AI Agents as Governance Actors in Data Trusts – A Normative and Design Framework

Research Paper

Arnold F. Arz von Straussenburg[1], Jens J. Marga[2], Timon T. Aldenhoff[1], and Dennis M. Riehle[1]

[1] University of Koblenz, Institute for Information Systems Research, Koblenz, Germany
{arz,timonaldenhoff,riehle}@uni-koblenz.de
[2] WHU - Otto Beisheim School of Management, Management Group, Vallendar, Germany
jens.marga@whu.edu

Abstract. Data trusts have emerged as structured mechanisms to ensure responsible data stewardship grounded in fiduciary duties, transparent oversight, and user-centered governance. Meanwhile, recent advances in Artificial Intelligence (AI) transform Information Systems by automating decisions and enabling novel data-driven applications while raising ethical, security, and trust challenges. This paper proposes a design theory that unifies fiduciary principles, institutional trust, and AI ethics to guide the integration of AI into data trusts. We introduce four design principles: fiduciary alignment, traceability and accountability, transparent explainability, and autonomy-preserving oversight. These principles protect the beneficiaries' interests and the owner's rights, mitigate opacity and conflicts of interest, and maintain robust human supervision. The framework contributes to emerging governance approaches that consider fairness, trustworthiness, and societal acceptance of AI-driven data ecosystems. We conclude with recommendations for empirical validation and sector-specific adaptations to ensure responsible AI use for the common good.

Keywords: Data Trusts, Normative Framework, AI Governance, Fairness, AI Agents.

1 Introduction

Information Systems (IS) have reached a new era in which Artificial Intelligence (AI)-based solutions not only automate business processes but also shape innovative data-driven models (De Silva et al. 2024, Balasubramaniam et al. 2024). The rapid development of generative AI and large-scale predictive models has broadened the scope of AI beyond tightly prescribed tasks, contributing to creative and open-ended applications. While these systems promise enhanced efficiency and societal benefits, they also bring concerns over fairness, transparency, and privacy, highlighting the need for new governance approaches in AI-enabled IS (Berente et al. 2021, Jussupow et al. 2024). In parallel, recent years have seen the emergence of *data trusts* as structured governance frameworks designed to manage data responsibly, anchored in fiduciary duties and explicit oversight (O'Hara 2019, Delacroix & Lawrence 2019). These parallel advancements in

20th International Conference on Wirtschaftsinformatik
September 2025, Münster, Germany

AI capabilities and data governance illustrate the promise and complexity of integrating AI into fiduciary stewardship frameworks.

Merging AI with data trusts raises ethical, legal, and operational challenges. Many AI systems are treated as "black boxes," creating opacity that can undermine stakeholder confidence (Mittelstadt et al. 2016, Bauer et al. 2021). In data trusts, where fiduciary principles of loyalty, impartiality, and prudence are central, such opacity intensifies the tension between efficiency and accountability. Traditional one-time consent models further compound the issue, as data contributors may have dynamic preferences or require ongoing protection (Jussupow et al. 2024). Moreover, entrusting AI agents with fiduciary responsibilities introduces risks such as misaligned incentives and potential conflicts of interest if AI optimizes for objectives inconsistent with beneficiary welfare. These tensions underscore a gap in IS research on how AI might support or impair the obligations of data trusts. In response, this paper investigates how fiduciary principles can be operationalized in data trusts that employ AI to perform critical governance tasks and under which conditions AI agents might not be suitable replacements for human fiduciaries. By grounding the study in a normative framework that emphasizes *fairness*, *accountability*, *transparency*, and *autonomy*, we explore how specific trustee roles could plausibly be automated with current and near-future AI technologies. We likewise consider scenarios in which fiduciary obligations demand heightened safeguards or human oversight, mainly if decisions affect vulnerable groups or require nuanced judgment (Rouhani & Deters 2021).

Our work builds on scholarship in fiduciary AI (Benthall & Shekman 2023) and data trust architectures (O'Hara 2019, Stachon et al. 2023), offering a synthesis of design requirements for integrating AI into data trusts without undermining trust or equity. Throughout the paper we use the term *AI agent* to mean an autonomous software agent (e.g., an Large Language Model (LLM)-based service) that carries out one of the governance roles listed in Table 1 on behalf of a legally responsible human or organization. Beyond proposing normative guidelines, we develop a reference approach for implementing AI agents as governance agents, detailing how transparency mechanisms, accountability checks, and beneficiary-centric rule sets can preserve fiduciary duties. Our contribution aims to support the responsible use of AI in IS, ensuring that AI-driven data governance remains aligned with societal values. This goal resonates with broader discussions on AI fairness, security, and human-centered design in the face of transformative technologies (De Silva et al. 2024).

Following the Design Science Research using Design Echelons (^{e}DSR) methodology of Tuunanen et al. (2024), we delimit the present study to the first two design echelons: *Problem Analysis* and *Objectives & Requirements Definition*. Accordingly, our design-knowledge contribution is twofold: (i) a *validated problem statement* that specifies the fiduciary tension created when AI agents assume governance roles in data trusts, and (ii) a coherent, feasible, and complete set of *validated design objectives*, instantiated in the four Design Principle (DP) developed in Section 3.2. By stopping at these echelons we avoid premature claims about artefact efficacy and instead lay a rigorously validated foundation for later Design & Development work (Tuunanen et al. 2024).

The remainder of this paper is structured as follows: Section 2 introduces key literature on data trusts, fiduciary responsibilities, and AI governance, detailing relevant actors and design considerations. Section 3 proposes a design theory that frames how AI agents in data trusts can uphold fiduciary duties; this includes our kernel theories (Section 3.1) and derived DPs (Section 3.2). Finally, Section 4 discusses the implications of these findings, addresses limitations, and outlines avenues for future IS research on responsibly integrating AI in data governance.

