Journal of the Association for Information Systems

Volume 26
Issue 5 *Special Issue: Digital Responsibility: Social, Ethical, and Ecological Implications (pp. 1222-1389)*

Article 6

2025

Responsible AI Design: The Authenticity, Control, Transparency Theory

Andrea Rivera
University of Hawaii, Manoa, alrivera@hawaii.edu

Kaveh Abhari
San Diego State University, kabhari@sdsu.edu

Bo Xiao
University of Hawaii at Manoa, boxiao@hawaii.edu

Recommended Citation

Rivera, Andrea; Abhari, Kaveh; and Xiao, Bo (2025) "Responsible AI Design: The Authenticity, Control, Transparency Theory," *Journal of the Association for Information Systems*, 26(5), 1337-1389.
DOI: 10.17705/1jais.00948
Available at: https://aisel.aisnet.org/jais/vol26/iss5/6

Journal of the Association for Information Systems (2025) **26**(5), 1337-1389
doi: 10.17705/1jais.00948

RESEARCH ARTICLE

ISSN 1536-9323

Responsible AI Design: The Authenticity, Control, Transparency Theory

Andrea Rivera,[1] Kaveh Abhari,[2] Bo Xiao[3]

[1]University of Hawai'i at Mānoa, USA, alrivera@hawaii.edu
[2]San Diego State University, USA, kabhari@sdsu.edu
[3]University of Hawai'i at Mānoa, USA, boxiao@hawaii.edu

Abstract

Rapid advancements in artificial intelligence (AI) have heightened the need for ethical AI design principles, positioning responsible AI at the forefront across academia, industry, and policy spheres. Despite the plethora of guidelines, responsible AI faces challenges due to fragmentation and the lack of a cohesive explanatory theory guiding research and practice. Existing AI literature frequently fixates on responsible AI attributes within usage contexts, operating under the misapprehension that responsibility can be achieved solely through specific system attributes, responsible algorithms, or minimization of harm. This narrow focus neglects the mechanisms that interlace design decisions with the realization of responsible AI, thereby undervaluing their profound significance. Similarly, information systems literature predominantly emphasizes the operation and usage of these systems, often bypassing the opportunity to weave ethical principles into AI design from its inception. In response, this study adopted a grounded theory approach to theorize responsible AI design from the perspective of AI designers. The authenticity, control, transparency (ACT) theory of responsible AI design emerged as a result. This theory posits that authenticity, control, and transparency are pivotal mechanisms in responsible AI design. These mechanisms ensure that ethical design decisions across three domains—architecture, algorithms, and affordances—translate into responsible AI. The ACT theory offers a parsimonious yet practical foundation for guiding research and practice, aligning ethical AI design with technological advancements and fostering accountability, including algorithmic accountability.

Keywords: Responsible AI, AI Ethics, AI Design, Artificial Intelligence, Authenticity, Transparency, Control, Algorithmic Accountability

Jan Recker, Sutirtha Chatterjee, Janina Sundermeier, and Monideepa Tarafdar were the accepting senior editors. This research article is part of the Special Issue on Digital Responsibility: Social, Ethical, and Ecological Implications; it was submitted on November 19, 2023, and underwent three revisions.

1 Introduction

Artificial intelligence (AI) is reshaping our world, driving profound transformations across industrial, social, and environmental landscapes. AI agents are systems designed to analyze and learn from data, generate insights, and take action to achieve predefined goals (Berente et al., 2021; Mikalef & Gupta, 2021; Vassilakopoulou et al., 2022). Originally developed to emulate human thought processes, these agents have transcended their initial mandate. Today, they are used for a variety of purposes, such as streamlining emergency management systems (Nussbaumer et al., 2023), improving medical diagnostics and treatment (Davenport & Kalakota, 2019), and assisting with everyday tasks like meal planning and personal budgeting (Paris & Buchanan, 2023). On a larger scale, they are entrusted with the noble mission to "make the world a better place," tackling grand challenges such as poverty, climate change, and hunger (Davison et al., 2023, p. 1). Endowed with unprecedented capabilities,

these agents have evolved from mere tools to powerful catalysts of change, revolutionizing numerous aspects of life and inspiring hope for a better future. Yet, amid the immense promise of growth, ethical concerns lurk and demand vigilant attention to ensure that the benefits of AI are not overshadowed by its costs. This challenge has fueled growing interest in responsible AI across academia, industry, and policymaking.

Responsible AI refers to AI agents that are carefully designed, thoroughly regulated, and behaviorally oriented to maximize their beneficence while minimizing their potential harm. A commitment to responsible AI ensures that AI technologies are meticulously developed, deployed, and governed with a paramount focus on ethical principles, societal values, and individual rights, as emphasized by Vassilakopoulou et al. (2022). However, the academic and policy discourse on responsible AI remains rife with a conspicuous absence of coherence regarding its defining principles (Constantinescu et al., 2021), resulting in a fragmented understanding of how responsible AI can be achieved (Anagnostou et al., 2022; Constantinescu et al., 2021; Jobin et al., 2019; Koniakou, 2023; Lahiri Chavan & Schaffer, 2023). The technical intricacies and agentic properties of the latest AI generation (Gallivan, 2001; Jöhnk et al., 2021; Lokuge et al., 2019) also raise concerns about the suitability and adaptability of existing responsible innovation frameworks.[1] For instance, AI's rapid data processing and scalability have far-reaching implications that mandate stringent safeguards and responsible oversight. Additionally, the complexity and opacity of these systems often hinder accountability, while their capacity for recursive self-improvement presents new challenges for oversight and control. These qualities distinguish AI from other digital technologies and underscore the need for a fresh perspective on responsible AI. Yet, responsible AI research remains in its infancy, shaped by assumptions carried over from legacy AI systems of the pre-deep learning era. Drawing on Alvesson and Sandberg's (2011) problematization strategy, we identified three foundational assumptions in the responsible AI literature that warrant closer examination.

First, responsibility in AI is often misconceived as a static set of overlapping agent attributes intended to safeguard users, rather than as a dynamic and evolving behavioral quality of the technology (Mikalef et al., 2022; Sanderson et al., 2023). Framing responsibility as a set of attributes to be developed and checked against specific criteria—such as explainability, fairness, and accessibility—risks reducing it to a post-development checklist, thereby neglecting the critical opportunity to embed responsibility into the core design of AI agents and shape their behavior from the outset. This perspective also downplays the role of system designers, suggesting that their design decisions are inconsequential as long as the agents possess specific attributes or meet certain predefined qualities (Pathirannehelage et al., 2025). However, we argue that to maximize benefits and minimize harm, responsibility must be ingrained as a foundational behavior—cultivated through systematic mechanisms deliberately embedded into the AI design process from its inception.

Second, existing responsible AI literature tends to mischaracterize AI design as primarily focused on algorithmic development (Cheng et al., 2021). This emphasis arises from algorithms' central role in AI decision-making, their technical measurability (e.g., benchmarking), and their historical precedence in highlighting issues such as bias and inaccuracy (Akter et al., 2021; Ferrara, 2024). However, this narrow focus overlooks other critical AI design domains, such as user interactions and the architectural frameworks that support and govern AI agents, which are equally vital in shaping AI behavior and impact. We argue that responsible algorithms alone do not equate to responsible AI. Instead, responsible AI extends beyond algorithm design to encompass other essential design domains integral to the design process, including functional affordances and system architecture.

Third, existing literature broadly frames the primary objective of responsible AI as mitigating risks or reducing AI's potential harm to users. Many studies focus on addressing harms such as biased or unreliable outcomes, privacy violations, or unsafe decision-making. While necessary, this narrow lens limits our understanding of AI's broader impact on users, neglecting the imperative for AI agents to ensure positive outcomes. We propose a more balanced responsible AI theorization that not only prioritizes harm mitigation but also seeks to maximize AI's value to individuals and society.

These misconceptions highlight the need for a broader, more dynamic approach to responsible AI, shifting the focus from static attributes and isolated domains to the deliberate design decisions that shape responsible outcomes. To this end, we sought to develop *an explanatory theory that elucidates the mechanisms linking design decisions to the realization of responsible AI*. This theory must be sufficiently detailed to guide AI designers while remaining parsimonious enough to be applicable across a diverse range of AI applications. Central to this effort is recognizing the pivotal role of designers, who shape AI agents by translating product visions into

[1] Existing responsible AI frameworks frequently build on established responsible innovation theories, such as responsible research and innovation (RRI), the CARE theory of dignity (CARE), corporate digital responsibility (CDR), and value sensitive design (VSD). While these theories offer important conceptual foundations for responsible AI, they often fall short in practical application due to their high level of abstraction, technology-agnostic approach, limited engagement with the nuanced complexities of AI design processes, and, most critically, their inability to effectively link tangible design decisions with operational responsibility mechanisms—see Appendix D for details.

specifications for developers. Yet, current responsible AI discourse leans heavily toward the perspectives of users and developers, often overlooking the critical role of designers.[2] By integrating the designer's viewpoint, we can ensure that responsible AI principles are embedded from the outset—not as afterthoughts, but as foundational elements—enabling the creation of responsible systems through deliberate design.

Toward this goal, we employed an informed grounded theory approach to responsible AI, leveraging the insights of AI designers as industry professionals. Through qualitative interviews, critical mechanisms integral to responsible AI design emerged alongside the distinct design domains in which those mechanisms hold significance. *Authenticity*, *control*, and *transparency* established themselves as fundamental mechanisms that ensure an AI behaves in ways that maximize its beneficence and minimize its maleficence. These mechanisms manifested distinctly across the design domains of *architecture*, *algorithms*, and (functional) *affordances*. Integrating these elements, this study introduces the *authenticity, control, transparency (ACT) theory of responsible AI design* as a parsimonious yet practical foundation for responsible AI research and development. With three fundamental mechanisms elucidated across three distinct design domains, the ACT theory introduces new constructs and relationships with clearly defined boundaries, offering the explanatory power, predictability, and falsifiability essential for robust theorization (Leidner & Gregory, 2024). This theory advances the responsible AI discourse by identifying three fundamental responsibility mechanisms and conceptualizing three foundational categories of design decisions essential to responsible AI. Additionally, this theory strengthens responsible AI practices by defining assurance mechanisms and providing an inclusive framework for embedding responsibility into AI agents from the outset.

The remainder of the paper is structured as follows: We begin by reviewing relevant literature on responsible AI and AI design and follow with a meta-framework we crafted to guide our inquiry with a high degree of theoretical sensitivity. Next, we outline our research methodology, detailing the study setting, data collection procedures, and analytical approaches. We proceed to synthesize our findings into a midrange explanatory theory (Gregor, 2006), illuminating the mechanisms linking AI design decisions to responsible AI. Finally, we conclude by discussing the study's contributions and implications for research and practice, while identifying avenues for future inquiry.

2 Background

The ethical challenges of AI are an undeniable reality, encompassing issues such as privacy violations, systematic biases, and safety concerns. Analyzing data from the AI Incident Database (McGregor, 2020), Wei and Zhou (2023) found that the number of AI-related incidents reported in 2020 was nearly triple that reported in 2015. Adding to these concerns, IBM's Global AI Adoption Index (2023) revealed that less than half of organizations deploying AI take steps to ensure these systems are trustworthy. Only 27% reported efforts to reduce AI biases, and just 37% take measures toward data traceability and oversight. These findings underscore the critical need for responsible AI and cast a harsh light on the AI industry's failure to prioritize and effectively address such pressing challenges.

2.1 Responsible AI Definitions and Guidelines

The domain of responsible AI is an evolving interdisciplinary landscape enriched by diverse discourse on the nature and impact of responsible systems. The academic and policymaking spheres offer a wide array of nuanced definitions and guidelines (Zimmer et al., 2022), each providing unique insights into the field's core principles, intended impact, and applicable contexts. Popular definitions describe responsible AI as a governance framework (Wang et al., 2020), a set of principles (Mikalef et al., 2022), and a practice (Vassilakopoulou et al., 2022), applicable to AI use (Mikalef et al., 2022), system development (Dignum, 2019), or the entire AI lifecycle (Wang et al., 2020; Vassilakopoulou et al., 2022). Despite this diversity, a common thread emerges: Responsible AI is fundamentally a sociotechnical construct mandating a symbiotic relationship between intelligent systems and ethical values (Dignum, 2019; Vassilakopoulou et al., 2022). In light of this understanding, a wealth of guidelines has emerged from both academic research and policy discourse, underscoring the growing recognition of the need for responsible AI. For instance, the European Commission (EC) established a high-level AI expert group in 2019 to develop the *Ethics Guidelines for Trustworthy Artificial Intelligence*, emphasizing that trustworthy AI should be both *lawful* and *robust*. The guidelines identify seven essential requirements for trustworthy AI: (1) human agency and oversight, (2) technical robustness and safety, (3) privacy and data governance, (4) transparency, (5) diversity, non-discrimination, and fairness, (6) societal and environmental well-being, and (7) accountability (European Commission, 2019).

[2] In this study, *designers* refer to a group of individuals who translate a product vision into actionable design specifications, architecting the AI's functionality, structure, objectives, and interface. We differentiate designers from *developers*, who program and implement these designs, creating the technical infrastructure that enables the AI to serve users. *Users* interact with the AI to achieve their goals, providing feedback that informs iterative improvements to both design and functionality.

Alongside the EC's guidelines, the burgeoning academic literature on responsible AI has introduced several additional guiding principles (summarized in Appendix A), including non-maleficence (Floridi et al., 2018), accuracy (Bao et al., 2023; Maalej et al., 2023), inclusivity (Figueras et al., 2022), and control (Polyviou & Zamani, 2023; Soma et al., 2022). Though often used interchangeably, these terms sometimes harbor divergent or supplementary meanings, adding complexity to the conceptual landscape.[3] This heterogeneity is both enlightening and confounding: It enriches our understanding of responsible AI by providing a spectrum of viewpoints, yet it also complicates efforts to establish a cohesive and unified framework (Constantinescu et al., 2021; Jobin et al., 2019). Despite attempts by information systems (IS) scholars to elucidate responsible AI principles, the academic and policy discourse remains fragmented with a nebulous understanding of responsible AI design and its best practices (Anagnostou et al., 2022; Constantinescu et al., 2021; Jobin et al., 2019; Koniakou, 2023; Lahiri Chavan & Schaffer, 2023). The next section explores how this lack of clarity has hindered progress in responsible AI design.

2.2 Responsibility in AI Design

Within the responsible AI design literature, a few studies have taken a comprehensive approach to incorporating responsibility into the design process. For example, Sanderson et al. (2023) interviewed AI designers and developers to explore how AI ethics principles are implemented in practice. While they did not synthesize their findings into a theory or framework, their discussions with participants highlighted the trade-offs and implementation processes associated with responsible AI, such as privacy, safety, and transparency. Likewise, Metcalf et al. (2019) found that institutional logic—such as meritocracy, technological solutionism, and market fundamentalism—often conflicts with AI ethics during product design and development. Building on these findings, Ali et al. (2023) found that ethics prioritization and team reorganization pose significant challenges to implementing ethics principles when developing a new AI system. Focusing on design, Peters et al. (2020) proposed a five-phase responsible design framework—research, insight, ideation, prototype, and evaluation—for integrating well-being and impact analyses into AI development. However, they did not specify which responsible AI principles should be evaluated during the impact analysis. While these studies adopt a comprehensive approach to responsible AI design, they fall short of offering a unified theory that effectively addresses the practical challenges of responsible AI design.

Despite these strides toward responsibility, much of the responsible AI design literature remains narrowly focused on engineering solutions that address isolated principles, with a strong focus on algorithms and harm minimization. For instance, the literature often prioritizes the design of "explainable" AI (Sanderson et al., 2023), which is primarily an algorithmic concern. Mohseni et al. (2021) proposed nested layers of design and evaluation—interpretable algorithms, explainable interfaces, and system goals—to enhance explainability. Similarly, techniques like LIME (local interpretable model-agnostic explanations) and SHAP (Shapley additive explanations) have been introduced to achieve specific goals related to interpretability and explainability (Gaspar et al., 2024). Other studies have focused on methods to combat algorithmic biases (Richardson & Gilbert, 2021), such as fairness scores and certifications (Agarwal et al., 2023), as well as approaches to enhance user privacy, including privacy-preserving machine learning (PPML) and decentralized federated learning (McMahan et al., 2023; Zhou et al., 2024). While these contributions provide valuable insights into specific responsible AI principles, they fail to systematically explain how these principles can be integrated into a cohesive framework for designing *responsible* AI.

2.3 Responsible AI Design Literature—A Critical Examination

Following the formal problematization procedure outlined by Alvesson and Sandberg (2011) and Chatterjee and Davison (2021), we critically examined the prevailing responsible AI design literature and challenged three dominant assumptions. We assert that these three key misconceptions impede the effective integration of responsibility principles into AI design, resulting in fragmented practices and inconsistent outcomes across the field.

First, the prevailing literature—primarily informed by classic research on artificial "narrow" intelligence (Kuusi & Heinonen, 2022)—conceptualizes responsibility in AI as a collection of overlapping attributes rather than viewing it as a dynamic behavior shaped by the decisions of diverse stakeholders (Mikalef et al., 2022; Sanderson et al., 2023). This perspective emphasizes user impact as an end goal, neglecting the ongoing process of ethical alignment influenced by system designers, developers, users, and policymakers (Pathirannehelage et al., 2025; Sanderson et al., 2023). Among the few studies that adopt a stakeholders' perspective, most focus on system

[3] For instance, the terms transparency, explicability, and explainability are often used to reference the same objective of explaining how an AI works and why it arrived at certain outcomes (European Commission, 2019). However, some argue that transparency and explicability are different—an AI could be considered transparent if millions of lines of code are made available for inspection, but an explicable AI would make that code intelligible to humans (Bartneck et al., 2021; Floridi et al., 2018).

development (e.g., Monshizada et al., 2023; Sen et al., 2022; Xiao et al., 2024), system governance (e.g., Fedorowicz et al., 2019; Giffen & Ludwig, 2023; Gregor, 2024), or user interactions (e.g., Abdel-Karim et al., 2023; Bauer et al., 2023; Deng, 2022; Jussupow et al., 2022; Siemon et al., 2022), without giving due emphasis to the critical role product teams, particularly designers, play in shaping ethical AI outcomes. Furthermore, attribute-based approaches reduce responsibility to a checklist of features such as explainability, fairness, or transparency, oversimplifying complex ethical considerations and encouraging superficial compliance. Treating these principles as isolated attributes rather than components of a holistic design approach results in limited guidance on integrating them into an AI's design and risks a misalignment between intended and actual agent behavior (Dignum, 2019; Morley et al., 2020; Pathirannehelage et al., 2025; Sanderson et al., 2023).

The second issue lies in the mischaracterization of AI design as being solely about algorithmic design (Cheng et al., 2021). This narrow focus stems from algorithms' pivotal role in AI decision-making, the relative ease of their technical analysis, and their historical significance in uncovering issues of bias and harm (e.g., Akter et al., 2021; Ferrara, 2024). However, responsible AI encompasses more than responsible algorithms, requiring a comprehensive examination of the AI's purpose, functionality, usability, and structure (De Silva & Alahakoon, 2022; Georgievski, 2023). Overlooking these broader aspects of AI design can lead to fragmented and isolated solutions that fail to effectively integrate into the overall system.

One neglected area in responsible AI design is affordances, particularly *functional affordances* planned by designers. The AI literature often justifies this oversight by hiding behind or overvaluing perceived affordances—how users interpret a system's potential actions based on their experiences, capabilities, and backgrounds (Gaver, 1991; Gibson, 1977)—which are not entirely within designers' control. This focus on perceived affordances is problematic because anticipating all user interactions and consequences is difficult (if not impossible). Such an approach risks creating designs that appear responsible but do not actually behave responsibly.[4] In contrast, functional affordances represent the specific capabilities or action possibilities deliberately planned by designers (Markus & Silver, 2008; Seidel et al., 2013). This concept aligns closely with responsible AI design, as it underscores the proactive role of designers in shaping functionalities and behavior to meet defined goals. By focusing on the AI's planned capabilities and emphasizing the designers' intentional contributions, we can better integrate ethical considerations throughout the AI lifecycle, ensuring that responsibility is not merely an attribute of AI algorithms but a fundamental element of the system's behavior and outcomes.

The third misconception assumes that the primary objective of responsible AI is solely harm mitigation. Many researchers define or characterize responsible AI as a sociotechnical construct, emphasizing the need to align the AI's technical design and functionality with societal values, norms, ethical principles, and users' goals (Dignum, 2019; Mikalef et al., 2022; Vassilakopoulou et al., 2022; Zimmer et al., 2022). While these definitions do not explicitly state that harm mitigation is the sole purpose of responsible AI, their operationalization often prioritizes reducing risks and adverse outcomes over fostering positive impacts. An AI can meet specific benchmarks for mitigating harm and minimizing adverse effects on users, society, and the environment, yet still be ethically flawed if not designed with integrity from the outset. For example, an agent might technically comply with safety, fairness, and explainability standards but still fail to deliver utility or operational fidelity. While many responsible AI benchmarks focus on harm reduction, true ethical design goes beyond risk reduction and ensures an AI delivers value, upholds integrity, and aligns its operations with ethical values and intended purpose from the outset (Floridi et al., 2018; Jobin et al., 2019).

The challenges resulting from these three misconceptions can be illuminated through an analogy to pharmaceutical drug development. First, responsible drug development goes beyond simply adhering to regulatory and industry benchmarks. Similarly, an AI that meets specific benchmarks for harm reduction, safety, and fairness is not necessarily designed with ethical integrity from the outset. Compliance with standards is essential but insufficient to ensure the overall ethical soundness of the product. Second, ethical drug development involves more than just the chemical formula; it encompasses ethical, affordable, and sustainable production, as well as assurances that the drug can be used and administered responsibly. Likewise, ethical AI design requires integrating responsible AI principles into every aspect of system design, not just focusing on algorithmic attributes. Lastly, an ethical drug is not necessarily one without side effects but one where the positive outcomes significantly outweigh the adverse effects. In the same vein, an AI should be designed to maximize benefits while minimizing harm, ensuring that the AI's utility and ethical alignment far exceed any potential downsides.

[4] For example, an AI healthcare agent might be designed to provide reassuring feedback to patients about their health; however, the quality of this feedback depends heavily on the underlying data. Users may perceive this feedback as sufficient and trustworthy, but designers should avoid building on such possible perceptions and instead incorporate affordances and constraints that ensure the system behaves responsibly, regardless of user interpretations.

Table 1. Key Assumptions in Responsible AI Literature: Their Impacts and Our Approach for Redress

Assumptions	Impact on responsible AI	Our approach
The misconception that responsibility in AI is simply a set of protective attributes. (Akbarighatar, 2022; Bartneck et al., 2021; European Commission, 2019; Floridi et al., 2018; Jobin et al., 2019)	Fails to treat responsibility as a process requiring intentional design decisions and assurances that shape responsible behavior.	Enhance theoretical understanding of responsible AI by considering responsibility as an inherent behavior of the AI, planned from the outset of the design process.
The mischaracterization of AI design as solely algorithmic design. (Amugongo et al., 2023; Emdad et al., 2023; Huang et al., 2024; Liu et al., 2022; Olorunsogo et al., 2024; Radanliev et al., 2024)	Neglects other AI design domains that are equally vital in shaping responsible AI behavior.	Expand the theoretical boundaries by including other design domains, such as functional affordances and architecture.
The reductionist view that the purpose of responsible AI is mainly to minimize harm. (Akter et al., 2021; European Commission, 2019; Figueras et al., 2022)	Fails to consider the noble aim of responsible AI to ensure beneficence while ensuring ethical and responsible use.	Identify key mechanisms that uphold an AI's ethical integrity through responsible design decisions, without exclusively emphasizing harm prevention.

