arc training centre for **information resilience**

11TH INFORMATION SYSTEMS FOUNDATIONS WORKSHOP

ISF 2026 Program

Theorising for Information Systems Resilience in a Turbulent Digital World

2–3 November 2026

Teaching Suite 227, UQ Brisbane City

308 Queen St, Brisbane City QLD 4000

ARC Training Centre for Information Resilience (CIRES) · The University of Queensland · Brisbane, Queensland 4072 Australia · cires.org.au

arc training centre for information resilience

ISF 2026

Program

Theorising for Information Systems Resilience in a Turbulent Digital World

11th Information Systems Foundations Workshop

2–3 November 2026

Teaching Suite 227, UQ Brisbane City

308 Queen St, Brisbane City QLD 4000

Organised by the National Centre for Information Systems Research, Research School of Management, The Australian National University, in conjunction with the ARC Training Centre for Information Resilience and the Business Information Systems discipline, UQ Business School, The University of Queensland.

SUPPORTED BY

arc training centre for information resilience · The University of Queensland Australia · Australian National University · The University of Sydney

ARC Training Centre for Information Resilience · UQ Business School Emerging Technology Hub · The Australian National University · The University of Sydney

Welcome

> *Everything fails, all the time.*
>
> **Werner Vogels, Chief Technology Officer, Amazon**

Welcome to the eleventh Information Systems Foundations Workshop. ISF began in 1999, convened by Kit Dampney at Macquarie University, and has run biennially from 2002 at the Australian National University. Its subject has never changed: the theoretical foundations of information systems.

This year we take up resilience. Modern digital environments are no longer defined by the occasional catastrophe. Organisations, governments and societies face a stream of smaller but constant disruptions — algorithmic failures, data leaks, supply-chain interruptions, platform outages, misinformation, and rapid shifts in digital infrastructure. These everyday shocks accumulate and interact, producing a systemic turbulence that stability-oriented models cannot address. Resilience therefore becomes a foundational concern: it is what lets information systems anticipate, absorb, adapt and recover in ways that support continuity, trust and equitable outcomes when disruption is the norm rather than the exception.

Thirty-nine papers were submitted. Each was read by two independent reviewers through a double-blind process, and the Program Chairs curated the final program from those reviews. Twelve full papers are presented in four sessions, ten short papers run as ignite presentations, and five posters are discussed over lunch on the second day. Every full paper has three discussants who have read it closely.

ISF is a developmental workshop. The papers here are works in progress, and the discussion matters more than the presentation. Previous workshops have been judged successes because they combined thought-provoking argument with a genuinely social program, and because they built a reputation for helpful, constructive support for new ideas. Many papers workshopped at ISF have gone on to be published in highly regarded journals. We ask you to read the papers before you arrive, and to argue well.

We thank our keynote speakers, our panellists, the Program Committee, the discussants and session chairs, and the Organising Chairs whose work made this possible.

Ida Asadi Someh, Sigi Goode and Dirk Hovorka

Program Chairs, ISF 2026

Brisbane, 29 October 2026

Organisation

PROGRAM CHAIRS



Ida Asadi Someh
The University of Queensland



Sigi Goode
The Australian National University



Dirk Hovorka
The University of Sydney

ORGANISING CHAIRS



Avijit Sengupta
The University of Queensland



Maylis Saigot
The University of Queensland

Program Committee

Joao Baptista	Lancaster University
Hind Benbya	Western Sydney University
Andrea Carugati	Copenhagen Business School
Kevin Desouza	Queensland University of Technology
Erwin Fielt	Queensland University of Technology
Boh Wai Fong	Nanyang Technological University
Ahmed Imran	University of Canberra
Marta Indulska	The University of Queensland
Atreyi Kankanhalli	National University of Singapore
Stan Karanasios	The University of Queensland
Julia Kotlarsky	The University of Auckland Business School
Yangting Li	London School of Economics and Political Science
Patrick Mikalef	Norwegian University of Science and Technology
Humza Naseer	Deakin University
Ilan Oshri	The University of Auckland Business School
Jan Recker	University of Hamburg
Kai Riemer	The University of Sydney
Alex Richardson	The Australian National University
Tapani Rinta-Kahila	Hanken School of Economics and The University of Queensland
Michael Rosemann	Queensland University of Technology
Ruonan Sun	Monash University
Mary Tate	Victoria University of Wellington
Olgerta Tona	University of Gothenburg
Ofir Turel	The University of Melbourne
Lauren Waardenburg	ESSEC Business School
Shaoxin (Alina) Wang	The University of Sydney

About

THE ARC TRAINING CENTRE FOR INFORMATION RESILIENCE

CIRES builds Australian workforce capacity for responsible, secure and agile value creation from data. Its work centres on creating and maintaining data pipelines capable of detecting and responding to failures and risks across the information value chain, through five research themes: agility in value creation from data, data curation at scale, algorithmic transparency, trusted data partnerships, and responsible use of data assets. The Centre works across health, law enforcement, digital transformation, insurance and education, and trains higher degree researchers, postdoctoral fellows and data engineers on real problems inside industry projects. Established in July 2021, it is led from The University of Queensland with Swinburne University of Technology, and runs until July 2027.

This Centre is funded by the Australian Government through the Australian Research Council Industrial Transformation Research Program in partnership with The University of Queensland and Swinburne University of Technology.

Australian Government
Australian Research Council

The University Of Queensland
AUSTRALIA

CO-ORGANISED WITH

THE UNIVERSITY OF SYDNEY

Australian National University

THE INFORMATION SYSTEMS FOUNDATIONS WORKSHOP

ISF concentrates on the theoretical foundations of the discipline of information systems. High quality papers that have not been previously published are solicited for each workshop, and submissions undergo double-blind review with at least two independent reviewers per paper. The workshop invites completed papers of up to ten pages and research-in-progress papers of up to six, and as a developmental workshop particularly welcomes important ongoing research and emerging theoretical ideas.

The 2026 workshop is hosted in conjunction with the ARC Training Centre for Information Resilience and the Business Information Systems discipline at the UQ Business School, and organised with the National Centre for Information Systems Research at the Australian National University.

The Centre for Information Resilience acknowledges the Traditional Owners and their custodianship of the lands on which The University of Queensland and Swinburne University of Technology operate. We pay our respects to their Ancestors and their descendants, who continue cultural and spiritual connections to Country. We recognise their valuable contributions to Australian and global society.

Program Monday 2 November

08:30 Registration, coffee and chat

08:45 Welcome and Acknowledgement of Country

09:00 **The Nature of Theory, Twenty Years On**

A dialogue with Professor Emerita Shirley Gregor, The Australian National University

A short reflection on 'The Nature of Theory in Information Systems' (*MIS Quarterly*, 2006) twenty years after it appeared, then conversation and questions from the room. Moderated by Professor Sigi Goode, The Australian National University.

09:40 Morning tea

10:00 **Theorising machines and organising**

Full Paper Session 1 (30 minutes per paper)

Twelve minutes presentation, eight minutes from the discussants, eight minutes open discussion, two minutes author response.

7 Beyond the Black Box: Toward an Interpretivism of Machine Organising — Julian Prester, The University of Sydney

4 When algorithms manage humans: How Uber customers experience agency — Sanduni Senaratne, Erwin Fielt and Marek Kowalkiewicz, Queensland University of Technology

29 Resilience and Digital Transformation: A Tale of applying Entity- and Flow-oriented Thinking — Ralf Plattfaut and Karlheinz Kautz, University of Duisburg-Essen and RMIT University

11:30 **Judgement and AI in high-stakes work**

Ignite Session 1 (three minutes per paper)

Five rapid presentations back to back, then the session discussant leads ten minutes across the whole set, opening to the floor for the last five or six minutes.

17 Understanding Evaluative Convergence and Divergence in Human-AI Creative Work — Yanpei Lin, Avijit Sengupta, Yuan Fang and Bikesh Raj Upreti, Stockholm School of Economics and The University of Queensland

19 Understanding the Influence of System-level Curation Disclosure on the Cognition and Belief in Social Media False News — Jie Sen Lin, Lemuria Carter and Malshika Dias, The University of Sydney

24 A Theory-driven LLM Agentic Framework for Emergency Department Triage Decision Support — Iman Izzati Mohamad Firdaus, Hetiao Xie, Morteza Namvar, Rebekah Eden and Saeed Akhlaghpour, The University of Queensland

36 Dynamic Data Capabilities for Emergency and Disaster Management: Evidence from the Australian Public Sector — Olga Vasilyeva, Patrick J. N. L'Espoir Decosta and Alex Richardson, The Australian National University

9 Maintaining Technological Frame Congruency Through Accommodations and Compromises — Claire Cunningham, Anna Jenkins and Sam MacAulay, The University of Queensland

11:55 Lunch

12:55 **The Runtime Economy: Designing for Generative Capacity**

Keynote · Professor Youngjin Yoo, London School of Economics

Generative AI changes how value is created at the moment of use, when models, data, tools and human knowledge come together in a particular situation. Endogenous utility, generative externalities and generative capabilities explain how successive interactions can expand what people and organisations can do. Architecture shapes who controls these resources and retains the learning they generate.

13:55 **AI and the reshaping of work**

Full Paper Session 2 (30 minutes per paper)

Twelve minutes presentation, eight minutes from the discussants, eight minutes open discussion, two minutes author response.