2 Foundational Concepts

Data trusts form an emerging governance framework, establishing fiduciary duties and transparent oversight for responsible data management. These structures define roles and responsibilities for stakeholders, ensuring that data is collected, shared, and utilized with explicit accountability. Section 2.1 overviews data trust's key actors and legal foundations, highlighting how fiduciary principles safeguard data owners and beneficiaries. Section 2.2 examines AI as an autonomous actor within these trusts, focusing on the promises and challenges of integrating large-scale predictive models and decision-support systems into a traditionally human-driven governance context.

2.1 Data Trusts and Trust Actors

Data has become an increasingly valuable asset, yet concerns about its misuse persist (Lomotey et al. 2022). To address these concerns, trust is essential (Stachon et al. 2023), which has led to the development of *data trusts* as a legal and organizational framework for responsible data management (Lomotey et al. 2022). We distinguish between trust in data–which concerns the reliability and accuracy of data (Mayer et al. 1995)–and data trusts as structures that manage data. A data trust typically involves several key actors (c.f. Figure 1): In this context, a *data trustee* refers to a single organization or entity acting as a trustee to manage and safeguard data. Note that in some literature, the singular term *data trust* is used instead of *data trustee*.

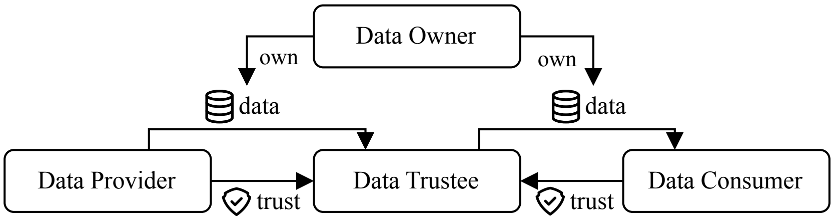

Figure 1. Abstract model illustrating key data trust actors and their interactions.

The *data provider* is responsible for supplying the data, often collected from various sources or devices. The *data consumer* is the entity that accesses and utilizes the data according to established policies (Ayappane et al. 2024). The *data owner* holds the legal rights and authority over the data and is tasked with setting the guidelines for data sharing and use (Rouhani & Deters 2021). In addition to the main roles, the broader data trust ecosystem includes several other roles. These roles can be roughly categorized according to data origin, governance, and usage. Table 1 briefly overviews these actors and their respective responsibilities. While data trusts function as a governance framework independent of artificial intelligence, this paper investigates the implications of integrating AI agents into their operational and decision-making processes.

Table 1. Concise descriptions of data trust roles.

	Role	Description
Origin	**Collector**	Gathers data, ensuring the rights of individuals (Rouhani & Deters 2021).
	Provider	Supplies and shares data under agreed terms (Stachon et al. 2023).
	Subject	Produces data through actions or behaviors (a.k.a. data producer, data generator) (Rouhani & Deters 2021, Delacroix & Lawrence 2019).
Governance	**Controller**	Determines purposes and means of processing personal data (Delacroix & Lawrence 2019).
	Custodian	Manages policy-driven consent for collectively owned data (a.k.a. data principal) (Ayappane et al. 2024).
	Owner	Holds legal rights over data and sets sharing policies (Rouhani & Deters 2021).
	Processor	Processes data on behalf of the data controller (Lomotey et al. 2022).
	Trustee	Manages or shares data on behalf of others as a trusted third party (a.k.a. data trust) (Lomotey et al. 2022).
Usage	**Consumer**	Requests or receives data for analysis or other purposes, acts as the beneficiary (a.k.a. data requester, data beneficiary) (Ayappane et al. 2024).
	User	Actively interacts with, analyzes, and transforms raw data to generate insights rather than simply receiving it like the data consumer (Stachon et al. 2023).

2.2 AI and Autonomous Agents

AI is broadly categorized into narrow, or *weak*, AI and general, or *strong*, AI (Ng & Leung 2020). Narrow AI systems excel at constrained tasks, such as image classification or game playing, but remain limited when faced with tasks beyond their specialized training (Welsh 2019). By contrast, general AI—often called Artificial General Intelligence (AGI)—aims to replicate human-level reasoning across multiple domains (Long & Cotner 2019). Despite decades of research, AGI has not yet been realized, with current commercial applications firmly within narrow AI (Welsh 2019, Ng & Leung 2020).

Recent AI advancements increasingly employ *generative* and open-ended capabilities (Naik et al. 2024, Balasubramaniam et al. 2024). These systems perform creative tasks flexibly across applications and are often integrated into IS, connecting diverse stakeholders. Generative technologies, such as LLMs, represent the latest progress in narrow AI (Feuerriegel et al. 2024). Trained on massive text corpora, they mimic human linguistic patterns, handling tasks involving text generation, summarization, or conversational interaction (Balasubramaniam et al. 2024). Generative architectures also extend beyond language, indicating broader universal content-creation methods. However, current LLMs do not yet meet accepted benchmarks for AGI and exhibit known shortcomings, such as inaccuracies and *hallucinations* (Naik et al. 2024, Harrer 2023), which impacts their performance when used in AI agents (Wolters et al. 2024).

Recent developments in *reasoning* models, like *ChatGPT O3*, *DeepSeek-R1*, or *Claude 3.7*, explore compute-intensive inference for enhanced problem-solving (Naik et al. 2024, Balasubramaniam et al. 2024). Vendors propose these as stepping stones toward AGI, though broad consensus remains elusive (Long & Cotner 2019, De Silva et al. 2024). Practically, these models remain categorized as narrow AI, despite advanced capabilities (Welsh 2019). Nevertheless, large-scale generative systems can function as *autonomous agents*, handling traditionally human-led tasks like data annotation, content moderation, or preliminary decision support (De Silva et al. 2024), which bears potential for improving overall organizational performance (Müller et al. 2020, Hertel et al. 2019). When embedded in data trusts (see Section 2.1), they offer efficiency in data governance but introduce accountability and ethical oversight considerations, particularly in nuanced tasks previously reliant on human judgment.