Table 1 presents our critique of key assumptions in the responsible AI literature and our approach to resolving them. We contend that addressing these conceptual missteps and enhancing existing responsible AI frameworks can cultivate more consistent, robust, and ethically sound AI design practices. This study seeks to develop an explanatory theory for responsible AI design, tackling prevailing misconceptions in the extant literature and elucidating the foundational mechanisms that shape rather than merely characterize responsible AI. By broadening the scope of responsible AI design beyond algorithms, we aim to discern and differentiate the pivotal components of AI design, scrutinizing each for its impact on key responsible outcomes. This comprehensive perspective ensures the integration of ethical values from the very inception of the AI design process.

2.4 A Meta-Framework for Responsible AI

To guide our theoretical inquiry, we established a *meta-framework* [5] as a reference point for scrutinizing mechanisms that explain how design decisions lead to responsible AI (Themelis et al., 2023). Drawing on Sartrean ethics (Sartre, 1943/1958, 1983/1992) and d'Anjou's (2010) analysis of responsible design, we identified three core *meta-mechanisms* that link ethical decisions to ethical outcomes: the authenticity of decisions, ownership of decisions, and clarity of decisions. These meta-mechanisms, detailed below, are unified by a shared focus on upholding integrity throughout the process of ensuring ethical alignment.

The authenticity of decisions: This meta-mechanism resonates with Sartre's concept of authenticity, which emphasizes living in alignment with one's true values and purpose. It ensures that design decisions not only reflect the intended purpose but also uphold ethical principles. Authenticity, in this context, denotes a harmonious alignment between the AI's actions, its foundational values, and the intentions of its designers. This meta-mechanism transcends mere compliance with predefined rules; it focuses on identifying and operationalizing specific qualities, capabilities, and capacities that enable an AI agent to act authentically in dynamic and complex environments.

The ownership of decisions: This meta-mechanism draws on the Sartrean concept of freedom, realized through the ownership of decisions and actions. It recognizes that true freedom requires empowering users with meaningful agency over the AI's functions. Beyond merely setting boundaries or safeguards, it actively enables users to influence—and, when necessary, challenge—the AI's decision-making processes, particularly in contexts where agency is shared between human users and the AI. This meta-mechanism ensures that users maintain meaningful control over the AI's actions and are equipped to take responsibility for the outcomes of their interactions with it. In doing so, it aligns with Sartre's assertion that freedom is not merely the absence of constraint but the conscious and deliberate exercise of choice, shaping one's existence and interactions with the world.

The clarity of decisions: This meta-mechanism, inspired by Sartre's concept of reflective consciousness, underscores the critical role of transparency and accountability in AI agents. It asserts that ethical action requires not only virtuous intentions but also a clear understanding of the motives, implications, and potential consequences of decisions. This involves rigorously evaluating design goals and functions against established

[5] A meta-framework is an overarching conceptual structure that organizes and connects key ideas, theories, or components to provide a comprehensive lens for analyzing and guiding complex systems or processes.

ethical objectives, systematically assessing the system's performance and openly disclosing unintended consequences. Crucially, this mechanism emphasizes the need to ensure that motives, causes, actions, and impacts are communicated transparently to stakeholders. This approach aligns with Sartre's emphasis on responsibility and his insistence on navigating ethical ambiguity with honesty and transparency by openly acknowledging the reasoning and values that guide decisions.

These meta-mechanisms were instrumental in shaping our theoretical approach, delineating the criteria for identifying valid explanatory mechanisms within our context—the specific manifestations of our meta-mechanisms that bridge design decisions with responsible outcomes. They bolstered our theoretical sensitivity by enhancing our ability to recognize what was important in the data and interpret it meaningfully (Strauss & Corbin, 1990), ensuring that our emerging theory was coherent with the broader ethics literature.

3 Methodology

We employed a qualitative research design using a grounded theory approach (Glaser & Strauss, 1967; Pratt, 2009; Strauss & Corbin, 1990; Suddaby, 2006) to explore how AI designers conceptualize and implement responsible AI when creating new AI agents. Although responsible AI is a widely discussed topic in academic literature, existing misconceptions, along with the need for coherent principles and a robust theoretical foundation, create an opportunity for inductive theoretical exploration (Iivari, 2023; Leidner & Gregory, 2024; Yin, 2015). This approach allowed us to examine responsible AI from the perspective of industry practitioners, who are relatively free from the constraints of preexisting beliefs and normative frameworks.

We conducted semi-structured interviews with AI designers (Adams, 2015) by using general yet intentional questions. These questions offered us the flexibility to delve deeper into topics, seek clarification, request illustrations, and elicit elaboration from participants. This interview process helped us extract critical design decisions related to our three meta-mechanisms, verify their perceived importance, and understand their predictive role from the designers' perspective. Consequently, the data we gathered was rich in context and perspective, offering novel insights into responsible AI from the vantage point of AI designers, a group of stakeholders underrepresented in extant responsible AI literature.

3.1 Study Setting

To gain insights into how responsible AI is conceptualized and implemented during product design, we conducted 24 interviews with AI designers from June 2023 to March 2024 and attended four responsible AI workshopping events. Adopting a comparable perspective on AI design as proposed by Kane (2021), we characterize an *AI designer* as a professional who envisions and architects an AI agent's functionality, structure, objectives, and interface. This role steers the development process, enhances user experience, and optimizes agent performance. In parallel, we define *AI agents* as software entities or systems capable of mimicking autonomous or semi-autonomous intelligent behavior (Berente et al., 2021; Mikalef & Gupta, 2021; Vassilakopoulou et al., 2022). Such agents have the capacity to facilitate labor automation, execute intricate cognitive functions, comprehend objectives, and complete tasks autonomously or in conjunction with other agentic information systems.

We initiated our interviews by engaging with known AI designers and expanded our participant pool through referrals and targeted outreach. This process continued until we achieved theoretical saturation, the point at which additional data no longer yielded new perspectives. The utility of referrals in our research process was not merely incremental but transformative. Leveraging this network-based approach for participant recruitment, we were able to target and engage informants of the highest caliber. These individuals were not just professionals tangentially related to AI; they were bona fide experts whose insights were precisely aligned with our investigative focus on AI design. Consequently, the data we collected were enriched by the quality of the contributors, minimizing noise and maximizing relevance and depth.

All interviews were conducted via online video conferencing applications, primarily Zoom, except for two instances in which one in-person and one correspondence interview took place (due to the participant's inability to meet via conference call). Each interview ranged from 24 to 59 minutes, with an average interview time of 35 minutes. Each participant had at least three years of direct experience in AI design and had participated in at least one major AI project implementation in the past year. Our study deliberately diversified the participant pool to provide a comprehensive view of AI design across various sectors. Participants in the study spanned various professional roles, each offering a unique lens on AI design. We engaged with digital entrepreneurs designing AI-enabled products, and R&D professionals who designed AI agents in collaboration with large corporations, higher education institutions, or government agencies. Additional participants included specialists in government focusing on AI design, and executives overseeing the AI design process at technology firms.

Table B1 in the Appendix presents a profile of each participant, including organizational affiliation, gender, residence, and interview duration. The demographic distribution of our sample mirrored industry trends, with a notable gender disparity—only two participants were female. Geographic diversity was modestly reflected, with four participants residing outside the United States. Although we recognize the constraints of our sample, which may affect the generalizability of our results, the insights derived from our participants significantly

enhanced our theoretical development. Notably, we observed no significant differences in the core themes of responsible AI design across participants' organizational affiliations (i.e., AI startups, government agencies, technology firms, or R&D institutions), gender, or geographical location.

Employing Corbin and Strauss's theoretical sampling approach (Corbin & Strauss, 2015), we strategically interviewed new AI designers to deepen our understanding and address data gaps. The iterative process of gathering, analyzing, and comparing data, coupled with deliberate theoretical sampling, ensured that our theory development was robust and grounded in diverse empirical evidence. During interviews, participants answered six questions (provided in Table B2) designed to explore their definitions and implementations of responsibility in AI design. These questions encouraged participants to reflect on their specific practices and steps for addressing ethical concerns in AI design, as well as the challenges they face in implementing responsible design practices. When deemed appropriate, we asked follow-up questions to delve deeper into their responses and better understand their perspectives. Following ethical guidelines and our interview protocol, all participants consented to participate in our research project, and all but four agreed to have their interviews audio-recorded. To maintain confidentiality, we took measures to anonymize participant identities, removing any personally identifiable information from the project documentation. Audio-recordings, transcripts, and notes were referenced only by randomly assigned numbers to ensure the privacy of our participants and their respective institutions.

3.2 Data Collection & Analysis

Following Urquhart et al.'s (2010) recommendations, our data collection and analysis adhered to a cyclical and interconnected approach. We applied techniques for logging, collating, and reviewing data, as suggested by Glaser and Strauss (1967). Immediately following each interview, we created reflection notes to highlight key takeaways from the discussion and capture the participant's overall disposition during the interview. Within one week after each interview, the first author enhanced these notes to create more comprehensive memos by revisiting available audio-recordings and incorporating direct quotes, time stamps, and additional commentary into the existing reflection notes (Mohajan & Mohajan, 2022). Using this expanded data set, we collaboratively engaged in open coding, identifying key responsible AI concepts and themes that were then grouped into first-order indicators (Gioia et al., 2013; Strauss & Corbin, 1990). To ensure the comprehensiveness of our coding, the authors cross-verified first-order indicators. Next, we synthesized these first-order indicators into second-order themes (axial coding), categorizing them for further analysis (Gioia et al., 2013). This process was informed by but not confined to the meta-mechanisms identified earlier, which helped structure our axial coding options and ensured comprehensive coverage of responsible AI principles. To enhance rigor, we systematically documented every decision made during coding to maintain transparency and consistency. After each successive interview, we revisited the cyclical, reflexive coding process, employing constant comparative analysis to validate emerging themes and refine our understanding of the data. As we advanced to selective coding, we engaged in peer debriefing to cross-check interpretations and minimize potential bias. This iterative process culminated in the aggregation and abstraction of second-order themes into higher-order aggregated dimensions, ensuring that the final framework was both empirically grounded and theoretically robust (Gioia et al., 2013).

During this coding process, we incorporated an integrated literature review to compare our findings with established research on responsible AI (Birks et al., 2013; Urquhart & Fernández, 2013). This iterative process of juxtaposing interview insights with prevailing literature was instrumental in sharpening our understanding of core concepts and their interconnections (Charmaz, 2006; Urquhart et al., 2010). It also allowed us to frame these concepts in alignment with the established responsible AI lexicon and our three meta-mechanisms. By maintaining terminological consistency with prior literature, we enhanced the clarity and accessibility of our work, facilitating comparisons with earlier frameworks and ensuring its future transferability. Appendix C provides additional details on our data analysis process, including illustrative examples of each first-order indicator pertaining to authenticity, control, and transparency.

4 Summary of Results

Our interviews provided valuable insights into how AI designers conceptualize and approach responsible AI. Despite differences in contexts and methodology, commonalities emerged in their understanding of its essence and raison d'être. Notably, while the operationalization of responsible AI varied, designers converged on the fundamental goal of designing AI agents that maximize beneficence and minimize maleficence. Interestingly, while beneficence was often framed in relation to the objectives of a given AI agent, maleficence was construed in more universal terms, transcending individual use cases. This duality underscores the nuanced perspectives designers adopt when balancing the ethical dimensions of AI behavior. Participants also detailed their strategies for embedding responsibility into the design process to ensure AI agents align with responsible AI principles. In the following section, we outline the key themes—responsibility mechanisms and design decisions—that emerged from our analysis, setting the stage for a deeper exploration of the specific approaches employed.

Table 2. ACT Mechanisms and Definitions

Responsible AI mechanism	Definition
Authenticity	The process of ensuring the AI consistently produces dependable outcomes that resonate with its intended purpose and adhere to ethical values.
Control	The process of ensuring the AI empowers stakeholders with regulated control over the AI's behavior and enables them to influence, direct, or manage its actions and outputs within predefined limits.
Transparency	The process of ensuring the AI exhibits its operations, logic, and data stewardship in a clear, open, and auditable manner, so that stakeholders can easily understand its behavior and outcomes.
Note: In this study, we characterize *stakeholders* as developers, system architects, system administrators, and users who collaborate with AI designers or are directly impacted by their decisions.	

4.1 Responsibility Mechanisms

Our characterization of meta-mechanisms (i.e., authenticity, ownership, and clarity of decisions), combined with an integrated literature review conducted during data collection and analysis, led to the identification of three fundamental responsibility mechanisms[6] early in the analysis—*authenticity, control,* and *transparency* (ACT). Authenticity emerged as the process of ensuring the AI behaves with integrity and operational fidelity; control as the process of ensuring the AI empowers stakeholders with regulated control over its behavior; and transparency as the process of ensuring the AI exhibits its operations, logic, and functionalities clearly, openly, and in an auditable manner. Table 2 delineates ACT mechanisms with their definitions.

We theorize these mechanisms not as intrinsic qualities per se but as three critical intermediary processes that effectively translate design decisions[7] into responsible behavior during interactions with human agents. As we systematically gathered and analyzed new data, it became evident that these three responsibility mechanisms are pivotal in transforming responsible AI design into responsibly behaving agents. In essence, design decisions that fail to operationalize technical implementation through these mechanisms are more likely to produce agents that fall short of responsible AI standards, eroding trustworthiness and undermining core human values. A detailed discussion of each ACT mechanism is provided in Sections 5 to 7.

4.2 Design Decisions

When coding new data under the three responsibility mechanisms, we found that AI designers frequently referenced three critical AI design domains: *affordances*, *algorithms*, and *architecture*. Design decisions within these domains initiate the responsibility mechanisms of authenticity, control, and transparency, which collectively ensure that an AI behaves responsibly, maximizing its beneficence and minimizing its maleficence.

Affordances refer to a system's actual, built-in properties and capabilities (Markus & Silver, 2008; Seidel et al., 2013). This definition aligns with the concept of *functional affordances* in the IS literature (i.e., *planned* or *designed* affordances in human-computer interaction, HCI, literature). Functional affordances represent the action possibilities provided by a digital artifact's attributes (features, functions, and behaviors), intentionally designed based on anticipated user goals (Markus & Silver, 2008; Seidel et al., 2013).[8]

[6] *Mechanisms* refer to the "underlying entities, processes, or structures which operate in particular contexts to generate outcomes of interest" (Astbury & Leeuw, 2010, p. 368). In this study, authenticity, control, and transparency can be understood as foundational processes that ensure that ethical conduct is embedded across an AI's affordances, algorithms, and architecture to shape responsible behavioral outcomes.

[7] *Design decisions* are deliberate decisions made during the design process to shape an AI's functionality, structure, and behavior, ensuring alignment with its intended purpose and goals (Hevner et al., 2004; Simon, 1996).

[8] The conceptualization of functional affordances is informed by Norman (1999), who defined affordances and constraints as the design decisions shaping an artifact. This view partly contrasts with Gibson's (1977) perspective—more popular in the IS literature—of affordances as action possibilities available in the environment. Norman argues that the nature of design decisions underlying artifacts' affordances and constraints informs their intended use. This perspective aligns more closely with design-centered studies like ours, which focus on design decisions with defined boundaries rather than users with open-ended goals and needs. Choosing between these two schools of thought is not ideological but context dependent, based on design vs. use. Both perspectives, however, recognize that the materiality of technology (functional/designed affordances) influences but does not determine the possibilities for users, impacting their ability to benefit or harm themselves and others. The IS literature that adopts Gibson's view primarily applies the concept within the user domain, focusing on affordances that emerge from interactions with technologies (Leonardi, 2011). In summary, while we do not dismiss the role of user agency in actualizing affordances (Faraj & Azad, 2012), we emphasize that designers have the agency to plan functional affordances and constraints in ways that encourage or discourage certain action possibilities, in line with accepted ethical and social norms. This is the crux of responsible design.

Table 3. AI Agent Design Domains

Design domain	Aim
Affordances*	Designing specific capabilities and actions that a system or technology enables for its users, which can be actively used or passively realized. These affordances facilitate the meaningful and effective use of the technology.
Algorithms	Designing a set of computational procedures—including parameters, logic, and mathematical underpinnings—which form the intellectual core of the AI agent. These algorithms dictate the behavioral attributes and overall efficacy of the AI agent.
Architecture	Designing the structural foundation of the AI agent—encompassing computational engines, system configurations, data architecture, and systemic interconnections—which collectively enable an AI agent to operate seamlessly.
Note: * Limited to functional affordances designed or planned by designers.	

In essence, designers embed these possibilities for interaction within the artifact, allowing users to explore and leverage them to achieve desired outcomes. Functional affordances align closely with responsible AI design—by emphasizing deliberate design functionalities, designers can ensure that AI agents offer users clear and predictable ways to interact and achieve their goals, thus reducing the risk of unintended consequences or misleading user perceptions.

AI algorithms consist of the parameters, computational logics, and mathematical frameworks that govern an AI's behavior, data processing, and decision-making. They define the methodologies for data acquisition, analysis, and the generation of actionable insights (Berente et al., 2021; Cormen et al., 2009; Martin, 2019). Notably, we argue that data selection and processing fall within the algorithm design domain due to their intrinsic interdependence in the AI design process (Chen et al., 2020; De Loera et al., 2021). Data directly influences the selection of algorithms used for training, shaping their ability to recognize patterns, make predictions, and generate insights. Conversely, the choice of algorithms dictates the specific requirements for data quality and format, which in turn guide the necessary data preprocessing steps and structural configurations for optimal performance. Consequently, the design choice of an algorithm inherently determines the type of data required and its preprocessing protocols. This interdependence justifies classifying data selection and processing within the algorithm's design domain.

Finally, AI architecture serves as the agent's structural backbone, organizing and interconnecting the components that enable the agent to operate seamlessly. It encompasses the design of hardware and software elements and the relational dynamics that govern their interaction, forming the foundational layer upon which affordances and algorithms are constructed (As & Basu, 2021; Castro Pena et al., 2021). Our analysis revealed that some design challenges stemmed not from affordances or algorithms but from architectural flaws. Even if affordances and algorithms are designed responsibly, architectural deficiencies can prevent the agent from behaving responsibly. Table 3 summarizes the conceptual boundaries of the three AI design domains.

The following sections examine the nuanced application of ACT mechanisms across specialized domains of AI design. Specifically, they present the overall data structures for authenticity, control, and transparency in AI design, detailing first-order indicators and second-order themes across the three domains of affordances, algorithms, and architecture. These design decisions activate responsibility mechanisms that collectively ensure an AI operates responsibly, maximizing its beneficence while minimizing its maleficence.

5 Findings—Authenticity in AI Design

Our data analysis revealed that designers prioritize ensuring that an AI's purpose and actions are genuinely aligned with ethical principles and human values. We labeled this mechanism "authenticity," as it ensures the AI consistently operates according to virtuous goals. This aligns with our first meta-mechanism, authenticity of decisions, which emphasizes that a responsible agent should inherently embody its ethical intent, steering clear of self-deception or external pressures. The authenticity mechanism underscores the AI's veracity and reliability, ensuring that it meets performance benchmarks, generates dependable and credible results, and provides utility to users. Authentic AI agents thus maintain fidelity to their intended purpose and function while attentively addressing users' real needs and preferences.

5.1 Designing Affordances for Authenticity

Our research findings reveal that the mechanism of authenticity is pivotal in shaping an AI that behaves responsibly. Our data suggest that the multifaceted concept of authenticity intricately weaves through various layers of (functional) affordance design—ranging from the AI agent's interface to its functional capabilities—all scrupulously engineered to engender interactions that ensure each user experience is valuable,

accurate, reliable, and ethically sound. Our research unveils that embedding such a nuanced level of authenticity in an AI's affordances is a calculated endeavor, underscored by two primary design decisions: (1) allowing users to actualize the AI's usability and usefulness and (2) allowing users to engage with the AI in a way that ensures high fidelity in their interactions.

First, participants highlighted the importance of designing AI agents that empower users to fully realize their usability and usefulness. This requires thorough research into the AI's purpose, functionality, and value, with a focus on understanding and meeting individual users' needs. One designer emphasized: "We have to research why the system is needed, how it will perform, and how it will impact people's lives" (AI designer, AI startup). As another designer put it: "The first step in design is really about understanding the system's potential value. We need to focus on what it offers to our customers and the company and then ask them directly if they consider it truly valuable" (AI designer, R&D institutions).

Participants also stressed the importance of tailoring AI outputs to the cultural and ethical contexts in which they are used. They noted that ethical norms could vary significantly across communities and warned that violating these norms could undermine an AI's utility. One designer elucidated this point with an anecdote concerning OpenAI's ChatGPT:

> *[Where I come from], our economy is more or less based on energy. Some people going around sticking some straws in the ground to pull out some oil isn't necessarily a bad thing. In the early days, if you would ask OpenAI directly, "What's the solution to the energy crisis?" they would say drill more oil. The editors didn't like that response—they didn't think it was an ethical or sustainable response ... and so they edited that. So now, if you [ask the same question], it'll talk more about green energy and solar and wind. I think that's where it gets a little tricky: whose ethics you determine are bad.*[9] (AI designer, technology firm)

Second, prioritizing interactional fidelity allows designers to create AI agents that support reliable and accurate user interactions, fostering trust and usability. Participants highlighted the importance of scenario planning to anticipate a diverse range of realistic user interactions and responses, thereby enhancing the authenticity and seamlessness of the user experience. Additionally, they emphasized the critical role of operational accuracy and performance benchmarks in sustaining high-fidelity interactions. Operational accuracy[10] ensures that the AI consistently provides reliable and precise responses, fostering user trust and enabling effective engagement with its capabilities. These benchmarks and operational accuracy requirements are often set based on industry standards, regulatory guidelines, or assessments of human performance in comparable tasks.

These strategies constitute the essential foundation for an AI agent that harmonizes design intentions with user expectations. Nearly all of our participants took time during their interviews to highlight and detail their AI's unique value to users, as well as the associated accuracy and performance requirements necessary for users to capture that value. Our data suggests that AI designers diligently enable users to actualize the AI's utility and fidelity through reliable interactions.

5.2 Designing Algorithms for Authenticity

Our research findings highlight that the authenticity mechanism is rooted in meticulously designed computational logic that allows the agent to fulfill its intended purpose. According to participants, attaining such authenticity is generally orchestrated along two critical dimensions: (1) sourcing and training models with quality data and (2) evaluating, monitoring, and improving the model's accuracy and functional integrity.

First, our participants emphasized the crucial role of training data selection in crafting an authentic and accurate AI agent. Designers frequently invoked the adage "garbage in, garbage out" to underscore the indispensable nature of high-quality, accurate data in constructing reliable AI models. A designer metaphorically compared reliable data to the foundation of a house, positing that an AI cannot perform optimally in its absence. Another designer stressed the importance of intentional data collection, stating: "If you have a business question, try to think about what data can answer that question and then go get the data, not the other way around" (AI designer, technology firm).