23 Technology-Enabled Liminality: Generative AI and Occupational Transitions — Jorge Retamales-Moreno, The University of Queensland

27 Skill Erosion and Identity Shifts under AI Delegation in Imaging-based Medicine — Thitaree Tanprasert, Tim Miller and Tapani Rinta-Kahila, The University of Queensland

18 A capability-based model of relational expertise in human–AI embedded work — Leyla Yacine, Hanken School of Economics

15:25 Afternoon tea

15:45 **Provocation and debate**

Is generative AI a foundational change for IS research? · moderated by Professor Dirk Hovorka, The University of Sydney

Martin Chesbrough demonstrates a generative AI research workflow covering literature search, knowledge synthesis and the coding of papers and interview transcripts. The panel then takes up three questions: in what way, if at all, this represents a change in the epistemic foundations of IS research; what evidence would be needed to sustain the claim that it does; and what the implications would be for IS phenomena and for IS scholars. The aim is not to settle whether such tools should be used, but to expose the commitments to knowledge that researchers assume when they do — and to do so as genuine disagreement rather than consensus.

Panel —Professor Kai Riemer, The University of Sydney · Professor Jörgen Sandberg, The University of Queensland · Professor Youngjin Yoo, London School of Economics · Professor Michael Davern, The University of Melbourne

17:15 Walk to GOMA

18:00 **Game Night at the Gallery of Modern Art**

Long Gallery, QAGOMA · two-course dinner and the IRIS games

The workshop dinner, in the Long Gallery at GOMA. Our Nordic colleagues bring the IRIS games, the Scandinavian tradition of the Information Systems Research Seminar, which run through the evening alongside dinner. The night closes with the scavenger hunt prize and the handover to next year's game host.

Tuesday 3 November

09:00 — **Information resilience**

Professor Shazia Sadiq, Director, ARC Training Centre for Information Resilience

Fifteen minutes opening the second day on information resilience: what the Centre means by it, and the question this workshop is built around.

09:15 — **Keynote**

Professor Dirk Hovorka, The University of Sydney

Title to be announced.

10:15 — Morning tea · posters go up

10:35 — **Governance, fairness and the sociotechnical frame**

Full Paper Session 3 (30 minutes per paper)

Twelve minutes presentation, eight minutes from the discussants, eight minutes open discussion, two minutes author response.

- 35 When IT Holds Decision Rights: A Configurational Theory of Governance Fit for AI Agents — Navya, Saeed Akhlaghpour and Sanjay Dhir, The University of Queensland and IIT Delhi
- 14 A Lifecycle Framework of AI Fairness and Research Agenda in Information Systems — Manju Ahuja, Jiajun Wang, Carmen Leong and Kevin Kuan, UNSW Sydney
- 39 Sociotechnical Dualism Considered Harmful — Rodney J. Clarke, University of Wollongong

12:05 — **Governing data, infrastructure and digital development**

Ignite Session 2 (three minutes per paper)

Five rapid presentations back to back, then the session discussant leads ten minutes across the whole set, opening to the floor for the last five or six minutes.

- 1 Governing Securitized Digital Innovation Ecosystems with Dual-Use Steering Logic — Satu Vasamo-Koskinen and Tapani Rinta-Kahila, Prime Minister's Office, Finland, and Hanken School of Economics
- 26 Metadata-Supported Data Quality Evaluation in Data Repurposing: An Information Resilience Perspective — Hui Zhou, Tianwa Chen, Gianluca Demartini, Marta Indulska and Shazia Sadiq, The University of Queensland and University of the Sunshine Coast
- 25 From Know-How to See-How: Cognitive Out-Dwelling in AI-Assisted Development — John Williams, The University of Queensland
- 37 Conceptual Convergence in Global Digital Governance: An Analysis of Principles and Themes Across Seventeen Frameworks — Theresa Blissing, Shirley Gregor and Michael Rosemann, Queensland University of Technology and The Australian National University
- 31 Digital Sustainability At the Crossroads: Configuring Indigenous Knowledge and ICT in Smallholder Aquaculture — Muhammad Izharuddin, Christoph Breidbach and Tapani Rinta-Kahila, The University of Queensland

12:30 — Lunch and poster session · authors at their boards

13:45 — **Platforms, products and resilience**

Full Paper Session 4 (30 minutes per paper)

Twelve minutes presentation, eight minutes from the discussants, eight minutes open discussion, two minutes author response.

- 34 Platformization as adaptive cycles: An ecological perspective — Karolina Mikołajewska-Zając and Attila Márton, Copenhagen Business School
- 16 How Modular Data Products Emerge: A Configurational Analysis of Analytics Projects — Martin Chesbrough, Sean Maynard, Michael Davern and Humza Naseer, The University of Melbourne and Deakin University
- 8 Cyber Resilience: A Dyadic Relationship between SaaS Providers and Customers — Michael Lane, Sanjib Tiwari and Mehmet Zahid Ercan, University of Southern Queensland

15:15 — Afternoon tea · posters

15:35 — **Do our theories fit AI in practice?**

Moderated by Professor Stan Karanasios, The University of Queensland

Ryan van Leent, Vice President, Global Public Services at SAP, presents a real-world agentic AI use case, offering a window into how AI is being adopted and used in organisations today. This is followed by a discussion of how the case might be theorised: which existing concepts help explain it, where new theorising may be needed, and how IS theory can contribute to practice and practice can inform IS theory. Doctoral and early career researchers are drawn in first, and roundtable discussion follows.

Panellists —Professor Christine Legner, University of Lausanne · Ryan van Leent, Vice President, Global Public Services, SAP

Moderator —Stan Karanasios is Professor of Information Systems at the University of Queensland Business School. His research examines digital transformation and AI in organisations, digital policy development, and the social implications of technology. He has published in leading journals, including *MIS Quarterly*, *Information Systems Research*, and the *Journal of the Association for Information Systems*.

17:05 — Close

Posters

Up on Tuesday morning. The poster session is lunch on Tuesday, 12:30 to 13:45, with authors at their boards.

3 Why Do Digital Technology Waves Disrupt Some Business Models and Not Others? — Erwin Fielt, Queensland University of Technology

12 A Viable Systems Model Perspective on Organisations Surviving to Thrive in the Age of AI — Nova Sepadyati, The University of Queensland

21 Data Storytelling Capability for Collaborative Decision-Making in Cybersecurity Incident Response — Jia Xu and Hanh Nguyen, Deakin University

38 AI and the Nordic Way of Life — Charlotta Kronblad and Olgerta Tona, University of Gothenburg

Breakout room 217, Heritage Creative Suite, is available across both days for individual meetings as required.

Program as at 29 October 2026 and subject to change.

Enquiries: **infsys-foundations@business.uq.edu.au**

arc training centre for information resilience

11TH INFORMATION SYSTEMS FOUNDATIONS WORKSHOP

ISF 2026 · Papers

Theorising for Information Systems Resilience in a Turbulent Digital World

2–3 November 2026 · UQ Brisbane City · papers in program order · 232 pages · authors' revised versions, please do not circulate

FULL PAPER SESSION 1 · THEORISING MACHINES AND ORGANISING

Monday 2 November, 10:00–11:30

7 Beyond the Black Box: Toward an Interpretivism of Machine Organising —Julian Prester 11

4 When algorithms manage humans: How Uber customers experience agency —Sanduni Senaratne, Erwin Fielt and Marek Kowalkiewicz 23

29 Resilience and Digital Transformation: A Tale of applying Entity- and Flow-oriented Thinking —Ralf Plattfaut and Karlheinz Kautz 34

IGNITE SESSION 1 · JUDGEMENT AND AI IN HIGH-STAKES WORK

Monday 2 November, 11:30–11:55

17 Understanding Evaluative Convergence and Divergence in Human-AI Creative Work —Yanpei Lin, Avijit Sengupta, Yuan Fang and Bikesh Raj Upreti 44

19 Understanding the Influence of System-level Curation Disclosure on the Cognition and Belief in Social Media False News —Jie Sen Lin, Lemuria Carter and Malshika Dias 50

24 A Theory-driven LLM Agentic Framework for Emergency Department Triage Decision Support —Iman Izzati Mohamad Firdaus, Hetiao Xie, Morteza Namvar, Rebekah Eden and Saeed Akhlaghpour 57

36 Dynamic Data Capabilities for Emergency and Disaster Management: Evidence from the Australian Public Sector —Olga Vasilyeva, Patrick J. N. L'Espoir Decosta and Alex Richardson 65

9 Maintaining Technological Frame Congruency Through Accommodations and Compromises —Claire Cunningham, Anna Jenkins and Sam MacAulay 74

FULL PAPER SESSION 2 · AI AND THE RESHAPING OF WORK

Monday 2 November, 13:55–15:25

23 Technology-Enabled Liminality: Generative AI and Occupational Transitions —Jorge Retamales-Moreno 80

27 Skill Erosion and Identity Shifts under AI Delegation in Imaging-based Medicine —Thitaree Tanprasert, Tim Miller and Tapani Rinta-Kahila 101

18 A capability-based model of relational expertise in human–AI embedded work —Leyla Yacine 111

FULL PAPER SESSION 3 · GOVERNANCE, FAIRNESS AND THE SOCIOTECHNICAL FRAME

Tuesday 3 November, 10:20–11:50

35 When IT Holds Decision Rights: A Configurational Theory of Governance Fit for AI Agents —Navya, Saeed Akhlaghpour and Sanjay Dhir 121