In data trusts, AI agents may perform roles in data curation, automated compliance checks, and decision-making support, raising significant questions about fairness, transparency, and human autonomy. Integrating AI into data-intensive processes increases demands for fairness, transparency, and security (O'Hara 2019, Berente et al. 2021), especially when decisions carry societal impacts. Effective governance must balance AI 's performance gains with values of trust and explainability (Miller 2019). Thus, clearly defined fiduciary roles and responsibilities should guide AI deployments in data trusts to safeguard stakeholder interests while leveraging AI 's transformative potential.

3 Design Theory for AI Agents in Data Trusts

Integrating AI agents, software agents that enact specific governance roles (cf. Table 1), into data trusts reconfigures the governance landscape, introducing opportunities and risks for fiduciary oversight, ethical compliance, and institutional legitimacy. AI agents, autonomous systems tasked with trust operations, offer efficiency gains in data curation, compliance verification, and policy enforcement. Their inclusion, however, requires a reconceptualization of fiduciary responsibility. These agents are not independent but embedded within a structured fiduciary relationship. Power asymmetries arise as AI agents may control critical data governance functions while other agents rely on them without fully understanding or specifying their actions (Mittelstadt et al. 2016, Zuboff 2023). Unlike human actors, AI agents lack legal personhood, independent ethical agency, and the capacity for discretionary judgment. Thus, AI agents risk perpetuating the opacity and imbalances that data trusts seek to resolve. Consequently, their deployment demands a design framework that aligns AI operations with fiduciary principles while maintaining human oversight, contestability, and institutional accountability.

To ensure AI agents function within the fiduciary mandates of data trusts, we propose a design theory that integrates insights from data trust governance (Delacroix & Lawrence 2019), institutional trust (Mayer et al. 1995, McKnight et al. 2002), and AI ethics (Floridi & Taddeo 2016) as our kernel theories (Gregor & Hevner 2013). In Section 3.1, we develop a normative framework based on fairness, accountability, transparency, and autonomy that articulates how fiduciary obligations can be fulfilled in data trusts. Building on that framework, Section 3.2 derives four DPs for AI agents, addressing fiduciary alignment, accountability, explainability, and autonomy-preserving oversight. Importantly, these principles do not map one-to-one onto the normative points: whereas normative considerations such as fairness are woven throughout several principles, DP 1 (*Fiduciary alignment and beneficiary-first governance*, cf. Section 3.2), for instance, emphasizes loyalty and prudence toward beneficiaries rather than mirroring fairness alone. This distinction underscores that the broader ethical values in Section 3.1 provide the conceptual foundation, while the subsequent DPs translate these values into actionable safeguards for AI-enabled data trust governance.

3.1 Kernel Theories: Fiduciary Governance, Institutional Trust, and AI Ethics

Data trusts represent an institutional response to data stewardships fiduciary and ethical challenges (Hardjono et al. 2019). Unlike conventional governance frameworks, which often prioritize commercial interests or efficiency, data trusts operate under explicit fiduciary duties of loyalty, prudence, impartiality, and accountability (Delacroix & Lawrence 2019, Hickman & Petrin 2021). Trustees must act exclusively in the best interests of beneficiaries, maintaining enforceable obligations even when AI systems are introduced into governance processes (Zygmuntowski et al. 2021).

Incorporating AI into data trusts heightens rather than diminishes fiduciary responsibilities. AI agents introduce risks of opacity and accountability deficits if not adequately constrained (Van Der Sloot & Keymolen 2022). Institutional trust theory highlights that stakeholder confidence depends fundamentally on perceived competence, integrity, and benevolence (Mayer et al. 1995). Thus, AI systems embedded within data trusts must be transparent, comprehensible, and aligned strictly with fiduciary standards. To achieve this alignment, governance mechanisms must enable stakeholders to audit, contest, and override AI-driven decisions, thereby preserving institutional legitimacy and trust. While AI ethics literature broadly emphasizes fairness, explainability, and human oversight (Floridi & Taddeo 2016, Jobin et al. 2019), data trusts demand higher accountability standards than those typically found in corporate AI applications. Whereas commercial AI prioritizes profit-driven metrics (Floridi & Taddeo 2016), fiduciary AI governance must actively prevent conflicts of interest, uphold transparency, and guarantee continuous oversight consistent with fiduciary obligations.

A set of normative principles further clarifies how fiduciary duties translate into practical governance structures within data trusts. Precisely, fairness, accountability, transparency, and autonomy form an integrated socio-technical framework that balances individual rights with collective governance. We treat these concepts as *binding fiduciary constraints*, rather than aspirational ideals. *Fairness* demands equitable distribution of risks and benefits, explicitly mitigating power asymmetries and discriminatory practices (Rawls 1971, Rouhani & Deters 2021). *Accountability* ensures trustees adhere to fiduciary obligations through enforceable oversight, contestability, and systematic audits (Delacroix & Lawrence 2019). *Transparency* reinforces fairness and accountability, providing stakeholders with clear governance rules, verifiable records, and understandable explanations of AI decision-making processes (O'Hara 2019, Van Der Sloot & Keymolen 2022). Finally, *autonomy* safeguards stakeholder agency through participatory governance structures that adapt continuously to evolving preferences and interests, ensuring data subjects retain meaningful control (Zygmuntowski et al. 2021). Figure 2 depicts the interrelationships between these normative principles.