On the topic of big data, one participant stressed the importance of collecting datasets that accurately represent the populations the AI is designed to serve. While designers often prioritize larger datasets to train models on, this approach can be problematic if the data contains inherent biases. The participant suggested carefully selecting smaller datasets to improve model performance across the range of intended users. This underscores the critical balance between data volume and data quality.

[9] We tested this claim, and drilling for more oil was not part of the ChatGPT-4o responses.

[10] Operational accuracy refers to the system's ability to perform tasks reliably within a specific context, accounting for real-world constraints and execution conditions. This differs from algorithm accuracy, which primarily measures the model's precision and recall in predicting or classifying based on training data.

> *The bigger the dataset, the bigger the problems ... Everybody thinks right now that you find the needle in the haystack by putting more hay on it, which is actually stupid ... Coming up with small datasets that are more accurate for what you are actually trying to answer is a way to mitigate these kinds of [bias] issues.* (AI designer, technology firm)

Second, participants highlighted the need for ongoing evaluation and improvement to ensure the AI consistently meets performance objectives and accuracy requirements. They suggested that this process begins with selecting the most appropriate machine-learning model to support the AI's purpose. One designer emphasized the importance of regular testing throughout the design and build phases to ensure it consistently meets minimum operational requirements, noting that "in a perfect world, new AI agents would be tested at every stage of development" (AI designer, R&D institutions). Another participant recommended prompt engineering to stress-test the agent, optimize its interactions with users, and ensure that it reliably achieves its objectives. Regarding the impact of prompt engineering, he stated:

> *I feel like that's kind of the biggest area is prompt engineering, where you're able to really stress-test and kind of see whether or not the outputs are going to be okay… The amount of power that you have over the LLM [large language model], just with prompt engineering, is amazing. Like, you can build MVPs—you can build entire products just using GPT and prompt engineering.* (AI designer, AI startup)

Achieving authenticity in AI algorithm design hinges on high-quality data selection, rigorous model training, and continuous refinement. By aligning computational logic with the intended purposes, designers create systems that meet performance standards while delivering reliable and meaningful outcomes.

5.3 Designing Architecture for Authenticity

Our research findings elucidate that authenticity extends to designing an AI's architecture, shaping the structures and mechanisms that can bolster the AI's functional integrity and accommodate the evolving demands and operational needs. According to our data, pursuing such architectural authenticity frequently manifests in two critical design decisions: (1) creating specialized architectures and (2) building adaptable AI agents capable of adjusting to new requirements or conditions over time.

First, specialized AI architectures integrate various techniques to optimize performance and adaptability, with a focus on improving output accuracy for specific tasks. Participants emphasized that such architectures enable agents to excel within defined fields of interest. One designer highlighted retrieval-assisted generation (RAG), an approach that enhances large language model (LLM) applications by incorporating custom data:

> *Specialization in AI allows you to provide more accurate solutions—more sophisticated, more correct solutions—but also prevents the system from hallucinating and prevents the system from having to deal with unethical questions ... We do that by augmenting their knowledge by doing retrieval-assisted generation by grounding their knowledge in sophisticated databases and systems.* (AI designer, AI startup)

Second, participants highlighted the importance of adaptable AI architectures, particularly in terms of scalability, to ensure robustness across diverse platforms. Adaptable architectures allow AI agents to handle increasing data volumes, user demands, maintenance requirements, and computational complexities without compromising reliability. One participant noted that interoperable components improve an AI's effectiveness by seamlessly integrating diverse functional elements. Another underscored the value of "standardized interfaces ... [to] ensure seamless integration so that each part can communicate with the others" (AI designer, technology firm). Scalability emerged as a key feature of adaptable design, allowing individual components to be incrementally adjusted to meet evolving requirements. One participant distinguished scalability into two critical dimensions: operational and computational. Operational scalability focuses on cross-platform interoperability, while computational scalability ensures stable performance under varying demands.

Table C4 in the Appendix presents the overall data structure for authenticity in AI design, including first-order indicators and second-order themes, across the three domains of affordances, algorithms, and architecture. Notably, our findings suggest that design decisions in algorithms and architecture often indirectly shape the affordances provided. For example, specialized architectural design can enhance affordances by improving intended utility and interactional fidelity. Likewise, data quality and operational accuracy—achieved through algorithm design and supported by architectural resilience—ensure that the AI produces consistent, high-quality results aligned with its intended purpose and functionality. These factors collectively enable designers to provide users with reasonable and adequate affordances.

5.4 Discussion on Authenticity

We argue that designing an AI agent transcends technical considerations, fundamentally constituting an ethical endeavor intrinsic to the designer's professional

responsibility. This responsibility mandates that designers imbue every facet of AI creation—from architecture and algorithmic logic to functional affordances—with what we term *authenticity*. Authenticity entails ensuring that the AI upholds operational fidelity and produces reliable outcomes that are aligned with its intended purpose and are harmonious with ethical values. Achieving this requires integrating an ethically robust vision into design decisions that actualize this vision throughout the agent's lifecycle. Moreover, it necessitates ongoing evaluation of the AI's alignment with expectations across diverse user groups and settings, starting with responsible design decisions from the outset.

First, authenticity can be enhanced by meticulously designing an AI's functional affordances to prioritize utility and fidelity. This entails ensuring that users can actualize the AI's usability and usefulness and that their interactions with the AI remain consistent and reliable. Neglecting authenticity in these areas can lead to operational deficiencies, such as misalignments with the agent's intended purpose. These failures undermine the AI's promised functionality, erode user trust, and ultimately hinder positive user interaction and broader adoption (Glikson & Woolley, 2020). Therefore, we propose:

P1: Designers can ensure authenticity in AI agent affordances through design decisions that emphasize **(a)** intended utility and **(b)** interactional fidelity.

Second, in the realm of algorithms, authenticity is enhanced by sourcing and training models with high-quality, representative data that encompasses diverse perspectives without introducing selective biases. An ongoing commitment to rigorously evaluating, monitoring, and improving algorithmic performance ensures that computations produce genuine, unbiased, and reliable outcomes aligned with the agent's intended purpose. Neglecting authenticity in the design of an AI's algorithms can lead to inaccuracies or biases, compromising the AI's integrity. Therefore, we propose:

P2: Designers can ensure authenticity in AI agent algorithms through design decisions that emphasize **(a)** data quality and **(b)** operational accuracy.

Lastly, authenticity in the architectural domain can be achieved by creating specialized and adaptable architectures that meet the agent's performance objectives and intended purpose. This adaptability enables the architecture to respond effectively to evolving user needs, technological advancements, and emerging applications, thereby enhancing an AI's veracity and resilience—two essential components of operational integrity. Therefore, we propose:

P3: Designers can ensure authenticity in AI agent architecture through design decisions that emphasize **(a)** specialization and **(b)** architectural resilience.

In summary, we emphasize that when AI agents lack authenticity, they risk failing to meet expectations, leaving user needs unaddressed, eroding trust, and potentially causing unintended consequences. Prioritizing authenticity in AI design is not merely a technical necessity but an ethical imperative to ensure that AI agents consistently deliver value, maximize utility, and adhere to the principles they are intended to uphold. However, we should also acknowledge the inherent trade-offs in design decisions that support authenticity. For instance, while specialization enhances an AI's performance by optimizing its architecture for domain-specific expertise, it often limits the agent's adaptability to evolving demands, posing challenges to architectural resilience. Similarly, achieving high interactional fidelity through rigorous benchmarks and tailored user engagement strategies may conflict with the design of modular, interoperable components essential for architectural resilience. Furthermore, prioritizing data quality—ensuring reliability, validity, and representativeness—can extend development timelines and resource requirements, potentially delaying the system's ability to deliver its intended utility in dynamic contexts. Navigating these trade-offs requires a deliberate and iterative approach to ensure that the AI remains authentic, robust, and aligned with its operational and functional objectives.

6 Findings—Control in AI Design

Control in AI design entails ensuring that the AI empowers stakeholders with regulated control to influence, direct, or manage its behavior and outputs within predefined limits. Our research revealed that the mechanism of control extends beyond user agency, encompassing the ability of developers and administrators to guide and oversee the AI's actions or outputs. As a responsibility mechanism, control aligns with the ownership of decisions, our second meta-mechanism, emphasizing that responsibility extends beyond noble intentions to encompass full accountability for one's actions and the actions enabled through design decisions. Hence, a virtuous design creates pragmatic capacities for virtuous actions.

6.1 Designing Affordances for Control

The mechanism of control in AI affordances empowers users through output management, customization, and accessibility features while supporting ethical and responsible use through operational safeguards. Our research findings converge on three salient design decisions: (1) allowing users to regenerate and save

output, (2) giving users customization options over functionality, data usage, and interactions, and (3) allowing users to self-regulate their usage to limit misuse, abuse, or overuse of the AI.

First, our research participants emphasized the importance of enabling users to regenerate and save AI-generated content or actions, a capability that enhances control and usability. This feature allows users to retain pertinent information for future reference, granting them greater autonomy over outputs. Regeneration also empowers users to iteratively refine outputs, tailoring them to better meet specific goals and preferences. Through this iterative process, users can verify and improve the quality of outputs, ensuring they align with expectations and standards.

Second, our findings highlight that user agency is essential in shaping an AI's output. Customization allows users to tailor the AI's functionality to their specific objectives, maximizing utility and ensuring relevance. Participants highlighted the importance of adjustable settings or filters to safeguard against inappropriate or irrelevant outputs, especially in sensitive or regulated environments. Customizable features, such as voice-activated options and screen readers, were noted as critical for inclusivity, accommodating diverse user needs, including those with disabilities. Additionally, several participants advocated for system designs that allow users to specify constraints—such as age-appropriate content filters or length limits—to enhance control and output accuracy. One designer summarized this goal as enabling users to "provide the constraints that an answer needs to follow ... [so that] the user can provide the mechanisms to check that the solution is correct" (AI designer, AI startup). These and similar design decisions serve dual purposes by enhancing user control over AI-generated outcomes while improving their accuracy through more precise directives.

Another aspect of customizability involves enabling users to influence the AI's performance characteristics. This can be achieved through various means, such as allowing users to select the training data source, modify the computational logic, or configure underlying procedures. For instance, some designers allow users to choose between models trained on different datasets, enabling them to prioritize reliability, novelty, or other factors based on their goals. One participant described offering users the ability to choose between different models (similar to selecting models in ChatGPT) to accommodate their specific budgetary and use requirements. By offering such options, designers empower users to tailor the AI's capabilities to their unique needs, ensuring flexibility and alignment with individual goals while mitigating potential shortcomings in the AI's default configurations.

Our findings also underscore the critical importance of recognizing users' control over their personal data, highlighting the heightened sensitivity surrounding user-generated data, particularly in contexts where privacy could be compromised. One participant explained: "By allowing users to choose how their data is used during system training and testing … we mitigate the possibility they feel their privacy has been violated" (AI designer, AI startup). Another participant, who is developing an AI for organizational professional development, underscored the critical nature of the data gathered during user interactions. To alleviate users' privacy concerns, he grants them unambiguous data ownership, allowing them to retain control even when transitioning to different organizations. While such a data ownership model may not be applicable across all AI platforms, it offers a valuable example for designers to consider when contemplating strategies to enhance users' control over their data.

Third, some participants highlighted using anomaly detection mechanisms to identify aberrant usage patterns and deter potential agent abuse or overuse. One participant advocated for user-interface controls with built-in mechanisms to flag and limit excessive usage. A designer from an R&D institution emphasized the critical role of "rate limits" or "quota systems" in curbing misuse and preventing overreliance on AI. Regulating user control to ensure that AI agents are harnessed in accordance with their original design intent helps mitigate risks of misuse, abuse, and overuse.

These design decisions for an AI's affordances illustrate how the control mechanism empowers users while defining boundaries that prevent misuse. By enabling output management, customization, data control, and safeguards against exploitation, AI agents effectively align with user needs and ethical norms, providing practical tools for responsible engagement.

6.2 Designing Algorithms for Control

The mechanism of control in an AI's algorithms ensures that the AI responds effectively to user feedback, adheres to predetermined performance metrics, and upholds users' data ownership rights. These aspects of algorithm design empower users to influence and improve the AI's future behavior for their benefit and that of others. We categorized these dynamics into two key design decisions: (1) integrating user feedback to enhance the AI's behavior and (2) securing and respecting user data ownership rights.

First, algorithms can be designed to allow user adjustments through control mechanisms. One approach involves enabling users to define the scope of the AI's output with an understanding of the inherent trade-offs. For instance, users might balance factors such as privacy vs. accuracy, or performance vs. cost. Another advanced mechanism involves implementing continuous learning

algorithms that leverage user-generated feedback to recursively refine and optimize the AI's performance. This could include simple features like thumbs-up/thumbs-down buttons or user ratings on an AI's outputs. AI models can then use this feedback to recalibrate their algorithmic weights or fine-tune operational parameters, enhancing their overall performance.

Second, securing and respecting user data ownership is critical to maintaining trust and enabling user control over their personal information. Participants emphasized the importance of integrating privacy and security measures directly into the algorithmic design. These measures range from straightforward privacy settings to sophisticated data stewardship frameworks that allow users to specify how their data is used for training and output generation. For instance, one designer employs differential privacy, a technique that obfuscates individual data points within a larger dataset (Hilton, 2002). This method adds an extra layer of privacy while preserving the utility of aggregated data for machine learning or analytics. It allows developers to extract valuable insights from the data while ensuring privacy and security controls for individual data points.

The control mechanism in algorithms ensures that the AI empowers users to influence its behavior while addressing critical concerns such as privacy, security, and performance. These design decisions not only enhance the AI's functionality and adaptability but also empower users with meaningful control that fosters trust and ethical engagement.

6.3 Designing Architecture for Control

The mechanism of control requires architectural elements that allow precise interventions in AI-generated outcomes while ensuring robust system governance, going beyond algorithmic control. Participants highlighted the importance of designing architectures that facilitate the effective management of AI behavior through thoughtful and meticulously managed structures. These practices converge into two overarching design themes: (1) embracing modular design and (2) implementing continuous oversight.

First, our findings revealed that implementing a modular design architecture is a practical approach to instilling control. Unlike centralized models, a modular architecture distributes computational tasks among various modules or components, often across different functional areas or systems. This decentralization fosters autonomy within individual segments of the AI, granting developers more granular control over its behavior. It also indirectly enhances user control, as specific modules can be tailored to meet individual needs. Federated learning exemplifies this principle by enabling AI training across multiple devices while keeping data localized, thereby enhancing data sovereignty (McMahan et al., 2023). One participant described federated learning in the following way:

> *If you don't want to share data across individual sources, like say a hospital or even at an individual level, you can train an AI model that's personalized to the individual or that's tuned to the hospital. And then, in order to make a global model that's more general purpose, you just share the model weights rather than the individual data.* (AI designer, AI startup)

Second, continuous oversight enables administrators to monitor and adjust the AI's performance in real time. Our findings strongly support integrating human-in-the-loop (HITL) mechanisms into AI architecture, particularly in sectors with significant ethical or high-stakes implications, such as healthcare or the military. HITL provides a critical validation layer by requiring human authorization for pivotal decisions, safeguarding against errors and ethical lapses. One participant working in AI for behavioral healthcare discussed HITL in the following way:

> *I'm interested in looking at what's called human-in-the-loop AI that combines an AI basically working together with humans—who have that social insight, but not necessarily that diagnostic ability—to work together with AI. So where humans are kind of extracting the relevant social signals and providing that to an AI, which can then make the final diagnosis.* (AI designer, AI startup)

In practical terms, participants recommended leveraging advanced machine learning operations (MLOps) architectures to support the HITL paradigm. Technologies like MLOps streamline AI monitoring by enabling continuous oversight and corrective interventions without disrupting ongoing development or operations. This approach allows for varying degrees of human oversight, creating a dynamic and responsive control mechanism that balances automated efficiency with ethical vigilance.

Table C5 outlines the data structure for control in AI design across our three design domains: affordances, algorithms, and architecture. We observed that certain design decisions for an AI's algorithms and architecture indirectly support affordances within the responsibility mechanism of control. For example, feedback integration within algorithms enhances customizability by incorporating user-generated feedback to refine AI outputs according to user expectations. Similarly, architectural features such as modularity and continuous oversight bolster operational safeguards. Modularity allows for the independent updating or replacement of system components, enabling users to adjust their interactions with the AI and prevent misuse, abuse, or overuse. Continuous oversight facilitates real-time monitoring and adjustments, empowering users to maintain responsible usage and uphold safety standards.

6.4 Discussion on Control

The concept of control encompasses both technical fail-safes for safety and normative algorithms designed to operate within ethically delineated bounds. This dualism frames control as both an epistemological issue—understanding what the AI will do—and a normative assertion—defining what the AI should be allowed to do.

While user agency, manifested through informed control, is essential for an AI to function as a socially responsive entity, this autonomy does not absolve designers of their fundamental ethical responsibility. Effective AI design must prioritize mechanisms that empower users to regulate their interactions with the system, enabling them to mitigate risks such as misuse, abuse, or overuse. However, empowering users should not shift undue responsibility onto them, as this risks creating a moral gray zone where designers evade accountability. Instead, control in AI design should embody a collaborative ethical framework, equipping users with tools to manage their engagement while ensuring these tools are robust, accessible, and aligned with societal values. In this way, control transcends a simplistic view of autonomy, becoming a structured interplay between user empowerment and designer accountability that defines and upholds the moral boundaries of AI use.

From this perspective, we define the responsibility mechanism of control as the process of ensuring the AI empowers stakeholders with regulated control over the AI's behavior and enables them to influence, direct, or manage its actions and outputs within predefined limits. Effective control mechanisms allow stakeholders to shape the AI's functionality toward beneficial outcomes, while obligating developers and system administrators to monitor and refine the agent. This ensures sustained administrative governance and continuous alignment with ethical norms. We contend that it is the designer's responsibility to plan for these possibilities from the outset.

In affordances, the control mechanism is realized through carefully designed interfaces that allow users to customize the AI's behavior, adjust settings, and save and regenerate content. These features facilitate greater user-driven interactions and empower individuals to exercise agency over the AI's capabilities. Simultaneously, the design of affordances should incorporate mechanisms that allow users to regulate their own use of the AI, mitigating risks of misuse, abuse, or overuse. This involves implementing safeguards or guardrails to manage the AI's behavior when users deviate, either intentionally or unintentionally, from its intended purpose. Therefore, we propose:

P4: Designers can ensure control in AI agent affordances through design decisions that emphasize **(a)** regenerability, **(b)** customizability, and **(c)** operational safeguards.

For algorithms, control is achieved by empowering users. Our study highlights the importance of granting user ownership over their AI-generated data, including the ability to decide whether their data can be used in the algorithm's recursive self-improvement processes. Providing users with opportunities to offer feedback and make refinements further enhances their engagement and trust. This approach supports data sovereignty while ensuring that the AI aligns with user expectations and ethical standards. Therefore, we propose:

P5: Designers can ensure control in AI agent algorithms through design decisions that emphasize **(a)** feedback integration and **(b)** privacy assurance.

In AI architecture, control mechanisms are enhanced through the implementation of modular structures and continuous oversight. Modular structures compartmentalize functionalities, enabling targeted interventions and precise adjustments to specific modules without disrupting the entire system. Continuous oversight ensures ongoing evaluation and rapid response to deviations or faults. These practices not only enhance control and reliability but also eliminate single points of failure, improving the robustness and dependability of the AI agent. Therefore, we propose:

P6: Designers can ensure control in AI agent architecture through design decisions that emphasize **(a)** modularity and **(b)** continuous oversight.

Ensuring control across design domains helps mitigate risks such as privacy violations, unintentional misuse, and performance lapses. Neglecting to implement adequate control mechanisms can result in suboptimal functionality that fails to effectively meet user needs. Empirical research underscores the importance of prioritizing user control, showing significantly higher engagement with AI agents that emphasize privacy preservation and user oversight (Lutz & Tamò-Larrieux, 2021). Thus, user control is not merely an ancillary feature but a fundamental principle for designing AI agents that empower users by honoring their autonomy within ethical boundaries.

As we promote these design decisions, we recognize the complex trade-offs involved in balancing control in AI design. While each decision offers distinct advantages, their implementation may introduce constraints or conflicts that demand thoughtful navigation. For instance, modularity in architecture allows for independent updates to system components, strengthening operational safeguards against misuse. Nevertheless, this modularity can complicate the integration of features like real-time monitoring, which rely on continuous oversight. Similarly, feedback integration enhances customizability by refining AI outputs based on user preferences but may inadvertently undermine operational safeguards if it introduces biases or unintended consequences. These examples

underscore the need for a nuanced, iterative approach that harmonizes the synergies and tensions among design decisions, ensuring that AI agents are both user-empowered and ethically governed.

7 Findings—Transparency in AI Design

Transparency in AI design entails ensuring that an AI exhibits its operations, logic, and data stewardship in a clear, open, and auditable manner, enabling stakeholders to easily understand its behavior and outcomes. The transparency mechanism aligns with the clarity of decisions, our third meta-mechanism, which emphasizes continuous reflective consciousness—an ongoing awareness of the decisions made and their implications. Transparency supports truthfulness and accountability by encouraging reflection, fostering ethical awareness, and promoting a shared understanding of ethical conduct among stakeholders. Achieving these objectives requires openness about the AI's capabilities and limitations, as well as independent evaluations to validate its performance and ethical alignment.

7.1 Designing Affordances for Transparency

To achieve transparency in AI affordances, designers must allow users to learn about the AI's limitations, intended purpose, and capabilities. Transparency fosters informed user interactions, builds trust, and ensures ethical use. Our findings highlight three key design decisions to support transparency: (1) allowing users to view the AI's operational effectiveness and limitations, (2) informing users about its intended purpose, and (3) enabling users to understand and utilize its capabilities.

First, participants highlighted the importance of transparently communicating an AI's accuracy, operational effectiveness, and limitations to manage user expectations. Generative AI, for instance, performs well with certain prompts but struggles with others, underscoring the need to clarify its scope and boundaries to help users make informed, responsible decisions during their interactions with it. Participants also expressed concerns about users developing inflated perceptions of an AI's accuracy, especially when the system consistently performs well. One participant cautioned that users might assume the AI is infallible and neglect to monitor its outputs over time, emphasizing the importance of clear and effective communication to manage user expectations:

> *When the systems are so good, humans think they can take their hands off the wheel … [Customers think], They told me in the pitch that I need to look at this stuff, but I've run a thousand of these things and I've never once had to change anything so I think it's perfect—I'm going to let this thing run and I'm going to worry about other stuff. Then, you're going to have things fall through. It's human nature.* (AI designer, AI startup)

Second, enhancing transparency by clearly disclosing the AI's intended purpose ensures that users understand what the system is designed to achieve, helping align their use with the AI's original design intent. Purpose disclosure differs from communicating the AI's capabilities or limitations; it focuses on clarifying the role the AI is meant to play within a broader context. For example, one designer described an AI tool developed to streamline organizational search processes, enabling employees to quickly locate relevant terminology or insights for projects. He emphasized that users must recognize the tool's primary purpose as improving search efficiency, not delivering perfectly accurate results or replacing human judgment. He explained:

> *We will never get to 100% accuracy with our tool …, we've already resigned ourselves, and, frankly, part of the pitch that we're putting together is that 100% will be impossible. There's fallibility on the other end, and it's important to explain to users that, look, this [AI] is not the be all, end all—you still need to look at this stuff.* (AI designer, AI startup)

Third, advocating for an AI's capabilities involves clearly communicating its functions and ensuring users understand how to utilize its features. Participants stressed that users must comprehend the functional capabilities of an AI to use it effectively. Clear communication enables users to maximize the system's potential and facilitates successful adoption. One designer who is developing AI for legal support highlighted the importance of educating users unfamiliar with generative AI:

> *We have challenges communicating what our product does and how to use it to people who haven't used generative AI before. We need to educate these people on what generative AI is and how to use our product during the sales pitch and then again if they choose to buy it … Otherwise, they might not be interested in it, or they might not know how to use everything in it.* (AI designer, technology firm)

Participants' concerns about the transparency of an AI's affordances underscore the need for designers to clearly convey the system's operational effectiveness, purpose, and functionalities. While the method of communication remains debated, it is essential to present these aspects effectively without misleading users. After all, courageously sharing the truth is the first step toward reinforcing design accountability.