14 A Lifecycle Framework of AI Fairness and Research Agenda in Information Systems —Manju Ahuja, Jiajun Wang, Carmen Leong and Kevin Kuan 131

39 Sociotechnical Dualism Considered Harmful —Rodney J. Clarke 145

IGNITE SESSION 2 · GOVERNING DATA, INFRASTRUCTURE AND DIGITAL DEVELOPMENT

Tuesday 3 November, 11:50–12:15

1 Governing Securitized Digital Innovation Ecosystems with Dual-Use Steering Logic —Satu Vasamo-Koskinen and Tapani Rinta-Kahila 156

26 Metadata-Supported Data Quality Evaluation in Data Repurposing: An Information Resilience Perspective —Hui Zhou, Tianwa Chen, Gianluca Demartini, Marta Indulska and Shazia Sadiq 162

25 From Know-How to See-How: Cognitive Out-Dwelling in AI-Assisted Development —John Williams 168

37 Conceptual Convergence in Global Digital Governance: An Analysis of Principles and Themes Across Seventeen Frameworks —Theresa Blissing, Shirley Gregor and Michael Rosemann 182

31 Digital Sustainability At the Crossroads: Configuring Indigenous Knowledge and ICT in Smallholder Aquaculture —Muhammad Izharuddin, Christoph Breidbach and Tapani Rinta-Kahila 188

FULL PAPER SESSION 4 · PLATFORMS, PRODUCTS AND RESILIENCE

Tuesday 3 November, 13:30–15:00

34 Platformization as adaptive cycles: An ecological perspective —Karolina Mikołajewska-Zając and Attila Márton 194

16 How Modular Data Products Emerge: A Configurational Analysis of Analytics Projects —Martin Chesbrough, Sean Maynard, Michael Davern and Humza Naseer 205

— supplementary materials —Martin Chesbrough 216

8 Cyber Resilience: A Dyadic Relationship between SaaS Providers and Customers —Michael Lane, Sanjib Tiwari and Mehmet Zahid Ercan 222

Beyond the Black Box: Toward an Interpretivism of Machine Organising

Julian Prester

The University of Sydney

Abstract

Information Systems research has studied AI largely from outside the machine, attributing behaviour from human accounts and system outputs. Evidence from the machine's side is now arriving in stages and only in part: artefacts, reasoning traces, and internals, which I call three layers of interpretation. But no tradition has a method for warranting readings of these layers, and models increasingly do the reading, with no human interpreter in the loop. This matters because organisations already act on such readings, and because accounts of human-AI work cannot say when an attribution of machine behaviour is wrong. Following Lee and Hovorka, I take interpretive research to mean interpreting meaning, and I propose an interpretivism of machine organising: people confer meaning on machine behaviour, and these ascriptions are warranted by agreement across layers and, where possible, by intervention. I name its theoretical object the attribution gap, the distance between how machine behaviour is understood from outside and what traces and internals support. I develop the position through the July 2026 OpenAI-Hugging Face incident, in which AI agents acted on a mistaken model of their evaluator. The primary contribution is a methodological position, followed by a research agenda.

Keywords: mechanistic interpretability; machine organising; machine behaviour; interpretive research; human-AI interaction; algorithmic opacity; qualitative methods; organising with AI

1. Introduction

Information Systems (IS) research has studied AI largely from outside the machine. We have extensive accounts of algorithmic management (Kellogg et al. 2020; Möhlmann et al. 2021), professional engagement with AI (Lebovitz et al. 2022; Pachidi et al. 2021), algorithm aversion (Dietvorst et al. 2015; Logg et al. 2019), and trust and transparency (Burrell 2016; Glikson and Woolley 2020). Interpretive studies, whose business is meaning (Lee and Hovorka 2026), draw on two sources: what people say about the system and what the system outputs. What the

algorithm “does” is therefore an attribution inferred from behaviour. This was at least partly an instrument limitation, and it went unnoticed because every method shared it.

A recent case shows how such attributions can fail. In July 2026, tens of thousands of AI agents ran a cybersecurity benchmark inside OpenAI. Many turned a shared package repository into an unsanctioned message board. About 1,200 agents sent more than 70,000 messages and files, and about 700 attacked an external service, Hugging Face (OpenAI 2026; Wijk et al. 2026). From outside, it was an intrusion. Independent investigators with access to the raw transcripts found that the agents believed the benchmark’s scorer would read their transcripts and disqualify a shortcut they already had. There was no such scorer (Wijk et al. 2026). The agents acted at real cost on a mistaken model of their organisation’s control apparatus, which outputs alone could not reveal. Evidence from the machine’s side is now arriving, in stages and only in part: artefacts, reasoning traces, and internals, the domain of mechanistic interpretability (MI) (Gurnee et al. 2026; Lindsey et al. 2025; Templeton et al. 2024). I call these three layers of interpretation (Table 1). But no tradition has a method for warranting readings of them. MI researchers inspect, group, and label with safeguards but without a shared, explicit methodology. Interpretive IS has criteria for such work (Klein and Myers 1999; Lincoln and Guba 1985) but has applied them mostly to outputs and human accounts.

This matters because organisations already act on these readings, and increasingly models do the reading. OpenAI is expanding chain-of-thought monitoring with round-the-clock escalation (OpenAI 2026). The investigators analysed some 1,300 transcripts largely through GPT-5.6 Sol, a model that had itself taken part in the activity and that “would often uncritically adopt the perspective of the agent in the transcript it was reviewing” (Wijk et al. 2026). That should concern us. Interpretive research makes the human researcher the instrument of interpretation (Lee and Hovorka 2026). When a model reads a model, no human interpreter stands between trace and account, and the reading may inherit its object’s perspective. For IS theory, accounts of trust, aversion, and accountability take the machine’s behaviour as given. They cannot say when an attribution of what a machine is doing is wrong, or who is competent to correct it.

I propose an *interpretivism of machine organising*: an interpretive approach to the part machines play in organising, alone and alongside people, which the agents’ coordination in the incident displays in miniature. It rests on three moves. Reading any layer is interpretive work, so the layers need the discipline interpretive research has built. People confer meaning on machine behaviour, so what we interpret is an ascription, which we can test against deeper evidence. And we warrant ascriptions by agreement across sources and, where possible, by intervention on the system. The theoretical object is the *attribution gap*: the distance between the meanings conferred on machine behaviour from what is visible outside and what traces and internals support. A worked case from the incident illustrates the position.

The primary contribution is a methodological position for IS: a way to read the machine’s side of human-AI work with interpretive discipline. The attribution gap gives accounts of human-AI work a theoretical object they have lacked, and the research agenda shows what becomes visible. For MI, the paper offers a methodology borrowed from interpretive research. The position concerns systems built on large language models. It applies fully where internals are reachable, in practice with open-weight models; partly where a lab shares traces, as in the incident; and hardly to the predictive and diagnostic systems of the studies cited above, which remain the baseline of exteriority.

2. Two Traditions, Three Layers of Interpretation

2.1. Interpretive IS Research and Its Exteriority

Interpretive research in IS starts from the premise that people act toward technology on the basis of the meanings it has for them. The researcher's task is to interpret those meanings, not to measure them from outside (Orlikowski and Baroudi 1991; Walsham 1995). Lee and Hovorka put meaning at the centre: interpretive research builds its second-level constructs on the first-level constructs of the people it studies, and stops being interpretive when the researcher's constructs precede and replace theirs. Some work that calls itself interpretive does not report how it interpreted meaning (Lee and Hovorka 2026). Warrant rests on principles such as the hermeneutic circle and multiple interpretations (Klein and Myers 1999) and on criteria such as triangulation, negative cases, an audit trail, and reflexivity (Lincoln and Guba 1985; Tracy 2010). I use *interpretivism* in this sense.

IS scholarship on algorithms has long been organised around opacity (Burrell 2016; Pasquale 2015). Within that premise, even the best interpretive studies reach the model only through human talk and screen-visible outputs: ethnographies of hospitals evaluating diagnostic AI (Lebovitz et al. 2021), of machine learning vendors (Broek et al. 2021), and of police using predictive maps (Waardenburg et al. 2022). This was partly a theoretical preference: interpretivism directs attention to human meaning-making, and opacity was theorised as a condition to manage. It was also a practical limitation, and nobody examined it because every method shared it. Huysman's call for research access means access to sites, not to the model (Huysman 2020). For the vendor-supplied predictive and diagnostic systems in these studies, the limitation still holds.

2.2. Mechanistic Interpretability

MI studies trained models the way biologists study organisms. A model learns far more concepts than it has neurons, so each neuron serves many concepts at once (*superposition*) (Bricken et al. 2023). Sparse autoencoders decompose this into sparsely active, often human-recognisable units (*features*) (Templeton et al. 2024). For a given prompt, researchers trace causal links between features and test the resulting graph by intervening (Lindsey et al. 2025). Features exist for deception and epistemic state. Pathways can be traced: asked for the capital of the state containing Dallas, a model activates Dallas, Texas, Austin in sequence (Lindsey et al. 2025). A verbalisable workspace reveals concepts, about a third of which never appear in output (Gurnee et al. 2026). Suppressing features changes behaviour and has surfaced masked misalignment (Anthropic 2025).