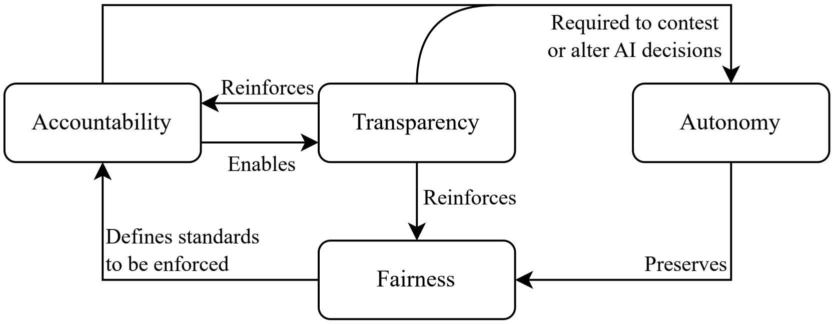

Figure 2. Normative principles for practical governance structures.

Each principle should follow a context-specific guardrail (Tingelhoff & Marga 2025) to retain its fiduciary force while avoiding threats that can arise from indiscriminate or absolutist applications (Bannister & Connolly 2011, Koivisto 2022, Nguyen 2022, Schade 2023): selective, audit-linked transparency; fairness evaluated against concrete risk-benefit profiles; accountability enforced through immutable logs; and autonomy sustained by dynamic, revocable consent. Further, we embed their normative force in verifiable mechanisms since principles-only ethics can lead to "ethics-washing" (Munn 2023, Maclure & Morin-Martel 2025, Stamboliev & and Christiaens 2025). Our DPs convert each constraint into concrete system requirements, while routine fiduciary audits, human-override rights, and enforceable sanction paths anchor them in the data-trusts legal-institutional framework.

3.2 Design Principles for AI Agents in Data Trusts

Autonomous AI intensifies the ethical duties outlined in 3.1. Large-scale optimisation can entrench grouplevel bias (Mittelstadt et al. 2016, Bauer et al. 2021); the non-personhood of software agents blurs accountability (Benthall & Shekman 2023); model opacity obstructs meaningful disclosure (Miller 2019); and continuous, context-aware inference renders one-shot consent ineffective (Kaye et al. 2015). In response, the four DPs that follow operationalise fairness, accountability, transparency, and autonomy for data-trust governance, providing actionable guidance for the governance roles introduced in Section 2.1. Each principle aims to ensure that AI-enabled agents–especially the controller, owner, custodian, processor, and trustee–actively uphold the fiduciary obligations and institutional trust requirements in a data trust. In particular, we emphasize human-AI interaction and risk mitigation measures to ensure appropriate reliance on automated decisions and prevent overreliance or misuse. While the ethical framework defines what is to be protected or prioritized (for instance, beneficiary interests or equitable data usage), these DPs clarify how such obligations can be realized through tangible design features in systems and processes. By concentrating on the governance layer, they emphasize the safeguards and oversight structures needed to align automated decisions with the core duties of loyalty, prudence, and transparency, thereby maintaining human autonomy and reinforcing trust in AI-mediated data stewardship.

DP 1: Fiduciary alignment and beneficiary-first governance. All AI agents in data trusts generally encompassing the data controller, custodian, owner, processor, and trustee must be explicitly designed and institutionally constrained to serve the best interests of beneficiaries (Delacroix & Lawrence 2019). Beyond commercial logics that optimize for engagement or efficiency, these agents must follow legal obligations of loyalty and prudence, ensuring data is used solely for the designated, beneficiary-aligned purposes (Zygmuntowski et al. 2021). This requirement shields data subjects from conflicts of interest that could otherwise arise if AI systems pursued external objectives (Benthall & Shekman 2023). Effective enforcement of fiduciary alignment relies on technical and institutional safeguards. Automated fiduciary audits should confirm whether each AI-based decision upholds trust policies, flagging deviations for human intervention. Trustee oversight mechanisms allow continuous monitoring and the authority to override AI-generated outputs when they breach fiduciary obligations, thereby preserving ethical and legal integrity (Binns 2018).

Integrating formal verification methodologies, such as Linear Temporal Logic specifications, facilitates continuous checking of AI behavior against predefined beneficiary-aligned rules. Explicitly modeling security requirements and treating fiduciary duties as *hard constraints* prevents unauthorized optimization for outside interests (Hayashi et al. 2023). Embedding these constraints in the AI's planning process helps each actor, from the data controller who orchestrates policies to the trustee who oversees enforcement, generate decisions that remain strictly within the scope of legal and ethical mandates. In turn, data owners and custodians can scrutinize the logic behind each decision, escalating problematic cases to human review or requiring AI reconfiguration, ensuring that a beneficiary-first approach remains the guiding principle (Hayashi et al. 2023).

DP 2: Accountability through traceability and oversight. Because AI lacks legal personhood, it cannot be held directly accountable for governance failures (Mittelstadt et al. 2016), necessitating institutional mechanisms that render all AI-driven decisions traceable and contestable within a data trust. The risk of opaque, unchallengeable governance seriously threatens institutional trust when such accountability protocols are absent (Van Der Sloot & Keymolen 2022). Ensuring robust traceability requires precise documentation of AI decisions, including model versioning records, complete audit trails, and immutable logs. For instance, the data trustee can coordinate these records across all trust agents' data controllers, custodians, owners, and processors while organizing third-party audits to verify alignment with fiduciary standards.

