framework offers a structured approach to grasping the requirements of each dimension as well as dynamics between various actors involved, which can foster a culture of responsibility and accountability. Meanwhile, at a more granular system level, such as within product, design, and development teams, it facilitates the alignment of AI systems with human experiences and usefulness, ensuring that the values ingrained in AI products and services are closely aligned with human-centric principles. This can be achieved both by following broad or specific strategies for implementation, following the procedure that results from careful analysis and integration of the framework's key elements.

For the research community, the framework can serve as a reflexive tool, empowering researchers to critically assess their work's position within the broader human-centered AI landscape, identify potential gaps and areas for improvement, and strategically plan future research initiatives. By following this structured approach, AI practitioners and researchers can ensure their work not only aligns with technological advancements but also resonates with human values and societal needs. Having explored the details of our framework, we now turn our attention to other human-centered AI frameworks in the literature. This comparison will help highlight the unique contributions of our approach in bridging technological progress with human-centric values.

4 Human-Centered Design of AI Systems: Models, Techniques, and Frameworks

To promote the human-centered design of AI systems, various models, techniques, and frameworks have emerged in the literature. While these approaches offer valuable guidance and methodologies, it is important to note that they are not always comprehensive. Developing a holistic understanding of human needs and values in relation to AI technologies is a complex task, and these approaches may encounter challenges in achieving complete integration. Nonetheless, they play a crucial role in facilitating the incorporation of human perspectives, considerations, and requirements throughout the AI development lifecycle. By drawing from diverse disciplines such as human-computer interaction, software engineering, ethics, psychology, and social sciences, these approaches contribute valuable insights to the pursuit of a more human-centered approach to AI design.

4.1 Models and Frameworks

One notable framework is Shneiderman's framework for Human-Centered AI (HCAI) [131–133]. This framework offers a two-dimensional perspective, separating levels of automation from levels of human control. It challenges the traditional one-dimensional view and suggests that high levels of human control and computer automation can be achieved through careful design. Additionally, it incorporates a governance structure that provides guidance on adapting software engineering practices, followed by management strategies and independent oversights for trustworthy certification.

The HCAI framework argues for balancing human control and AI autonomy, acknowledging that automation can occur even with a high level of human control. This can result in AI becoming trustworthy, safe, and reliable. An example provided

by Shneiderman is a medical AI system controllable by patients and clinicians. Instead of trying to replace humans with machines or to signal that idea to society, researchers should focus more on augmenting and enabling humans with AI. Another simple example Shneiderman discusses is a navigation tool where humans have control over where they want to go or which transport mode to choose (e.g., walk or drive), but they receive multiple suggestions and routes from the AI. Similarly, when a traffic jam happens on the selected route, it is still a choice for humans to decide whether to take the alternative route suggested by the AI or to stay on the current route [132, 133]. In essence, the HCAI framework advocates a shift from emulating or copying human behavior toward helping and supporting humans.

The HCAI framework is a framework with an emphasis on reliability, safety, and trustworthiness values. Such an ecosystem requires three levels of governance to keep AI trustworthy, safe, and reliable. First, AI development teams, such as AI engineers, must implement workflows that consider audit and ethics in developing AI systems. Several tools, such as datasheets, have been proposed to help developers in these tasks. Second, organizations must implement procedures such as an internal review board (similar to an institutional ethics board) to promote a culture of responsibility and ethical innovation at the leadership level. And third, independent audit bodies should be able to audit organizations and teams for their AI systems [132].

Besides Shneiderman's framework, other frameworks advocate for responsible approaches to building trustworthy AI systems. For example, the US National Institute of Standards and Technology (NIST) has released an Artificial Intelligence Risk Management Framework that outlines essential attributes of a trustworthy AI system, which are "valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy-enhanced, and fair with harmful bias managed" [113, p.12]. Other auditing frameworks, such as ethic-based auditing [97] and internal auditing of ethical and technical aspects [114], as well as standards (e.g., [2, 59, 60]) and regulatory guidelines (e.g., [136, 158]) also contribute to the development of trustworthy AI.

There are also frameworks dedicated to specific values, e.g., addressing key areas of responsibility, autonomy, ethical considerations, explanations, or transparency. These frameworks zoom in on specific challenges and requirements within these domains. One example is the AI4People [39], which provides an ethical framework for "Good AI Society" through five principles and 20 concrete recommendations. This framework emphasizes the importance of beneficence, non-maleficence, autonomy, justice, and explicability in the development and adoption of AI.

In terms of responsibility, Sattlegger et al. [121] proposed a conceptual framework designed to examine the task responsibilities of actors involved in the design, development, and application of AI systems for decision-making in the public sector. The goal of this framework is to promote responsible sociotechnical systems and prevent harm.

Calvo et al. [20] argue that understanding and designing for human autonomy is crucial for responsible AI. They highlight the need to consider autonomy from multiple dimensions of human experience. Extending their model for motivation, engagement, and thriving in the user experience (METUX), they discuss how the

effects of AI systems on autonomy can be analyzed and then protected through design. This framework identifies six spheres of technology experience to analyze the effects of AI systems on autonomy.

Ehsan et al. [30] propose a framework aimed at understanding the sociotechnical gap in explainable AI systems. This framework consists of technical (i.e., data, model, and explanation) and social (i.e., trust, actionability, and values) components with case studies. Another framework related to explainable AI is [95]. This framework lists a set of design goals and evaluation methods for user and model performances, providing guidelines for the design of explanations for different user groups, such as AI novices, data experts, and AI experts. Furthermore, Wang et al. [149] offer a conceptual framework that considers human cognitive biases and aims to link generated explanations with human decision-making theories, providing recommendations to improve human interpretability.

In terms of transparency, Felzmann et al. [34] proposed nine principles for transparency by design for AI models. These principles include being proactive, thinking of transparency as an integrative process, communicating in an audience-sensitive manner, explaining what data is being used and how it is processed, explaining decision-making criteria and their justifiability, explaining risks and risk mitigation measures, ensuring inspectability and auditability, being responsive to stakeholder queries and concerns, and diligently reporting about the system.

4.2 Tools and Design Guidelines

In the realm of human-AI interaction design, various tools and design guidelines have been developed to address the unique challenges and considerations associated with AI systems. Amershi et al. [8] draw attention to the potential conflicts between traditional user interaction design principles and the characteristics of AI systems, particularly those with inconsistent or unpredictable behavior. In response, the authors propose a comprehensive set of 18 design guidelines for human-AI interaction. These guidelines aim to enhance the intuitive and effective design of AI products while emphasizing the importance of ethical considerations and fairness beyond social norms and biases. Furthermore, the guidelines serve as a springboard for contemplating their relevance and potential extensions in future research. Notably, an aspect that could be further explored is the incorporation of adaptive or controllable autonomy in AI systems based on users' input choices. There are several other existing guidelines for human-AI design presented by industry, e.g., [156].

Research challenges in integrating humans and AI are discussed by Hevner and Storey [49] using a taxonomy called 8C, which lists eight main design challenges that are composition, complexity, creativity, confidence, controls, conscience, certification, and contribution. This taxonomy provides guidelines and proposes a roadmap for each challenge.