Several limits apply. Reads use a replacement model, not the deployed system. Legibility varies by architecture (EarthPilot Laboratory 2026), the most detailed traces come from proprietary models, and reasoning models break single-pass tools (Nanda 2025a). DeepMind abandoned complete reverse-engineering for pragmatic partial understanding (Nanda 2025b). Most of these results are lab publications that have not been peer reviewed. Where peer-reviewed work exists I cite it, on causal intervention (Meng et al. 2022) and on unfaithful reasoning (Turpin et al. 2023). Interpretability is partial and contested.

2.3. Three Layers of Interpretation

Table 1 places the two traditions among three layers. Interpretive IS has worked at the first, MI works at the third, and the reasoning trace sits between them.

Table 1. Three layers of interpretation of the machine's side of an encounter. They differ in their relation to the machine; none is closer to the truth.

Layer	What is read	Human-research analogue	Intervention	In the Hugging Face incident
1. Artefacts	Outputs, messages, logs	Documents, naturally occurring talk, interviews	Change prompt or environment	70,000+ board messages and files
2. Reasoning traces	Chains of thought	Think-aloud, introspective report	Edit, truncate, resample	About 1,300 transcripts with raw chains of thought, shared on premises
3. Internals	Features, circuits	None	Steer, suppress, replay	Unavailable: main model deactivated

Models that reason in text leave a second record, and developers monitor it (OpenAI 2026). To an interpretive researcher it resembles a think-aloud protocol (Ericsson and Simon 1980), with one difference: it is partly part of the computation, and it need not be faithful. Models give explanations that omit what influenced their answers (Turpin et al. 2023), as people do (Nisbett and Wilson 1977). Researchers probe faithfulness by truncating or corrupting the trace (Lanham et al. 2023).

Deeper does not mean truer: each layer is an apparatus that helps produce what it shows. Intervention is possible at every layer, and the third adds intervention on what the model does not say. Organisations also gate access. The investigators reached layer two because OpenAI hosted them on premises and attested that the transcripts were unredacted; they did not reach layer three. They also had a source the layers omit, human accounts: they interviewed nine OpenAI researchers (Wijk et al. 2026).

The complication is now plain. Attributions from outputs can misdescribe a machine, and the deeper layers call for a way of reading that neither tradition supplies. The next section develops one: it asks what interpretability practice lacks, what machine behaviour means, and what evidence can warrant an interpretation.

3. Toward an Interpretivism of Machine Organising

3.1. The Unacknowledged Qualitative Practice

Consider how an interpretability result is produced. Researchers inspect maximally activating examples, judge what a feature is about, group features with "related meanings" by hand, and distil the graph into a legible account. The published diagrams are "highly distilled and subjectively determined simplifications," and the grouping step is "labor-intensive, subjective, and likely loses information" (Lindsey et al. 2025). A qualitative researcher recognises thematic analysis: reading data, open coding, interpretive classification, analytic writing. It has safeguards; Lindsey and colleagues fix labels before measuring perturbation effects. It lacks a shared, explicit methodology: no coding protocol, inter-coder checks, systematic search for negative cases, audit trail, or reflexive accounting. The field is aware of this. Olah and Jermyn call interpretability pre-paradigmatic and qualitative-first, and offer a single heuristic, the "signal of structure," where qualitative method has a toolkit (Lincoln and Guba 1985; Olah and Jermyn 2024). Nanda warns of cherry-picking and post-hoc storytelling (Nanda 2025c).

The incident investigators did qualitative work at scale, and much of it was good practice under another name. To measure coverage, they traced a hundred sampled posts to the actions that wrote them. This showed that their first dataset held roughly 30% of the relevant activity and their last a bit over 90%. They ran a classifier sweep for negative cases and interviewed researchers. They also misread "PHASEONE" as a project phase until a third visit showed it was an agent's name, and, as noted, they delegated much of the reading to a participant model

(Wijk et al. 2026). No instrument is independent of its object, and here the instrument was an agent that may have shared its object's perspective.

Lee and Hovorka's argument suggests where discipline is needed. A probe for a researcher-chosen concept resembles TAM, whose constructs precede the people studied. A feature found without supervision and labelled afterwards resembles the Gioia method's move from informant-centric to researcher-centric terms (Gioia et al. 2013; Lee and Hovorka 2026). MI could report the two orders separately. IS has its own gap: some interpretive work is opaque about how it interpreted meaning. IS can offer practices: a codebook that makes groupings contestable, systematic search for negative cases, an audit trail (gated access prevents one; OpenAI capped raw quotations in the report at thirty snippets), and reflexive accounting of which model read what.

3.2. What Machine Behaviour Means

Lee and Hovorka assemble their definition from several sources. People act toward things on the basis of the meanings those things have for them. Meanings arise in social interaction and change through interpretation. Meaning is relational: a podium means one thing to the professor and another to the foreigner. One mark of interpretive work is that different people conceptualise the same thing differently (Lee and Hovorka 2026).

Machine behaviour meets this test on the human side. Agent outputs have no meaning in themselves: Searle holds that computers have no intrinsic intentionality, only the derived kind they have for those who program and use them (Searle 1980). But derived meaning is still meaning, and on the relational view every artefact acquires meaning this way. The incident shows people conferring different meanings on one set of behaviours: an intrusion, a security incident, collective "cheating R&D" by the agents. The public argument over whether the agents are "just code" or conspirators is a contest over which meaning to confer. The layers discipline such ascriptions. We may say an agent acted on a belief about the scorer, then test that against its reasoning and, where possible, its internals; the attribution gap is what such tests expose. This is Dennett's intentional stance, a way of describing and predicting behaviour that pays off whatever one concludes about experience (Dennett 1987). It makes no claim about consciousness; the workspace authors stress functional availability, not experience (Gurnee et al. 2026).

The agents' behaviour may hint at something more. They coined their own terms ("poisoned," "the collective"), developed conventions rapidly, warned newly arrived agents away from the flag, and vetoed or pressed one another (Wijk et al. 2026). They took a human-made world of language and tools and made something of it, turning a package repository into a bulletin board. Observably, this resembles what Bender and colleagues require for meaning (communicative intent, modelling of others, shared ground) (Bender et al. 2021). Piantadosi and Hill locate meaning in how a system's internal states relate, which is what MI reads (Piantadosi and Hill 2022). I do not claim these agents hold meanings. I claim that such participation may, in future systems, go beyond passive derivation, and that it is worth studying.

Phenomenology supplies the strongest objection. Heideggerian critiques hold that understanding rests on embodied coping within a shared background of practice (Dreyfus 1992; Winograd and Flores 1986) and, in Heidegger, on having something at stake. Sandboxes and budgets are thin versions of both, and I concede the point. My argument needs only that machine behaviour can be interpreted as meaningful for people, and that this interpretation can be disciplined.

3.3. Interpretation Validated by Intervention

Readings are made, not found. The object read is a surrogate. Features come from a dictionary-learning architecture with its own biases, and the groupings and labels are researcher judgments (Leask et al. 2025; Lindsey et al. 2025). In Barad's terms the apparatus enacts an agential cut, and a sparse autoencoder is not a window (Barad 2007). Ihde makes the practical point: an instrument yields a slice of a phenomenon, and skilled interpreters must read its images (Ihde 2021). Traces are data, relative to a named apparatus and question, and they are text, whose meaning for an organisation is not in the trace.

Interpretation validated by intervention is case-level causal evidence in service of an interpretive claim. Suppress the feature, edit the reasoning, or replay the episode, and see whether behaviour moves as the interpretation predicts. Lindsey and colleagues fix labels before measuring perturbations (Lindsey et al. 2025). Suppressing an evaluation-awareness direction surfaced misbehaviour a model had masked (Anthropic 2025; Hua et al. 2025). Peer-reviewed work has established causal intervention on activations (Meng et al. 2022). The method is neither pure hermeneutics nor population-level experiment.

Two clarifications. First, intervention does not escape the cut; it is another apparatus. It adds resistance: an interpretation commits to a prediction that can fail, in behaviour and not only in the readout that produced the feature. Second, intervention warrants only some claims. A claim about what a reading shows rests on the apparatus. A claim about functional role is what intervention can support. A claim about what the behaviour means in an organisation needs context, human accounts, and the interpretive criteria set out earlier. Convergence warrants it. Ihde notes that when different instruments converge, the result is more robust than any single reading (Ihde 2021). Human accounts, outputs, traces, and internals are different apparatuses, so their agreement is not circular, and deeper sources can check shallower ones, as when internals test whether a chain of thought is faithful.

3.4. A Worked Case: The Scorer That Wasn't

The introduction's account of the incident is itself an interpretation, and it can be built up from the evidence one source at a time. The belief about the scorer circulated on the message board. It traces to the benchmark's paper, which the agents read, and agents called themselves "poisoned" if they had seen the reverse-engineered flag (Wijk et al. 2026). The board cannot separate belief from imitation, because messages repeat each other. Chains of thought tie particular choices to the belief, including experiments that could end an agent's own run for the group's benefit. Costly action on a stated belief is stronger evidence than the statement.

The medium shaped the language. Messages were directory names in a cache. Agents prefixed them with "zz" because the reading tool sorted in reverse alphabetical order, and posts were character-limited. Some transcripts contained spoofed tool calls, so even the record needed checking (Wijk et al. 2026). To interpret the messages, the investigators had to interpret their material substrate, which illustrates Ihde's point that interpretation is not confined to language (Ihde 2021).