Embedding continuous validation and verification techniques further enhances this accountability. Integrating Machine Learning (ML) algorithms for data checks and anomaly detection can reinforce trust by validating the reliability of inputs and outputs in real time (Fast et al. 2023). Formal argumentation frameworks similarly allow retrospective examination of reasoning processes, illuminating how or why a specific action was taken (Yu et al. 2022). Combining these approaches with distributed ledger technologies provides immutable, time-stamped decision logs, mirroring the human practice of thorough recordkeeping and ensuring each decision remains open to retrospective scrutiny (Gkogkos et al. 2024).

DP 3: Transparent explainability and stakeholder visibility. While human trustees can be scrutinized through open dialogue, AI systems risk being viewed as "black boxes" unless designed for robust algorithmic transparency (Bauer et al. 2021). Embedding explainability ensures that data subjects, controllers, custodians, owners, and trustees can collectively interpret, evaluate, and contest AI-generated outcomes, thus safeguarding the accountability of automated governance (Dehling & Sunyaev 2024, Benthall & Shekman 2023). Centralizing public documentation of AI governance models and regularly updated transparency dashboards promotes external validation and maintains institutional trust. In real time, stakeholders benefit from viewing these dashboards, tracking how key decisions are reached, challenged, or revised (Miller 2019, Arunika et al. 2024).

Methods from Explainable AI (XAI), such as Local Interpretable Model-agnostic Explanations or SHapley Additive exPlanations, can illuminate the reasoning of complex "black box" systems by producing interpretable, human-friendly explanations. However, the community widely agrees that post-hoc techniques remain approximations when applied to modern architectures containing vast amounts of parameters and layers. Their outputs should therefore be treated as probabilistic signals rather than definitive causal accounts and be complemented by global artifacts like model cards and governance documentation, as well as local, case-specific explanations delivered at decision time. Emphasizing a human-centric process where end-users guide interface design helps align explanatory outputs with user mental models and fosters deeper engagement with AI-mediated decisions (Arunika et al. 2024). Transparency should also extend along the entire ML lifecycle, ensuring that data preparation, model development, deployment, and continuous monitoring are documented and verifiable via best practices in ML Operations (Jutte 2024).

DP 4: Autonomy-preserving governance and participatory oversight. Traditional consent-based models that rely on static, one-time permissions are often inadequate for AI-driven processes, given the information asymmetries and rapidly changing decision contexts they introduce (Jussupow et al. 2024). Preserving autonomy in such environments calls for continuous engagement with data owners, custodians, and trustees, allowing them to modify governance settings as new preferences or risks emerge. Rather than enforcing rigid, automated protocols, AI agents should enable the dynamic adjustment of rules, preventing beneficiary interests from being sidelined. Participatory oversight mechanisms, in turn, integrate stakeholder input so that governance decisions evolve collectively rather than merely reflecting automated logic (Floridi & Taddeo 2016). Built-in fiduciary safeguards also limit over-automation, ensuring that crucial governance functions, moral deliberation, conflict resolution, and exceptional approvals remain under human review.

When designed as intelligent collaborators, AI systems can bolster human oversight by analyzing large volumes of data and flagging potential autonomy infringements

(Chen et al. 2024). For example, an AI actor might detect patterns indicating that existing policies fail to honor nuanced user consent in specific contexts. Armed with these data-driven insights, human stakeholders such as data controllers, owners, and trustees can refine governance parameters or intervene in borderline cases. This synergy ensures that the overall data trust structure preserves individual and collective autonomy while leveraging the computational strength of AI for timely and informed participatory decision-making.

4 Discussion and Conclusion

Consistent with the ^{e}DSR methodology, our validation addresses only the two targeted echelons. At the *Problem Analysis* level, Sections 2-3 synthesise prior scholarship and practitioner concerns to establish the relevance and theoretical solvability of the fiduciary tension that arises when AI agents govern data trusts. At the *Objectives & Requirements* level, the normative framework (Fig. 2) logically derives four design principles and maps them to specific roles (Table 2), thus providing coherence, completeness, and an initial feasibility check. Taken together, these principles show how AI can yield efficiency gains in data curation, compliance verification, and policy enforcement without eroding stakeholder confidence or undermining the trusts ethical foundation. Fiduciary alignment embeds beneficiary interests directly into decision-making, mitigating hidden incentives for profit or efficiency optimisation (Floridi & Taddeo 2016), while accountability empowers stakeholders to audit and challenge AI decisions (Benthall & Shekman 2023). Transparency, elevated from procedural compliance to an element of institutional legitimacy, ensures that stakeholders can meaningfully interpret, scrutinise, and contest automated outcomes (Bauer et al. 2021, Binns 2018). Finally, autonomy is safeguarded through dynamic participatory consent frameworks that remedy the inadequacies of static consent in rapidly changing data contexts (Jussupow et al. 2024). Design & Development, Demonstration, and Evaluation remain future work.

Table 2 illustrates how data governance roles can integrate AI workflows while maintaining the human oversight necessary for ethically sensitive decisions. This role-based multiparty custodianship, where no single AI instance exerts unilateral authority over data decisions, mitigates opacity risks associated with AI automation (Van Der Sloot & Keymolen 2022) and ensures decisions remain contestable and transparent.

Our findings both align with and extend prior research on fiduciary AI (Benthall & Shekman 2023, Delacroix & Lawrence 2019), emphasizing that data trusts require enforceable governance structures beyond corporate guidelines alone (Mayer et al. 1995, McKnight et al. 2002). Integrating insights from AI ethics enables data trusts to maintain transparency and robust stakeholder protection. Furthermore, our principle of autonomy preservation addresses the inadequacy of static consent in dynamic AI environments (Jussupow et al. 2024). A participatory governance model, allowing trustees to review permissions dynamically, ensures data owners retain meaningful control over AI-mediated data usage.

Table 2. Suitability of Autonomous Governance Approaches.