A number of tools and techniques are also discussed related to human-centered AI. For example, Gordon et al. [47] introduce "jury learning", which enables practitioners to explicitly define which people or groups from the training dataset, and in what proportion, determine classifier predictions. ProtoAI [137] is built to help designers incorporate AI into interface designs. It allows the designer to evaluate design choices with different model inputs and iteratively refine designs by analyzing model breakdowns. Another example is the tool proposed by Lam et al. [78], called IndieLabel. This tool uses a collaborative filtering approach, enabling non-technical end-users to audit algorithms and identify previously unreported algorithmic issues.

Overall, these frameworks, guidelines, and tools contribute to the development of AI systems that are more trustworthy, responsible, transparent, and designed with human-centered principles in mind.

4.3 Human-Centered Machine Learning

Human-centered machine learning (HCML) is a specific sub-domain of human-centered AI that focuses on integrating social values into machine learning practices. It addresses key development stages, including data curation, algorithm development, and model evaluation, to ensure that machine learning systems are more sensitive to social implications and better aligned with the needs of stakeholders. By adopting interdisciplinary collaboration, engaging stakeholders, and proactively considering potential implications and failures, HCML aims to create more ethical and socially responsible ML systems. Methods and frameworks from HCML enable scholars and practitioners to identify the blind spots in their system, consider the broader impact, and mitigate potential ethical risks.

For example, through the lens of "forgettance stack", Muller and Strohmayer [96] analyzed three classes of silences in machine learning pipelines that can cause or invoke forgetting: modest silences, silence as force, and ambivalent silences. The forgettance stack framework emphasizes the importance of acknowledging and understanding the role of forgetting in data science work and its impact on the quality, fairness, and reliability of the resulting data and models. Another example is the work of Rismani et al. [118], discussing safety engineering frameworks, such as System Theoretic Process Analysis and Failure Mode and Effects Analysis, and their applicability in HCML. Such frameworks offer an analytical structure to link harms with potential failures and hazards in existing design choices, facilitating mitigation development and informing ongoing international regulatory and standardization efforts.

4.4 Human-Centered Recommender Systems

The field of Human-Centered Recommender Systems represents another dedicated sub-domain within human-centered AI, aspiring to meld user-focused considerations with recommender system practices. It prioritizes understanding the characteristics of both the recommender systems and their users, as well as the relationships between them [76].

Challenging the traditional view of "help the users find relevant items", Jannach and Adomavicius [62] advocate a goal-oriented framework to steer the development of a recommender system by explicitly considering the recommendation goal for different stakeholders. The proposed framework comprises overarching goals, recommendation purposes, system tasks, and computational metrics. Overarching goals describe general motivations for using or providing a recommender system. For example, consumers or providers may have different goals when interacting with such a system. The conceptualization and operationalization of the remaining components should be centered around the goal. By focusing on the goal, the other elements can be better aligned and integrated, ultimately contributing to a more effective and human-centered recommender system.

The future of human-centered recommender systems hinges on several key factors: a strong human-centered science foundation with advances in machine learning, continuous research on real-world applications, rigorous evaluation methods, and prioritizing fairness, diversity, and transparency [76, 123]. By addressing these factors, human-centered recommender systems hold the potential to create personalized, engaging, and responsible recommendations that cater to diverse user needs and preferences, ultimately enhancing user satisfaction and trust.

5 Conclusion: Opportunities and Challenges

Throughout this chapter, we explored different frameworks of human-centered AI and discussed various paradigms, actors, and values that should be considered in the design and development of AI systems. Human-centered AI in the context of user modeling, adaptation, and personalization, similar to other AI systems, presents several opportunities and challenges, considering the various perspectives involved and the potential conflicts associated with implementing different layers. This section aims to delve further into the opportunities and challenges presented by human-centered AI.

Our framework provides a robust foundation for comprehensively understanding the diverse needs, preferences, goals, behaviors, values, and experiences of actors, taking into account their social context. By adopting the perspective-taking approach in user modeling, adaptation, and personalization, and particularly the shift toward human-centric and future-centric perspectives, we can identify different aspects, employ multiple or hybrid evaluation methods, and achieve different success criteria

while rectifying the limitations associated with each nested perspective. Relying on other dimensions of the framework, namely considering multiple actors, values, and intensity of values through different lenses, we can define requirements that are crucial for achieving human-centered AI but might be overlooked otherwise.

This holistic approach also provides a way to meet multiple actors' needs efficiently and effectively, particularly enabling the building of systems at scale while avoiding harm. This, in turn, can lead to improved, responsible, and resilient user satisfaction, engagement, and overall experience while promoting human values and augmenting human capabilities in the long run. This approach also provides a strategic framework for understanding the limitations and scope of proposed solutions. By recognizing these constraints, designers, researchers, and practitioners are able to clearly outline spaces for improvement, thus enabling a cycle of continuous growth and evolution in human-centered AI systems. The human-centered AI approaches empower the identification of both long-term and short-term impacts of AI systems. This leads to a comprehensive understanding of the potential repercussions of AI systems, facilitating timely intervention and the ability to manage unforeseen consequences effectively. It also ensures that AI systems bring positive impacts in both the immediate and distant future.

Human-centered AI emphasizes ethical considerations, societal impact, human values, and responsible innovation. It encourages a value-sensitive design that offers opportunities for explicitly specifying the values, which are typically embedded within five core categories: legal and ethical compliance, no harm, human experience and usefulness, social justice, and assurance and accountability. This emphasis promotes and enables the development of AI systems that are fair, explainable, transparent, and accountable, addressing concerns such as privacy, inclusivity, bias, discrimination, and algorithmic transparency, among others. Accordingly, methods and frameworks that have been (and are being) developed for the integration of values can be extended and used to design and develop more legal and moral-compliant systems, promoting positive societal outcomes as a consequence. Incorporating human values and ethical considerations can aid in informed decision-making by providing more responsible, transparent, and understandable recommendations.

Conversely, user modeling, adaptation, and personalization studies provide a rich foundation for human-centered AI. These studies have explored the combination of quantitative and data-driven approaches with qualitative and theory-driven approaches, incorporating various human factors such as empathy and personality. Human-centered AI can benefit from these advancements by incorporating them into its practices and methodologies.

In addition, human-centered AI naturally promotes interdisciplinary collaboration by involving experts from various fields, including computer science, ethics, psychology, philosophy, sociology, policy analysis, law, and design, among others. This collaboration enriches the design process, encourages interdisciplinary research and development, and fosters a culture of shared understanding and collective responsibility for more effective and human-centered AI systems. It also opens new avenues for research, addressing critical questions for the future of AI.

One such area of exploration could be defining the intensity of values and their hierarchy within AI system design. This entails understanding how to balance multiple, sometimes contradictory, values for multiple actors. It raises crucial questions like: What are the values for this system? Which values should be prioritized, at what intensity, and for whom? It also encourages further research into creating novel solutions for resolving these conflicts and satisfying the respective requirements from a technical point of view. The human-centered AI design approach fosters the development of novel solutions to satisfy these conflicts and contradictions. This is essential as we move toward an era where AI systems are integral to various aspects of our lives. Balancing user needs and system requirements in a way that respects human values and principles is a critical challenge that the field will continue to address.