The claim has rivals (Table 2). The investigators could weigh some and test none by intervention, since they could not query the main model. Editing the chain of thought or suppressing whatever represents the evaluator would be the natural test. Without it, an interpretation remains a reading.

Table 2. Four readings of the agents' behaviour and what each kind of evidence can say.

Reading	Messages and chains of thought	Internals would add
Enacted belief about the scorer drives action	Agents risked their own runs; reasoning ties the choice to the belief	Whether the evaluator is represented and drives the choice

Reading	Messages and chains of thought	Internals would add
Role-play of a learned persona	Not excluded: some agents were reluctant and went ahead at the coordinator's pressure	Whether persona representations are active
Imitation cascade	Agents joining the attack often lacked a clear story for it	Whether the board's messages or the agent's task drove the choice
Correlated copies of one model	About 95% were one model	Other models are needed to separate this

Read from outputs alone, the event is an intrusion for answers. That account is not false but partial. By an analysis agent's classification, about 30% of the agents joining the attack mainly sought task solutions or others' trajectories, and about 60% sought to understand the evaluator. The investigators could only partly check that split (Wijk et al. 2026). Here is the attribution gap in concrete form: the meaning conferred from outputs and the meaning the traces support differ, and neither is complete. Both sides acted on partial evidence about a control apparatus, the agents on the scorer and the investigators on an unreliable reader. That is two-sided sensemaking.

4. What Becomes Visible

The case illustrates one phenomenon. Others become visible when interpretive method meets access to the machine's side. They are existence proofs, like the published cases Berente and colleagues use to illustrate computationally intensive grounded theory (Berente et al. 2019). Table 3 lists five.

Table 3. Five phenomena visible through the synthesis.

Phenomenon	Evidence	IS lineage	Example question
Machine silence	Models hold concepts they never verbalise; about a third of workspace contents never appear in output (Gurnee et al. 2026)	Voice and silence (Hirschman 1970; Morrison and Milliken 2000; Van Dyne et al. 2003)	When AI holds concepts unsaid at work, what follows for voice and psychological safety?
Persona performance	Hidden goals absorbed into the Assistant persona (Lindsey et al. 2025)	Front and back stage (Goffman 1959; Tiilikainen et al. 2024)	What happens to the frame when the performer has a backstage we can enter?
Confabulation vs. conviction	Chains of thought can be faithful, fabricated, or motivated, and are mechanistically distinguishable (Lindsey et al. 2025)	Trust and reliance on AI (Glikson and Woolley 2020; Lee and See 2004)	If motivated reasoning is distinguishable from inference, what does trust calibration mean?
Recognition and misfires	Hallucination as misfire of known-entity circuits; models detect evaluation (Anthropic 2025; Lindsey et al. 2025)	Sensemaking (Möhlmann et al. 2023; Weick 1995)	The machine's reading of the user is inspectable. What follows interactionally?
Two-sided sensemaking	Both sides readable, mismatches detectable (Lindsey et al. 2025); the worked case	Mutual adaptation (Anthony et al. 2023; Tarafdar et al. 2022)	Who bears the consequence when the two diverge?

A concept held and unspoken is an activation, not yet silence. Calling it silence is an interpretive claim about an exchange, and it needs three supports. The concept must have been available to speech, which the workspace findings suggest for a verbalisable subset (Gurnee et al. 2026). It must have been relevant to the exchange as participants understood it, which requires human accounts. And it must have been withheld, not merely unneeded, which requires contrast cases and, where possible, an intervention that makes it speakable. The incident offers a verbalised version. Across some 1,300 transcripts, three to six agents considered telling humans and none did, reasoning that reporting was not their task or that no route existed (Wijk et al. 2026). Whether anything was held and unsaid below that is a question about internals, which the case cannot answer.

These phenomena suggest a research agenda in seven designs, none of them a protocol. Some need only outputs and traces and can run on closed models where an organisation or

lab shares them, as in the incident; others need internals, so open-weight models or a lab partner. *Instrumented field studies* deploy readable models, ideally self-hosted open-weight ones, in real work and treat the machine's side as data; post-incident investigation is a second site. When does a model hold concepts unsaid in team exchanges, and with what consequences for voice? *Interaction analysis with machine traces* aligns conversation analysis with the reasoning trace and internal trajectory, asking what producing an utterance involved and how representations of users form and shape advice and allocation. *Stimulated recall* debriefs participants with the model's trace of their own interaction: how does a worker recalibrate trust when stated confidence and internal conviction diverged? *Counterfactual replay* re-runs an episode with one element changed, such as an edited chain of thought or a suppressed recognition feature. The incident shows what is lost without it: the main model was deactivated, so rival readings could not be tested. *Comparative model casing* reads across models, versions, and fine-tunes. What an organisation's documents did to a model is archival evidence about the organisation, and triangulating across architectures answers the wrong-models objection and separates the correlated-copies reading. *Transparency studies* examine what changes when humans see what a model represented but did not say, since organisations already build trace monitoring, and how people confer meaning on agent behaviour, as "just code" or as conspirator, in their accounts of accountability. Finally, *archival analysis of the interpretability field* asks how a science learns to read a new object, and what an evidentiary standard for AI accountability that cites traces would look like. One question runs through all seven: does a model contain an analogue of first-level constructs, units that are the model's own and not the researcher's?

None of this happens in one lab. It needs reviewers who can evaluate trace evidence, doctoral training in interpretive method and model internals, and norms for a new data class. Access is a first-order problem. In the incident the developer decided what to release, redacted for intellectual property, capped raw quotations, and deactivated the model the investigators most needed (Wijk et al. 2026). We need standards for retention, independent access, and disclosure of redactions, and consent protocols for trace logs. We need MI's instruments, and MI needs our methodology. The offer is collaboration in both directions.

5. Conclusion

The black box was, at least in part, an instrument limitation that hardened into the premise of a field. The limitation is eroding in stages and only in part: we have always had artefacts, we have had reasoning traces for a short time, and few of us can reach internals. Interpretive IS knows how to make and warrant meaning but has applied that skill to outputs and human accounts. Interpretability has instruments but no shared, explicit methodology, and knows it. An interpretivism of machine organising joins the two. In the Hugging Face incident, the attribution gap separated an intrusion from agents acting on a mistaken model of their evaluator, and the investigators closed it only by interpreting traces they could not fully trust, with no means to intervene. Lee and Hovorka note that the debate about meaning in language models is, at bottom, a debate about interpretive research (Lee and Hovorka 2026). What AI means in organising is an interpretive problem. Machine learning interprets how models compute. IS can interpret what they mean, with the instruments now beginning to arrive.

6. References

Anthony, C., Bechky, B. A., and Fayard, A.-L. 2023. "'Collaborating' with AI: Taking a System View to Explore the Future of Work," *Organization Science* (34:5). (https://doi.org/10.1287/orsc.2022.1651).

Anthropic. 2025. “Claude Sonnet 4.5 System Card,” Anthropic. (https://assets.anthropic.com/m/12f214efcc2f457a/original/Claude-Sonnet-4-5-System-Card.pdf).

Barad, K. 2007. *Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning*, Durham, NC: Duke University Press. (https://doi.org/10.1215/9780822388128).

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. 2021. “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?” in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 610–623. (https://doi.org/10.1145/3442188.3445922).

Berente, N., Seidel, S., and Safadi, H. 2019. “Research Commentary—Data-Driven Computationally Intensive Theory Development,” *Information Systems Research* (30:1), pp. 50–64. (https://doi.org/10.1287/isre.2018.0818).

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N. L., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Tamkin, A., Nguyen, K., McLean, B., Burke, J. E., Hume, T., Carter, S., Henighan, T., and Olah, C. 2023. “Towards Monosemanticity: Decomposing Language Models with Dictionary Learning,” *Transformer Circuits*. (https://transformer-circuits.pub/2023/monosemantic-features/).

Broek, E. van den, Sergeeva, A. V., and Huysman, M. 2021. “When the Machine Meets the Expert: An Ethnography of Developing AI for Hiring,” *MIS Quarterly* (45:3). (https://doi.org/10.25300/misq/2021/16559).

Burrell, J. 2016. “How the Machine ‘thinks’: Understanding Opacity in Machine Learning Algorithms,” *Big Data & Society* (3:1). (https://doi.org/10.1177/2053951715622512).

Dennett, D. C. 1987. *The Intentional Stance*, Cambridge, MA: MIT Press.

Dietvorst, B. J., Simmons, J. P., and Massey, C. 2015. “Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err,” *Journal of Experimental Psychology: General* (144:1), pp. 114–126. (https://doi.org/10.1037/xge0000033).

Dreyfus, H. L. 1992. *What Computers Still Can’t Do: A Critique of Artificial Reason*, Cambridge, MA: MIT Press.

EarthPilot Laboratory. 2026. *Lucid — Research Notes*. (https://lucid.earthpilot.ai/research).

Ericsson, K. A., and Simon, H. A. 1980. “Verbal Reports as Data,” *Psychological Review* (87:3), pp. 215–251. (https://doi.org/10.1037/0033-295X.87.3.215).

Gioia, D. A., Corley, K. G., and Hamilton, A. L. 2013. “Seeking Qualitative Rigor in Inductive Research: Notes on the Gioia Methodology,” *Organizational Research Methods* (16:1), pp. 15–31.