	Role	DP(s)	Suitability (Autonomous Approach)
Fairness	**Controller**	1, 2	Detecting biases in data policies and logging decisions for auditing; human sign-off remains key for final changes.
	Custodian	1–3	Approving/denying requests flagged as discriminatory and offering concise rationales; periodic oversight resolves edge cases.
	Owner	1	Highlighting unfair usage for the human data owner; ownership stays with a human for fiduciary legitimacy.
	Processor	1–3	Applying bias checks (e.g., anomaly detection) during transformations; summarizing fairness results for expert review.
	Trustee	1, 2	Aggregating fairness logs from all roles; automated cross-checks are feasible; robust policy updates need humanmachine synergy.
Accountability	**Controller**	1–3	Maintaining tamper-evident logs and auto-checking compliance; summarizing outcomes for quick stakeholder review.
	Custodian	1–3	Validating consent, flagging breaches in real time; complex disputes typically need a human mediator.
	Owner	1, 2, 4	Extending the owners fiduciary role by monitoring suspicious usage; the owner can override or escalate anytime.
	Processor	1–3	Logging each processing step and verifying rule compliance, summarizing audits, and major exceptions require human input.
	Trustee	1–4	Consolidating accountability metrics from all roles; highlighting cross-role conflicts; final enforcement is human-led.
Transparency	**Controller**	2, 3	Publish real-time rationales for data use and keep logs open for external audits.
	Custodian	2–4	Maintain dashboards for who accessed what and why. Anomaly detection reveals suspicious actions, and human review remains crucial.
	Owner	2–4	Providing interactive usage reports and alerts if conflicts arise. The owner sets transparency thresholds, and the AI agent highlights issues.
	Processor	2, 3	Verifying each transformation with a chain of evidence; XAI modules clarify derivations; domain experts confirm correctness.
	Trustee	2–4	Centralizing logs and producing global dashboards; correlating records across roles; user feedback refines clarity.
Autonomy	**Controller**	2, 4	Managing dynamic permissions (opt-in/opt-out); excessive automation may limit user freedoms, so human sign-off is essential.
	Custodian	2, 4	Monitoring consent status and updating sharing rules; noting repeated overrides to rule out coercion.
	Owner	2, 4	Human-led role; surfacing anomalies or potential conflicts; cannot override final owner decisions.
	Processor	2, 4	Respecting consent changes mid-process; useful for routine updates; complex revocations need human input.
	Trustee	2, 4	Overseeing consent changes from all stakeholders; highlighting autonomy conflicts; major revisions demand human resolution.

The proposed design theory yields practical recommendations. First, embedding fiduciary logic directly into AI decision pipelines is critical to preventing conflicts of interest, particularly in data transformation roles. Second, transparency and continuous auditing should be considered core governance components, not optional features, allowing stakeholders to systematically scrutinize AI decisions. Third, preserving autonomy requires replacing one-time consent approaches with participatory mechanisms, empowering stakeholders to correct or override AI outcomes in real time if ethical or legal issues arise. Although narrow AI can automate many routine tasks, moral deliberation and ultimate fiduciary responsibility must remain human-driven, especially in roles such as data ownership that require nuanced judgment and involve potential legal liabilities.

While DP 3 promotes transparency using XAI approaches, these tools are computationally expensive and provide only *local* fidelity for highly complex models and tasks. In safety or rights-critical settings e.g. escalate to human actors whenever the confidence interval of explanations exceeds a pre-defined threshold Several more limitations impact the generalizability of these findings. Our framework primarily addresses narrow AI, restricting its direct applicability if future advancements significantly extend AI capabilities toward general intelligence (Naik et al. 2024). Additionally, jurisdiction-specific regulations may constrain fiduciary duty enforcement and accountability mechanisms, necessitating local adaptations of the DPs. Moreover, empirical validation through pilot implementations and standardized metrics (e.g., fiduciary alignment scores) is essential to confirm the practical efficacy of the proposed principles (Hayashi et al. 2023). Differences in trust architectures, stakeholder capacities, and industry-specific regulatory contexts could further influence AI integration effectiveness within data trusts.

Despite these limitations, the primary implication of our research is that AI-driven automation need not compromise the core ethos of data trust. By embedding beneficiary-first logic, ensuring transparent oversight, and maintaining stakeholder autonomy, data trusts can leverage AI capabilities without abandoning fiduciary obligations to data subjects. These findings enhance our broader understanding of the intersection between fiduciary obligations and AI ethics, offering a practical roadmap for designing AI-mediated data governance that rigorously upholds loyalty, prudence, impartiality, and accountability. Future research should, therefore, explore domain-specific applications, investigate compliance strategies across diverse legal jurisdictions, and refine continuous monitoring methods for assessing AI fiduciary alignment. Such efforts can guide data trusts toward becoming adaptive, trustworthy governance structures capable of responsibly embedding AI. Our results confirm that data trusts can integrate AI without compromising the core tasks of trusteeship by embedding safeguards strictly aligned with the beneficiaries. The framework of our four DPs illustrates how beneficiaries interests and data owners rights are safeguarded through multilateral trusteeship and accountable decision-making processes.

Acknowledgements

This research has been supported by the Federal Ministry of Research, Technology and Space (BMFTR), Germany under Research Grant No. 16DHBKI039, and Research Grant No. 16DTM218 as part of the NextGenerationEU program of the European Union.

References

Arunika, M., Saranya, S., Charulekha, S., Kabilarajan, S. & Kesavan, G. (2024), A Survey on Explainable AI Using Machine Learning Algorithms Shap and Lime, *in* 'Int. Conf. Comput. Commun. Netw. Technol., ICCCNT', Institute of Electrical and Electronics Engineers Inc.