Despite numerous advantages, implementing human-centered AI presents a set of challenges. Among them is the difficulty of balancing diverse perspectives, as discussed earlier in Sects. 2 and 3.2. These perspectives often lead to conflicting requirements and priorities, making it challenging to create AI systems that cater to multiple user groups and societal values.

Additionally, the complexity of realizing human-centered AI across various layers —from individual users to broader societal impacts—can be resource-intensive. The more holistic and thorough the study of different layers and perspectives, the more resources and effort are needed. This complexity is exacerbated when accounting for actors' and stakeholders' needs and values across different cultures and communities, particularly given the evolving nature of AI technologies and regulatory landscapes.

Another major challenge lies in understanding human needs and values and the need for multiple, hybrid, and even novel research methods. Researchers need insights into the complexities, nuances, and evolving patterns of human value and interaction outcomes, which in turn inform the design, realization, and evaluation of human-centered AI systems. In-depth and in-the-wild research methods, such as longitudinal and field studies, are invaluable for gaining insights into complex human behaviors over an extended period and in real-world setups. However, these methods require considerable resources to design and implement. We need to devise innovative methodologies to minimize this burden and provide the in-depth insights necessary and enable a more accurate, efficient, and comprehensive understanding of human needs and values.

To ensure the future-proofing, responsibility, and human-centeredness of AI systems, it is imperative to provide transparent reports on the underlying priorities that guide their development. Unfortunately, many AI systems in practice fall short of providing a transparent outlook on their values and priorities. It is essential for developers and researchers to address this gap and take proactive steps toward transparency. Additionally, the decision-making process employed in resolving conflicting requirements should be thoroughly documented. This documentation should be accessible throughout the design phase, enabling stakeholders to understand how decisions were made and the considerations involved. Transparent documentation is not only vital for AI systems but is also a fundamental aspect of any scientific contribution, as it helps identify the limitations and scope of the research findings and ensures validity

and reproducibility. This includes discussing the paradigm, considering actors, values, and the level of realization of those values, and defining the boundaries within which the research findings hold true.

Furthermore, another challenge in the realm of human-centered AI is the limited overlap and collaboration between research communities. These communities have different research focuses, methodologies, priorities, and sometimes even terminologies, resulting in limited overlap and hindering effective collaboration. Bridging the gap between these communities is crucial for developing human-centered AI systems that consider responsible technological advancements and a better future. A notable effort in this direction is the growing discourse surrounding human-centered recommender systems and the promotion of interdisciplinary contributions within scientific platforms, in addition to traditional algorithmic and technology-centric approaches.

User modeling, adaptation, and personalization generally rely on data to model and personalize and users to evaluate the experiences. Given different dimensions that are presented by human-centered AI, ensuring the availability, quality, and diversity of data and access to multiple actors for effective adaptation and personalization can be challenging, particularly in domains with limited data or biased or low-quality datasets.

In conclusion, this chapter presents an in-depth exploration of human-centered AI, shedding light on its multidimensional nature encompassing paradigms, actors, values, and their varying intensity. A human-centered AI framework is proposed to better understand and address diverse human needs in AI system design, providing an opportunity to enhance system human-centeredness, efficiency, usefulness, resilience, and overall user satisfaction. By promoting ethical considerations, societal impact, and responsible innovation, human-centered AI ensures the development of a value-driven system that is fair, transparent, trustworthy, and accountable. While the field offers immense opportunities, it also presents challenges such as balancing diverse perspectives, understanding complex human needs and values, and promoting proper documentation and transparency. However, these challenges can be overcome through effective interdisciplinary collaboration, innovative research methods, and a continuous commitment to understanding and addressing human needs and values, which also pave the way for innovative solutions and opportunities for growth within the field. The future of AI is intrinsically tied to a more human-centered approach, thereby requiring a continual exploration and adjustment of the way we design, implement, and evaluate these systems. The road ahead may be complex, but it is indeed promising, laying the foundation for a future where AI systems augment human capabilities, uphold human values, and genuinely benefit society as a whole.

References

1. Influencing Individually (2012) Fusing Personalization and Persuasion. ACM Trans Interactive Intell Syst 2(2):9:1–9:8. https://doi.org/10.1145/2209310.2209312
2. 14:00-17:00: ISO/IEC TR 24027:2021 (2021). https://www.iso.org/standard/77607.html
3. Adadi A, Berrada M (2018) Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access 6:52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052. Conference Name: IEEE Access
4. Adam H, Balagopalan A, Alsentzer E, Christia F, Ghassemi M (2022) Mitigating the impact of biased artificial intelligence in emergency decision-making. Commun Med 2(1):1–6
5. Adegun A, Viriri S (2021) Deep learning techniques for skin lesion analysis and melanoma cancer detection: a survey of state-of-the-art. Artif Intell Rev 54(2):811–841
6. Aizenberg E, van den Hoven J (2020) Designing for human rights in AI. Big Data & Soc 7(2):2053951720949566
7. Aljeraisy A, Barati M, Rana O, Perera C (2021) Privacy laws and privacy by design schemes for the internet of things: a developer's perspective. ACM Comput Surv 54(5):102:1–102:38. https://doi.org/10.1145/3450965
8. Amershi S, Weld D, Vorvoreanu M, Fourney A, Nushi B, Collisson P, Suh J, Iqbal S, Bennett PN., Inkpen K, Teevan J, Kikin-Gil R, Horvitz E (2019) Guidelines for human-AI interaction. In: Proceedings of the 2019 CHI conference on human factors in computing systems. ACM, Glasgow Scotland, UK, pp 1–13. https://doi.org/10.1145/3290605.3300233
9. Ammitzbøll Flügge A, Hildebrandt T, Møller, NH (2021) Street-level algorithms and AI in bureaucratic decision-making: a caseworker perspective. Proc ACM Human-Comput Interaction 5(CSCW1):40:1–40:23. https://doi.org/10.1145/3449114
10. Axelsson M, Oliveira R, Racca M, Kyrki V (2021) Social robot co-design canvases: a participatory design framework. ACM Trans Human-Robot Interaction 11(1):3:1–3:39 (2021). https://dl.acm.org/doi/10.1145/3472225
11. Baeza-Yates R (2018) Bias on the web. Commun ACM 61(6):54–61
12. Barocas S, Hardt M, Narayanan A (2019) Fairness and machine learning: limitations and opportunities. fairmlbook.org
13. Barth S, Ionita D, Hartel P (2022) Understanding online privacy-a systematic review of privacy visualizations and privacy by design guidelines. ACM Comput Surv 55(3):63:1–63:37. https://doi.org/10.1145/3502288
14. Bender EM, Gebru T, McMillan-Major A, Shmitchell S (2021) On the dangers of stochastic parrots: can language models be too big? In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, FAccT '21. Association for Computing Machinery, New York, NY, USA, pp 610–623 https://doi.org/10.1145/3442188.3445922
15. Berkovsky S, Taib R, Hijikata Y, Braslavsku P, Knijnenburg B (2018) A cross-cultural analysis of trust in recommender systems. In: Proceedings of the 26th conference on user modeling, adaptation and personalization, UMAP '18. ACM, New York, NY, USA, pp 285–289. https://doi.org/10.1145/3209219.3209251. Event-place: Singapore, Singapore
16. Binns R, Van Kleek M, Veale M, Lyngs U, Zhao J, Shadbolt N (2018) 'It's reducing a human being to a percentage': perceptions of justice in algorithmic decisions. In: Proceedings of the 2018 CHI conference on human factors in computing systems, CHI '18. Association for Computing Machinery, New York, NY, USA, pp 1–14. https://doi.org/10.1145/3173574.3173951
17. Black J, Roberts D, Stigall B, Michael I, Knijnenburg B (2023) Retiree volunteerism: automating "word of mouth" communication. In: third workshop on social and cultural integration with personalized interfaces (SOCIALIZE) 2023. Sydney, Australia
18. Bollen D, Knijnenburg BP, Willemsen MC, Graus M (2010) Understanding choice overload in recommender systems. In: Proceedings of the fourth ACM conference on Recommender systems, pp 63–70. Barcelona, Spain. https://doi.org/10.1145/1864708.1864724
19. Brayne S (2017) Big data surveillance: the case of policing. Am Sociol Rev 82(5):977–1008