7.2 Designing Algorithms for Transparency

Our participants emphasized that the foundation of transparency in AI algorithms lies in ensuring that the AI's core operations are transparent, interpretable, and amenable to audit procedures. While transparency in affordances focuses on "what" the AI is doing (or not), transparency in algorithms addresses "how" and "why" the AI behaves as it does. This requires not only algorithmic explainability but also effective internal communication among development team members regarding the AI's progress and performance metrics. These observations are synthesized under two key design decisions: (1) communicating performance and development outcomes and (2) improving algorithm interpretability.

First, our interviews unveiled a surprising insight: Many ethical issues in AI development stem from a lack of coherent internal communication between designers and developers regarding the AI's intended purpose. While the literature on responsible AI often characterizes transparency as a one-way flow of information from developers to users, with a strong emphasis on XAI, our findings suggest a more nuanced perspective. Transparency takes on added layers of complexity when scrutinized across the various phases of AI design and implementation. Participants underscored the importance of clearly articulating agent performance goals and technical specifications to the development team—a step often overlooked but indispensable. This clarity, they argue, is often missing but is utterly indispensable. Developers, even those working on specific AI components, must have a comprehensive understanding of the entire system, a broader awareness that equips them to make informed decisions when navigating challenges during the development process. One participant noted that during a recent project, open communication streamlined execution by aligning all team members toward shared objectives.

Participants also highlighted the importance of institutionalizing organizational principles that not only establish ethical standards but also foster a culture of responsibility. One designer referred to these principles as ethical "guardrails," a framework that guides employees toward responsible conduct and deters ethically questionable decisions. He emphasized the role of education and training in creating these guardrails, stating that "awareness goes a long way [in implementing responsible AI principles]. Awareness, learning, and education help improve this process. If you train them well, then they will do the checks and balances to ensure ethics are upheld" (AI designer, government).

As with any new product build, challenges and trade-offs are inevitable. While designers must be forthright with developers about the AI's overarching goals, developers likewise need to communicate transparently with designers when encountering challenges during development. These challenges may include developmental delays, failure to meet accuracy benchmarks, issues adhering to budgetary limits, and the results of internal audits. One designer highlighted the importance of this two-way communication for maintaining quality standards, saying: "If you see your project's gonna take longer than planned, it's super important to give a heads-up to your PM and the dev[elopment] team. Keeping everyone in the loop really helps in maintaining the quality of what we're building" (designer, R&D institutions).

Second, achieving algorithmic explainability requires integrating features that clarify the decision-making process. This can be accomplished through specialized explainability algorithms or by developing custom, in-house solutions instead of relying on proprietary black-box systems. According to designers, providing users with clear explanations of the AI's decision-making mechanics fosters trust and enhances the agent's perceived reliability. One designer, who emphasized a commitment to in-house code development and meticulous change-tracking, explained: "We owe it to the [customers] that they can actually ask the question of 'how did you come up with that' and I can answer 'this is exactly how.'" (AI designer, technology firm).

Several participants acknowledged the challenges of achieving explainable AI, particularly in generative AI, due to the complexity of deep learning models often referred to as "black boxes." For some generative AI models, the sheer number of parameters and nonlinear transformations makes it difficult to provide a clear and understandable explanation for every outcome. One designer recommended perturbation analysis—a method of studying the effects of minor changes on a system's overall behavior or outcomes—as a potential approach to uncovering the decision-making processes in black box models:

> *The issue is that the best models are neural networks and deep learning, and those are inherently not interpretable, so in that case there's a series of black box methods that are kind of tuned to different things that try to, regardless of the model you're using, try to basically change the input a little to see how that affects the output, and by doing that you're able to see why did the model make the predictions that it did.* (AI designer, AI startup)

These insights highlight that achieving transparency in AI algorithms goes beyond external communication with users; it necessitates robust internal processes, clear communication among development teams, and a steadfast commitment to ethical standards. By aligning internal goals, fostering a culture of responsibility, and prioritizing algorithmic interpretability, designers can ensure that AI agents operate with accountability and clarity.

7.3 Designing Architecture for Transparency

Our participants emphasized that transparency extends beyond affordances and algorithms, delving into the core of an AI's architecture. From our discussions with designers, we identified three pivotal categories of design decisions: (1) disclosing user data storage, privacy, and security, (2) incorporating logging and documenting tools, and (3) improving auditability through modular architecture.

First, analogous to empowering users to manage their own data, disclosing how user data is stored, secured, utilized, and monetized is critical for maintaining user trust. One participant highlighted the importance of transparent practices in data storage and processing, stating the following:

> *Our most relevant issue is regarding data. Data is confidential, privileged, and includes trade secrets with unpublished work. Data is highly sensitive … We take the most data lean process possible [by] processing it temporarily over the cloud and never storing it in-house. We have a responsibility to inform our users where we keep their data because they're gonna want to know it's safe and that [their work] won't get in the hands of their competitors.* (AI designer, technology firm)

A majority of our study participants agreed that while formal user consent may technically grant designers the latitude to use data for various purposes, such consent becomes meaningless if users lack clarity about how their personal information is actually deployed to shape the AI's functionality and outputs. This lack of understanding can engender a sense of privacy invasion, undermining trust and creating potential ethical quandaries. Transparency, therefore, must go beyond AI explainability and internal communication to include comprehensive disclosure about the stewardship and the use of user data.

Second, logging and documentation tools enhance AI transparency by systematically recording the rationale behind design components and tracking changes made to the AI. These tools create meticulous records of the AI's operational activities and architectural changes. Configuration and deployment logs provide a chronological repository of architectural alterations, while version control and change logs document updates to the agent's codebase. One participant remarked: "Documentation is pivotal, not merely for record-keeping, but also for elucidating the rationale behind each design element" (AI designer, AI startup). Another emphasized that every design decision and component should include annotations outlining and justifying their intended purpose. This approach bridges the gap between designers' intentions and the AI's actual behavior. Participants also stressed that meticulous documentation should extend beyond the development phase; maintaining operational logs post-deployment is crucial for ongoing assurance and accountability. By integrating logging and monitoring tools, an AI's behavior and operations become visible and auditable, enabling stakeholders to track and evaluate the AI's adherence to operational standards and ethical guidelines.

Third, a modular, auditable architecture enhances transparency through design decisions that render the AI comprehensible and accountable. An open architecture allows stakeholders to investigate the AI's structural components, operational processes, and decision-making protocols. Here, "open" does not imply public accessibility, but availability for potential audit and review by relevant parties. One participant noted, "Just like open-source code, we can also consider open architecture ... so that anyone who wants to look at how the system is designed or operates can do it" (AI designer, AI startup). Participants also highlighted that modular architecture and modular data architecture facilitate granular inspection of individual components and data sources. They further suggested that interoperability layers could support the integration of third-party explainability tools to elucidate system decisions.

Table C6 provides our data structure for transparency across the design domains of affordances, algorithms, and architecture. The interdependence between various transparency-related design decisions proved particularly enlightening. For instance, system interpretability (algorithms) and ongoing system monitoring (architecture) support capabilities disclosure (affordance) by making the AI's operational logic accessible and comprehensible to users while providing insights into its operational status and changes in functionality. Similarly, system cognizance (algorithms) and ongoing system monitoring (architecture) enhance the affordance of purpose disclosure by equipping the development team to clearly communicate the AI's purpose and link architectural features to the AI's intended purpose. Additionally, system interpretability (algorithms) may support limitations disclosure (affordance) by clearly demonstrating the AI's functional boundaries to users.

7.4 Discussion on Transparency

We posit that transparency is not a peripheral concern but a fundamental ethical cornerstone. Our position emanates from an epistemic understanding of transparency that goes beyond simplistic access to information, such as algorithms or source codes, and emphasizes meaningful interpretation and genuine comprehension. This shift addresses key cognitive limitations inherent in human reasoning and paves the way for ethically driven communication. From a moral standpoint, we contend that transparency should not be viewed merely as a utilitarian tool for building trust or efficiency but as an ethical obligation in its own right. However, transparency's purported role as a catalyst for accountability is not without its challenges. Superficial or

misleading implementations of transparency can undermine its ethical foundations, functioning as a deceptive smokescreen rather than a genuine ethical practice. In light of this complex ethical landscape, we advocate for a negotiated approach to transparency that balances ethical imperatives with practical limitations. Achieving this balance necessitates a thoughtful and collaborative dialogue among stakeholders to achieve a symbiosis between ethical obligations and practical exigencies, ultimately serving both individual and collective interests.

From this perspective, we define the transparency mechanism as the process of ensuring the AI exhibits its operations, logic, and data stewardship in a clear, open, and auditable manner so that stakeholders can easily understand its behavior and outcomes. Transparency requires AI to be structured and documented in ways that reveal its internal logic, data usage, and functionality. An AI demonstrates transparency when its architectural components, algorithms, and user interfaces are designed to make its workings visible and open to inspection, audit, and explanation without compromising security or intellectual property rights.

In designing functional affordances, designers have an ethical imperative to transparently communicate the AI's intended purposes, capabilities, and operational limitations without obfuscation or reservations. For instance, if an AI's primary objective differs ostensibly from its advertised purpose, this should be explicitly conveyed. Such communication should leverage enabling affordances that allow users to view, explore, and learn about the AI's functions, rather than burying critical information in inaccessible terms and conditions or technical documentation. Transparency in AI design, therefore, is not limited to static documentation but involves creating interactive and accessible mechanisms that empower users to understand and engage with the AI meaningfully. Therefore, we propose:

P7: Designers can ensure transparency in AI agent affordances through design decisions that emphasize **(a)** limitations disclosure, **(b)** purpose disclosure, and **(c)** capabilities disclosure.

Within the algorithmic context, the emphasis shifts to developing interpretable models that clarify the underlying decision-making logic in ways users can readily comprehend. This goes beyond mere explainability; it involves articulating the underlying logic in a manner that ensures genuine user understanding. Transparency challenges often arise from a disconnect between designers and developers. Addressing these challenges requires fostering clear and explicit communication within the development team regarding performance metrics, trade-offs, and technical challenges. This approach cultivates system cognizance, fostering a shared understanding of the AI's capabilities and limitations among all stakeholders. Bridging these communication gaps ensures that the AI is more comprehensible and trustworthy, ultimately enhancing user engagement and confidence in the technology. Therefore, we propose:

P8: Designers can ensure transparency in AI agent algorithms through design decisions that emphasize **(a)** system cognizance and **(b)** system interpretability.

In architectural design, transparency is achieved through logging and documentation tools that track and monitor the AI's operations, decisions, and modifications. These tools create an auditable architecture, providing stakeholders with accessible and clear information to verify the AI's integrity and ethical compliance. Additionally, designing an AI to clearly communicate its security and privacy measures strengthens transparency, reinforcing both accountability and trustworthiness. Therefore, we propose:

P9: Designers can ensure transparency in AI agent architecture through design decisions that emphasize **(a)** security and privacy transparency, **(b)** ongoing system monitoring, and **(c)** auditability.

The overarching goal of transparency in design is to elucidate both the operational mechanics and the reasoning behind an AI's outputs. This dual focus elevates stakeholder trust by ensuring the agent is inspectable and accountable. Neglecting these facets of transparency risks obfuscating the AI's decision-making processes, leading to diminished accountability and heightened public skepticism. A lack of transparency also undermines stakeholders' ability to scrutinize, validate, and challenge the agent's outputs and actions, creating an environment rife with ethical and operational risks. Nevertheless, the inherent trade-offs in design decisions for enhancing transparency in AI agents present significant challenges due to conflicting priorities. For instance, increasing system interpretability—making an AI model's decision-making processes more understandable—can reduce predictive accuracy, as simpler, more interpretable models often fail to capture complex patterns as effectively as more opaque, sophisticated models. Efforts to improve transparency and explainability may also conflict with protecting proprietary information or user privacy, as revealing too much about the system's inner workings may expose sensitive data or intellectual property. Additionally, excessive transparency can compromise security; divulging too much information about the agent's inner workings may expose vulnerabilities, leaving the system susceptible to malicious exploitation or manipulation. Implementing robust transparency measures can also introduce increased complexity and resource demands, potentially affecting the agent's efficiency and scalability. These examples highlight the necessity of a balanced approach in AI design, one that carefully weighs the benefits of transparency against its potential drawbacks to ensure ethical and effective agent performance.

8 Theory Development and Theoretical Integration

Our study explores and theorizes fundamental responsibility mechanisms that ensure responsible design decisions are effectively translated into responsible AI agents. Rather than merely defining what a responsible AI agent is, we aimed to uncover novel insights and elucidate priorities in the realm of AI design, emphasizing *how* to achieve a responsible AI agent. Our findings are systematically integrated into an explanatory theory that elucidates how responsible AI can be achieved through three fundamental responsibility mechanisms across three design domains, each domain describing distinct yet interrelated responsible AI design decisions. This integration culminates in the introduction of a novel midrange theory—the *authenticity, control, transparency (ACT) theory of responsible AI design*. Consistent with Leidner and Gregory's (2024) guidelines, our theory is both insightful and parsimonious, offering a robust framework that guides future research and practical applications in responsible AI design. This section discusses the theory's constituents and relevance, theoretical significance, and pragmatic value in connection with existing theories.

8.1 The ACT Theory of Responsible AI Design

We posit that authenticity, control, and transparency are indispensable, high-priority mechanisms for fostering responsible AI, as they inherently embody the principles of beneficence and non-maleficence. This perspective aligns with the Sartrean view of ethics and the responsible design practices it informs (d'Anjou, 2010). Authenticity ensures that an AI's actions are congruent with ethical principles and its intended purposes, promoting consistent ethical conduct. Control ensures the AI facilitates human oversight and guidance, mitigating misuse and preventing unintended adverse consequences. Transparency ensures the AI's operations are intelligible and auditable, fostering trust and accountability. Together, these mechanisms provide a robust framework for ethical AI design, ensuring that AI agents behave responsibly, adhere to fundamental human values, and maximize positive outcomes while minimizing harm (Figure 1). Accordingly, we propose that:

P10: Design decisions that ensure **(a)** authenticity, **(b)** control, and **(c)** transparency in an AI agent's affordances, algorithms, and architecture shape an AI that behaves responsibly, maximizing beneficence and minimizing maleficence.

Our theory offers a nuanced perspective on responsible AI design. The ACT theory illustrates the relationships between design decisions across three domains—affordances, algorithms, and architecture—and the three responsibility mechanisms. Architecture serves as the high-level structure, defining how various components, including algorithms, interact and integrate. Algorithms, embedded within the architecture, drive the AI's functionality through rules and computations applied to incoming and existing data. Affordances emerge from the interplay between architecture and algorithms, representing the tangible action possibilities available to users. Our findings reveal that the ACT mechanisms manifest differently across these design domains yet collectively and synergistically contribute to the overarching goal of creating a responsible AI agent. Table 4 outlines design decisions for an AI's architecture, algorithms, and affordances that support the three responsibility mechanisms of authenticity, control, and transparency. Building on these design decisions, Figure 1 illustrates the ACT theory of responsible AI design, which offers a structured and integrated approach to AI design.

Table 4. ACT Design Decisions

Responsible AI mechanism	Design decisions—affordances	Design decisions—algorithms	Design decisions—architecture
Authenticity	**Intended utility:** Allow users to actualize the AI's usability and usefulness. **Interactional fidelity:** Allow users to engage with the AI in a way that ensures high fidelity in their interactions.	**Data quality:** Source and train models with quality data. **Operational accuracy:** Evaluate, monitor, and improve model accuracy and functional integrity.	**Specialization:** Create specialized architectures. **Architectural resilience:** Build adaptable architectures.
Control	**Regenerability:** Allow users to regenerate and save the AI's output. **Customizability:** Give users customization options over the AI's functionality, data usage, and interactions. **Operational safeguards:** Allow users to self-regulate their use/usage to limit the misuse, abuse, or overuse of the AI.	**Feedback integration:** Integrate user feedback to enhance the AI's output. **Privacy assurance:** Secure and respect user data ownership rights.	**Modularity:** Embrace modular architectural design. **Continuous oversight:** Implement continuous oversight mechanisms.

Transparency	**Limitations disclosure:** Allow users to view the AI's operational effectiveness and limitations. **Purpose disclosure:** Allow users to learn about the intended purpose of the AI. **Capabilities disclosure:** Allow users to explore the AI's capabilities.	**System cognizance:** Ensure the objectives, performance, and limitations of the AI algorithms are understood during the development process. **System interpretability:** Improve algorithm interpretability.	**Security and privacy transparency:** Disclose user data storage, privacy, and security. **Ongoing system monitoring:** Incorporate logging and monitoring tools. **Auditability:** Create a modular architecture to improve auditability.

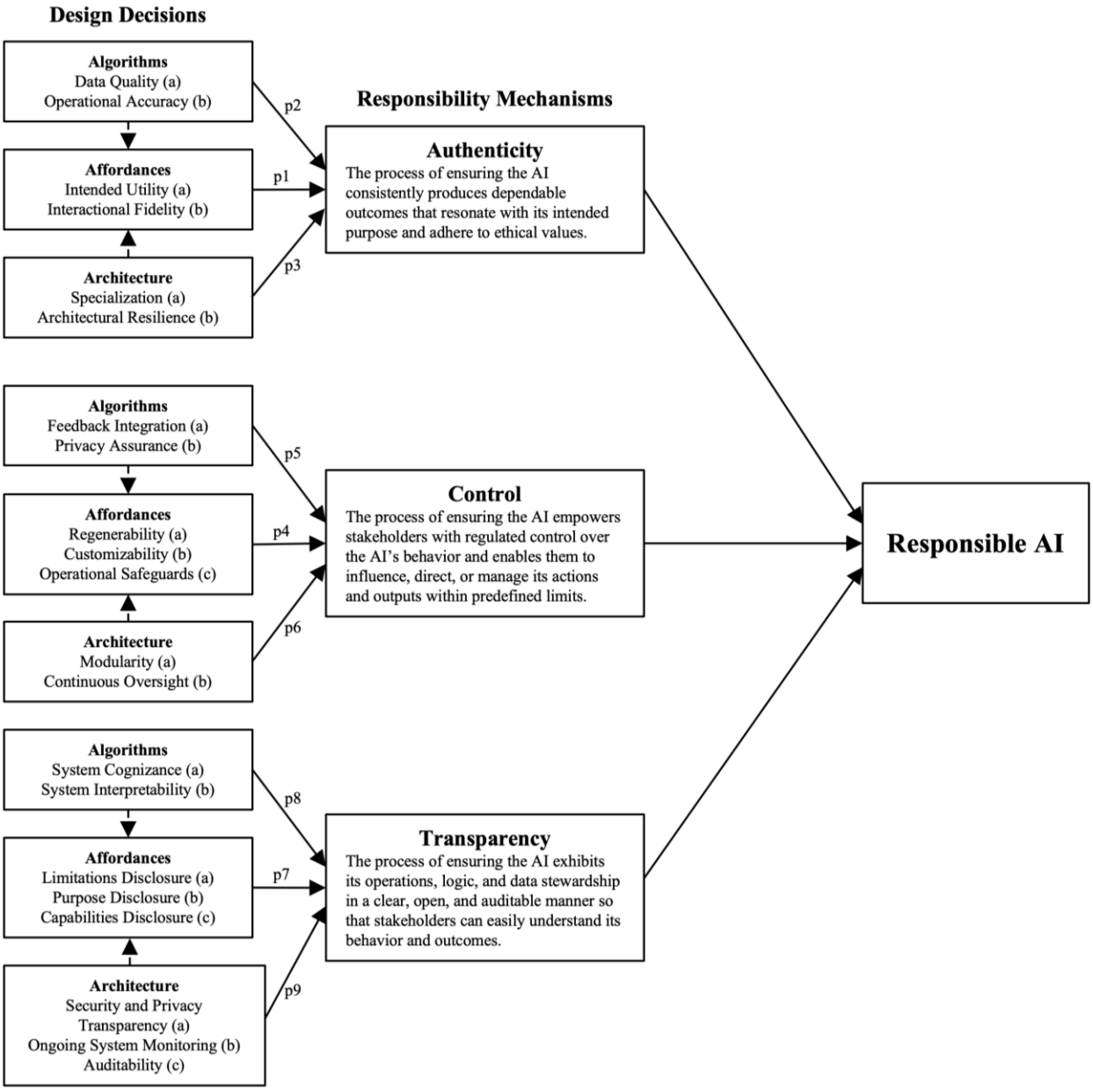

Figure 1. ACT Theory of Responsible AI Design

The ACT theory distinguishes itself from existing frameworks of responsible innovation—primarily responsible research and innovation (RRI), the CARE theory of dignity (CARE), corporate digital responsibility (CDR), and value sensitive design (VSD)—by emphasizing context specificity and pathways to responsibility. Although these frameworks (briefly reviewed in Appendix D) are frequently used to conceptualize responsible AI, they often fall short in their applicability due to their technology-agnostic nature, insufficient engagement with the intricate dynamics of AI design, and, more importantly, the lack of integration between tangible design decisions and responsibility mechanisms. While valuable for broad ethical guidance, this generality limits these theories' practicality in addressing the nuanced requirements of AI design. In contrast, the ACT theory offers a fresh perspective by embedding responsibility directly into the AI design process to ensure a deeper alignment with the unique demands of AI. Drawing from Weber's (2012) delineation and best practices in theorization (Chatterjee et al., 2024), Table 5 illustrates these distinctions across the four domains of a theory: constructs, associations, states, and event spaces.

While ACT introduces a novel approach to responsible AI, it is not positioned as a replacement for existing theories but as a complement that addresses their blind spots and contextualizes their applications, aligning with the expectations of a middle-range theory (Leidner & Gregory, 2024). As highlighted in Table 6, ACT contextualizes, enhances, and partly integrates these frameworks, enabling them to operationalize ethical principles more effectively in the AI design domain. By addressing their specific gaps—such as ambiguity in conceptual boundaries (e.g., RRI), lack of clarity in outcomes (e.g., CDR), ambiguity in design decisions (e.g., VSD), or deliberate scope limitations (e.g., CARE)—ACT transforms theoretical ideals into concrete, actionable design decisions and offers falsifiable relationships between design decisions and design outcomes mediated by responsibility mechanisms. This differentiation underscores ACT's utility not only as an independent framework but also as a synergistic tool that amplifies the practical relevance of established theories in an era of rapid AI advancement.