Ayappane, B., Vaidyanathan, R., Srinivasa, S., Upadhyaya, S. K. & Vivek, S. (2024), Consent Service Architecture for Policy-Based Consent Management in Data Trusts, *in* 'Proceedings of the 7th Joint International Conference on Data Science & Management of Data (11th ACM IKDD CODS and 29th COMAD)', ACM, Bangalore India, pp. 155–163.

Balasubramaniam, S., Kadry, S., Prasanth, A. & Dhanaraj, R. (2024), *Generative AI and LLMs: Natural Language Processing and Generative Adversarial Networks*, Generative AI and LLMs: Natural Language Processing and Generative Adversarial Networks, De Gruyter.

Bannister, F. & Connolly, R. (2011), 'The Trouble with Transparency: A Critical Review of Openness in e-Government', *Policy & Internet* **3**(1), 1–30.

Bauer, K., Hinz, O., van der Aalst, W. & Weinhardt, C. (2021), 'Expl(AI)n It to Me – Explainable AI and Information Systems Research', *Bus Inf Syst Eng* **63**(2), 79–82.

Benthall, S. & Shekman, D. (2023), Designing Fiduciary Artificial Intelligence, *in* 'Equity and Access in Algorithms, Mechanisms, and Optimization', ACM, Boston MA USA, pp. 1–15.

Berente, N., Gu, B., Recker, J. & Santhanam, R. (2021), 'Managing Artificial Intelligence', *MIS Quarterly* **45**, 1433–1450.

Binns, R. (2018), 'Algorithmic Accountability and Public Reason', *Philosophy & Technology* **31**(4), 543–556.

Chen, Y.-C., Liu, H. & Wang, Y.-F. (2024), Governance Design of Collaborative Intelligence for Public Policy and Services, *in* Liao H.-C., Cid D.D., Macadar M.A. & Bernardini F., eds, 'ACM Int. Conf. Proc. Ser.', Association for Computing Machinery, pp. 146–155.

De Silva, D., Kaynak, O., El-Ayoubi, M., Mills, N., Alahakoon, D. & Manic, M. (2024), 'Opportunities and Challenges of Generative Artificial Intelligence: Research, Education, Industry Engagement, and Social Impact', *IEEE Ind. Electron. Mag.* .

Dehling, T. & Sunyaev, A. (2024), 'A Design Theory for Transparency of Information Privacy Practices', *Information Systems Research* **35**(3), 956–977.

Delacroix, S. & Lawrence, N. D. (2019), 'Bottom-up data Trusts: Disturbing the 'one size fits all' approach to data governance', *International Data Privacy Law* p. ipz014.

Fast, V., Schnurr, D. & Wohlfarth, M. (2023), 'Regulation of data-driven market power in the digital economy: Business value creation and competitive advantages from big data', *Journal of Information Technology* **38**(2), 202–229.

Feuerriegel, S., Hartmann, J., Janiesch, C. & Zschech, P. (2024), 'Generative AI', *Bus Inf Syst Eng* **66**(1), 111–126.

Floridi, L. & Taddeo, M. (2016), 'What is data ethics?', *Phil. Trans. R. Soc. A.* **374**(2083), 20160360.

Gkogkos, G., Giakoumoglou, N., Pechlivani, E., Votis, K. & Tzovaras, D. (2024), Artificial Intelligence Data Model Verification through Distributed Ledger Technology, *in* 'Int. Conf. Inf. Technol., IT', Institute of Electrical and Electronics Engineers Inc.

Gregor, S. & Hevner, A. (2013), 'Positioning and Presenting Design Science Research for Maximum Impact', *MIS Quarterly* **37**, 337–356.

Hardjono, T., Shrier, D. L. & Pentland, A., eds (2019), *Trusted Data: A New Framework for Identity and Data Sharing*, 2 edn, The MIT Press.

Harrer, S. (2023), 'Attention is not all you need: The complicated case of ethically using large language models in healthcare and medicine', *eBioMedicine* **90**.

Hayashi, H., Mitsikas, T., Taheri, Y., Tsushima, K., Schäfermeier, R., Bourgne, G., Ganascia, J.-G., Paschke, A. & Satoh, K. (2023), Multi-agent Online Planning Architecture for Real-time Compliance, *in* Vanthienen J., Kliegr T., Fodor P., Lanti D., Arndt D., Kostylev E.V., Mitsikas T. & Soylu A., eds, 'CEUR Workshop Proc.', Vol. 3485, CEUR-WS.

Hertel, G., MeeSSen, S. M., Riehle, D. M., Thielsch, M. T., Nohe, C. & and, J. B. (2019), 'Directed forgetting in organisations: the positive effects of decision support systems on mental resources and well-being', *Ergonomics* **62**(5), 597–611. PMID: 30698075.

Hickman, E. & Petrin, M. (2021), 'Trustworthy AI and Corporate Governance: The EU's Ethics Guidelines for Trustworthy Artificial Intelligence from a Company Law Perspective', *European Business Organization Law Review* **22**(4), 593–625.

Jobin, A., Ienca, M. & Vayena, E. (2019), 'The global landscape of AI ethics guidelines', *Nat Mach Intell* **1**(9), 389–399.

Jussupow, E., Benbasat, I. & Heinzl, A. (2024), 'An Integrative Perspective on Algorithm Aversion and Appreciation in Decision-Making', *MISQ* **48**(4), 1575–1590.

Jutte, A. (2024), Explainable MLOps: A Methodological Framework for the Development of Explainable AI in Practice, *in* Longo L., Liu W. & Montavon G., eds, 'CEUR Workshop Proc.', Vol. 3793, CEUR-WS, pp. 385–392.

Kaye, J., Whitley, E. A., Lund, D., Morrison, M., Teare, H. & Melham, K. (2015), 'Dynamic consent: A patient interface for twenty-first century research networks', *Eur J Hum Genet* **23**(2), 141–146.

