Bus Inf Syst Eng 68(5):971–976 (2026)
https://doi.org/10.1007/s12599-026-01017-w

EDITORIAL

Value-Centric AI – Designing Sociotechnical Systems Before Legitimacy Fails

Alexander Richter · Jella Pfeiffer · Oliver Hinz · Hajo Reijers

Published online: 17 August 2026

1 Introduction

Marshallese stick charts are traditional navigational knowledge tools made from wooden or plant strips and shells (for an example see Fig. 1). Different chart types represented swell patterns, islands, and routes; many were studied and memorised on land rather than carried to sea. During a voyage, navigation depended on situated experience, sensory judgement, and responsibility for the course taken.

What can we learn from Marshallese stick charts for discussing artificial intelligence (AI)?

AI has become increasingly capable of identifying patterns, generating text, producing code, synthesising evidence, and recommending action. Yet its value does not emerge from computational output alone. It emerges through human interpretation, organisational use, and responsible action in context, all of which are shaped by cultural, ethical, and organisational values. When AI is understood as part of a sociotechnical system, an important challenge is not only how to keep humans in the loop, but to keep them in practice: exercising interpretation, contestation, ethical reasoning, and accountability.

For the Business and Information Systems Engineering (BISE) community, this is a defining moment. Beyond asking what tasks AI can perform, and how well, we need to design AI-enabled sociotechnical systems that remain aligned with cultural and organisational values.

2 From Tool Logic to Collaborator Logic

With the introduction of generative AI alongside predictive AI, the language used in organisations has shifted quickly. AI has been increasingly described as a "copilot", "assistant", or "agent" (Baird and Maruping 2021; Cranefield et al. 2023). Such framing can be useful because it acknowledges that AI systems increasingly shape work rather than merely execute commands. Yet, it also obscures an important difference between human and algorithmic participation. Human teammates can form intentions, negotiate commitments, and be held accountable through socially recognised relations. AI can influence team outcomes and can be considered an integral part of teams (Richter and Schwabe 2025), but it does not enter responsibility relations in the way humans do (Floridi and Sanders 2004).

Five characteristics help explain how contemporary AI can shape values.

1. *Delegated agency* Choices that were once made explicitly by humans may now be framed, ranked or narrowed by AI. Humans may retain formal authority,

A. Richter
The University of Auckland Business School, Auckland, New Zealand

J. Pfeiffer (✉)
Institute for Information Systems (WIN), Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
e-mail: pfeiffer@kit.edu

O. Hinz
Faculty of Economics and Business Administration, Goethe University Frankfurt, Frankfurt am Main, Germany
e-mail: ohinz@wiwi.uni-frankfurt.de

H. Reijers
Department of Information and Computing Sciences, Utrecht University, Utrecht, The Netherlands
e-mail: h.a.reijers@uu.nl

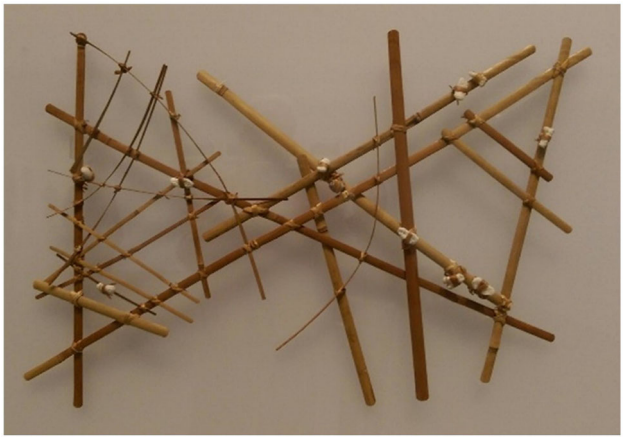

Fig. 1 A Marshallese stick chart. Photo: Jim Heaphy, Cullen328, CC BY-SA 3.0 https://creativecommons.org/licenses/by-sa/3.0, via Wikimedia Commons

while practical agency shifts earlier in the workflow (Faraj et al. 2018).