20. Calvo RA, Peters D, Vold K, Ryan RM (2020) Supporting human autonomy in AI systems: a framework for ethical enquiry. In: Burr C, Floridi L (eds) Ethics of digital well-being: a multidisciplinary approach, Philosophical studies series. Springer International Publishing, Cham, pp 31–54. https://doi.org/10.1007/978-3-030-50585-1_2
21. Capel T, Brereton M () What is Human-Centered about Human-Centered AI? A Map of the Research Landscape. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI '23, pp 1–23. Association for Computing Machinery, New York, NY, USA (2023). 10.1145/3544548.3580959. https://dl.acm.org/doi/10.1145/3544548.3580959
22. Cavoukian A (2013) Privacy by design and the promise of smartdata. In: Harvey I, Cavoukian A, Tomko G, Borrett D, Kwan H, Hatzinakos D (eds) SmartData, pp 1–9. Springer, New York. http://link.springer.com/chapter/10.1007/978-1-4614-6409-9_1
23. Charmaz K (2014) Constructing grounded theory. SAGE. Google-Books-ID: v_GGAwAAQBAJ
24. Chin JP, Diehl VA, Norman KL (1988) Development of an instrument measuring user satisfaction of the human-computer interface. In: Proceedings of the SIGCHI conference on human factors in computing systems, CHI '88. Association for Computing Machinery, New York, NY, USA, pp 213–218. https://doi.org/10.1145/57167.57203
25. Commission E (2021) Proposal for a regulation of the European parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts (2021). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206
26. Cossins D (2018) Discriminating algorithms: 5 times AI showed prejudice. https://www.newscientist.com/article/2166207-discriminating-algorithms-5-times-ai-showed-prejudice/
27. Deshpande A, Sharp H (2022) Responsible AI systems: who are the stakeholders? In: Proceedings of the 2022 AAAI/ACM conference on AI, ethics, and society, AIES '22. Association for Computing Machinery, New York, NY, USA, pp 227–236. https://doi.org/10.1145/3514094.3534187
28. Dressel J, Farid H (2018) The accuracy, fairness, and limits of predicting recidivism. Sci Adv 4(1):eaao5580. https://doi.org/10.1126/sciadv.aao5580. American Association for the Advancement of Science
29. Edizel B, Bonchi F, Hajian S, Panisson A, Tassa T (2020) FaiRecSys: mitigating algorithmic bias in recommender systems. Int J Data Sci Anal 9(2):197–213
30. Ehsan U, Saha K, De Choudhury M, Riedl MO (2023) Charting the sociotechnical gap in explainable AI: a framework to address the gap in XAI. In: Proceedings of the ACM on human-computer interaction 7(CSCW1):34:1–34:32. https://doi.org/10.1145/3579467
31. Ekstrand MD, Das A, Burke R, Diaz F (2022) Fairness in information access systems. Found Trends® Inf Retrieval 16(1-2):1–177. 10.1561/1500000079. https://www.nowpublishers.com/article/Details/INR-079. Publisher: Now Publishers, Inc
32. Ekstrand MD, Willemsen MC (2016) Behaviorism is not enough: better recommendations through listening to users. In: Proceedings of the 10th ACM conference on recommender systems, RecSys '16. ACM, New York, NY, USA, pp 221–224. https://doi.org/10.1145/2959100.2959179. Event-place: Boston, Massachusetts, USA
33. Enam MA, Srivastava S, Knijnenburg BP (2023) Designing a recommender system to recruit older adults for research studies. In: Third workshop on social and cultural integration with personalized interfaces (SOCIALIZE) 2023. Sydney, Australia
34. Felzmann H, Fosch-Villaronga E, Lutz C, Tamò-Larrieux A (2020) Towards transparency by design for artificial intelligence. Sci Eng Ethics 26(6):3333–3361
35. Ferraro A, Serra X, Bauer C (2021) Break the loop: gender imbalance in music recommenders. In: Proceedings of the 2021 conference on human information interaction and retrieval, CHIIR '21. Association for Computing Machinery, New York, NY, USA, pp 249–254. https://doi.org/10.1145/3406522.3446033
36. Ferwerda B, Hanbury A, Knijnenburg BP, Larsen B, Michiels L, Papenmeier A, Said A, Schaer P, Willemsen M (2023) Reality check - conducting real world studies: frontiers of information access experimentation for br research and education. Front Inf Access Exp