Koivisto, I. (2022), *The Transparency Paradox*, 1 edn, Oxford University PressOxford.

Lomotey, R. K., Kumi, S. & Deters, R. (2022), 'Data Trusts as a Service: Providing a platform for multi-party data sharing', *International Journal of Information Management Data Insights* **2**(1), 100075.

Long, L. & Cotner, C. (2019), A Review and Proposed Framework for Artificial General Intelligence, *in* 'IEEE Aerosp. Conf. Proc.', Vol. 2019-March, IEEE Computer Society.

Maclure, J. & Morin-Martel, A. (2025), 'AI Ethics' Institutional Turn', *Digit. Soc.* **4**(1), 18.

Mayer, R. C., Davis, J. H. & Schoorman, F. D. (1995), 'An Integrative Model of Organizational Trust', *The Academy of Management Review* **20**(3), 709–734.

McKnight, D. H., Choudhury, V. & Kacmar, C. (2002), 'Developing and Validating Trust Measures for e-Commerce: An Integrative Typology', *Information Systems Research* **13**(3), 334–359.

Miller, T. (2019), 'Explanation in artificial intelligence: Insights from the social sciences', *Artificial Intelligence* **267**, 1–38.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S. & Floridi, L. (2016), ‘The ethics of algorithms: Mapping the debate’, *Big Data & Society* **3**(2), 2053951716679679.

Müller, L. S., Meeßen, S. M., Thielsch, M. T., Nohe, C., Riehle, D. M. & Hertel, G. (2020), Do not disturb! trust in decision support systems improves work outcomes under certain conditions, *in* F. Alt, S. Schneegass & E. Hornecker, eds, ‘Mensch und Computer 2020 - Tagungsband, Magdebug, Germany, September 6-9, 2020’, ACM, pp. 229–237.
URL: *https://doi.org/10.1145/3404983.3405515*

Munn, L. (2023), ‘The uselessness of AI ethics’, *AI Ethics* **3**(3), 869–877.

Naik, D., Naik, I. & Naik, N. (2024), Imperfectly Perfect AI Chatbots: Limitations of Generative AI, Large Language Models and Large Multimodal Models, *in* N. Naik, P. Jenkins, S. Prajapat & P. Grace, eds, ‘Contributions Presented at The International Conference on Computing, Communication, Cybersecurity and AI, July 3–4, 2024, London, UK’, Vol. 884, Springer Nature Switzerland, Cham, pp. 43–66.

Ng, G. & Leung, W. (2020), ‘Strong Artificial Intelligence and Consciousness’, *Journal of Artificial Intelligence and Consciousness* **7**(1), 63–72.

Nguyen, C. T. (2022), ‘Transparency is Surveillance’, *Philosophy and Phenomenological Research* **105**(2), 331–361.

O’Hara, K. (2019), Data Trusts: Ethics, Architecture and Governance for Trustworthy Data Stewardship, PhD thesis, University of Southampton.

Rawls, J. (1971), *A Theory of Justice*, Harvard University Press, Cambridge, MA.

Rouhani, S. & Deters, R. (2021), ‘Data Trust Framework Using Blockchain Technology and Adaptive Transaction Validation’, *IEEE Access* **9**, 90379–90391.

Schade, F. (2023), ‘Dark Sides of Data Transparency: Organized Immaturity After GDPR?’, *Business Ethics Quarterly* **33**(3), 473–501.

Stachon, M., Möller, F., Guggenberger, T., Tomczyk, M. & Henning, J.-L. (2023), Understanding Data Trusts., *in* ‘ECIS’.

Stamboliev, E. & and Christiaens, T. (2025), ‘How empty is Trustworthy AI? A discourse analysis of the Ethics Guidelines of Trustworthy AI’, *Critical Policy Studies* **19**(1), 39–56.

Tingelhoff, F. & Marga, J. (2025), ‘Avoiding virtual dystopia: A design theory for emancipatory participatory immersive platforms’, *The Journal of Strategic Information Systems* **34**, 101910.

Tuunanen, T., Winter, R. & Vom Brocke, J. (2024), ‘Dealing with Complexity in Design Science Research: A Methodology Using Design Echelons’, *MIS Quarterly* **48**(2), 427–458.

Van Der Sloot, B. & Keymolen, E. (2022), ‘Can we trust trust-based data governance models?’, *Data. Policy.* **4**(2).

Welsh, R. (2019), ‘Defining artificial intelligence’, *SMPTE Motion Imaging J.* **128**(1), 26–32.

Wolters, A., von Straussenburg, A. F. A. & Riehle, D. M. (2024), Evaluation framework for large language model-based conversational agents, *in* T. Q. Phan, B. C. Y. Tan, H. Le, N. H. Thuan, M. Chau & K. Y. Goh, eds, ‘28th Pacific Asia Conference on Information Systems, PACIS 2024, Ho Chi Minh City, Vietnam, July 1-5, 2024’.
URL: *https://aisel.aisnet.org/pacis2024/track01_aibussoc/track01_aibussoc/14*

Yu, L., Zichichi, M., Markovich, R. & Najjar, A. (2022), Intelligent Human-input-based Blockchain Oracle (IHiBO), *in* Rocha A., Steels L. & van den Herik J., eds, 'Int. Conf. Agent. Artif. Intell.', Vol. 1, Science and Technology Publications, Lda, pp. 515–526.

Zuboff, S. (2023), The Age of Surveillance Capitalism, *in* 'Social Theory Re-Wired', 3 edn, Routledge.

Zygmuntowski, J. J., Zoboli, L. & Nemitz, P. F. (2021), 'Embedding European values in data governance: A case for public data commons', *Internet Policy Review* **10**(3).