2. *Hidden AI* AI is widely accessible and has become part of many work systems without organisational authorisation or visibility. This resembles Shadow IT (Haag and Eckhardt 2017), but it changes the governance problem because machine-generated advice may enter decisions without being recorded as a source, or even without users recognising how extensively they relied on it.
3. *Opacity of values* Beyond technical opacity, users may not see the values embedded in how AI defines desirable work. For example, users might not realise when an AI implicitly prioritises efficiency over organisational values like fairness, privacy, or collaborative teamwork. A productivity assistant may implicitly privilege uninterrupted individual work over meetings and collaboration (Cranefield et al. 2023).
4. *Speed and concentration* AI operates at a scale and pace that allow both beneficial and harmful effects to travel rapidly across processes and institutions. This is intensified when many organisations depend on a small number of dominant models and infrastructures (Korinek and Vipra 2025).
5. *Feedback effects* AI-mediated decisions can create self-reinforcing loops. Outputs influence expectations, data, and later judgments, which may then be treated as evidence of the patterns the system initially produced (Ensign et al. 2018; Bauer et al. 2024; Glickman and Sharot 2025).

These characteristics make AI's influence on organisational values difficult to identify, contest, and govern. Their consequences become especially significant when AI moves beyond isolated tasks to shape priorities, workflows, and consequential judgments.

In hiring, machine learning systems may be developed as if they could replace expert judgment, yet their practical value depends on hybrid practices in which recruiters, developers, and data infrastructures learn from one another (van den Broek et al. 2021). In clinical decision support, AI advice can improve performance while also creating susceptibility to erroneous recommendations and automation bias (Goddard et al. 2014; Gaube et al. 2021), particularly for novice physicians (Jussupow et al. 2021). In each case, humans may remain formally accountable while the origin and framing of judgment become increasingly difficult to locate.

Immersive environments can intensify this collaborator logic. When AI is paired with spatial, embodied, or socially present interfaces, the AI's trustworthiness might increase further, although it is still unclear whether this translates into behaviour (Felnhofer et al. 2024). As interfaces become more compelling, it may become more difficult for users to maintain appropriate distance from recommendations, simulated presence, or agentic cues, making interaction design and governance design inseparable (Slater 2018; Richter et al. 2025). We use value-centric AI to refer to the design and governance of the entire sociotechnical system in which cultural, ethical, and organisational values are considered in shaping roles, work practices, and infrastructure. We elaborate on characteristics of contemporary AI that help explain recurring governance failures and propose steps for preserving responsible human practice in AI-enabled organisations.

3 Why Performance-Centric Governance is Insufficient

Organisations have been increasingly introducing model evaluations, oversight procedures, fairness metrics, ethical guardrails, risk classifications, and compliance routines. The EU AI Act, for example, makes human oversight, risk management, transparency, and governance obligations central to high-risk AI systems (European Parliament and Council 2024). Fairness metrics and bias audits can reveal discriminatory patterns (Mehrabi et al. 2021), and interpretability debates have sharpened the question of how AI systems should be designed for high-stakes decisions (Rudin 2019). Other debates have pointed to the trade-offs arising from contradicting design goals (Sunyaev et al. 2025).

These measures are necessary, but remain insufficient when governance is reduced to model performance, formal compliance, or nominal human oversight. A company might adopt an AI system that evaluates employees based on quantitative metrics (response times, collaboration scores from team software) instead of human judgment of managers and colleagues. A system can have strong model performance and still optimise the wrong proxy. A hiring

system may rank candidates by similarity to past successful employees when the organisational value at stake is also fair access to employment opportunities. Trade-offs between accuracy and fairness or privacy, for example, might be hard to quantify in an objective function (Sunyaev et al. 2025) and individual circumstances of applicants will hardly be used by AI to balance trade-offs. A clinical system may improve average diagnostic performance while weakening the professional judgment needed to challenge an incorrect recommendation. These examples illustrate failures of direction, interpretation, and accountability. In each case, core organisational values, whether these concern fair access to employment, the privacy of sensitive data, or the retention of human ethical responsibility, are compromised by a narrow focus on technical accuracy.