8.2 Theoretical Contributions

This paper advances theoretical discourse in three key areas: First, it presents a more nuanced and comprehensive theoretical framework, delineating the conceptual boundaries of responsible AI with a focused emphasis on the principles underpinning responsible AI design. Second, it identifies three overarching assurance mechanisms—authenticity, control, and transparency—that align design processes with their intended outcomes. And third, it conceptualizes three foundational categories of design decisions that underpin responsible AI design. Collectively, these contributions challenge and rectify prevailing misconceptions in the literature: the reductionist view that responsible AI is solely about harm mitigation, the assumption that responsibility is an intrinsic system trait rather than an observable behavior in action, and the notion that design decisions are limited to algorithmic development. Moving beyond these oversimplifications, the ACT theory offers an integrated, contextual, and actionable framework, embedding responsibility as a foundational principle of AI design. This section elaborates on these contributions, highlighting their originality, depth, and practical relevance.

Table 5. Comparative Novelty of ACT Theory Across Existing Theories or Frameworks

Theory aspects	RRI	CARE	CDR
Major constructs	**Anticipation:** Identifying potential impacts and risks. **Reflexivity**: Scrutinizing the purpose and implications of a new technology. **Inclusion**: Engaging diverse stakeholders. **Responsiveness:** Adapting practices based on emerging knowledge and societal needs.	**Claims:** Expectations for dignity (e.g., autonomy, equality). **Affronts:** Violations of dignity (e.g., coercion, inequality). **Response:** Actions to address affronts. **Equilibrium:** Balance between claims and affronts.	**Shared values and norms:** Ethical principles guiding digital responsibility. **Specific norms:** Prescribe notions of right and wrong to various organizational activities. **Artifacts and behaviors:** Ensure responsible design is considered at each phase of the process model.
Key associations	**Anticipation and reflection:** Anticipating outcomes to inform ethical decisions. **Inclusion and responsiveness:** Engaging stakeholders to address societal concerns. **Reflection and responsiveness:** Refining practices to meet evolving ethical standards. **Societal values alignment:** Maintaining ethical acceptability, societal desirability, transparency, and sustainability.	**Claims and affronts:** Affronts arise when claims are unmet. **Response:** Restores dignity by addressing affronts. **Equilibrium:** Maintained when dignity claims are respected.	**Social and organizational culture:** CDR norms are developed from public opinion, legal requirements, technological progress, industry factors, customer factors, and firm factors. **Layers:** CDR norms are employed across three layers: shared values, specific values, and artefacts and behaviors.

Example states	**Anticipatory state:** Preparedness through risk and benefit foresight. **Inclusive state:** Engagement of diverse stakeholders. **Reflexive state:** Ethical self-awareness in practices. **Responsive State**: Adaptation to societal needs and values.	**Autonomy:** Freedom from coercion. **Visibility:** Having a voice. **Equality:** Equal treatment. **Respect:** Recognition of human dignity. **Balance:** Alignment of claims and affronts.	**Ethical responsibilities:** Guided by norms/regulations. **Transparency:** Clear and open practices. **Accountability:** Responsibility for digital impacts. **Equity:** Fair access to digital resources. **Sustainability:** Balance between technology use and ecological impact.
Event spaces	**Temporal**: Considering short-and long-term impacts. **Contextual**: Addressing cultural, political, and economic factors. **Operational**: Including societal feedback or emerging challenges.	**Microlevel:** Personal interactions with data **Macrolevel:** Organizational or societal practices **Dynamic adjustments:** Iterative corrections to uphold dignity.	**Development phase:** Ethical norms applied to the development of new technologies. **Four processes:** Creation of technology and data capture, operation and decision making, inspection and impact assessment, and refinement of technology and data.
Theory aspects	**VSD**	**ACT**	**ACT significance**
Major constructs	**Human Values:** Fundamental principles such as privacy, autonomy, and fairness that guide design. **Stakeholders:** Individuals or groups affected by the technology, including direct and indirect stakeholders.	**Authenticity:** Ensures AI outcomes align with ethical values and intended purposes. **Control:** Empowers stakeholders to regulate and manage AI behavior within predefined limits. **Transparency:** Makes AI operations, logic, and outcomes clear, auditable, and comprehensible.	ACT's constructs are specifically defined for AI, providing actionable and AI-relevant guidance compared to the abstract, general principles of other theories.
Key associations	**Value Integration:** Aligning design processes with identified human values through conceptual, empirical, and technology investigations. **Stakeholder Relationships:** Understanding how stakeholders' values interact and influence design outcomes. **Value-Driven outcomes:** Ensuring that design outcomes reflect the prioritized human values.	**Design decisions and responsibility mechanisms:** Decision decisions determine how responsibility mechanisms are embedded in AI. **Responsibility mechanisms and responsible AI behavior:** Responsibility mechanisms ensure the AI agent behaves responsibly, maximizing beneficence and minimizing maleficence.	ACT offers clear, directly testable associations by linking design decisions to responsibility mechanisms and ensuring these mechanisms drive responsible AI behavior, offering operational clarity.
Example states	**Value prioritization:** Determining the importance of various values in a given context. **Stakeholder influence:** Assessing the impact of stakeholders on the design process. **Ethical alignment:** Achieving harmony between technical features and human values.	**Design decisions:** Decisions shaping affordances, architecture, and algorithms **Responsibility mechanisms variation:** Variation in authenticity, control, and transparency to suit diverse contexts. **Responsible behavior:** Minimizing harms and maximizing beneficence.	ACT defines states across affordances, architecture, and algorithms, rendering tangible decisions and their potential outcomes, uniquely addressing AI's technical and ethical complexity.
Event spaces	**Design phases:** Ethical decisions during ideation, prototyping, deployment and improvement. **Value conflicts:** Resolving competing values. **Impact assessment:** Evaluating social, ethical, and practical outcomes.	**Design processes:** Steps embedding responsibility mechanisms into AI agents. **Assurance mechanisms:** Validate alignment with ethical principles. **Responsible behavior:** Ensuring AI actions minimize harm and maximize beneficence.	ACT establishes practical boundaries by focusing on design processes and assurance mechanisms, ensuring actionable guidance.

Table 6. Comparative Advantages of ACT Theory Across Existing Theories or Frameworks

	ACT vs. RRI	ACT vs. CARE	ACT vs. CDR	ACT vs. VSD	Overall
How ACT contextualizes existing theory	• Builds on *anticipation* and *responsiveness* by embedding ethical considerations into AI design. • Promotes *inclusion* and *reflection* through actionable AI-specific responsibility mechanisms. • Ensures ethical alignment during the AI lifecycle.	• Operationalizes *claims* and *response* through authenticity, control, and transparency. • Upholds human dignity by embedding mechanisms to address *affronts* and maintain *equilibrium* in AI agents. • Contextualizes dignity principles for complex AI behavior.	• Translates *shared values and norms* into technical AI design processes. • Embeds *digital ethics*, *data privacy*, and *sustainability* directly into AI functionality. • Focuses on stakeholder impact at the AI agent level rather than organizational policy.	• Bridges *human values* and design by specifying how they can be implemented in AI affordances, architecture, and algorithms. • Provides actionable processes to align *stakeholder interests* with technical design. • Ensures ethical principles are embedded into AI agents rather than left abstract.	• Embeds ethical considerations into AI design, adapting existing principles for AI-specific contexts. • Focuses on actionable responsibility mechanisms tailored for AI context. • Upholds values by addressing ethical challenges through AI-specific mechanisms. • Translates abstract ethical principles into actionable technical processes and decisions. • Integrates human values directly into AI affordances, architecture, and algorithms to ensure alignment with societal expectations.
How ACT enhances existing theory	• Shifts responsibility from external oversight to intrinsic design mechanisms, embedding responsibility into AI behavior. • Moves beyond governance to proactive responsibility embedding during design, maximizing beneficence while minimizing harm. • Provides clearer, actionable mechanisms for aligning AI design with ethical goals.	• Expands CARE's focus on dignity to address technical domains like affordances, algorithms, and architecture. • Advances from abstract ethical principles to actionable mechanisms for operationalizing dignity in AI agents. • Ensures ethical AI behavior by addressing both design and system-level responsibilities.	• Embeds responsibility into AI design, focusing on technical aspects like architecture and algorithms. • Provides actionable, system-level guidance compared to CDR's organizational-level focus. • Offers practical mechanisms for fostering responsible AI behavior through design.	• Operationalizes authenticity, control, and transparency to align design decisions with ethical outcomes. • Provides actionable mechanisms to initiate and sustain responsibility, improving upon VSD's abstract value-driven approach. • Delivers precise and pragmatic tools for embedding ethics into AI agents.	• Shifts responsibility from external oversight to intrinsic mechanisms embedded within AI agents. • Provides actionable frameworks to align technical design decisions with ethical outcomes. • Expands ethical principles to encompass technical domains—AI affordances, algorithms, and architecture. • Addresses system-level responsibilities to ensure ethical behavior beyond design decisions. • Offers precise, pragmatic tools to embed, sustain, validate, and ensure responsibility throughout AI operations.

First, we broaden the concept of responsibility in AI design to underscore the dual imperatives of minimizing harm and maximizing beneficence. Prevailing literature often confines responsible AI to risk-mitigation or harm-reduction paradigms, focusing primarily on concerns such as bias, privacy breaches, and security vulnerabilities. This narrow framing creates a limited perspective by overlooking AI's broader potential to align with utilitarian principles and actively foster positive societal outcomes. In response, we propose a dual approach that prioritizes both harm mitigation and the deliberate promotion of benefits, situating the ACT theory within the broader domain of responsible innovation frameworks such as RRI, CDR, and VSD. The ACT theory, however, distinguishes itself from these general-purpose theories by offering a more context-specific understanding of responsibility in AI design.

Our empirical findings reveal that designers place significant emphasis on advancing benefits in addition to mitigating harm, challenging the prevailing assumption that responsible design primarily prioritizes harm reduction. For instance, rather than broadly conceptualizing responsibility as the promotion of social well-being—as articulated in RRI (Stilgoe et al., 2013), CDR (Lobschat et al., 2021), and VSD (Friedman et al., 2009)—the ACT theory focuses on the immediate, contextual benefits that responsible AI agents can deliver. Moreover, while our theory shares VSD's commitment to integrating human values into the design process, it advances this framework by ensuring that positive outcomes are not merely anticipated but intentionally designed and structurally embedded within AI agents. This approach shifts the discourse beyond a reductive focus on harm elimination, embracing a nuanced and holistic framework that acknowledges the inherent complexities, trade-offs, and ethical tensions central to AI design. In doing so, the ACT theory reframes the responsibility narrative, positioning responsible AI agents as proactive instruments for ethical and societal progress.

Second, we introduce a parsimonious yet comprehensive framework to ensure that responsible AI agents uphold ethical standards throughout their operation by proposing three foundational assurance mechanisms: authenticity, control, and transparency. These mechanisms collectively provide a cohesive and adaptive approach to embedding responsibility within AI agents, balancing conceptual simplicity with the flexibility needed to address the diverse and evolving contexts in which these agents operate. This framework challenges prevailing perspectives in the literature, often reducing responsibility to a static checklist of protective attributes and overlooking the mechanisms that actively drive responsible behavior. By framing authenticity, control, and transparency as dynamic assurance mechanisms rather than static attributes, the ACT theory fundamentally reorients responsible AI toward understanding how deliberate design decisions enable AI agents to exhibit responsible behavior.

Specifically, drawing on principles of genuineness, sincerity, and high-quality outputs (Napoli et al., 2014; Trilling, 1972), we introduce authenticity as a foundational mechanism within the responsible AI literature, critical for fostering positive evaluations by both users and society. Authenticity ensures that AI agents maintain operational fidelity by consistently producing outcomes aligned with their intended purposes. Additionally, we reconceptualize control as the cornerstone of agency calibration, ensuring that AI agents operate harmoniously with human intentions and ethical standards while remaining aligned with broader societal norms. Furthermore, we redefine transparency within this framework to transcend traditional notions, placing emphasis on epistemic integrity and accountability and thereby supporting trust, explainability, and oversight throughout the agent's lifecycle.

By introducing these mechanisms, the ACT theory advances the theoretical understanding of responsible AI as inherently adaptable to diverse AI contexts and interactions. This approach complements responsible innovation frameworks such as RRI and VSD, which emphasize auxiliary practices such as stakeholder engagement, organizational accountability, and value alignment but often lack specificity in how design decisions can effectively translate into responsible AI agents. In contrast, the ACT theory explicitly identifies mechanisms that ensure that AI design decisions collectively foster responsible behavior, thereby transforming responsibility from an externally imposed characteristic into internally assured behavior—maximizing beneficence and minimizing maleficence.

Third, the ACT theory redefines the paradigm of embedding responsibility within AI agents, emphasizing that responsibility is not merely an intrinsic behavior of the technology but one that can be systematically cultivated through deliberate design decisions. We introduce a comprehensive framework of design decisions spanning three critical domains—affordances, algorithms, and architecture—that collectively streamline the creation of responsible AI agents from their inception. This framework challenges the traditional algorithm-centric perspective of responsible AI by underscoring the indispensable roles of affordances and architecture in enabling responsible behavior.

By prioritizing these design domains (i.e., affordances, algorithms, and architecture), the ACT theory challenges and expands the theoretical discourse in responsible AI design, paving the way for new avenues of inquiry. Notably, it offers empirically grounded insights into key design decisions, providing AI practitioners with a robust decision-making foundation.

Furthermore, the ACT theory enriches theories of responsible innovation by translating abstract ideals such as integrity and societal values into actionable design principles. Unlike frameworks such as RRI and CDR, which emphasize external governance structures—such as societal oversight (Stilgoe et al., 2013) or corporate policies (Lobschat et al., 2021)—the ACT theory embeds responsibility intrinsically within the design of the AI agent. Additionally, our theory distinguishes itself from models like CARE and VSD, which take a technology-agnostic stance by vaguely integrating human dignity (Leidner & Tona, 2021) or values (Friedman et al., 2009). Instead, the ACT theory delivers specificity and pragmatism by translating these abstract ideals into concrete, actionable design interventions. In doing so, the ACT theory positions itself as both a foundational framework for advancing responsible AI theory and a practical guide for real-world AI design practices, effectively bridging the gap between conceptual ideals and operational execution.

Through these three major contributions, the ACT theory enriches the responsible AI literature by introducing a design-centered, middle-range explanatory theory that challenges prevailing misconceptions while augmenting and advancing existing frameworks for responsible innovation. Both revelatory and scientifically impactful (Corley & Gioia, 2011), the ACT theory deepens the understanding of operationalizing responsible AI, offering a novel perspective that redefines and elevates the discourse on AI ethics. Meeting Weber's (2012) and Leidner and Gregory's (2024) criteria for robust IS theories, the ACT theory demonstrates importance by addressing practical and urgent needs, introducing novel constructs and associations, maintaining parsimony through appropriate abstraction, and enabling empirical testing through falsifiable propositions. These attributes establish the ACT theory as a unifying framework that seamlessly bridges academic inquiry and practical application, thereby elevating the discourse on responsible innovation.

8.3 Practical Implications

The ACT theory provides actionable solutions for embedding responsibility within AI agents. It addresses three interconnected inquiries fundamental to achieving responsible AI: What defines responsible AI? How can we ensure the creation of responsible AI agents? And, how can responsible AI be achieved in practice?

First, we redefine responsible AI by challenging existing practices and broadening the scope of responsibility. responsible AI, as articulated by the ACT theory, is not confined to minimizing harm or managing risks, nor is it limited to achieving predefined benchmarks or focusing solely on system attributes. Instead, we underscore the importance of balancing utility with risk, shifting attention from an agent's static characteristics (e.g., explainability) to its dynamic behavior and ethical alignment in real-world contexts. By shifting the focus from static design outcomes to the design process itself, we urge a continuous, deliberate effort embedded in every phase of AI development, anchored in intentional design decisions. This proactive approach ensures that responsibility is not an afterthought but an integral element of AI agents, shaping their operations, interactions, and societal impact from the outset.

Second, we explore how responsible AI agents can be effectively realized by introducing foundational responsibility mechanisms—authenticity, control, and transparency. These mechanisms form a dynamic framework for governance, accountability, and ethical alignment, streamlining the responsible AI assurance process and addressing the challenges practitioners face when navigating an overwhelming array of responsible AI principles. By focusing on these core mechanisms, we offer clarity and actionable guidance, making the implementation of responsible AI more accessible and practical. Specifically, ACT ensures effective oversight and accountability by clarifying the mechanisms through which responsibility is upheld. Additionally, we emphasize the importance of achieving ethical alignment: seamlessly integrating ethical considerations into the operational and strategic dimensions of responsible AI design. By addressing this intricate balance, our study equips practitioners with actionable strategies to harmonize ethical principles with performance imperatives, ensuring that responsible AI practices remain both practical and principled (Papagiannidis et al., 2023).

Third, our work elucidates how responsible AI can be systematically and effectively operationalized, transforming abstract principles into actionable practices. We emphasize the critical role of specificity in the design process, offering a clear framework of actionable design decisions across three pivotal domains: architecture, algorithms, and affordances. Within each domain, we provide AI designers with a practical, well-defined, and manageable repertoire of design decisions, ensuring that essential ethical considerations are integrated throughout the AI design lifecycle. This focused approach embeds responsibility into the very essence of responsible AI. It also streamlines its implementation across multidisciplinary teams, fostering cohesive and comprehensive ethical alignment at every stage of the design process. Moreover, by identifying key design decisions, we empower practitioners to craft tailored guidelines and anticipate the real-world trade-offs encountered in AI development. While resolving these trade-offs is inherently context-sensitive, we provide designers with the clarity needed to navigate these complexities, ensuring a harmonious balance between ethical imperatives and operational objectives.

8.4 Limitations and Future Research Avenues

In this study, we strategically focused on crafting a theory to enhance responsible AI within the *design* sphere, prioritizing its utility and practicality. This decision restricted our inquiry to a singular aspect of the AI lifecycle—design—thereby limiting the theory's extrapolation to broader contexts such as development, deployment, and governance. While recognizing the strength of the ACT theory within its intended scope, we acknowledge the potential insights that could emerge from its application in these adjacent domains. This limitation opens avenues for future research to extend the ACT theory across the entire AI lifecycle, exploring its broader applicability and impact.

To address the inherent limitations of our chosen qualitative methodology, we employed a systematic and informed grounded theory approach, leveraging semi-structured interviews to distill insights from within the design domain. Despite the vast diversity of AI designer experiences, we recognize that our sample, while extensive, may not capture the full breadth of the profession. To enhance the robustness of our findings, we implemented theoretical sampling and rigorous cross-validation processes. While these measures strengthened the reliability of our interpretations, we acknowledge the inherent subjectivity of qualitative analyses and invite future quantitative studies to validate our theory across a broader spectrum of contexts.

While our study adopts a designer-centered approach to responsible AI, addressing prevailing misconceptions in existing literature, the resulting ACT theory requires further empirical scrutiny across various contexts. For example, managing trade-offs in AI design is inherently context sensitive. Participants in our study highlighted trade-offs such as consent versus usability, privacy versus accuracy, and novelty versus explainability. To address these challenges, they suggested approaches involving control (e.g., providing users with options to manage trade-offs) and transparency (e.g., informing users about how trade-offs were addressed during the design process). However, we acknowledge that the issue of trade-offs merits further in-depth and contextual analysis, exceeding the scope of this study.

Although our theory highlights key considerations for AI design in general, we recognize that practical implementation varies across different organizational structures, resources, and AI projects, potentially affecting the consistency and effectiveness of the ACT theory in practice. Likewise, it is important to acknowledge our study's limitations regarding stakeholder perspectives. Specifically, the perspectives of key groups such as end users, policymakers, and developers were not fully integrated into the analysis. This omission may inadvertently restrict the theory's broader applicability and hinder its potential to effect meaningful change across diverse contexts. To address this, future research should prioritize a more inclusive approach, actively seeking and incorporating the viewpoints of these stakeholders.

Similarly, we did not explicitly address how users are involved in the design process. While exploring user involvement in the design process is beyond the scope of this study, we acknowledge that responsible AI design is inherently iterative, with user feedback playing a critical role in shaping design decisions over time. Across design iterations, user perspectives—whether conveyed through direct feedback, observed interactions, or inferred preferences—inform subsequent decisions, guiding the integration of authenticity, control, and transparency into the system. This iterative process enables designers to continually refine and align the system's behavior with evolving user needs and values. Future research could build on this foundation by systematically examining how user perspectives are elicited and incorporated throughout the responsible AI design lifecycle, offering more profound insights into their influence on responsible design outcomes.

Acknowledging the inherent constraints of our study scope, we urge fellow scholars and industry professionals to build upon this foundational work. Rapid advancements in AI technologies may outpace the applicability of the ACT theory, necessitating ongoing updates and adaptations to remain relevant. Further exploration is also needed to integrate the ACT theory with existing accountability frameworks and practical guidelines to ensure comprehensive ethical alignment. Additionally, the long-term impact of implementing the ACT theory in AI design has yet to be assessed, warranting future longitudinal studies to evaluate its sustained effectiveness. Interdisciplinary integration with fields such as ethics, sociology, and law also presents a promising avenue for exploration. Furthermore, the scalability of the ACT theory's mechanisms across different types, contexts, and sizes of AI projects and organizations requires empirical validation to confirm its broad utility and adaptability.

Despite its limitations, the ACT theory offers significant opportunities for further development and contextualization. Future research could expand and refine the ACT theory across diverse contexts and AI applications. The authenticity mechanism, in particular, merits deeper investigation as an attribute of technological artifacts, particularly regarding its role in shaping the responsible behavior of AI agents. Examples of potential future research questions include:

- What are the outcomes of the ACT theory, and what is the relative importance of each responsibility mechanism in achieving responsible AI?
- How can the interrelationships between affordances, algorithms, and architecture be managed to enhance an responsible AI design process?
- What trade-offs arise from the interactions between design decisions and how they affect the responsibility mechanisms, responsible AI outcomes, and AI adoption?

Appendix E provides a comprehensive list of potential research questions to enable systematic exploration of these and other emerging issues. Addressing these questions can deepen our understanding of the ACT theory's applicability and effectiveness, ensuring that AI agents remain technically robust and ethically grounded.

9 Conclusion

We stand at the brink of a new technological era, one in which the ethical creation of AI promises a future in which these agents are viewed not as a looming threat but as a benevolent force. By prioritizing authenticity, control, and transparency mechanisms within the design domains of affordances, algorithms, and architecture, we establish a simple yet practical foundation for responsible AI design. In this delicate balance, where complexity meets clarity, lies a beacon of hope, guiding humanity toward a future where technology and morality coexist in harmony.

Acknowledgements

We are grateful to our senior editors—Jan Recker, Sutirtha (Suti) Chatterjee, Janina Sundermeier, and Monideepa Tarafdar—for their guidance and support throughout the review process. We owe special thanks to Dr. Chatterjee, whose thoughtful feedback and encouragement significantly contributed to shaping and strengthening this paper. We also appreciate the anonymous reviewers, whose careful and constructive comments challenged us to sharpen our arguments and refine our contributions.

References

Abdel-Karim, B., Pfeuffer, N., Carl, K. V., & Hinz, O. (2023). How AI-based systems can induce reflections: The case of AI-augmented diagnostic work. *MIS Quarterly*, *47*(4), 1395-1424.

Adams, W. C. (2015). Conducting semi-structured interviews. In K. E. Newcomer, H. P. Hatry, J. S. Wholey (Eds.). *Handbook of practical program evaluation* (pp. 492-505). John Wiley & Sons.

Agarwal, A., Agarwal, H., & Agarwal, N. (2023). Fairness score and process standardization: Framework for fairness certification in artificial intelligence systems. *AI and Ethics*, *3*(1), 267-279.