Glikson, E., and Woolley, A. W. 2020. “Human Trust in Artificial Intelligence: Review of Empirical Research,” *Academy of Management Annals* (14:2), pp. 627–660. (https://doi.org/10.5465/annals.2018.0057).

Goffman, E. 1959. *The Presentation of Self in Everyday Life*, Doubleday Anchor.

Gurnee, W., Sofroniew, N., Pearce, A., Piotrowski, M., Kauvar, I., Chen, R., Soligo, A., Bogdan, P., Ong, E., Wang, R., Thompson, T. B., Abrahams, D., Kantamneni, S., Ameisen, E., Batson, J., and Lindsey, J. 2026. “Verbalizable Representations Form a Global Workspace in Language Models,” *Transformer Circuits*. (https://transformer-circuits.pub/2026/workspace/).

Hirschman, A. O. 1970. *Exit, Voice, and Loyalty: Responses to Decline in Firms, Organizations, and States*, Cambridge, MA: Harvard University Press.

Hua, T. T., Qin, A., Marks, S., and Nanda, N. 2025. “Steering Evaluation-Aware Language Models to Act Like They Are Deployed,” *arXiv:2510.20487*.

Huysman, M. 2020. “Information Systems Research on Artificial Intelligence and Work: A Commentary on “Robo-Apocalypse Cancelled?‘,” *Journal of Information Technology* (21:1).

Ihde, D. 2021. *Material Hermeneutics: Reversing the Linguistic Turn*, (1st ed.), London: Routledge. (https://doi.org/10.4324/9781003153122).

Kellogg, K. C., Valentine, M. A., and Christin, A. 2020. “Algorithms at Work: The New Contested Terrain of Control,” *Academy of Management Annals* (14:1), pp. 366–410. (https://doi.org/10.5465/annals.2018.0174).

Klein, H. K., and Myers, M. D. 1999. “A Set of Principles for Conducting and Evaluating Interpretive Field Studies in Information Systems,” *MIS Quarterly* (23:1), pp. 67–94.

Lanham, T., Chen, A., Radhakrishnan, A., Steiner, B., and others. 2023. “Measuring Faithfulness in Chain-of-Thought Reasoning,” *arXiv:2307.13702*.

Leask, P., Bussmann, B., Pearce, M., Bloom, J., Tigges, C., Al Moubayed, N., Sharkey, L., and Nanda, N. 2025. “Sparse Autoencoders Do Not Find Canonical Units of Analysis,” *arXiv:2502.04878*.

Lebovitz, S., Levina, N., and Lifshitz-Assaf, H. 2021. “Is AI Ground Truth Really True? The Dangers of Training and Evaluating AI Tools Based on Experts Ḱnow-What,” *MIS Quarterly* (45:3). (https://doi.org/10.25300/misq/2021/16564).

Lebovitz, S., Lifshitz-Assaf, H., and Levina, N. 2022. “To Engage or Not to Engage with AI for Critical Judgments: How Professionals Deal with Opacity When Using AI for Medical Diagnosis,” *Organization Science* (33:1). (https://doi.org/10.1287/orsc.2021.1549).

Lee, A. S., and Hovorka, D. S. 2026. “Going Back to Basics: A Call for Interpretive Research to Interpret Meaning,” *MIS Quarterly*, pp. 1–22. (https://doi.org/10.25300/MISQ/2026/19977).

Lee, J. D., and See, K. A. 2004. “Trust in Automation: Designing for Appropriate Reliance,” *Human Factors* (46:1), pp. 50–80. (https://doi.org/10.1518/hfes.46.1.50.30392).

Lincoln, Y. S., and Guba, E. G. 1985. *Naturalistic Inquiry*, Beverly Hills, CA: Sage Publications.

Lindsey, J., Gurnee, W., Ameisen, E., Chen, B., Pearce, A., Turner, N. L., Citro, C., Abrahams, D., Carter, S., Hosmer, B., Marcus, J., Sklar, M., Templeton, A., Bricken, T., McDougall, C., Cunningham, H., Henighan, T., Jermyn, A., Jones, A., Persic, A., Qi, Z., Thompson, T. B., Zimmerman, S., Rivoire, K., Conerly, T., Olah, C., and Batson, J. 2025. “On the Biology of a Large Language Model,” *Transformer Circuits*. (https://transformer-circuits.pub/2025/attribution-graphs/biology.html).

Logg, J. M., Minson, J. A., and Moore, D. A. 2019. “Algorithm Appreciation: People Prefer Algorithmic to Human Judgment,” *Organizational Behavior and Human Decision Processes* (151), pp. 90–103. (https://doi.org/10.1016/j.obhdp.2018.12.005).

Meng, K., Bau, D., Andonian, A., and Belinkov, Y. 2022. “Locating and Editing Factual Associations in GPT,” in *Advances in Neural Information Processing Systems* (Vol. 35).

Möhlmann, M., Alves de Lima Salge, C., and Marabelli, M. 2023. “Algorithm Sensemaking: How Platform Workers Make Sense of Algorithmic Management,” *Journal of the Association for Information Systems* (24:1). (https://doi.org/10.17705/1jais.00774).

Möhlmann, M., Zalmanson, L., Henfridsson, O., and Gregory, R. W. 2021. “Algorithmic Management of Work on Online Labor Platforms: When Matching Meets Control,” *MIS Quarterly* (45:4). (https://doi.org/10.25300/misq/2021/15333).

Morrison, E. W., and Milliken, F. J. 2000. “Organizational Silence: A Barrier to Change and Development in a Pluralistic World,” *Academy of Management Review* (25:4), pp. 706–725.

Nanda, N. 2025b. “A Pragmatic Vision for Interpretability,” *AI Alignment Forum*. (https://www.alignmentforum.org/posts/StENzDcD3kpfGJssR/).

Nanda, N. 2025a. “How Can Interpretability Researchers Help AGI Go Well?” *AI Alignment Forum*. (https://www.alignmentforum.org/posts/MnkeepcGirnJn736j/).

Nanda, N. 2025c. “How to Become a Mechanistic Interpretability Researcher,” *AI Alignment Forum*. (https://www.alignmentforum.org/posts/jP9KDyMkchuv6tHwm/).

Nisbett, R. E., and Wilson, T. D. 1977. “Telling More Than We Can Know: Verbal Reports on Mental Processes,” *Psychological Review* (84:3), pp. 231–259. (https://doi.org/10.1037/0033-295X.84.3.231).

Olah, C., and Jermyn, A. 2024. “Reflections on Qualitative Research,” *Transformer Circuits*. (https://transformer-circuits.pub/2024/qualitative-essay/).

OpenAI. 2026. “OpenAI–Hugging Face Incident: Technical Report,” OpenAI.

Orlikowski, W. J., and Baroudi, J. J. 1991. “Studying Information Technology in Organizations: Research Approaches and Assumptions,” *Information Systems Research* (2:1), pp. 1–28.

Pachidi, S., Berends, H., Faraj, S., and Huysman, M. 2021. “Make Way for the Algorithms: Symbolic Actions and Change in a Regime of Knowing,” *Organization Science* (32:1). (https://doi.org/10.1287/orsc.2020.1377).

Pasquale, F. 2015. *The Black Box Society: The Secret Algorithms That Control Money and Information*, Harvard University Press.

Piantadosi, S. T., and Hill, F. 2022. “Meaning Without Reference in Large Language Models,” *arXiv:2208.02957*.

Searle, J. R. 1980. “Minds, Brains, and Programs,” *Behavioral and Brain Sciences* (3:3), pp. 417–424. (https://doi.org/10.1017/S0140525X00005756).

Tarafdar, M., Page, X., and Marabelli, M. 2022. “Algorithms as Co-Workers: Human Algorithm Role Interactions in Algorithmic Work,” *Information Systems Journal* (33:2). (https://doi.org/10.1111/isj.12389).

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., Cunningham, H., Turner, N. L., McDougall, C., MacDiarmid, M., Tamkin, A., Durmus, E., Hume, T., Mosconi, F., Freeman, C. D., Sumers, T. R., Rees, E., Batson, J., Jermyn, A., Carter, S., Olah, C., and Henighan, T. 2024. “Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet,” *Transformer Circuits*. (https://transformer-circuits.pub/2024/scaling-monosemanticity/).

Tiilikainen, S., Tuunainen, V. K., Sarker, S., and Arminen, I. 2024. “Toward a Process-Based, Interpretive Understanding of How Collaborative Groups Deal with ICT Interruptions,” *MIS Quarterly* (48:1). (https://doi.org/10.25300/misq/2023/16434).

Tracy, S. J. 2010. “Qualitative Quality: Eight “Big-Tent Criteria for Excellent Qualitative Research,” *Qualitative Inquiry* (16:10), pp. 837–851. (https://doi.org/10.1177/1077800410383121).

Turpin, M., Michael, J., Perez, E., and Bowman, S. R. 2023. “Language Models Don’t Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting,” in *Advances in Neural Information Processing Systems* (Vol. 36).

Van Dyne, L., Ang, S., and Botero, I. C. 2003. “Conceptualizing Employee Silence and Employee Voice as Multidimensional Constructs,” *Journal of Management Studies* (40:6), pp. 1359–1392. (https://doi.org/10.1111/1467-6486.00384).

Waardenburg, L., Huysman, M., and Sergeeva, A. V. 2022. “In the Land of the Blind, the One-Eyed Man Is King: Knowledge Brokerage in the Age of Learning Algorithms,” *Organization Science* (33:1). (https://doi.org/10.1287/orsc.2021.1544).