We conceptualise three recurring sociotechnical governance failures that can arise when AI shapes practical judgment: accountability drift, capability erosion, and legitimacy gaps.

3.1 Accountability Drift

As AI recommendations become embedded in workflows, responsibility can become diffuse. Users may overtrust, defer, or treat outputs as neutral facts, particularly when systems are framed as agentic. Interestingly, even in moral contexts, humans delegate the decision more often when they can delegate to AI rather than to a human counterpart (Hüholt and Szech 2026). Without deliberate role and boundary design, humans may remain accountable for outcomes over which their practical agency has diminished. Accountability cannot be repaired only at the end of the process. It must be designed into the distribution of roles, evidence, escalation rights, and audit trails.

3.2 Capability Erosion

If professionals routinely rely on AI to draft, classify, summarise, diagnose, or recommend, they may gradually stop practising the very judgment needed to challenge its outputs (Bauer et al. 2026). A nominal human-in-the-loop can then become an empty ritual. The human signs off, but the system has already performed the consequential framing. This concern is consistent with evidence that overreliance can persist even when users receive explanations, and that interventions may be needed to compel more deliberate engagement with AI advice (Buçinca et al. 2021).

3.3 Legitimacy Gaps

Accuracy does not equal legitimacy. Stakeholders may reasonably ask whether a decision was reached under acceptable rules, with appropriate regard for affected values, and by actors who can explain, defend, and revise it. Technical performance alone cannot answer these

Table 1 Suggested steps from a sociotechnical design stance

Step	Goal	Actions
1. Values articulation and trade-off authority	Make non-negotiable values explicit for each use context	Identify likely trade-offs early, including cultural and domain-specific differences, and define who has authority to resolve them (Sadek et al. 2024; Spiekermann et al. 2022)
2. Directional alignment and information integrity	Ensure systems pursue intended organisational goals and value commitments, not only narrow performance targets	Preserve provenance, assumptions, confidence boundaries, and contextual metadata across human-AI handoffs so outcomes remain reconstructable and contestable (Gebru et al. 2021)
3. Role and boundary design	Specify the role assigned to AI in each workflow: advisory, assistive, or decision-executing within defined limits	Define escalation paths, override rights, stop conditions, responsibilities for operators, approvers, and owners, and contestation rights for affected stakeholders. (Baird and Maruping 2021; European Parliament and Council 2024)
4. Futures-oriented design	Rehearse possible value conflicts before infrastructures harden	Future narratives can make boundary transformations visible: where AI frames action, where humans can still contest it, and where responsibility may silently shift (Hovorka and Mueller 2025; Richter et al. 2025)
5. Digital sovereignty and accountability architecture	Design for organisational control over critical data, models, vendors, and infrastructures	Implement safeguarding practices, such as reversibility and exit-by-design, to reduce dependency (Wang et al. 2026). Separate user-facing explanation from governance-grade auditability, and ensure that accountability follows practical control rather than formal titles alone
6. Feedback monitoring and continuous adaptation	Allow systems and governance to evolve as capabilities, risks, regulations, and public expectations shift	Build contestability channels, monitoring of longer-term feedback effects, and review cycles (Glickman and Sharot 2025)

questions. In public and regulated settings, legitimacy must therefore be treated as a design requirement rather than an afterthought (Habermas 1991; European Parliament and Council 2024).

A governance approach focused on preserving human practice therefore complements, rather than replaces, technical control. The central concern is not to prevent AI from contributing to decisions. It is to ensure that AI does not silently displace the capacities and responsibilities on which legitimate decisions depend. If we continue to evaluate AI primarily by efficiency gains and benchmark performance, we risk building high-performing systems that undermine responsibility, legitimacy, and trust in everyday organisational practice. Moreover, beyond identifying risks, it is essential that their escalation reaches an actor who can intervene even when doing so conflicts with delivery targets or operational priorities.