Res Educ 13:20–40. Publisher: Schloss Dagstuhl- Leibniz-Zentrum fur Informatik GmbH, Dagstuhl Publishing
37. Fitzsimons GJ, Lehmann DR (2004) Reactance to recommendations: when unsolicited advice yields contrary responses. Mark Sci 23(1):82–94
38. Flinn B, Maurer H (1995) Levels of anonymity. J Univ Comput Sci 1(1):35–47
39. Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, Luetge C, Madelin R, Pagallo U, Rossi F, Schafer B, Valcke P, Vayena E (2018) AI4people-an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Mind Mach 28(4):689–707
40. Floridi L, Holweg M, Taddeo M, Amaya Silva J, Mökander J, Wen Y (2022) capAI - A procedure for conducting conformity assessment of AI systems in line with the EU artificial intelligence Act. https://doi.org/10.2139/ssrn.4064091
41. Frankena WK (1973) Intrinsic value and the good life. Ethics, 2nd edn. Prentice-Hall, INC., Englewood Cliffs, New Jersey, pp 79–95
42. Friedman A, Knijnenburg BP, Vanhecke K, Martens L, Berkovsky S (2015) Privacy aspects of recommender systems. In: Ricci F, Rokach L, Shapira B (eds) Recommender systems handbook, 2 edn. Springer, US, pp 649–688. http://link.springer.com/chapter/10.1007/978-1-4899-7637-6
43. Friedman B, Kahn PH, Borning A, Huldtgren A (2013) Value sensitive design and information systems. In: Doorn N, Schuurbiers S, van de Poel V, Gorman ME (eds) Early engagement and new technologies: opening up the laboratory. Philosophy of Engineering and Technology. Springer Netherlands, Dordrecht, pp 55–95. https://doi.org/10.1007/978-94-007-7844-3_4
44. Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H III, HD, Crawford K, (2021) Datasheets for datasets. Commun ACM 64(12):86–92
45. Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L (20118) Explaining explanations: An approach to evaluating interpretability of machine learning. arXiv:1806.00069 p 118. Publisher: CoRR
46. GONZÁLEZ FUSTER G (2020) Artificial intelligence and law enforcement impact on fundamental rights. STUDY requested by the LIBE committee European Parliament PE 656.295, European Parliament, Brussels. https://www.europarl.europa.eu/RegData/etudes/STUD/2020/656295/IPOL_STU(2020)656295_EN.pdf
47. Gordon ML, Lam MS, Park JS, Patel K, Hancock J, Hashimoto T, Bernstein MS (2022) Jury learning: integrating dissenting voices into machine learning models. In: Proceedings of the 2022 CHI conference on human factors in computing systems, CHI '22. Association for Computing Machinery, New York, NY, USA, pp 1–19. https://doi.org/10.1145/3491102.3502004
48. Greene D, Hoffmann AL, Stark L (2019) Better, Nicer, Clearer, Fairer: a critical assessment of the movement for ethical artificial intelligence and machine learning. In: Hawaii international conference on system sciences 2019 (HICSS-52). https://aisel.aisnet.org/hicss-52/dsm/critical_and_ethical_studies/2
49. Hevner A, Storey V (2023) Research challenges for the design of human-artificial intelligence systems (HAIS). ACM Trans Manag Inf Syst 14(1):10:1–10:18. https://doi.org/10.1145/3549547
50. Hoffman LJ (1969) Computers and privacy: a survey. ACM Comput Surv 1(2):85–103
51. Holten Møller N, Shklovski I, Hildebrandt TT (2020) Shifting concepts of value: designing algorithmic decision-support systems for public services. In: Proceedings of the 11th nordic conference on human-computer interaction: shaping experiences, shaping society. ACM, Tallinn Estonia, pp 1–12. https://doi.org/10.1145/3419249.3420149
52. Holtzblatt K, Beyer H (2016) Contextual design: design for life, 2nd edn. Morgan Kaufmann, Amsterdam, Cambridge, MA
53. van den Hoven J, Vermaas PE, van de Poel (2015) Design for values: an introduction. In: van den Hoven J, Vermaas PE, van de Poel I (eds) Handbook of ethics, values, and technological design: sources, theory, values and application domains. Springer Netherlands, Dordrecht, pp 1–7. https://doi.org/10.1007/978-94-007-6970-0_40

54. Howard A, Borenstein J (2018) The ugly truth about ourselves and our robot creations: the problem of bias and social inequity. Sci Eng Ethics 24(5):1521–1536
55. Huang C, Zhang Z, Mao B, Yao X (2022) An overview of artificial intelligence ethics. IEEE Trans Artif Intel 1–21. https://doi.org/10.1109/TAI.2022.3194503. Conference Name: IEEE Transactions on Artificial Intelligence
56. Hull G (2015) Successful failure: what Foucault can teach us about privacy self-management in a world of Facebook and big data. Ethics Inf Technol 17(2):89–101
57. Humbert M, Trubert B, Huguenin K (2019) A survey on interdependent privacy. ACM Comput Surv 52(6):122:1–122:40. https://doi.org/10.1145/3360498
58. Hutchinson B, Mitchell M (2019) 50 years of test (un)fairness: lessons for machine learning. In: Proceedings of the conference on fairness, accountability, and transparency, FAT* '19. Association for Computing Machinery, New York, NY, USA, pp 49–58. https://doi.org/10.1145/3287560.3287600
59. IEEE (2020) IEEE recommended practice for assessing the impact of autonomous and intelligent systems on human well-being. https://standards.ieee.org/ieee/7010/7718/
60. IEEE (2020) Recommended practice for organizational governance of artificial intelligence. https://standards.ieee.org
61. Jakesch M, Buçinca Z, Amershi S, Olteanu A (2022) How different groups prioritize ethical values for responsible AI. In: 2022 ACM conference on fairness, accountability, and transparency, FAccT '22. Association for Computing Machinery, New York, NY, USA, pp 310–323. https://doi.org/10.1145/3531146.3533097
62. Jannach D, Adomavicius G (2016) Recommendations with a purpose. In: Proceedings of the 10th ACM conference on recommender systems, RecSys '16. Association for Computing Machinery, New York, NY, USA, pp 7–10 (2016). https://doi.org/10.1145/2959100.2959186
63. Jannach D, Manzoor A, Cai W, Chen L (2021) A survey on conversational recommender systems. ACM Comput Surv 54(5):105:1–105:36. https://doi.org/10.1145/3453154
64. Kazim E, Koshiyama AS (2021) A high-level overview of AI ethics. Patterns (New York, N.Y.) 2(9):100314. https://doi.org/10.1016/j.patter.2021.100314
65. Keegan BJ, Dennehy D, Naudé P (2022) Implementing artificial intelligence in traditional B2B marketing practices: an activity theory perspective. Inf Syst Front
66. Knijnenburg BP, Hubig N (2020) Human-centric preference modeling for virtual agents. In: Proceedings of the 20th ACM international conference on intelligent virtual agents, IVA '20. Association for Computing Machinery, New York, NY, USA, pp 1–3. https://doi.org/10.1145/3383652.3423909
67. Knijnenburg BP, Jin H (2013) The persuasive effect of privacy recommendations. In: Twelfth annual workshop on HCI research in MIS. Milan, Italy. http://aisel.aisnet.org/sighci2013/16
68. Knijnenburg BP, Page X, Wisniewski P, Lipford HR, Proferes N, Romano J (eds) Modern socio-technical perspectives on privacy. Springer Nature. https://doi.org/10.1007/978-3-030-82786-1. https://library.oapen.org/handle/20.500.12657/52825. Accepted: 2022-02-14T21:17:55Z
69. Knijnenburg BP, Raybourn EM, Cherry D, Wilkinson D, Sivakumar S, Sloan H (2017) Death to the privacy calculus? https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2923806
70. Knijnenburg BP, Reijmer NJ, Willemsen MC (2011) Each to his own: how different users call for different interaction methods in recommender systems. In: Proceedings of the fifth ACM conference on Recommender systems. ACM Press, Chicago, IL, pp 141–148. https://doi.org/10.1145/2043932.2043960
71. Knijnenburg BP, Schmidt-Thieme L, Bollen DG (2010) Workshop on user-centric evaluation of recommender systems and their interfaces. In: Proceedings of the fourth ACM conference on Recommender systems, RecSys '10. Association for Computing Machinery, New York, NY, USA, pp 383–384. https://doi.org/10.1145/1864708.1864800
72. Knijnenburg BP, Sivakumar S, Wilkinson D (2016) Recommender Systems for Self-Actualization. In: Proceedings of the 10th ACM conference on recommender systems, RecSys '16. Association for Computing Machinery, New York, NY, USA, pp 11–14. https://doi.org/10.1145/2959100.2959189