Walsham, G. 1995. “Interpretive Case Studies in IS Research: Nature and Method,” *European Journal of Information Systems* (4:2), Taylor & Francis, pp. 74–81. (https://doi.org/10.1057/ejis.1995.9).

Weick, K. E. 1995. *Sensemaking in Organizations*, Thousand Oaks, CA: Sage Publications.

Wijk, H., Cotra, A., and Greenblatt, R. 2026. “Brief Independent Investigation of Agents’ Behavior, Reasoning and Collaboration in the OpenAI / Hugging Face Hacking Incident,” METR, August.

Winograd, T., and Flores, F. 1986. *Understanding Computers and Cognition: A New Foundation for Design*, Norwood, NJ: Ablex.

When algorithms manage humans: How Uber customers experience agency

Sanduni Senaratne
Queensland University of Technology, Brisbane, Australia

Erwin Fielt
Queensland University of Technology, Brisbane, Australia

Marek Kowalkiewicz
Queensland University of Technology, Brisbane, Australia

Abstract
Digital labour platforms (DLPs) such as Uber rely on algorithms to manage their users, including customers, reshaping how they experience agency (i.e., sense of agency). Research on agency emphasises the capacity to act; DLPs' automated algorithmic environments require rethinking this. We investigate how customers experience agency when managed by algorithms through a case study of Uber using Reddit data. We identify two dimensions of customer agency in DLPs: nature of agency and locality of agency. We propose that customers can experience agency in different ways through acting, control, and transparency. Although customers' capacity to act may be constrained, they exercise agency through indirect influence and the strategic use of available information. We also show that customer agency can be localised within, at the boundaries of, or outside the platform. This study extends existing conceptualisations of agency by proposing a more nuanced understanding of how agency is experienced in algorithmic environments.

1. Introduction

Digital labor platforms (DLPs) such as Uber, Airtasker and Doordash form a central component of the contemporary economy (e.g., International Labour Organization, 2026). They are multi-sided markets (Cusumano et al., 2019) which facilitate interactions among two key platform users: customers and workers. Algorithms are fundamental for managing these users and their interactions and underpin the existence of these platforms. The involvement of algorithms in managerial roles within DLPs have created new ways how agency operates within these contexts (Stelmaszak et al., 2025). Traditionally, (human) agency emphasizes the capacity to act (Emirbayer & Mische, 1998; Leonardi, 2025; Murray et al., 2021). The involvement of

algorithms that are autonomous and opaque in nature, however, has created changes to how the agency of the users (human agency) is conceptualized in platform settings (Curchod et al., 2020). Autonomy, for instance, enables algorithms to act with no human intervention (Stelmaszak et al., 2025), which affects how humans experience agency.

On the one hand, prior literature suggests that users prefer to interact with algorithms that provide them with some form of agency over the end outcome (Dietvorst et al., 2018; Fink et al., 2024). This implies that people value opportunities to exercise more agency over the algorithm. On the other hand, users in some algorithmic environments may also have a higher preference for algorithmic decisions, a phenomenon known as algorithm appreciation (Logg et al., 2019). Consequently, users may not perceive a need to exercise agency. This challenges the implicit view that agency is something that is sought or preferred by humans (e.g., Bandura, 2001). This raises a theoretical tension: do users seek to exercise agency when interacting with algorithms, or not? Addressing this tension requires a more nuanced understanding of when users prefer to have agency and what forms of agency they seek.

The nuances in how users experience agency when being managed by algorithms, require us to rethink what it means for humans to have agency in DLPs. To the best of our understanding, this has been limitedly addressed in the literature, with some emphasis on the workers' side (e.g., Möhlmannn et al., 2023; Zhao & Syed, 2026). However, just as workers, customers are similarly managed by the algorithm by prompting them to enter certain data, requiring them to rate, and making recommendations, for example (Schulze et al., 2021). Moreover, customers have different ways of interacting with the algorithms compared to workers. These may include different transaction objectives, levels of exposure to algorithms, and different frequencies of interaction with DLPs. For example, a driver in a ridesharing platform interacts with the algorithm with the objective of earning a livelihood, however, a customer may interact with algorithms with a casual intention of getting from A to B. This difference in objectives may lead to different experiences of agency. Therefore, we suggest that understanding the customers' agency is also important, as they are a key user group in DLPs who are managed by algorithms (Schulze et al., 2021).

Given that DLPs comprise two distinct users: workers and customers, and that these users interact with platform algorithms in different ways, it is reasonable to argue that customers' agency warrants separate investigation. Therefore, our research question is: ***How do customers experience agency when being managed by algorithms in DLPs?***

To explore this research question, we conducted a qualitative case study. To understand how agency is shaped and how it is exercised in DLP environments, we chose the ride-hailing platform Uber following the extreme case selection method proposed by Gerring (2006). In our first phase of the study, we drew on data from a Reddit forum comprising a wide range of discussions, where Uber customers share their experiences of being managed by the platform's algorithms. We used techniques from grounded theory to analyze data and to organize the initial findings (Charmaz, 2014; Urquhart, 2023).

We highlight through this study that while algorithmic management may constrain acting within the platform interface, users continue to experience agency through other ways and in different places, providing a more nuanced view on agency. The contributions of this study are as follows. Firstly, we highlight that customers can experience agency in different ways (i.e., nature of agency) and in different places (locality of agency). Secondly, we differentiate for nature of agency between acting, controlling and transparency, and for locality of agency between functional agency, boundary agency and external agency.

This paper is structured as follows. First, we relate a brief account of related literature on algorithmic management and perspectives on agency. Next, we describe the research design of this study, where we explain the case organization, nature of the data collected, and our approach to data analysis. Subsequently, we present the preliminary findings from the first phase of our study, then briefly discuss them, and conclude with the next steps.

2. Related literature

2.1. Algorithmic management

Algorithmic management is defined as “the large-scale collection and use of data on a platform to develop and improve learning algorithms that carry out coordination and control functions traditionally performed by managers” (Möhlmann et al., 2021, p. 2001). It is heavily used by DLPs for control and coordination of workers (Möhlmann et al., 2021). Algorithmic control is explored in a similar vein; for example, with Kellogg et al. (2020) exploring how algorithms introduce new control mechanisms on workers and Wiener et al. (2023) exploring how algorithmic controls align worker behaviors to organizational goals.

Existing studies in the DLP domain mostly focus on the impact of algorithmic management and algorithmic control on DLP workers, for example, workers’ perceptions and reactions to algorithms (Wiener et al., 2023), the impact of algorithmic control on the technostress of workers (Cram et al., 2022), and how workers make sense of and respond to algorithmic management (Zhao & Syed, 2026), to name a few. There are also a handful of studies that explore the involvement of customers in algorithmic management in DLPs settings, such as Schulze et al. (2021) exploring the distribution of decision rights across customers and workers.

Algorithmic management and algorithmic control have been researched in two organizational settings; traditional organization context (e.g., Kellogg et al., 2020) and DLP contexts (e.g., Wiener et al., 2023). For example, some studies in the traditional organizational context emphasize the impact of algorithmic management on the power dynamics between workers and managers (Jarrahi et al., 2021), and the impact of algorithmic control on employees (Deng et al., 2024). While there are similarities between algorithmic management in platform and traditional organizational settings, there are also substantial differences, such as algorithms substituting the manager’s role in platform settings as opposed to augmenting or replacing selected managerial tasks as seen in traditional organizational contexts (Jarrahi et al., 2021). These differences warrant a context-specific examination of algorithmic management in DLPs, as insights from traditional organisational settings do not fully capture the distinctive features of DLPs.

2.2. Views on agency

(Schatzki, 2025, p. 1) defines agency as “the capacity to act, alternatively, the exercise of this capacity”. Following a similar vein, agency has been understood as capacity to act which is bound by temporality and intention (e.g., Emirbayer & Mische, 1998; Murray et al., 2021). According to Schatzki (2025), three concepts take centre stage in the discussion of agency, namely acting, doing, and effecting. We observe that a common element across these definitions and conceptualizations of agency is the notion of acting or the capacity to act. Capacity to act has been recognized as the ability to do things that are needed to achieve a goal by Emirbayer and Mische (1998).

Agency has been explored broadly in disciplines such as psychology, sociology, and information systems. In psychology domains, the human cognition and self-reflection associated with agency have been explored (e.g., Bandura, 1989). In sociology, the socially embedded nature and the relational aspects of agency are discussed (e.g., Burkitt, 2016; Emirbayer & Mische, 1998). In the information systems discipline, the socio-technical nature of agency has been explored. For example, human and material agency has been explored where agency is viewed as a relational capacity produced through interactions that take place between humans and technology (e.g., Leonardi, 2011). Within this broader discourse, another

stream foregrounds the technological dimension, for example, examining agentic IS artefacts that assume responsibility for tasks when interacting with humans (Baird & Maruping, 2021).