4 Extending Value-Centric Design for AI

Existing traditions already provide important foundations. Value Sensitive Design has long argued that human values should be investigated and translated into technical design (Friedman et al. 2013; Sadek et al. 2024). Value-based engineering and the ISO/IEC/IEEE 24748-7000:2022 standard offer process-oriented ways to trace ethical values from early concepts into requirements and design decisions (Spiekermann and Winkler 2020; Spiekermann et al. 2022; ISO/IEC/IEEE 2022). Technology assessment contributes methods for systematically examining possible consequences and futures before technology is stabilised in use (Grunwald 2017). Recent discussions on process mindlessness similarly warn against applying AI without attention to the process it is supposed to improve (van der Aalst et al. 2025).

As a community we can extend and integrate these traditions for contemporary AI. This means extending value elicitation from individual systems to organisational and institutional design, and moving from compliance documentation to lived practice. As with other instances of digital work design, value-centric AI requires an agile, participative, and interdisciplinary process that puts human work practices and their specific contexts at the centre of technology development (Richter et al. 2018). By incorporating user feedback at all stages, organisations can ensure that human needs shape the sociotechnical system. Integrating participatory practices prevents value-centric design from becoming a static exercise.

It also means treating digital sovereignty as an emergent sociotechnical capacity. Organisations should exercise legitimate authority over their data, models, and technological infrastructures to remain capable of revising, contesting, and governing AI-enabled work without illegitimate external interference (Wang et al. 2026). Organisations cannot sustain their value commitments if they lack the authority to revise, contest, or exit critical data, model, infrastructure, and vendor arrangements.

Values are not abstract aspirations to be declared around deployment. They are design commitments that shape roles, interfaces, escalation pathways, data practices, procurement decisions, and organisational learning over time.

We suggest the following steps from a sociotechnical design stance that are displayed in Table 1.

5 Research Directions for Value-Centric AI

The Business and Information Systems Engineering (BISE) community can make a distinctive contribution by connecting technical design, organisational coordination, and societal responsibility.

Doing so requires research that asks not only what AI can do or how it is adopted, but what forms of human-AI collaboration organisations and societies want to create. Our suggestions are to …

> … study agency and interpretation at critical junctures. Examine when AI begins to frame action, when humans can meaningfully contest it, and whether those charged with oversight possess intervention rights and sufficient organisational standing.
> …measure preserved practice as well as performance. Evaluation should include whether people retain appropriate skills of interpretation, contestation, escalation, and accountable judgment.
> …measure legitimacy outcomes. Contestability, accountability clarity, trust calibration, capacity for sovereignty control, and defensibility under scrutiny should be treated as core outcomes alongside productivity and accuracy.
> …design for high-stakes contexts. Public-sector, health, employment, education, and other regulated contexts make governance quality especially visible and consequential.
> …build cumulative design knowledge. We should develop reusable principles, boundary conditions, methods, and reference architectures for AI as navigational scaffolding in organisations.

This agenda also calls for methodological pluralism. Ethnographic and process-oriented studies can reveal how AI changes work in practice. Design science can build and evaluate artefacts for value articulation, traceability, and contestability. Foresight and narrative methods can help researchers and practitioners explore future tensions before they become embedded in routines. Quantitative evaluation

can measure whether interventions preserve practice, improve calibration, and sustain legitimacy under scrutiny.

6 Conclusion

The Marshallese stick chart offers a timely reminder that a technology for orientation is valuable not because it removes the need for human judgment, but because it helps people develop and exercise that judgment in complex conditions. The governance paradox of human-AI collaboration is that AI can shape collective outcomes without entering collective accountability. That paradox cannot be resolved by better optimisation alone.