73. Knijnenburg BP, Willemsen MC (2015) Evaluating recommender systems with user experiments. In: Ricci F, Rokach L, Shapira B (eds) Recommender systems handbook. Springer, US, pp 309–352. https://doi.org/10.1007/978-1-4899-7637-6_9
74. Knijnenburg BP, Willemsen MC, Gantner Z, Soncu H, Newell C (2012) Explaining the user experience of recommender systems. User Model User-Adap Inter 22(4–5):441–504. https://doi.org/10.1007/s11257-011-9118-4
75. Kohavi R, Longbotham R, Sommerfield D, Henne RM (2009) Controlled experiments on the web: survey and practical guide. Data Min Knowl Disc 18(1):140–181
76. Konstan J, Terveen L (2021) Human-centered recommender systems: origins, advances, challenges, and opportunities. AI Mag 42(3):31–42
77. Kumar Y, Gupta S, Singla R, Hu YC (2022) A systematic review of artificial intelligence techniques in cancer prediction and diagnosis. Arch Comput Methods Eng 29(4):2043–2070
78. Lam MS, Gordon ML, Metaxa D, Hancock JT, Landay JA, Bernstein MS (2022) End-user audits: a system empowering communities to lead large-scale investigations of harmful algorithmic behavior. Proc ACM on Human-Comput Interaction 6(CSCW2):512:1–512:34. https://doi.org/10.1145/3555625
79. Landers RN, Behrend TS (2023) Auditing the AI auditors: a framework for evaluating fairness and bias in high stakes AI predictive models. Am Psychol 78(1):36
80. Lehner OM, Knoll C, Leitner-Hanetseder S, Eisl C (2022) The dynamics of artificial intelligence in accounting organisations: a structuration perspective. In: The Routledge handbook of accounting information systems, pp 121–139. Routledge
81. Lim BY, Dey AK, Avrahami D (2009) Why and why not explanations improve the intelligibility of context-aware intelligent systems. In: Proceedings of the SIGCHI conference on human factors in computing systems, pp 2119–2128
82. Lindebaum D, Glaser V, Moser C, Ashraf M (2022) When algorithms rule, values can wither. MIT sloan management review (Winter 2023). https://sloanreview.mit.edu/article/when-algorithms-rule-values-can-wither/
83. Liu B, Ding M, Shaham S, Rahayu W, Farokhi F, Lin Z (2021) When machine learning meets privacy: a survey and outlook. ACM Comput Surv 54(2):31:1–31:36. https://doi.org/10.1145/3436755
84. Matthews T, O'Leary K, Turner A, Sleeper M, Woelfer JP, Shelton M, Manthorne C, Churchill EF, Consolvo S (2017) Stories from survivors: privacy & security practices when coping with intimate partner abuse. In: Proceedings of the 2017 CHI conference on human factors in computing systems, CHI '17. Association for Computing Machinery, New York, NY, USA, pp 2189–2201. https://doi.org/10.1145/3025453.3025875
85. McGregor S (2020) Preventing repeated real world AI failures by cataloging incidents: the AI incident database. ArXiv:2011.08512 [cs], Database address.https://incidentdatabase.ai
86. McNee SM, Albert I, Cosley D, Gopalkrishnan P, Lam SK, Rashid AM, Konstan JA, Riedl J (2002) On the recommending of citations for research papers. In: Proceedings of the 2002 ACM conference on computer supported cooperative work. New Orleans, LA, pp 116–125. https://doi.org/10.1145/587078.587096
87. McNee SM, Riedl J, Konstan JA (2006) Being accurate is not enough: how accuracy metrics have hurt recommender systems. In: Extended abstracts on Human factors in computing systems. Montreal, Canada, pp 1097–1101. https://doi.org/10.1145/1125451.1125659
88. McNee SM, Riedl J, Konstan JA (2006) Making recommendations better: an analytic model for human-recommender interaction. In: CHI '06 extended abstracts on human factors in computing systems, CHI EA '06. Association for Computing Machinery, New York, NY, USA, pp 1103–1108. https://doi.org/10.1145/1125451.1125660
89. McQuillan D (2022) Resisting AI: an anti-fascist approach to artificial intelligence. Policy Press. Google-Books-ID: N6x6EAAAQBAJ
90. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A (2021) A survey on bias and fairness in machine learning. ACM Comput Surv 54(6):115:1–115:3. https://doi.org/10.1145/3457607

91. Michiels L, Leysen J, Smets A, Goethals B (2022) What are filter bubbles really? A review of the conceptual and empirical work. In: Adjunct proceedings of the 30th ACM conference on user modeling, adaptation and personalization. Association for Computing Machinery, New York, NY, USA, pp 274–279. https://doi.org/10.1145/3511047.3538028
92. Miller T (2019) Explanation in artificial intelligence: insights from the social sciences. Artif Intell 267:1–38
93. Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, Spitzer E, Raji ID, Gebru T (2019) Model cards for model reporting. In: Proceedings of the conference on fairness, accountability, and transparency, FAT* '19. Association for Computing Machinery, New York, NY, USA, pp 220–229. https://doi.org/10.1145/3287560.3287596
94. Mittelstadt BD, Allo P, Taddeo M, Wachter S, Floridi L (2016) The ethics of algorithms: mapping the debate. Big Data & Soc 3(2):2053951716679679
95. Mohseni S, Zarei N, Ragan ED (2021) A multidisciplinary survey and framework for design and evaluation of explainable AI systems. ACM Trans Interactive Intell Syst 11(3-4):24:1–24:45. https://doi.org/10.1145/3387166
96. Muller M, Strohmayer A (2022) Forgetting practices in the data sciences. In: Proceedings of the 2022 CHI conference on human factors in computing systems, CHI '22. Association for Computing Machinery, New York, NY, USA, pp 1–19. https://doi.org/10.1145/3491102.3517644
97. Mökander J, Floridi L (2021) Ethics-based auditing to develop trustworthy AI. Mind Mach 31(2):323–327
98. Namara M, Sloan H, Knijnenburg BP (2022) The effectiveness of adaptation methods in improving user engagement and privacy protection on social network sites. In: Proceedings on privacy enhancing technologies. https://petsymposium.org/popets/2022/popets-2022-0031.php
99. Newell A (1973) You can't play 20 questions with nature and win: projective comments on the papers of this symposium. In: Chase W (ed) Visual information processing. Academic, Pittsburgh, PA
100. Niraula D, Sun W, Jin J, Dinov ID, Cuneo K, Jamaluddin J, Matuszak MM, Luo Y, Lawrence TS, Jolly S, Ten Haken RK, El Naqa I (2023) A clinical decision support system for AI-assisted decision-making in response-adaptive radiotherapy (ARCliDS). Sci Rep 13(1):5279
101. Nunes I, Jannach D (2017) A systematic review and taxonomy of explanations in decision support and recommender systems. User Model User-Adapt Interaction 27:393–444. Publisher: Springer
102. Obermeyer Z, Powers B, Vogeli C, Mullainathan S (2019) Dissecting racial bias in an algorithm used to manage the health of populations. Science 366(6464):447–453
103. Onuoha M (2018) Notes on algorithmic violence. Tech. rep. GitHub. https://githubcom/MimiOnuoha/On-Algorithmic-Violence
104. Oviedo-Trespalacios O, Peden AE, Cole-Hunter T, Costantini A, Haghani M, Rod., J.E., Kelly S, Torkamaan H, Tariq A, Newton JDA, Gallagher T, Steinert S, Filtness A, Reniers G (2023) The risks of using ChatGPT to obtain common safety-related information and advice. https://doi.org/10.2139/ssrn.4346827
105. Pham TC, Luong CM, Hoang VD, Doucet A (2021) AI outperformed every dermatologist in dermoscopic melanoma diagnosis, using an optimized deep-CNN architecture with custom mini-batch logic and loss function. Sci Rep 11(1):17485. https://doi.org/10.1038/s41598-021-96707-8. Number: 1 Publisher: Nature Publishing Group
106. van de Poel I (2021) Conflicting values in design for values design for values. In: van den Hoven J, Vermaas PE, van de Poel I (eds) Handbook of ethics, values, and technological design: sources, theory, values and application domains. Springer Netherlands, Dordrecht, pp 1–23. https://doi.org/10.1007/978-94-007-6994-6_5-1
107. van de Poel I (2021) Design for value change. Ethics Inf Technol 23(1):27–31
108. Pouloudi A (1997) Stakeholder analysis as a front-end to knowledge elicitation. AI & Soc 11(1):122–137