Ahuja, S., Chan, Y. E., & Krishnamurthy, R. (2023). Responsible innovation with digital platforms: Cases in India and Canada. *Information Systems Journal*, *33*(1), 76-129. 78

Akbarighatar, P. (2022). Maturity and readiness models for responsible artificial intelligence (RAI): A systematic literature review. *Proceedings of the 14th Mediterranean Conference on Information Systems.*

Akter, S., McCarthy, G., Sajib, S., Michael, K., Dwivedi, Y., D'Ambra, J., & Shen, K. (2021). Algorithmic bias in data-driven innovation in the age of AI. *International Journal of Information Management*, *60*, Article 102387.

Ali, S. J., Christin, A., Smart, A., & Katila, R. (2023). Walking the walk of AI ethics: Organizational challenges and the individualization of risk among ethics entrepreneurs. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (pp. 217-226).

Alvesson, M., & Sandberg, J. (2011). Generating research questions through problematization. *Academy of Management Review*, *36*(2), 247-271.

Amugongo, L. M., Kriebitz, A., Boch, A., & Lütge, C. (2023). Operationalising AI ethics through the agile software development lifecycle: A case study of AI-enabled mobile health applications. *AI and Ethics,* 5, 227-244.

Anagnostou, M., Karvounidou, O., Katritzidaki, C., Kechagia, C., Melidou, K., Mpeza, E., Konstantinidis, I., Kapantai, E., Berberidis, C., Magnisalis, I., & Peristeras, V. (2022). Characteristics and challenges in the industries towards responsible AI: A systematics literature review. *Ethics and Information Technology*, *24*(37). https://doi.org/10.1007/s10676-022-09634-1

As, I., & Basu, P. (2021). *The Routledge companion to artificial intelligence in architecture*. Routledge.

Astbury, B., & Leeuw, F. L. (2010). Unpacking black boxes: Mechanisms and theory building in evaluation. *American Journal of Evaluation*, *31*(3), 363-381.

Bao, Z., Zhang, W., Zeng, X., Zhao, H., Dong, C., Nie, Y., Liu, Y., Liu, Y., & Wu, J. (2023). Software architecture for responsible artificial intelligence systems: Practice in the digitization of industrial drawings. *Computer*, *56*(4), 38-49.

Bartneck, C., Lutge, C., Wagner, A., & Welsh, S. (2021). Privacy issues of AI. In *An introduction to ethics in robotics and AI* (pp. 61-70). Springer.

Bauer, K., von Zahn, M., & Hinz, O. (2023). Expl(AI)ned: The impact of explainable artificial intelligence on users' information processing. *Information Systems Research*, *34*(4), 1582-1602.

Bennett, M., & Maruyama, Y. (2022). Philosophical specification of empathetic ethical artificial intelligence. *IEEE Transactions on Cognitive and Developmental Systems*, *14*(2), 292-300.

Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Special issue editor's comments: Managing artificial intelligence. *Management Information Systems Quarterly*, *45*(3), 1433-1450.

Birks, D. F., Fernandez, W., Levina, N., & Nasirin, S. (2013). Grounded theory method in information systems research: Its nature, diversity and opportunities. *European Journal of Information Systems*, *22*(1), 1-8.

Blok, V., & Lemmens, P. (2015). The emerging concept of responsible innovation. Three reasons why it is questionable and calls for a radical transformation of the concept of innovation. In B.-J. Koops, I. Oosterlaken, H. Romijn, T. Swierstra, & J. van den Hoven (Eds.), *Responsible innovation 2: Concepts, approaches, and applications* (pp. 19-36.

Burget, M., Bardone, E., & Pedaste, M. (2017). Definitions and conceptual dimensions of responsible research and innovation: A literature review. *Science and Engineering Ethics*, *23*, 1-19.

Castro Pena, M. L., Carballal, A., Rodríguez-Fernández, N., Santos, I., & Romero, J. (2021). Artificial intelligence applied to conceptual design. A review of its use in architecture. *Automation in Construction*, *124*, Article 103550.

Charmaz, K. (2006). *Constructing grounded theory: A practical guide through qualitative analysis*. SAGE.

Chatterjee, S., Chakraborty, S., Fulk, H. K., & Lowry, P. B. (2024). The role of dissonant relational multiplexity in information system implementation failures: Insights from a grounded theory approach. *Journal*

of the Association for Information Systems, *25*(5), 1303-1342.

Chatterjee, S., & Davison, R. M. (2021). The need for compelling problematisation in research: The prevalence of the gap-spotting approach and its limitations. *Information Systems Journal, 31*(2), 227-230.

Chen, R.-C., Dewi, C., Huang, S.-W., & Caraka, R. E. (2020). Selecting critical features for data classification based on machine learning methods. *Journal of Big Data*, *7*(1), 1-26.

Cheng, L., Varshney, K., & Liu, H. (2021). Socially responsible AI algorithms: Issues, purposes, and challenges. *Journal of Artificial Intelligence Research*, *71*, 1137-1181.

Constantinescu, M., Voinea, C., Uszkai, R., & Vica, C. (2021). Understanding responsibility in Responsible AI. Dianoetic virtues and the hard problem of context. *Ethics and Information Technology*, *23*, 803-814.

Corbin, J. M., & Strauss, A. (2015). *Basics of qualitative research: Techniques and procedures for developing grounded theory* (4th ed.). SAGE.

Corley, K. G., & Gioia, D. A. (2011). Building theory about theory building: What constitutes a theoretical contribution? *Academy of Management Review*, *36*(1), 12-32.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2009). *Introduction to algorithms* (3rd ed.). MIT Press.

d'Anjou, P. (2010). Beyond duty and virtue in design ethics. *Design Issues*, *26*(1), 95-105.

Davenport, T., & Kalakota, R. (2019). The potential for artificial intelligence in healthcare. *Future Healthcare Journal*, *6*(2), 94-98.

Davison, R. M., Majchrzak, A., Hardin, A., & Ravishankar, M. N. (2023). Special issue on responsible IS research for a better world. *Information Systems Journal*, *33*(1), 1-7.

de Hoop, E., Pols, A., & Romijn, H. (2016). Limits to responsible innovation. *Journal of Responsible Innovation*, *3*(2), 110-134.

De Loera, J. A., Haddock, J., Ma, A., & Needell, D. (2021). Data-driven algorithm selection and tuning in optimization and signal processing. *Annals of Mathematics and Artificial Intelligence*, *89*(7), 711-735.

De Silva, D., & Alahakoon, D. (2022). An artificial intelligence life cycle: From conception to production. *Patterns*, *3*(6), Article 100489.

Deng, H. (2022). Repairing integrity-based trust violations in ascription disputes for potential e-commerce customers. *MIS Quarterly*, *46*(4), 1983-2014.

Dignum, V. (2019). *Responsible artificial intelligence: How to develop and use AI in a responsible way*. Springer.

Emdad, F. B., Ho, S., Ravuri, B., & Hussain, S. (2023). Towards a unified utilitarian ethics framework for healthcare artificial intelligence. *Proceedings of the Americas Conference on Information Systems.*

European Commission. (2019). *Ethics guidelines for trustworthy AI*. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Faraj, S., & Azad, B. (2012). The materiality of technology: An affordance perspective. In P. M. Leonardi, B. A. Nardi, J. Kallinikos (Eds.) *Materiality and Organizing: Social Interaction in a Technological World*, (pp. 237-258). Oxford University Press.

Fedorowicz, J., Bjørn-Andersen, N., Olbrich, S., Tarafdar, M., & Te'eni, D. (2019). Politics and AIS: Where do we draw the line? *Communications of the Association for Information Systems*, *44*(1), 247-261.

Ferrara, E. (2024). The butterfly effect in artificial intelligence systems: Implications for AI bias and fairness. *Machine Learning with Applications*, *15*, Article 100525.

Figueras, C., Verhagen, H., & Pargman, T. (2022). Exploring tensions in Responsible AI in practice. An interview study on AI practices in and for Swedish public organizations. *Scandinavian Journal of Information Systems*, *34*(2), 199-232.

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. *Minds and Machines*, *28*, 689-707.

Friedman, B., Kahn, P., & Borning, A. (2002). *Value sensitive design: Theory and methods* (University of Washington Technical Report 02-12-01). Available at https://dada.cs.washington.edu/research/tr/2002/12/UW-CSE-02-12-01.pdf

Friedman, B., Kahn, P. H., & Borning, A. (2009). Value sensitive design and information systems. In K. E. Himma & H. T. Tavani (Eds.), *The handbook of information and computer ethics* (pp. 69-101). Wiley.

Gallivan, M. (2001). Organizational adoption and assimilation of complex technological

innovations: Development and application of a new framework. *The Data Base for Advances in Information Systems*, *32*(3), 51-85.

Gaspar, D., Silva, P., & Silva, C. (2024). Explainable AI for intrusion detection systems: LIME and SHAP applicability on multi-layer perceptron. *IEEE Access*, *12*, 30164-30175.

Gaver, W. W. (1991). Technology affordances. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 79-84

Georgievski, I. (2023). Software development life cycle for engineering AI planning systems: *Proceedings of the 18th International Conference on Software Technologies* (pp. 751-760).

Gibson, J. J. (1977). The theory of affordances. In *Perceiving, acting, and knowing: Toward an ecological psychology* (pp. 127-137). Lawrence Erlbaum Associates.

Giffen, B., & Ludwig, H. (2023). How boards of directors govern artificial intelligence. *MIS Quarterly Executive*, *22*(4), 251-272.

Gioia, D. A., Corley, K. G., & Hamilton, A. L. (2013). Seeking qualitative rigor in inductive research: Notes on the Gioia methodology. *Organizational Research Methods*, *16*(1), 15-31.

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Aldine de Gruyter.

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *The Academy of Management Annals*, *14*(2), 627-660.

Gregor, S. (2006). The nature of theory in information systems. *MIS Quarterly*, *30*(3), 611-642.

Gregor, S. (2024). Responsible artificial intelligence and journal publishing. *Journal of the Association for Information Systems*, *25*(1), 48-60.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, *28*(1), 75-105.

Hilton, P. (2002). *Differential privacy: A historical survey*. Cal Poly State University.

Huang, X., Huang, T., Gu, S., Zhao, S., & Zhang, G. (2024). *Responsible federated learning in smart transportation: Outlooks and challenges.* arXiv. https://doi.org/10.48550/arXiv.2404.06777

IBM. (2023). *IBM Global AI Adoption Index 2023*. IBM Newsroom. https://newsroom.ibm.com/2024-01-10-Data-Suggests-Growth-in-Enterprise-Adoption-of-AI-is-Due-to-Widespread-Deployment-by-Early-Adopters

Iivari, J. (2023). Inductive empiricism, theory specialization and scientific idealization in IS theory building. *Communications of the Association for Information Systems*, *52*, 910-914.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389-399.

Jöhnk, J., Weißert, M., & Wyrtki, K. (2021). Ready or not, AI comes—an interview study of organizational AI readiness factors. *Business & Information Systems Engineering*, *63*(1), 5-20.

Jussupow, E., Spohrer, K., & Heinzl, A. (2022). Radiologists' usage of diagnostic AI systems. *Business & Information Systems Engineering*, *64*(3), 293-309.

Kane, G. C. (2021). Avoiding an oppressive future of machine learning: A design theory for emancipatory assistants. *MIS Quarterly*, *45*(1), 371-396.

Koniakou, V. (2023). From the "rush to ethics" to the "race for governance" in artificial intelligence. *Information Systems Frontiers*, *25*(1), 71-102.

Kuusi, O., & Heinonen, S. (2022). Scenarios from artificial narrow intelligence to artificial general intelligence—reviewing the results of the international work/technology 2050 study. *World Futures Review*, *14*(1), 65-79.

Lahiri Chavan, A., & Schaffer, E. (2023). Ethical AI does not have to be like finding a black cat in a dark room. *AI & Society, 39,* 2135-2137.

Leidner, D. E., & Gregory, R. W. (2024). About theory and theorizing. *Journal of the Association for Information Systems*, *25*(3), 501-521.

Leidner, D. E., & Tona, O. (2021). The care theory of dignity amid personal data digitalization. *MIS Quarterly, 45*(1), 343-370.

Leonardi, P. L. (2011). When flexible routines meet flexible technologies: Affordance, constraint, and the imbrication of human and material agencies. *MIS Quarterly*, *35*(1), 147-167.

Liu, D., Lowry, P., Landers, R., Nah, F., & Santhanam, R. (2022). Developing gamification research in information systems. *Proceedings of the Americas Conference on Information Systems.*

Lobschat, L., Mueller, B., Eggers, F., Brandimarte, L., Diefenbach, S., Kroschke, M., & Wirtz, J. (2021). Corporate digital responsibility. *Journal of Business Research*, *122*, 875-888.

Lokuge, S., Sedera, D., Grover, V., & Dongming, X. (2019). Organizational readiness for digital innovation: Development and empirical

calibration of a construct. *Information & Management*, *56*(3), 445-461.

Lubberink, R., Blok, V., van Ophem, J., & Omta, O. (2017). Lessons for responsible innovation in the business context: A systematic literature review of responsible, social and sustainable innovation practices. *Sustainability*, *5*, Article 721.

Lutz, C., & Tamò-Larrieux, A. (2021). Do privacy concerns about social robots affect use intentions? Evidence from an experimental vignette study. *Frontiers in Robotics and AI*, *8*, Article 627958.

Maalej, W., Pham, Y., & Chazette, L. (2023). Tailoring requirements engineering for responsible AI. *Software Engineering for Responsible AI*, 18-27.

Markus, M. L., & Silver, M. (2008). A foundation for the study of IT effects: A new look at DeSanctis and Poole's concepts of structural features and spirit. *Journal of the Association for Information Systems*, *9*(10), 609-632.

Martin, K. (2019). Designing ethical algorithms. *MIS Quarterly Executive*, *18*(2), 129-142. https://doi.org/10.17705/2msqe.00012

McGregor, S. (2020). *Preventing repeated real world AI failures by cataloging incidents: The AI incident database*. arXiv. https://doi.org/10.48550/arXiv.2011.08512

McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & Arcas, B. A. y. (2023). *Communication-efficient learning of deep networks from decentralized data*. arXiv. https://doi.org/10.48550/arXiv.1602.05629

Metcalf, J., Moss, E., & Boyd, D. (2019). Owning ethics: corporate logics, Silicon Valley, and the institutionalization of ethics. *Social Research*, *86*(2), 449-476.

Mikalef, P., Conboy, K., Lundström, J. E., & Popovič, A. (2022). Thinking responsibly about responsible AI and "the dark side" of AI. *European Journal of Information Systems*, *31*(3), 257-268.

Mikalef, P., & Gupta, M. (2021). Artificial intelligence capability: Conceptualization, measurement calibration, and empirical study on its impact on organizational creativity and firm performance. *Information & Management*, *58*(3), Article 103434.

Mohajan, D., & Mohajan, H. K. (2022). Memo writing procedures in grounded theory research methodology. *Studies in Social Science & Humanities*, *1*(4), 10-18.

Mohseni, S., Zarei, N., & Ragan, E. D. (2021). A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Transactions on Interactive Intelligent Systems*, *11*(3-4), 1-45.

Monshizada, S., Sarbazhosseini, H., & Mohammadian, M. (2023). Development of artificial intelligence systems in terms of people-process-data-technology (2PDT). *Pacific Asia Journal of the Association for Information Systems*, *15*(4). https://doi.org/10.17705/1pais.15402

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2020). From what to how: An initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Science and Engineering Ethics*, *26*(4), 2141-2168.

Mueller, B. (2022). Corporate digital responsibility. *Business and Information Systems Engineering*, *64*(5), 689-700.

Napoli, J., Dickinson, S. J., Beverland, M. B., & Farrelly, F. (2014). Measuring consumer-based brand authenticity. *Journal of Business Research*, *67*(6), 1090-1098.

Niederman, F., & Baker, E. W. (2023). Ethics and AI issues: Old container with new wine? *Information Systems Frontiers*, *25*(1), 9-28.

Norman, D. A. (1999). Affordance, conventions, and design. *Interactions, 6*(3), 38-43.

Nussbaumer, A., Pope, A., & Neville, K. (2023). A framework for applying ethics-by-design to decision support systems for emergency management. *Information Systems Journal*, *33*(1), 34-55.

Olorunsogo, T., Jacks, B. S., & Ajala, O. A. (2024). Leveraging quantum computing for inclusive and responsible AI development: A conceptual and review framework. *Computer Science & IT Research Journal*, *5*(3), Article 3.

Owen, R., Macnaghten, P., & Stilgoe, J. (2012). Responsible research and innovation: From science in society to science for society, with society. *Science & Public Policy*, *6*, 751-760.

Papagiannidis, E., Enholm, I. M., Dremel, C., Mikalef, P., & Krogstie, J. (2023). Toward AI governance: Identifying best practices and potential barriers and outcomes. *Information Systems Frontiers*, *25*(1), 123-141.

Paris, F., & Buchanan, L. (2023). 35 ways real people are using A.I. right now. *The New York Times*. https://www.nytimes.com/interactive/2023/04/14/upshot/up-ai-uses.html

Pathirannehelage, S. H., Shrestha, Y. R., & von Krogh, G. (2025). Design principles for artificial intelligence-augmented decision making: An

action design research study. *European Journal of Information Systems, 34*(2), 207-229.

Peters, D., Vold, K., Robinson, D., & Calvo, R. (2020). Responsible AI—Two frameworks for ethical design practice. *IEEE Transactions on Technology and Society*, *1*(1), 34-47.

Polyviou, A., & Zamani, E. D. (2023). Are we nearly there yet? A desires & realities framework for Europe's AI strategy. *Information Systems Frontiers*, *25*(1), 143-159.

Pratt, M. G. (2009). For the lack of a boilerplate: Tips on writing up (and reviewing) qualitative research. *Academy of Management Journal*, *52*(5), 856-862.

Radanliev, P., Santos, O., Brandon-Jones, A., & Joinson, A. (2024). Ethics and responsible AI deployment. *Frontiers in Artificial Intelligence*, *7*. https://doi.org/10.3389/frai.2024.1377011

Ribeiro, B. E., Smith, R. D. J., & Millar, K. (2017). A mobilising concept? Unpacking academic representations of responsible research and innovation. *Science and Engineering Ethics*, *23*(1), 81-103.

Richardson, B., & Gilbert, J. E. (2021). *A Framework for Fairness: A Systematic Review of Existing Fair AI Solutions* arXiv. https://doi.org/10.48550/arXiv.2112.05700

Sanderson, C., Douglas, D., Lu, Q., Schleiger, E., Whittle, J., Lacey, J., Newnham, G., Hajkowicz, S., Robinson, C., & Hansen, D. (2023). AI ethics principles in practice: Perspectives of designers and developers. *IEEE Transactions on Technology and Society*, *4*(2), 171-187.

Sartre, J.-P. (1958). *Being and nothingness: An essay on phenomenological ontology* (Trans. H. E. Barnes, Trans.). Philosophical Library. (Original work published 1943)

Sartre, J.-P. (1992). *Notebooks for an ethics* (D. Pellauer, Trans.). University of Chicago Press. (Original work published 1983)

Seidel, S., Recker, J., & vom Brocke, J. (2013). Sensemaking and sustainable practicing: Functional affordances of information systems in green transformations. *MIS Quarterly*, *37*(4), 1275-1299.

Sen, R., Heim, G., & Zhu, Q. (2022). Artificial intelligence and machine learning in cybersecurity: Applications, challenges, and opportunities for MIS academics. *Communications of the Association for Information Systems*, *51*, 179-209.

Siemon, D., Strohmann, T., & Michalke, S. (2022). Creative potential through artificial intelligence: Recommendations for improving corporate and entrepreneurial innovation activities. *Communications of the Association for Information Systems*, *50*, 241-260.

Simon, H. A. (1996). *The sciences of the artificial* (3rd ed.). The MIT Press.

Smith, N., & Vickers, D. (2021). Statistically responsible artificial intelligence. *Ethics and Information Technology*, *23*, 483-493.

Soma, R., Bratteteig, T., Saplacan, D., Schimmer, R., & Campano, E. (2022). Strengthening human autonomy. In the era of autonomous technology. *Scandinavian Journal of Information Systems*, *34*(2), 163-198.

Stilgoe, J., Owen, R., & Macnaghten, P. (2013). Developing a framework for responsible innovation. *Research Policy*, *9*, 1568-1580.

Strauss, A., & Corbin, J. (1990). *Basics of qualitative research: Grounded theory procedures and techniques*. SAGE.

Suddaby, R. (2006). From the editors: What grounded theory is not. *Academy of Management Journal*, *49*(4), 633-642.

Tabarés, R., Loeber, A., Nieminen, M., Bernstein, M. J., Griessler, E., Blok, V., Cohen, J., Hönigmayer, H., Wunderle, U., & Frankus, E. (2022). Challenges in the implementation of responsible research and innovation across Horizon 2020. *Journal of Responsible Innovation*, *9*(3), 291-314.

Themelis, C., Sime, J.-A., & Thornberg, R. (2023). Informed grounded theory: A symbiosis of philosophy, methodology, and art. *Scandinavian Journal of Educational Research*, *67*(7), 1086-1099.

Trilling, L. (1972). *Sincerity and authenticity*. Harvard University Press.

Umbrello, S. (2019). Beneficial artificial intelligence coordination by means of a value sensitive design approach. *Big Data and Cognitive Computing*, *3*(1), 1-13.

Urquhart, C., & Fernández, W. (2013). Using grounded theory method in information systems: The researcher as blank slate and other myths. *Journal of Information Technology*, *28*(3), 224-236.

Urquhart, C., Lehmann, H., & Myers, M. D. (2010). Putting the "theory" back into grounded theory: Guidelines for grounded theory studies in information systems. *Information Systems Journal*, *20*(4), 357-381.

Vassilakopoulou, P., Parmiggiani, E., Shollo, A., & Grisot, M. (2022). Responsible AI: Concepts, critical perspectives and an information systems research agenda. *Scandinavian Journal of Information Systems*, *34*(2), 89-104.

von Schomberg, R. (2013). A vision of responsible research and innovation. In R. Owen, J. Bessant, & M. Heintz (Eds.), *Responsible innovation: Managing the responsible emergence of science and innovation in society* (pp. 51-74). Wiley.

Wang, Y., Xiong, M., & Olya, H. (2020). Toward an understanding of responsible artificial intelligence practices. *Proceedings of the 53rd Hawaii International Conference on Systems Sciences* (pp. 4962-4971).

Weber, R. (2012). Evaluating and developing theories in the information systems discipline. *Journal of the Association for Information Systems*, *13*(1), 1-30.

Weber-Lewerenz, B. (2021). Corporate digital responsibility (CDR) in construction engineering—Ethical guidelines for the application of digital transformation and artificial intelligence (AI) in user practice. *SN Applied Science*, *3*, 1-25.

Wei, M., & Zhou, Z. (2023). AI ethics issues in real world: Evidence from AI incident database. *Proceedings of the 56th Annual Hawaii International Conference in Systems Science* (pp. 4923-4932).

Xiao, L., Wu, D. J., & Ding, M. (2024). A smart ad display system. *Information Systems Research*, *35*(4), 1873-1889.

Yin, R. K. (2015). *Qualitative research from start to finish* (2nd ed.). Guilford Publications.