The contribution of the Business and Information Systems Engineering (BISE) community should be to design sociotechnical systems in which AI supplies navigational scaffolding while humans retain situated judgment, responsibility, and the ability to contest the route taken. If we get this right, AI will not simply scale decisions. It will help organisations navigate complexity without surrendering the values that make their decisions legitimate. For managers and policy leaders, the practical message of our deliberations is: Do not ask only whether an AI system works. Ask whether your organisation can still navigate responsibly with it.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Baird A, Maruping LM (2021) The next generation of research on IS use: a theoretical framework of delegation to and from agentic IS artifacts. MIS Q 45(1):315–341. https://doi.org/10.25300/MISQ/2021/15882

Bauer K, Heigl R, Hinz O, Kosfeld M (2024) Feedback loops in machine learning: a study on the interplay of continuous updating and human discrimination. J Assoc Inf Syst 25(4):804–866. https://doi.org/10.17705/1jais.00853

Bauer K, Nofer M, Henrich B, Drachsler H, Hinz O (2026) The dependency dilemma: how machine learning decision aids can undermine skill growth. Bus Inf Syst Eng, forthcoming

Buçinca Z, Malaya MB, Gajos KZ (2021) To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. In: Proc ACM Hum-Comput Interact, vol 5, no CSCW1, p 188. https://doi.org/10.1145/3449287

Cranefield J, Winikoff M, Chiu YT, Li Y, Doyle C, Richter A (2023) Partnering with AI: the case of digital productivity assistants. J R Soc N Z 53(1):95–118. https://doi.org/10.1080/03036758.2022.2114507

Ensign D, Friedler SA, Neville S, Scheidegger C, Venkatasubramanian S (2018) Runaway feedback loops in predictive policing. In: Proc 1st Conf Fairness Accountability Transparency. PMLR, vol 81, pp 160–171

European Parliament and Council (2024) Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence. ELI: http://data.europa.eu/eli/reg/2024/1689/oj

Faraj S, Pachidi S, Sayegh K (2018) Working and organizing in the age of the learning algorithm. Inf Organ 28(1):62–70. https://doi.org/10.1016/j.infoandorg.2018.02.005

Felnhofer A, Knaust T, Weiss L, Goinska K, Mayer A, Kothgassner OD (2024) A virtual character's agency affects social responses in immersive virtual reality: a systematic review and meta-analysis. Int J Hum-Comput Interact 40(16):4167–4182. https://doi.org/10.1080/10447318.2023.2209979

Floridi L, Sanders JW (2004) On the morality of artificial agents. Minds Mach 14:349–379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d

Friedman B, Kahn PH, Borning A, Huldtgren A et al (2013) Value sensitive design and information systems. In: Doorn N (ed) Early engagement and new technologies: opening up the laboratory. Springer, pp 55–95. https://doi.org/10.1007/978-94-007-7844-3_4

Gaube S, Suresh H, Raue M, Merritt A, Berkowitz SJ, Lermer E, Coughlin JF, Guttag JV, Colak E, Ghassemi M (2021) Do as AI say: susceptibility in deployment of clinical decision-aids. npj Digit Med 4:Article 31. https://doi.org/10.1038/s41746-021-00385-9

Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, Daumé H III, Crawford K (2021) Datasheets for datasets. Commun ACM 64(12):86–92. https://doi.org/10.1145/3458723

Glickman M, Sharot T (2025) How human-AI feedback loops alter human perceptual, emotional and social judgements. Nat Hum Behav 9:345–359. https://doi.org/10.1038/s41562-024-02077-2

Goddard K, Roudsari A, Wyatt JC (2014) Automation bias: empirical results assessing influencing factors. Int J Med Inform 83(5):368–375. https://doi.org/10.1016/j.ijmedinf.2014.01.001

Grunwald A (2017) Assigning meaning to NEST by technology futures: extended responsibility of technology assessment in RRI. J Responsible Innov 4(2):100–117. https://doi.org/10.1080/23299460.2017.1360719

Haag S, Eckhardt A (2017) Shadow IT. Bus Inf Syst Eng 59(6):469–473. https://doi.org/10.1007/s12599-017-0497-x