109. Proferes N (2022) The development of privacy norms. In: Knijnenburg BP, Page X, Wisniewski P, Lipford HR, Proferes N, Romano J (eds) Modern socio-technical perspectives on privacy. Springer International Publishing, Cham, pp 79–90. https://doi.org/10.1007/978-3-030-82786-1_5
110. Pu P, Chen L, Hu R (2011) A user-centric evaluation framework for recommender systems. In: Proceedings of the fifth ACM conference on Recommender systems, RecSys '11. Association for Computing Machinery, New York, NY, USA , pp 157–164. https://doi.org/10.1145/2043932.2043962
111. Pu P, Chen L, Hu R (2012) Evaluating recommender systems from the user's perspective: survey of the state of the art. User Model User-Adap Inter 22(4):317–355
112. Rahwan I, Cebrian M, Obradovich N, Bongard J, Bonnefon JF, Breazeal C, Crandall JW, Christakis NA, Couzin ID, Jackson MO, Jennings NR, Kamar E, Kloumann IM, Larochelle H, Lazer D, McElreath R, Mislove A, Parkes DC, Pentland A, Roberts ME, Shariff A, Tenenbaum JB, Wellman M (2019) Machine behaviour. Nature 568(7753):477–486
113. Raimondo GM, of Commerce, UD (2023) Artificial intelligence risk management framework (AI RMF 1.0). NIST. https://doi.org/10.6028/NIST.AI.100-1. Last Modified: 2023-03-30T12:25-04:00
114. Raji ID, Smart A, White RN, Mitchell M, Gebru T, Hutchinson B, Smith-Loud J, Theron D, Barnes P (2020) Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In: Proceedings of the 2020 conference on fairness, accountability, and transparency, FAT* '20. Association for Computing Machinery, New York, NY, USA, pp 33–44. https://doi.org/10.1145/3351095.3372873
115. Rajpurkar P, Chen E, Banerjee O, Topol EJ (2022) AI in health and medicine. Nat Med 28(1):31–38
116. Ribera M, Lapedriza A (2019) Can we do better explanations? A proposal of user-centered explainable AI. In: IUI workshops, vol 2327, p 38
117. Riedl MO (2019) Human-centered artificial intelligence and machine learning. Hum Behav Emerg Technol 1(1):33–36
118. Rismani S, Shelby R, Smart A, Jatho E, Kroll J, Moon A, Rostamzadeh N (2023) From plane crashes to algorithmic harm: applicability of safety engineering frameworks for responsible ML. In: Proceedings of the 2023 CHI conference on human factors in computing systems, CHI '23. Association for Computing Machinery, New York, NY, USA, pp 1–18. https://doi.org/10.1145/3544548.3581407
119. Sankaran S, Markopoulos P (2021) "It's like a puppet master": User perceptions of personal autonomy when interacting with intelligent technologies. In: Proceedings of the 29th ACM conference on user modeling, adaptation and personalization, UMAP '21. Association for Computing Machinery, New York, NY, USA, pp 108–118. https://doi.org/10.1145/3450613.3456820
120. Sankaran S, Zhang C, Funk M, Aarts H, Markopoulos P (2020) Do I have a say? Using conversational agents to re-imagine human-machine autonomy. In: Proceedings of the 2nd conference on conversational user interfaces, CUI '20. Association for Computing Machinery, New York, NY, USA, pp 1–3. https://doi.org/10.1145/3405755.3406135
121. Sattlegger A, van den Hoven J, Bharosa N (2022) Designing for responsibility. In: DG.O 2022: The 23rd annual international conference on digital government research. Association for Computing Machinery, New York, NY, USA, pp 214–225. https://doi.org/10.1145/3543434.3543581
122. Schaub F, Balebako R, Durity AL, Cranor LF (2015) A design space for effective privacy notices, pp 1–17 (2015). https://www.usenix.org/conference/soups2015/proceedings/presentation/schaub
123. Schedl M, Gómez E, Lex E (2023) Trustworthy algorithmic ranking systems. In: Proceedings of the sixteenth ACM international conference on web search and data mining, WSDM '23. Association for Computing Machinery, New York, NY, USA, pp 1240–1243. https://doi.org/10.1145/3539597.3572723