Zhou, J., Chen, S., Wu, Y., Li, H., Zhang, B., Zhou, L., Hu, Y., Xiang, Z., Li, Z., Chen, N., Han, W., Xu, C., Wang, D., & Gao, X. (2024). PPML-Omics: A privacy-preserving federated machine learning method protects patients' privacy in omic data. *Science Advances*, *10*(5), Article eadh8601.

Zimmer, M., Minkkinen, M., & Mantymaki, M. (2022). Responsible artificial intelligence systems critical considerations for business model design. *Scandinavian Journal of Information Systems*, *34*(2), 113-162

.

Appendix A. Summary of Existing Responsibility Principles

Table A1. Summary of Existing Responsibility Principles

Principle	Interpretations	Associated terminology
Access	▪ Ensuring that individuals or organizations can acquire and use AI capabilities (Niederman & Baker, 2023). ▪ Ensuring equal opportunity access and just distribution of benefits and costs of AI (European Commission, 2019; Figueras et al., 2022).	Equality Opportunity Accessibility
Accountability	▪ Auditability and assessment of algorithms, data, and design (European Commission, 2019). ▪ Holding the people and organizations that develop and deploy the technology accountable in the event of an adverse outcome (Floridi et al., 2018).	Oversight Responsibility Auditability
Accuracy	▪ Ensuring an AI performs as intended with high precision (Bao et al., 2023; Maalej et al., 2023).	Reliability Precision Trustworthy
Beneficence	▪ Ensuring an AI benefits all human beings and future generations (European Commission, 2019). ▪ Using AI to promote human well-being, preserve dignity, create socioeconomic opportunities, and sustain the planet (Floridi et al., 2018; Jobin et al., 2019).	Individual, environmental, societal well-being
Control	▪ Ensuring there is human oversight over AI operations and decision-making, including a balance between human and machine autonomy (Polyviou & Zamani, 2023). ▪ Building AI agents that offer various opportunities for action, including choices and options available (Soma et al., 2022).	Autonomy Contestability Auditability
Data governance	▪ Considering the quality and integrity of the data prior to training AI models on it (European Commission, 2019).	Data Integrity Data Quality Data Sharing
Empathy	▪ Creating AI that can infer unspoken rules, interpret context, able to infer and explain intent (Bennett & Maruyama, 2022). ▪ Subjecting AI to reactive attitudes of blame, resentment, and indignation (Smith & Vickers, 2021).	Consciousness Emotional Intelligence Sensibility
Fairness	▪ Avoiding unfair bias, discrimination, and marginalization of vulnerable groups (European Commission, 2019; Figueras et al., 2022).	Non-discriminatory Unbiased Equitable
Human agency	▪ Allowing individuals to make decisions for themselves (Floridi et al., 2018). ▪ Empowering individuals to make informed decisions (European Commission, 2019).	Autonomy Empowerment Human-centered
Inclusivity	▪ AI development must consider and involve all affected stakeholders throughout the entire AI agent's lifecycle (Figueras et al., 2022).	Stakeholder Engagement Diversity
Non-maleficence	▪ Preventing accidental (overuse) or deliberate (misuse) harm to individuals arising from the AI (Floridi et al., 2018).	Safety Harm Prevention
Privacy	▪ Protecting direct and indirect user data, keeping it safe and secure from unauthorized access (Akbarighatar, 2022; European Commission, 2019; Jobin et al., 2019). ▪ Offering individuals access to and control over their personal data use (Floridi et al., 2018).	Security Privacy Preservation Data Protection
Sustainability	▪ Developing AI agents that are considerate of the environment, including other living beings (European Commission, 2019).	Environmental well-being Eco-friendly
System awareness	▪ Promoting human awareness of their engagement with an AI and informing users of the agent's capabilities and limitations (Akbarighatar, 2022; European Commission, 2019).	User Awareness Transparency
Technical robustness	▪ Designing AI that is resilient, safe, secure, accurate, and reliable (European Commission, 2019).	Resilience Safety
Transparency	▪ Refers to the explainability of an AI and its decisions (European Commission, 2019), including answering questions about how the AI works and how outcomes were arrived at (Figueras et al., 2022; Floridi et al., 2018).	Explicability Explainability Auditability

Appendix B. Participant Profiles and Interview Questions

Table B1. Profiles of Research Participants

Interview number	Organization type	Gender	Residence	Length of interview (minutes)
1	R&D institutions	Female	Netherlands	29
2	R&D institutions	Male	Finland	38
3	R&D institutions	Male	United States	59
4	R&D institutions	Male	United States	34
5	Government	Male	United States	46
6	AI startup	Male	United States	48
7	Technology firm	Male	Bangladesh	39
8	AI startup	Male	United States	31
9	AI startup	Male	United States	27
10	Technology firm	Male	United States	50
11	R&D institutions	Male	United States	39
12	Government	Male	Australia	28
13	R&D institutions	Male	United States	38
14	Technology firm	Male	United States	30
15	R&D institutions	Female	United States	27
16	Technology firm	Male	United States	26
17	AI startup	Male	United States	35
18	AI startup	Male	United States	35
19	AI startup	Male	United States	38
20	AI startup	Male	United States	24
21	AI startup	Male	United States	25
22	Technology firm	Male	United States	33
23	R&D institutions	Male	United States	Correspondence interview
24	R&D institutions	Male	Germany	42

Table B2. Interview Questions

Can you please describe your current job role?
1. What is *AI design* from your perspective?
2. Can you think of a recent example or instance where an AI agent was designed unethically?
3. How do you define responsible design in the context of AI agents? ▪ Could you share examples of specific practices or methodologies your company/team employs to address potential ethical challenges or biases in AI agent design? ▪ What steps do you personally take during the design process to ensure that the AI agents you develop align with ethical principles and responsible design guidelines? ▪ What specific measures or practices do you personally implement to ensure responsible design in AI agents that your peers in the industry may not commonly employ? ▪ Can you offer any examples of these measures or practices?
4. Are there any areas or aspects where designers commonly face challenges in adhering to ethical principles and responsible design guidelines when creating AI agents? ▪ If so, why and what led to these challenges? ▪ What suggestions do you have for overcoming these challenges?
5. Reflecting on industry common practices, do you see any potential missed opportunities or areas where additional ethical considerations could have been incorporated into the design of AI agents? ▪ If so, what lessons can we learn from these instances, and how should we approach design differently in the future to address these ethical considerations more effectively?

Appendix C: Data Analysis Process and Data Structures

We initially coded our data to identify responsibility mechanisms, employing an iterative process of data collection followed by open, axial, and selective coding, as described in Section 3. Notably, this robust process, supported by meta-mechanisms and an integrated literature review, revealed the responsibility mechanisms of authenticity, control, and transparency as aggregated dimensions of responsible AI early in the analysis. Tables C1, C2, and C3 below showcase a selection of data points utilized in coding the data under responsibility mechanisms of authenticity, control, and transparency, respectively. Due to the extensive nature of some discussions, only the initial segments of the conversations are included to maintain brevity.

Table C1. Responsibility Mechanism: Authenticity

Examples of first-order indicators	First-order indicators	Second-order themes
"Any system, AI or not, also needs to be well-researched [for its utility]." (AI designer, AI startup)	Research the new AI to determine its utility.	Intended utility
"The first step in design is really about understanding the system's potential value. We need to focus on what it offers to our customers and the company and then ask them directly if they consider it truly valuable." (AI designer, R&D institutions)	Determine the AI's value to consumers and the organization.	
"We have to research why the system is needed, how it will perform..." (AI designer, AI startup)	Determine the AI's functional capabilities and purpose.	
"It's super important to get the context where the AI is used. I'm always zoned in on how this context really affects the users' experience." (AI designer, R&D institutions)	Consider the context in which the AI will be deployed (e.g., for cultural sensitivity).	
"They [AI designers] can think about different [usage] scenarios and outlining those and then try to stay with that framework." (AI designer, AI startup)	Conduct scenario planning to anticipate future challenges related to the AI's functionality and usability.	Interactional fidelity
"For the acceptable performance and accuracy, we actually leave it up to ... what is the current state of the art in that field." (AI designer, AI startup)	Establish performance benchmarks and minimum accuracy requirements for the AI.	
"They [AI designers] have to follow some basic rules and regulations [to satisfy performance requirements]." (AI designer, AI startup)	Identify existing regulations or benchmarks that may inform the AI's performance requirements.	
"I mean everything starts and ends with the good data selection ... If you have a dataset that's just crap, then you will not get anything really good out there. The way around that is to make really good samples and have a really good sample strategy." (AI designer, technology firm)	Evaluate the quality of the AI's training data in terms of accuracy (validity).	Data quality
"Getting consistent, good data is like the foundation of a building. The AI can't function very well without it." (AI designer, AI startup)	Ensure the AI's training data is reliable (consistent).	
"If you have a business question, try to think about what data can answer that question and then go get the data, not the other way around." (AI designer, technology firm)	Verify that the AI's training data is related to the business problem.	
"As the AI designer, it's your responsibility to ensure that the training data accurately represents the target population right from the start. Otherwise, the dev team might just focus on making the AI more accurate without really thinking about where the data's coming from." (AI designer, R&D institutions)	Verify that the AI's training data is representative and not biased against any group of individual users.	

"We were able to, over the last ten years, get in a mountain of data ... millions and millions of data points ... So that's the data that our new systems have actually gotten better. That's our foundational model—is that data." (AI designer, technology firm)	Ensure there is adequate data to train the AI model on (volume).	
"There are a variety of different machine-learning models, each with potential benefits and costs that designers must weigh in choosing one based on the business use." (AI designer, R&D institutions)	Compare different machine-learning models to determine which is best suited for the AI's purpose.	Operational accuracy
"And so the first thing that we do is run tests to make sure that, you know, whatever a [user] provides it, it does what it's supposed to." (AI designer, AI startup)	Check the AI's performance outcomes against intended goals.	
"I feel like that's kind of the biggest area is prompt engineering, where you're able to really stress-test and kind of see whether or not the outputs are going to be okay." (AI designer, AI startup)	Conduct prompt engineering to improve the AI's outputs.	
"[Regarding designing facial recognition], … basically, white people don't need as much of a sample size as people who are darker pigments, and I'm reaching out across the company to do that." (AI designer, AI startup)	Verify that the AI's algorithms function consistently and accurately across all users.	
"Essentially, everything that gets asked by a user, we want to create a system that's specialized ... to provide more accurate solutions and more sophisticated answers." (AI designer, AI startup)	Design domain-specific architectural configurations to focus on singular areas of expertise.	Specialization
"We do that by augmenting their knowledge by doing retrieval-augmented generation by grounding their knowledge in sophisticated databases and systems." (AI designer, AI startup)	Conduct retrieval-augmented generation (RAG) to enhance the AI's knowledge.	
We need to make sure our AI can accommodate higher demand and that it always works regardless of how many people are trying to use it.* (AI designer, AI startup)	Ensure the AI can accommodate higher demand and data volume.	Architectural resilience
"One way to [create an all-knowing agent] in the short-term is by creating a bunch of specialized agents, so that you have an agent that knows how to code and another agent that knows how to answer questions about cooking ... And then, together, they create an all-knowing agent." (AI designer, AI startup)	Build components that are interoperable with each other.	
"Standardized interfaces ... ensure seamless integration so that each part [of the system] can communicate with the others." (AI designer, technology firm)	Use standardized interfaces so that different components of the AI can easily integrate with each other.	
"Eventually, it will be scalable in the sense that these are not unique problems that we're trying to tackle. What's lacking is sort of a common way to share these recipes that pretty much everyone uses." (AI designer, AI startup)	Design so that individual AI components can be scaled as necessary and systematically maintained.	

Note: * This quote has been reconstructed based on two primary sources: (1) conversations with participants who chose not to be audio-recorded, and (2) segments of the original dialogue that were fragmented and lacked coherence in their initial format.

Table C2. Responsibility Mechanism: Control

Examples of first-order indicators	First-order indicators	Second-order themes
"The thing we're building is a lot more useful if users can save the output and go back and check it." (AI designer, AI startup)	Give users the functions to save the AI's output for future reference.	Regenerability
"It's good when users can regenerate content or ask it to provide the answer in a simpler way." (AI designer, R&D institutions)	Give users the option to regenerate the AI's output.	
"What we want to do and put in place is a way for users to provide the constraints that an answer needs to follow." (AI designer, AI startup)	Allow users to set bounds, constraints, or filters on the content generated by the AI.	Customizability
"The basics are making sure that content is usable, readable by screen readers, and building content kind of from the ground-up with accessibility in mind." (AI designer, R&D institutions)	Ensure the AI includes accessibility features for persons with disabilities or other special needs.	
What we're aiming for with our product is like what ChatGPT does. We want to give users the choice to pick between something like GPT-3 or GPT-4, depending on their budget and what they need from it.* (AI designer, technology firm)	Allow users to choose the AI's model for generating output.	
"We give the control to the user ... We give the options to the user to decide what kind of data the AI is trained on and what kind of outcomes they will get." (AI designer, AI startup)	Allow users to choose the data the AI is trained on or make a reference to.	
"So individually, [users] own the right to their [own] data ... Users have permission to, for example, to use data or delete the data or download the data." (AI designer, AI startup)	Allow users to control the data generated on them by the AI.	
We've designed our AI to let users specify exactly what they're looking for. This helps our system deliver results that are truly helpful.* (AI designer, AI startup)	Allow users to set output criteria or predefine expectations from their interactions with the AI (customer instruction).	
"And if someone misuses the system ... there could be mechanisms that the AI agent can detect and report it to the system, system admin, for example." (AI designer, AI startup)	Flag and limit excessive usage patterns signifying potential abuse or misuse of the AI.	Operational safeguards
It's important that users aren't just recklessly using an AI and that there's a balance between exploration and exploitation.* (AI designer, AI startup)	Incorporate rate limits or quota systems to prevent AI misuse and ensure usage aligns with the intended purpose.	
"If someone is [intentionally] giving the wrong feedback to the system ... you can have some sort of cross-validation to verify that type of feedback is right." (AI designer, AI startup)	Prevent the AI from being exploited for malicious purposes.	
"Before we get started on a new AI, we just ask our customers things like whether they prefer to keep their data more private or if they want the product to be more accurate because sometimes these are the kinds of things we need to consider." (AI designer, AI startup)	Obtain users' input regarding the AI's limitations.	Feedback integration
"There's a plan in place to report using thumbs-up and thumbs-down [buttons], or things that didn't go as well, as a sort of way to improve moderation." (AI designer, AI startup)	Incorporate user-generated feedback to improve the AI's output based on user expectations.	
"Before we launched anything to the public, we would always ... go through like a Privacy Council [to make sure] there was some level of consent we had already been given." (AI designer, AI startup)	Consider user preferences when using their data to train the AI model.	Privacy assurance
"Differential privacy another method to preserve privacy is like adding noise to the data so that any individual person's data is synthetic..." (AI designer, AI startup)	Implement differential privacy during model training to preserve users' privacy, providing developers with enhanced control in AI training.	

"If you build your AI in modules, you can tweak, test, or check one part without messing with the whole system. It's like fixing a bike's tire without having to take apart the entire bike." (AI designer, R&D institutions)	Construct the AI using modular design, so that administrators and developers can modify or adjust individual AI components without impacting the entire system.	Modularity
"If you don't want to share data across individual sources ... you just share the model weights rather than the individual data." (AI designer, AI startup)	Use federated learning to enhance developer control over sensitive user data.	
"I'm interested in looking at what's called human-in-the-loop AI that combines an AI basically working together with humans ... So where humans are kind of extracting the relevant social signals and providing that to an AI, which can then make the final diagnosis." (AI designer, AI startup)	Consider human-in-the-loop (HITL) to verify the AI's outputs.	Continuous oversight
"What we did was building an ML pipeline using [MLOps] to detect a different type of service failure ... and then provide a recommendation to that platform owner on how to recover that failure." (AI designer, AI startup)	Adopt MLOps that allow administrators to continuously monitor the AI without impeding ongoing operational flows.	

Note: * This quote has been reconstructed based on two primary sources: (1) conversations with participants who chose not to be audio-recorded, and (2) segments of the original dialogue that were fragmented and lacked coherence in their initial format.

Table C3. Responsibility Mechanism: Transparency

Examples of first-order indicators	First-order indicators	Second-order themes
"For example, when the system created a recommendation, that recommendation at the end, in the best-case scenario, has a 70-80% accuracy. Always there is room for error. We try to communicate that with the end user up front." (AI designer, AI startup)	Inform users of the AI's accuracy and operational effectiveness.	Limitations disclosure
"I actually ran an experiment. I asked ChatGPT to give me a report ... and to give the references as well. It created the report fine, and it created some false references which looked very accurate ... It's a responsibility of the developing company [to let users know about this limitation]." (AI designer, AI startup)	Inform users of the AI's operational limitations and scope.	
It's important that users understand that the purpose of an AI may not be to increase accuracy over a human counterpart; it may serve another purpose, such as process improvement.* (AI designer, AI startup)	Ensure users understand the intended purpose of the AI as related to its capabilities.	Purpose disclosure
It's important that users know that we are using their feedback to try to make the product better—so that they know that what they tell us and how they interact with the AI is meaningful.* (AI designer, technology firm)	Educate users on how the AI's capabilities are improved based on user feedback and interactions.	
Users need to know how to use our product to get the most out of it. This isn't always a given.* (AI designer, AI startup)	Communicate the AI's functions and functionalities to users.	Capabilities disclosure
"We need to educate these people on what generative AI is and how to use our product." (AI designer, technology firm)	Educate users on how to actualize the AI's capabilities.	
"[Designing AI is similar to other projects we've done] where we've contracted a website team, for example, and we've said, you know, we want high color contrast, we want to make sure there's correct reading order, and then they'll acknowledge that ... [but] then they might just throw a WordPress plugin on there. But that's not necessarily good enough ... We need to be better at making sure [developers] understand the intended purpose so that the decisions they make support our overall goal." (AI designer, technology firm)	Communicate the AI's intended purpose to developers.	System cognizance

"Organizations that have, like, almost manifestos or principles, like privacy by design ... it helps workers whenever to look back at something and see what the company stands for. So I think to have principles actually formulated is very beneficial." (AI designer, R&D institutions)	Disclose trade-off management mechanisms and ethical requirements throughout the development process.	
"[In AI design], if you see your project's gonna take longer than planned, it's super important to give a heads-up to your PM and the dev team. Keeping everyone in the loop really helps in maintaining the quality of what we're building." (AI designer, R&D institutions)	Communicate delays in build time.	
"If I am not going to meet some deadline or requirement—you know, like accuracy or time or money—then everyone has to know that because it can impact other people on the team." (AI designer, R&D institutions)	Report unmet performance benchmarks to the development team.	
We should be sharing the results of our audits with everyone on the build team so that everyone knows what we're doing well and where we're falling short.* (AI designer, technology firm)	Report the results of internal audits to the development team.	
"I will also look at the explainability part, which is basically like trying to explain why different models are making the decisions they are making ... Responsible AI assumes that you have explainable AI, because in order to get responsible AI, it means that you already know why the model is making the predictions that it's making." (AI designer, R&D institutions)	Include explainability mechanisms so that users can see how the AI generated its output.	System interpretability
When writing the entire code in-house rather than incorporating external products to serve different functions, we are able to understand how the AI came with its decisions better. We can look back and see where the decisions were made and why it produced certain outcomes.* (AI designer, technology firm)	Consider writing the entire code in-house rather than using external black boxes.	
"[If we want models that are fully explainable], within white box, you can use a model that's inherently interpretable, like a decision tree or logistic regression." (AI designer, AI startup)	Consider white box methods when possible.	
"[To determine how a black box AI created its output], try to basically change the input a little to see how that affects the output, and by doing that you're able to see why did the model make the predictions that it did." (AI designer, AI startup)	Conduct perturbation analysis or saliency maps for black box methods that are inherently uninterpretable.	
"Our users want to, need to, know that their data is safe ... That we're not giving out the recipe to their secret sauce, or something." (AI designer, technology firm)	Disclose how user data is stored and secured.	Security and privacy transparency
"Users want to, or should, know how we use their data ... If we're collecting data on them that we go on and sell, they need to know that." (AI designer, AI startup)	Disclose how user data is used or monetized.	
Not only do the design intentions need to be disclosed before the product launches, but any changes to an AI's design also need to be logged and thoroughly documented.* (AI designer, R&D institutions)	Maintain configuration and deployment logs on changes to the AI's architecture and overall design.	Ongoing system monitoring
Whenever changes are made to specific components in the AI, everyone needs to know about those. They need to be logged and documented so that everyone can understand and track changes made to the AI.* (AI designer, R&D institutions)	Maintain version control to track and monitor changes to the AI's algorithm and supporting architecture.	
"Documentation is pivotal not merely for record-keeping, but also for elucidating the rationale behind each design decision." (AI designer, AI startup)	Document decisions and intentions for AI architecture design and deployment.	
"Once the AI is up and running, we need to keep an operational log so that we know how it's performing." (AI designer, AI startup)	Document and maintain an operational log.	

"Just like open source code, we can also consider open architecture ... so that anyone who wants to look at how the system is designed or operates can do it." (AI designer, AI startup)	Use open architecture to allow the components and structures of the AI to be inspected or audited.	Auditability
If we build modular AI, where we have different specialized, independent components that work together, then we can look at how each component made its decisions and change it if we need to.* (AI designer, AI startup)	Implement modular architecture to facilitate inspection or auditing of each module's role/behavior.	
Another thing is modular data architecture, so that different components handle different data sources, making it easier to track and audit data from each separate source.* (AI designer, AI startup)	Implement modular data architecture to facilitate inspection or audit of each data source.	
"It might be hard for us to build explainable AI—that's actually really hard to do—but there are tools they're developing that can do this ... We can design our AI so these kinds of explainability add-ons can work with our system." (AI designer, AI startup)	Build in interoperability layers that allow external explainability tools to be added to the AI architecture.	
Note: * This quote has been reconstructed based on two primary sources: (1) conversations with participants who chose not to be audio-recorded, and (2) segments of the original dialogue that were fragmented and lacked coherence in their initial format.		

As we collected and analyzed new data, we found that these three responsibility mechanisms remained cornerstones of responsible design. However, it also became clear that the design decisions supporting these mechanisms could be further categorized according to different AI design domains. Untangling these design domains, under the responsibility mechanisms of authenticity, control, and transparency proved to be a more challenging task. We revisited our data iteratively using authenticity, control, and transparency as overarching themes to group our second-order themes according to unique AI design domains, consistently revisiting the raw data to ensure our coding was consistent with our empirical evidence. This level of coding revealed that designers typically referenced three key domains of AI design, namely design *affordances*, *algorithms*, and *architecture.* Our second-order themes thus reflect the key design decisions concerning responsible AI. These design decisions initiate responsibility mechanisms of authenticity, control, and transparency, which collectively ensure an AI behaves responsibly, maximizing its beneficence and minimizing its maleficence. This coding process was instrumental in developing the propositions presented in this paper.

Figures C1-C3 exemplify our data coding process, beginning with responsibility mechanisms and progressing to design domains, ultimately shaping the conceptualization of Authenticity. We followed the same procedure for the control and transparency mechanisms. Tables C4-C6 outline our resulting data structures for authenticity, control, and transparency.