Habermas J (1991) Erläuterungen zur Diskursethik. Suhrkamp

Hovorka DS, Mueller B (2025) Speculative foresight: a foray beyond digital transformation. Inf Syst J 35(1):140–162. https://doi.org/10.1111/isj.12530

Hüholt N, Szech N (2026) Trusting machines with morality-delegating moral decisions to AI. Eur Econ Rev 105255. https://doi.org/10.1016/j.euroecorev.2025.105255

ISO/IEC/IEEE (2022) ISO/IEC/IEEE 24748-7000:2022: systems and software engineering – Life cycle management – Part 7000: standard model process for addressing ethical concerns during system design. https://doi.org/10.1109/IEEESTD.2022.9967807

Jussupow E, Spohrer K, Heinzl A, Gawlitza J (2021) Augmenting medical diagnosis decisions? An investigation into physicians' decision-making process with artificial intelligence. Inf Syst Res 32(3):713–735. https://doi.org/10.1287/isre.2020.0980

Korinek A, Vipra J (2025) Concentrating intelligence: scaling and market structure in artificial intelligence. Econ Policy 40(121):225–256. https://doi.org/10.1093/epolic/eiae057

Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A (2021) A survey on bias and fairness in machine learning. ACM Comput Surv. https://doi.org/10.1145/3457607

Richter A, Schwabe G (2025) There is no "AI" in "TEAM"! Or is there? Towards meaningful human-AI collaboration. Australas J Inf Syst. https://doi.org/10.3127/ajis.v29.5753

Richter A, Heinrich P, Stocker A, Schwabe G (2018) Digital work design: the interplay of human and computer in future work practices as an interdisciplinary (grand) challenge. Bus Inf Syst Eng 60(3):259–264. https://doi.org/10.1007/s12599-018-0534-4

Richter A, Richter S, Mohammadhossein N (2025) A narrative exploration of the immersive workspace 2040. MIS Q Exec. https://doi.org/10.17705/2msqe.00121

Rudin C (2019) Stop explaining black box machine learning models for high-stakes decisions and use interpretable models instead. Nat Mach Intell 1(5):206–215. https://doi.org/10.1038/s42256-019-0048-x

Sadek M, Constantinides M, Quercia D, Mougenot C (2024) Guidelines for integrating value sensitive design in responsible AI toolkits. In: Proc 2024 CHI Conf Hum Factors Comput Syst. https://doi.org/10.1145/3613904.3642810

Slater M (2018) Immersion and the illusion of presence in virtual reality. Br J Psychol 109(3):431–433. https://doi.org/10.1111/bjop.12305

Spiekermann S, Krasnova H, Hinz O, Baumann A, Benlian A, Gimpel H, Heimbach I, Köster A, Maedche A, Niehaves B, Risius M, Trenz M (2022) Values and ethics in information systems: a state-of-the-art analysis and avenues for future research. Bus Inf Syst Eng 64:247–264. https://doi.org/10.1007/s12599-021-00734-8

Spiekermann S, Winkler T (2020) Value-based engineering for ethics by design. arXiv. https://arxiv.org/abs/2004.13676

Sunyaev A, Benlian A, Pfeiffer J, Jussupow E, Thiebes S, Maedche A, Gawlitza J (2025) High-risk artificial intelligence. Bus Inf Syst Eng 67(6):981–994. https://doi.org/10.1007/s12599-025-00942-6

van den Broek E, Sergeeva A, Huysman M (2021) When the machine meets the expert: an ethnography of developing AI for hiring. MIS Q 45(3):1557–1580. https://doi.org/10.25300/MISQ/2021/16559

van der Aalst WMP, Reijers HA, Hinz O, Pfeiffer J (2025) Process mindlessness: when we lose sight of what AI is supposed to improve. Bus Inf Syst Eng 67(6):771–775. https://doi.org/10.1007/s12599-025-00972-0

Wang B, Ciriello R, Richter A (2026) Digital sovereignty. Bus Inf Syst Eng, forthcoming