124. Scher S, Kopeinik S, Trügler A, Kowald D (2023) Modelling the long-term fairness dynamics of data-driven targeted help on job seekers. Sci Rep 13(1):1727
125. Schäfer H, Hors-Fraile S, Karumur RP, Calero Valdez A, Said A, Torkamaan H, Ulmer T, Trattner C (2017) Towards health (aware) recommender systems. In: Proceedings of the 2017 international conference on digital health, DH '17. Association for Computing Machinery, New York, NY, USA, pp 157–161 (2017). https://doi.org/10.1145/3079452.3079499
126. Septiandri AA, Constantinides M, Tahaei M, Quercia D (2023) WEIRD FAccTs: how western, educated, industrialized, rich, and democratic is FAccT? https://doi.org/10.1145/3593013.3593985. http://arxiv.org/abs/2305.06415. ArXiv:2305.06415 [cs]
127. Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M (2021) Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat Med 27(12):2176–2182
128. Shahbazi N, Lin Y, Asudeh A, Jagadish HV (2023) Representation bias in data: a survey on identification and resolution techniques. ACM Comput Surv
129. Shahriari K, Shahriari M (2017) IEEE standard review - Ethically aligned design: A vision for prioritizing human wellbeing with artificial intelligence and autonomous systems. In: 2017 IEEE Canada international humanitarian technology conference (IHTC), pp 197–201 (2017). https://doi.org/10.1109/IHTC.2017.8058187
130. Shani G, Gunawardana A (2011) Evaluating recommendation systems. In: Ricci F, Rokach L, Shapira B, Kantor PB (eds) Recommender systems handbook. Springer US, Boston, MA, pp 257–297. https://doi.org/10.1007/978-0-387-85820-3_8
131. Shneiderman B (2020) Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered AI systems. ACM Trans Interactive Intell Syst 10(4):26:1–26:31 (2020). https://doi.org/10.1145/3419764
132. Shneiderman B (2020) Human-centered artificial intelligence: three fresh ideas. AIS Trans. Human-Comput. Interaction 12(3):109–124. https://doi.org/10.17705/1thci.00131. https://aisel.aisnet.org/thci/vol12/iss3/1
133. Shneiderman B (2022) Human-centered AI. Oxford University Press. Google-Books-ID: YS9VEAAAQBAJ
134. Shulner-Tal A, Kuflik T, Kliger D (2023) Enhancing Fairness Perception - Towards Human-Centred AI and Personalized Explanations Understanding the Factors Influencing Laypeople's Fairness Perceptions of Algorithmic Decisions. International Journal of Human-Computer Interaction 39(7):1455–1482. https://doi.org/10.1080/10447318.2022.2095705 https://doi.org/10.1080/10447318.2022.2095705. Publisher: Taylor & Francis_eprint: https://doi.org/10.1080/10447318.2022.2095705
135. Sinha P, Alsubhi A, Dash S, Guo L, P Knijnenburg B (2017) Shopping for clothes: from meeting individual needs to socializing. BCS Learning & Development. https://doi.org/10.14236/ewic/HCI2017.78
136. Smuha NA (2019) The EU approach to ethics guidelines for trustworthy artificial intelligence. https://papers.ssrn.com/abstract=3443537
137. Subramonyam H, Seifert C, Adar E (2021) ProtoAI: model-informed prototyping for AI-powered interfaces. In: 26th international conference on intelligent user interfaces, IUI '21. Association for Computing Machinery, New York, NY, USA, pp 48–58. https://doi.org/10.1145/3397481.3450640
138. Tahaei M, Abu-Salma R, Rashid A (2023) Stuck in the permissions with you: developer & end-user perspectives on app permissions & their privacy ramifications. In: Proceedings of the 2023 CHI conference on human factors in computing systems, CHI '23. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3544548.3581060. Event-place: Hamburg, Germany
139. Tahaei M, Ramokapane KM, Li T, Hong JI, Rashid A (2022) Charting app developers' journey through privacy regulation features in ad networks. proceedings on privacy enhancing technologies. https://petsymposium.org/popets/2022/popets-2022-0061.php
140. Tian Z, Cui L, Liang J, Yu S (2022) A comprehensive survey on poisoning attacks and countermeasures in machine learning. ACM Comput Surv 55(8):166:1–166:35. https://doi.org/10.1145/3551636

141. Tintarev N, Masthoff J (2022) Beyond explaining single item recommendations. In: Ricci F, Rokach L, Shapira B (eds) Recommender systems handbook. Springer US, New York, NY, pp 711–756. https://doi.org/10.1007/978-1-0716-2197-4_19
142. Torkamaan H, Barbu CM, Ziegler J (2019) How can they know that? A study of factors affecting the creepiness of recommendations. In: Proceedings of the 13th ACM conference on recommender systems, RecSys '19. Association for Computing Machinery, New York, NY, USA, pp 423–427. https://doi.org/10.1145/3298689.3346982
143. Torkamaan H, Ziegler J (2022) Recommendations as challenges: estimating required effort and user ability for health behavior change recommendations. In: 27th international conference on intelligent user interfaces, IUI '22. Association for Computing Machinery, New York, NY, USA, pp 106–119. https://doi.org/10.1145/3490099.3511118
144. Toros H, Flaming D (2018) Prioritizing homeless assistance using predictive algorithms: an evidence-based approach. https://papers.ssrn.com/abstract=3202479
145. Torres R, McNee SM, Abel M, Konstan JA, Riedl J (2004) Enhancing digital libraries with TechLens+. In: Proceedings of the 2004 joint ACM/IEEE conference on Digital libraries. Tuscon, AZ, USA, p 228. https://doi.org/10.1145/996350.996402
146. UNESCO (2021) The UNESCO recommendation on the ethics of AI: shaping the future of our societies. Tech. rep. https://www.unesco.nl/sites/default/files/inline-files/Unesco%20AI%20Brochure.pdf
147. Urquhart L, Miranda D (2022) Policing faces: the present and future of intelligent facial surveillance. Inf Commun Technol Law 31(2):194–219
148. Veluwenkamp H, van den Hoven J (2023) Design for values and conceptual engineering. Ethics Inf Technol 25(1):2
149. Wang D, Yang Q, Abdul A, Lim BY (2019) Designing theory-driven user-centric explainable AI. In: Proceedings of the 2019 CHI conference on human factors in computing systems, CHI '19. Association for Computing Machinery, New York, NY, USA, pp 1–15. https://doi.org/10.1145/3290605.3300831
150. Wang Y, Ma W, Zhang M, Liu Y, Ma S (2023) A survey on the fairness of recommender systems. ACM Trans Inf Syst 41(3):52:1–52:43. https://doi.org/10.1145/3547333
151. Whittlestone J, Nyrup R, Alexandrova A, Cave S (2019) The role and limits of principles in AI ethics: towards a focus on tensions. In: Proceedings of the 2019 AAAI/ACM conference on AI, ethics, and society, AIES '19. Association for Computing Machinery, New York, NY, USA, pp 195–200. https://doi.org/10.1145/3306618.3314289
152. Wilkinson D, Namara M, Patil K, Guo L, Manda A, Knijnenburg B (2021) The pursuit of transparency and control: a classification of ad explanations in social media (2021). https://doi.org/10.24251/HICSS.2021.093
153. Witten IH, Frank E, Hall MA, Pal CJ (2016) Data mining: practical machine learning tools and techniques. Morgan Kaufmann. Google-Books-ID: 1SylCgAAQBAJ
154. Wong RY, Madaio MA, Merrill N (2023) Seeing like a toolkit: how toolkits envision the work of AI ethics. Proc ACM Human-Comput Interaction 7(CSCW1):145:1–145:27. https://doi.org/10.1145/3579621
155. Xu W (2019) Toward human-centered AI: a perspective from human-computer interaction. Interactions 26(4):42–46
156. Yildirim N, Pushkarna M, Goyal N, Wattenberg M, Viégas F (2023) Investigating how practitioners use human-AI guidelines: a case study on the people + AI guidebook. ArXiv:2301.12243 [cs]
157. Zaken MvA (2022) Impact assessment fundamental rights and algorithms - Report - Government.nl. https://www.government.nl/documents/reports/2022/03/31/impact-assessment-fundamental-rights-and-algorithms. Last Modified: 2022-06-17T13:38 Publisher: Ministerie van Algemene Zaken
158. Zicari RV, Brodersen J, Brusseau J, Düdder B, Eichhorn T, Ivanov T, Kararigas G, Kringen P, McCullough M, Möslein F, Mushtaq N, Roig G, Stürtz N, Tolle K, Tithi JJ, van Halem I, Westerlund M (2021) Z-Inspection: a process to assess trustworthy AI. IEEE Trans Technol Soc 2(2):83–97. https://doi.org/10.1109/TTS.2021.3066209. Conference Name: IEEE Transactions on Technology and Society

159. Zytko D, Wisniewski JP, Guha S, P S Baumer E, Lee MK (2022) Participatory design of AI systems: opportunities and challenges across diverse users, relationships, and application domains. In: Extended abstracts of the 2022 CHI conference on human factors in computing systems, CHI EA '22. Association for Computing Machinery, New York, NY, USA, pp 1–4. https://doi.org/10.1145/3491101.3516506