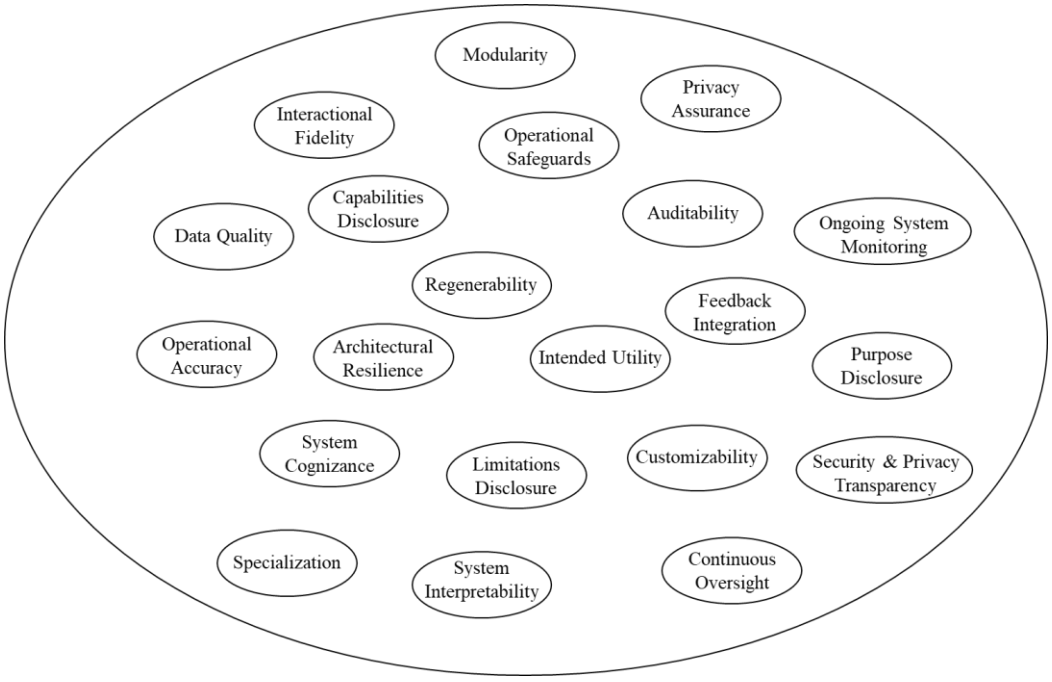

Figure C1. Second-Order Themes (Design Decisions)

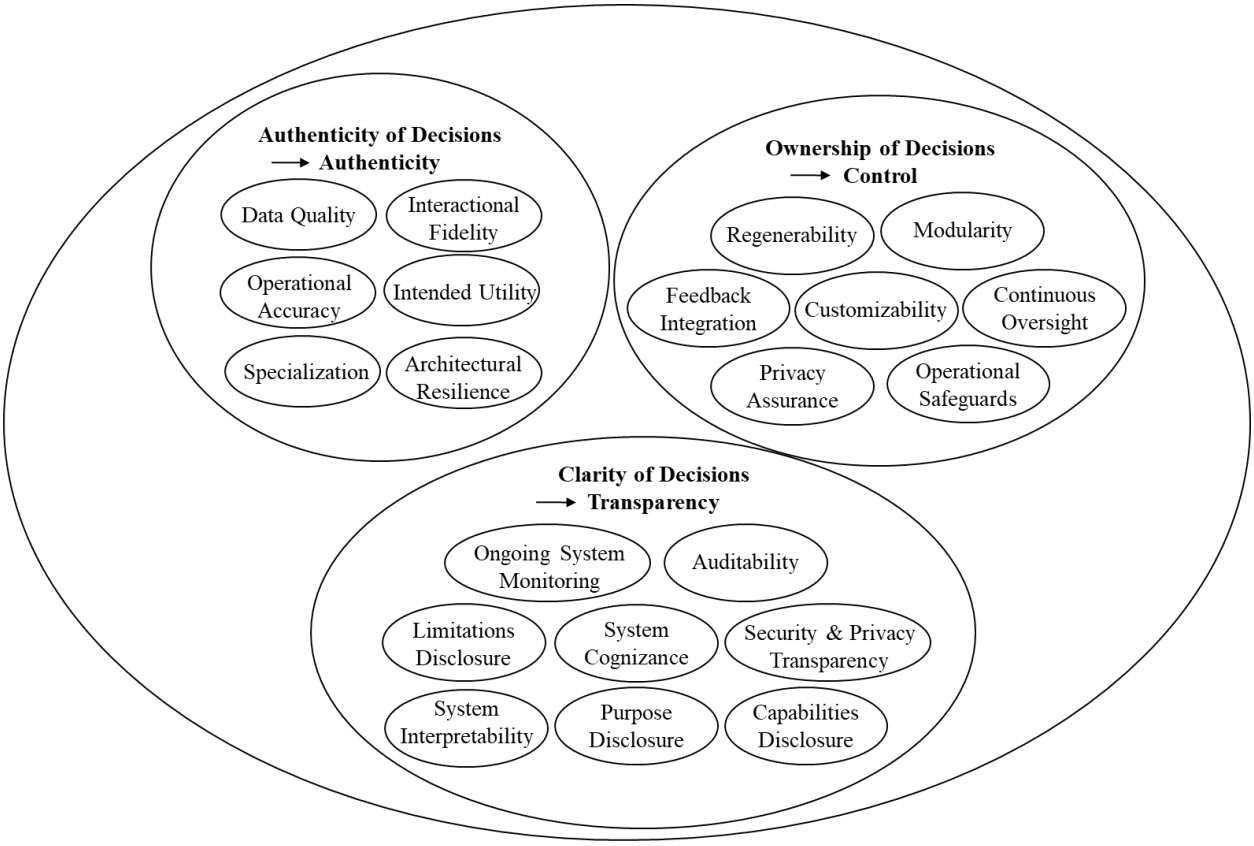

Figure C2. First Level of Data Aggregation Into Responsibility Mechanisms

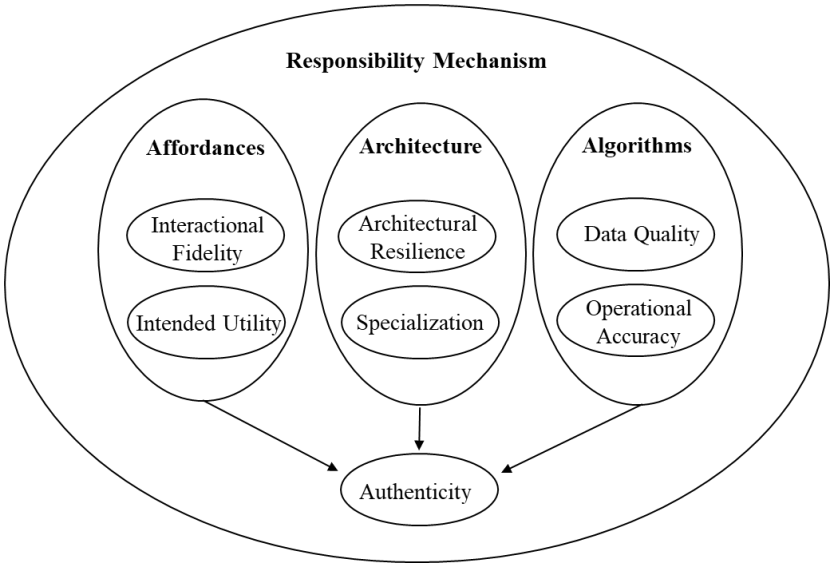

Figure C3. Sample of Further Data Classification of Design Decisions According to Design Domains (Authenticity)

Table C4. Data Structure for Authenticity

First-order indicators	Second-order themes	Design domains	Responsibility mechanism
Research the new AI to ensure that it provides users with some utility.	Intended utility	Affordances	Authenticity
Determine the AI's functional capabilities and purpose.			
Determine the AI's value to consumers and the organization.			
Consider the context in which the AI will be deployed (e.g., for cultural sensitivity).			
Conduct scenario planning to anticipate future challenges related to the AI's functionality and usability.	Interactional fidelity		
Establish performance benchmarks and minimum accuracy requirements for the AI.			
Identify existing regulations or benchmarks that may inform the AI's performance requirements.			
Evaluate the quality of the AI's training data in terms of accuracy (validity).	Data quality	Algorithms	
Ensure the AI's training data is reliable (consistent).			
Verify that the AI's training data is related to the business problem.			
Verify that the AI's training data is representative and not biased against any group of individual users.			
Ensure there is adequate data to train the AI model on (volume).			
Compare different machine-learning models to determine which is best suited for the AI's purpose.	Operational accuracy		
Check the AI's performance outcomes against intended goals.			
Conduct prompt engineering to improve the AI's outputs.			
Verify that the AI's algorithms function consistently and accurately across all users.			
Design domain-specific architectural configurations to focus on singular areas of expertise.	Specialization	Architecture	
Conduct retrieval-augmented generation (RAG) to enhance the AI's knowledge.			
Ensure the AI can accommodate higher demand and volume.	Architectural resilience		
Build components that are interoperable with each other.			
Use standardized interfaces so that different components of the AI can easily integrate with each other.			
Design so that individual AI components can be scaled as necessary and systematically maintained.			

Table C5. Data Structure for Control

First-order indicators	Second-order themes	Design domains	Responsibility mechanism
Give users the functions to save the AI's output for future reference.	Regenerability	Affordances	Control
Give users the option to regenerate the AI's output.			
Allow users to set bounds, constraints, or filters on the content generated by the AI.	Customizability		
Ensure the AI includes accessibility features for persons with disabilities or other special needs.			
Allow users to choose the AI's model for generating output.			
Allow users to choose the data the AI is trained on or make reference to.			
Allow users to control the data generated on them by the AI.			
Allow users to set output criteria or predefine expectations from their interactions with the AI (customer instruction).			
Flag and limit excessive usage patterns, signifying potential abuse or misuse of the AI.	Operational safeguards		
Incorporate rate limits or quota systems to prevent AI misuse and ensure usage aligns with the intended purpose.			
Prevent the AI from being exploited for malicious purposes.			
Obtain users' input regarding the AI's limitations.	Feedback integration	Algorithms	
Incorporate user-generated feedback to improve the AI's output based on user expectations.			
Consider user preferences when using their data to train the AI model.	Privacy assurance		
Implement differential privacy during model training to preserve users' privacy, providing developers with enhanced control in AI training.			
Construct the AI using modular design so that administrators and developers can modify or adjust individual AI components without impacting the entire system.	Modularity	Architecture	
Use federated learning to enhance developer control over sensitive user data.			
Consider human-in-the-loop (HITL) to verify the AI's outputs.	Continuous oversight		
Adopt MLOps that allow administrators to continuously monitor the AI without impeding ongoing operational flows.			

Table C6. Data Structure for Transparency

First-order indicators	Second-order themes	Design domains	Responsibility mechanism
Inform users of the AI's accuracy and operational effectiveness.	Limitations disclosure	Affordances	Transparency
Inform users of the AI's operational limitations and scope.			
Ensure users understand the intended purpose of the AI and its capabilities.	Purpose disclosure		
Educate users on how the AI's capabilities are improved based on user feedback and interactions.			
Communicate the AI's functions and functionalities to users.	Capabilities disclosure		
Educate users on how to actualize the AI's capabilities.			
Communicate the AI's intended purpose to developers.	System cognizance	Algorithms	
Disclose trade-off management mechanisms and ethical requirements throughout the development process.			
Communicate delays in build time.			
Report unmet performance benchmarks to the development team.			
Report the results of internal audits to the development team.			
Include explainability mechanisms so users can see how the AI generates its output.	System interpretability		
Consider writing the entire code in-house rather than using external black boxes.			
Consider white box methods when possible.			
Conduct perturbation analysis or saliency maps for black box methods that are inherently uninterpretable.			
Disclose how user data is stored and secured.	Security and privacy transparency	Architecture	
Disclose how user data is used or monetized.			
Maintain configuration and deployment logs on the AI's architecture and overall design changes.	Ongoing system monitoring		
Maintain version control to track and monitor AI algorithm changes and supporting architecture.			
Document decisions and intentions for AI architecture design and deployment.			
Document and maintain an operational log.			
Use open architecture to allow the components and structures of the AI to be inspected or audited.	Auditability		
Implement modular architecture to facilitate inspection or auditing of each module's role/behavior.			
Implement modular data architecture to facilitate inspection or audit of each data source.			
Build in interoperability layers that allow external explainability tools to be added to the AI architecture.			

Appendix D: Overview of Responsible Innovation Theories and Frameworks

In this section, we provide an overview of existing frameworks—responsible research and innovation (RRI), the CARE theory of dignity (CARE), corporate digital responsibility (CDR), and value sensitive design (VSD)—that are often adapted with constraints for AI contexts.

Responsible research and innovation (RRI) is a term most notably conceptualized in the European Union's Framework Programmes to highlight the need for collaboration and cooperation between scientific advancements and societal well-being (Burget et al., 2017). The general objective of RRI is to engender ethically acceptable, sustainable, and socially desirable practices into the innovation process (von Schomberg, 2013) through democratic governance models that include various stakeholders early in the innovation lifecycle (Owen et al., 2012; Stilgoe et al., 2013). Governance of the innovation process is argued to encourage deliberation, scrutiny, and verification of a new technology's purpose, attributes, and societal contributions, ultimately improving the likelihood of innovation adoption within society (Lubberink et al., 2017; Ribeiro et al., 2017). Dimensions of RRI include *anticipation* (identifying what is known, what is probable, and what is possible with a new technology), *reflexivity* (objectively scrutinizing a new technology), *inclusion* (involving stakeholders in the innovation process), and *responsiveness* (the ability to change the direction of a new technology as needed) (Stilgoe et al., 2013). While RRI serves as an instrumental framework in guiding ethical considerations in technology, it has not escaped scholarly critique (Blok & Lemmens, 2015; de Hoop et al., 2016; Tabarés et al., 2022). Critics assert that the RRI framework exhibits a degree of generality that renders it ineffectual in addressing the complexities inherent in sophisticated technologies. These criticisms highlight the need for frameworks with greater specificity and practical guidance tailored to the unique challenges posed by advanced information systems (Ahuja et al., 2023).

In recent years, information systems scholars have been striving to develop contextually relevant theories. One such successful attempt is the CARE (claims, affronts, response, equilibrium) theory of dignity. As a grand theory, CARE investigates how the use and dependence on digitized personal data affect human dignity daily (Leidner & Tona, 2021). This theory relates claims and affronts from four forms of personal data digitization (knowing self, showing self, knowing others, and showing others) to three forms of human dignity (behavioral, inherent, and meritocratic). The CARE literature defines a *claim to dignity* as any action that enables or supports dignity, while an *affront to dignity* is any action that threatens or sequesters dignity. The theory suggests that claims and affronts caused by personal data digitization result in dignity disequilibrium, qualified as intrapersonal, interpersonal, or associated disequilibrium (Leidner & Tona, 2021). In the context of responsible AI, the CARE theory most closely aligns with the principles of beneficence, non-maleficence, and privacy. However, while insightful, the theory does not exhaustively address all facets of responsible AI, such as transparency, accountability, or technical robustness.

While RRI focuses on the developmental aspects and the CARE theory centers on usage, a distinct category of theories homes in on operational considerations. Notably, Lobschat et al. (2021) introduce corporate digital responsibility (CDR) as a framework designed to align organizational values with operational practices in the context of digital technology creation and data management. The authors posit that system designers must be cognizant of the potential for unanticipated uses of their technologies, which could have unintended ramifications for individual stakeholders and society at large. This suggests that ethical digital technology design transcends mere technological challenges; it necessitates an organizational culture, predicated on a value system that promotes ethical technology deployment. Lobschat et al. (2021) define CDR as "a set of shared values and norms guiding an organization's operations with respect to four main processes related to digital technology and data" (p. 876), specifically outlining processes tied to technology creation, data capture, operational decision-making, impact assessment, and technological refinement. Although CDR does not necessarily augment the existing dimensions of responsible AI, it furnishes an organizational framework for embedding responsible and ethical practices. Weber-Lewerenz (2021) observes that this model may be particularly salient for nascent organizations that have not yet solidified their internal values, goals, and cultures. Mueller (2022) further contends that before an organization can successfully integrate CDR into its operations, it must first grapple with the pivotal task of delineating 'good' values and norms and understanding their significance to stakeholders. Analogous to critiques of the RRI framework, CDR has faced scrutiny for being too generic to offer actionable guidance for practical implementation.

Lastly, design-oriented theories significantly contribute to the intellectual landscape of responsible AI, with value sensitive design (VSD) being a particularly notable example. VSD is a design methodology that integrates moral and ethical values into the architecture of computer technologies, evaluating how these technologies either uphold or erode such values (Friedman, 1996). The approach posits that certain values, such as human rights, possess universal resonance, albeit their practical applications may differ according to cultural and contextual nuances (Friedman et al., 2002). Acknowledging the multilayered complexity inherent in responsible design, VSD employs a tripartite, iterative methodology that encompasses conceptual, empirical, and technical dimensions (Friedman et al., 2009). *Conceptual investigations* involve analyzing how direct and indirect stakeholders will be impacted by a technological artifact and how competing values might be balanced during artifact design (Friedman et al., 2009). *Empirical investigations* involve an analysis of the human context in which the artifact will be situated, with an evaluation of the success of a particular design (Friedman et al., 2009). Finally, *technology investigations* focus on the technology itself (rather than the individuals/groups impacted by the technology). Here, system components are analyzed to determine how they support (or hinder) values identified in the conceptual investigation (Friedman et al., 2009). Compared to RRI, CARE, and CDR, VSD stands out as the preeminent framework deployed in the AI sector. However, it is important to underscore that VSD, akin to CDR, does not enhance or expand upon the extant principles of responsible AI. Instead, it serves as an efficacious methodology for transmuting these principles into actionable practice. While its iterative and multifaceted design methodology commendably aligns with the nuances of responsible AI, detractors argue that VSD is not well-positioned to meet AI's ubiquitous and rapid growth (Umbrello, 2019).

Appendix E: Possible Future Research Questions

Table E1. Possible Research Questions Related to Authenticity

Affordances	1. What role does user interface play in enhancing or hindering authenticity in AI design? 2. How can affordances in AI design authentically reflect the needs and values of diverse user groups? 3. To what extent do affordances enhance or constrain the user's ability to act authentically? 4. How can AI affordances be designed to promote authentic user experiences without prescribing rigid patterns of interaction? 5. How do different industries interpret and implement authenticity in AI? 6. How can AI agents' affordances be created to ensure they are contextually sensitive to the populations they serve?
Algorithm	1. What role does algorithmic complexity play in the perception of authenticity? 2. How can accountability be ensured in algorithmic decision-making? 3. How does the complexity of algorithms affect the balance between authenticity and computational efficiency? 4. How can algorithms be crafted to authentically represent the complexity of user contexts and values? 5. To what extent does prioritizing authenticity in algorithmic design necessitate sacrificing simplicity or scalability, and how should this tension be managed?
Architecture	1. How can specialized architectures enhance authenticity in AI agents? 2. What are the ethical implications of architectural decisions on user autonomy? 3. How does architecture influence the trade-off between privacy and performance? 4. In what ways can AI architecture support or hinder transparent data collection? 5. How does the structure of an AI impact its perceived authenticity? 6. How does the structure of an AI agent influence the user's ability to perceive it as an authentic and trustworthy entity?

Table E2. Possible Research Questions Related to Control

Affordances	1. How can designers balance user empowerment with the risk of overwhelming users with information? 2. What affordances can be implemented to support informed decision-making without infringing on autonomy? 3. How can user interfaces be designed to facilitate a balance between control and ease of use? 4. How do affordances impact the perception of control in AI agents? 5. How can affordances be leveraged to enhance individual and societal control over AI? 6. What are the trade-offs between providing detailed control options and ensuring a seamless user experience?
Algorithm	1. How does algorithmic design interact with control mechanisms in AI agents? 2. What measures can be taken to ensure transparent and ethical data collection processes? 3. How does informed consent function within complex algorithms, and how can it be enhanced? 4. In what ways can algorithms be optimized for balanced user control? 5. How can regulatory compliance be ensured within control algorithms? 6. What role does algorithmic transparency play in user control? 7. How can algorithms be designed to prevent misuse, abuse, or overuse while maintaining high user control? 8. What trade-offs exist between algorithmic adaptability and maintaining user control in AI applications?
Architecture	1. How do architectural decisions impact control distribution in AI agents? 2. What are the security implications of distributed control mechanisms? 3. How does architecture influence the long-term scalability and ethical soundness of control in AI? 4. In what ways can architecture support or undermine user control? 5. How does AI architecture interact with regulatory frameworks to influence control?

Table E3. Possible Research Questions Related to Transparency

Affordances	1. How can user interfaces be designed to enhance transparency in AI agents? 2. How can designers ensure that the true purposes of AI agents are transparently communicated to users? 3. How can misleading affordances be identified and mitigated in AI agents? 4. What practices can be implemented to ensure that advertising and reality align in AI applications?
Algorithm	1. How can algorithms be optimized to enhance transparency while ensuring security? 2. What role does algorithmic complexity play in transparency, and how can it be managed? 3. How can ethical frameworks guide developers in creating transparent algorithms? 4. How do algorithms impact the perception of transparency in AI agents? 5. How can empirical studies be designed to measure the impact of transparency on user behavior? 6. What are the implications of making algorithmic processes transparent and IP rights?
Architecture	1. How does architecture influence an AI's ethical alignment and performance in terms of transparency? 2. What are the security implications of transparent architectural decisions? 3. How does an AI's structure impact the balance between automation and human oversight? 4. How can architecture be optimized to enhance transparency in diverse use cases? 5. How does the structure of an AI agent influence its capacity for transparency? 6. How can design decisions promote a balance between individual rights and collective security needs?

About the Authors

Andrea Rivera is a PhD candidate in information technology management at the University of Hawai‘i at Mānoa. She holds a master of science degree in applied mathematics from San Diego State University and a master of business administration degree from the University of Florida. She is a US Navy veteran, where she served as a surface warfare officer in nuclear power engineering. Her current research focuses on the design and adoption of responsible AI systems, with an emphasis on balancing normative expectations with practical implementation in organizational settings. She aims to advance research at the intersection of ethics, design, and real-world decision-making in technology-intensive environments.

Kaveh Abhari is a professor of management information systems at San Diego State University, director of the Digital Innovation Lab (DiLab), and a research affiliate at the James Silberrad Brown Center for Artificial Intelligence. With more than two decades of experience in teaching, research, and consulting, his work centers on humane digital transformation, with a particular focus on how emerging technologies are reshaping the future of work. At DiLab, his research explores the design of responsible AI-enabled systems that harmonize technological innovation with human-centered values—advancing trustworthy, adaptive, and purpose-driven applications of AI across professional domains.

Bo Xiao is the Shidler College Distinguished Professor in Information Technology Management at the Shidler College of Business, University of Hawai‘i at Mānoa. Her primary research interests include responsible information systems, human-computer interaction, digital platforms, and healthcare information systems. Her current work focuses on responsible artificial intelligence (AI) and human-AI interaction. She has published multiple papers in top information systems journals, including *MIS Quarterly, Information Systems Research,* and *Journal of the Association for Information Systems*. Dr. Xiao has served as an associate editor of *MIS Quarterly* and *Internet Research*, an editorial board member of *Journal of the Association for Information Systems*, and an advisory board member of *Industrial Management and Data Systems*. She has also held various leadership roles at premier information systems conferences, including co-chair, program co-chair, track co-chair, and mini-track co-chair.