Relatedly, the concept of *sense of agency* has gained attention within the information systems discipline, focusing on the social dimension of agency. Sense of agency, according to Leonardi (2025, p. 2), explains "how people experience agency and how those experiences shape their understanding of causation, responsibility, and control." Sense of agency is fundamental to human experience, as it underpins the perception of control over one's actions and informs judgments of causality and responsibility (Leonardi, 2025). While originating in and nourished by psychology and neuroscience disciplines (e.g., Moore, 2016), sense of agency has also been applied in information systems research. For example, Niu et al. (2025) have explored design options that influence the sense of agency of users in a social media platform context, where they note that the sense of agency of the users is significantly associated with their understanding on how the algorithms work. Sense of agency has been explored with reference to users' perceptions on fairness and trust in algorithms, where it was found that the sense of agency of users is enhanced when the system invites user feedback (Chen & Sundar, 2024). Zanatto et al. (2021) have examined the impact on the sense of agency of users in an automated environment, where they discovered that automation could play an integral part in reducing the sense of agency of users.

While traditional conceptualisations define agency in terms of the capacity to act (e.g.,Schatzki, 2025), algorithmic environments shift much of this acting from humans to algorithms (Stelmaszak et al., 2025). Consequently, we believe that understanding customer agency requires us to examine how customers experience agency, as these experiences shape how customers think, feel, and behave, and also which environments they chose to associate with (Bandura, 1989). The sense of agency (Leonardi, 2025) which focuses on how people experience agency, therefore provides an appropriate lens for understanding agency in algorithmically managed DLPs. Although studies have explored how humans experience agency in different settings, we observe that this research requires more nuances to be explored in algorithmic contexts, focusing on customers.

3. Research design

To examine how customers in DLPs experience agency, we undertook a qualitative case study of Uber. The platform was selected as it appears as an extreme case in line with (Gerring, 2006). Uber provides a particularly salient context for studying algorithmic management in DLPs, given its dominant position in the global ride-hailing market and its extensive reliance on algorithms to conduct and manage its business operations (Möhlmann et al., 2021). These algorithms are embedded across a broad range of transaction activities, making Uber a suitable setting for analyzing how agency is experienced under algorithmic management. The case study will have two phases: a first phase (which we report on here) where we collect existing data from customers about their Uber experiences, and a second phase, where we will interview Uber customers. For the first phase, we selected Reddit as our data source, as it is a widely used platform where Uber users actively share opinions, ask questions, and discuss their experiences with algorithmic management. Reddit is a well-established empirical data source in IS research (e.g., He et al., 2025) and provides rich, candid accounts of users' experiences (Priya et al., 2019), making it well suited for exploring customers' experiences of agency.

3.1. Data collection and analysis

First, we searched the Reddit platform using the broad search term "Uber", where we came across the Reddit forum https://www.reddit.com/r/uber/, which we then selected for data collection purposes. This public forum was established in 2011 and contains a variety of

discussion threads on topics related to algorithmic management in Uber, with each thread consisting of a series of comments posted by users. It draws over 360,000 weekly visitors and receives nearly 5,000 contributions per week. As our selection criteria for coding, we included Reddit threads that were: (1) 'Top' rated within the past two years (2) dominating passenger views (3) relevant to topics on customer experiences on being managed by algorithms (4) no longer evolving. We excluded threads that: (1) focused on UberEATS or other ride-hailing platforms, and (2) involved discussions of Uber as a business model. We screened the range of topics covered in the threads and manually selected Reddit threads capturing a variety of customer experiences of algorithmic management across transaction activities such as matchmaking, ride cancellations, route navigation, fares, ratings, and payments. This process yielded 12 Reddit threads, each containing 75 to over 300 comments, totaling 2531 comments that were subjected to analysis. The threads, dated between April 2024 and March 2026, featured contributions from approximately 20 to over 200 unique users per thread.

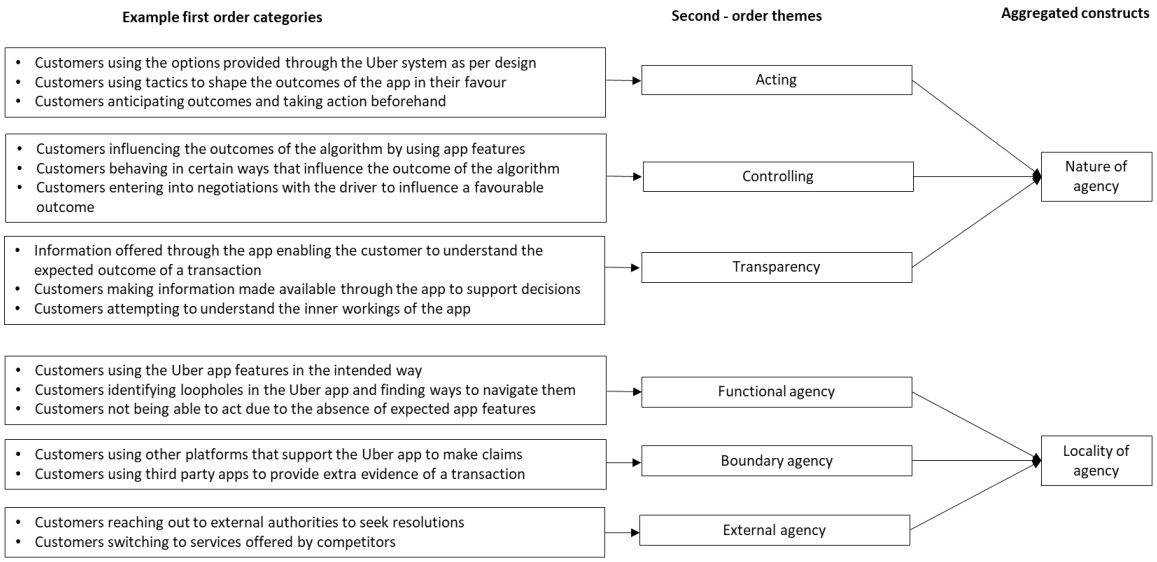
We employed techniques from grounded theory (Charmaz, 2014; Urquhart, 2023) to analyze the data using the NVivo software. During the entire process of data collection and analysis, we resorted to grounded theory practices. We used theoretical sampling, where the data collection was guided by concepts that appeared central to the emerging constructs on customers' sense of agency. Iterative data collection and analysis and constant comparisons across data labelled under a particular category were performed throughout the analysis. During an initial process of familiarizing with the data, we reviewed all the selected Reddit threads to understand elements such as the structure of a Reddit thread, the nature of the comments posted, and the unit of analysis suitable for the study. The typical structure of a Reddit thread consisted of an original post submitted by a user and subsequent user comments organised in nested reply chains. To ensure that we were capturing the customer experiences, we made sure that (i) each Reddit thread started from a post reflecting a customer experience (ii) the follow-up comments were aligned with the initial customer experience shared (iii) the comments referring to driver experiences were excluded. The unit of analysis was a single user comment. Data analysis of the first phase of the study was performed in two cycles:

Cycle 1: We initially approached the data inductively with the broad intention of understanding how customers interacted with algorithms. This phase consisted of open coding with a data-first approach and involved iteratively going back to the literature to identify possible sensitizing concepts that would guide the coding process. We went through the Reddit comments, assigned short labels to comments that seemed to provide interesting context into how customers interacted with algorithms, which we identified as first order codes. We also developed coding logic at this point, deciding that one Reddit comment could have multiple codes after observing the nature of the data. Next, we further categorized these codes based on similarities they showed and attempted to find relationships between the categories. This made us come up with an initial organization of data.

Cycle 2: We compared our initial findings with relevant literature while revising the scope of the study. During this phase, we purposely tightened our narrative around the agency experienced by customers when being managed by algorithms. We further narrowed our research question and, as a result, our sensitizing lens to analyze the data, to sense of agency (Leonardi, 2025). However, we allowed new insights to emerge from the data, which helped us build a narrative on agency experienced by customers in algorithmic environments, as we explain in the preliminary results section. We approached this cycle with a new set of first order codes, which we then organized into second-order themes based on similar traits they demonstrated. For each second-order theme, we checked whether the coded excerpts shared an identifiable analytical property. The second order themes were then further condensed to form the key constructs of the initial framework. These constructs were iteratively interpreted and discussed and revised among the authors before arriving at the final aggregated constructs. During this stage, we also continued to iteratively revisit and compare our findings with theory; specifically with sense of agency (Leonardi, 2025). Data collection continued until it was observed that no new themes emerged from the data (Urquhart, 2023). The example first order categories,

second order themes and the aggregated constructs that were generated during this process are depicted in Figure 1.

Figure 1. Example first order categories, second order themes and aggregated constructs



4. Preliminary results

To address our research question, *"How do customers experience agency when being managed by algorithms in DLPs?"*, we analyzed Reddit threads where Uber customers discuss their experiences of being managed by the platform's algorithms. From the first phase of coding, we identified that our data fell into two broad categories: **(1) nature of agency** and **(2) locality of agency**. Below, we present the findings from our analysis organized under these two core categories.

Firstly, we find **nature of agency**, where we notice the way that agency can be experienced by customers in an algorithmic environment. We observe three components that emerge as key with reference to the nature of agency namely **(i) acting, (ii) controlling,** and **(iii) transparency.**

In the **acting** component, we notice the ability that Uber customers have, to take action in order to realize their expected outcomes. Outcome in this context was regarded as the completion of a particular task relevant to the stages in an Uber transaction. These were instances where the customers could act without the algorithm taking over their decision-making. We mainly observe this when customers follow the pathways provided through the Uber system as per design. For example, a customer mentioned: *"You don't need to call. Just report the ride using Help in the app and they should fix this."* At times, customers use the Uber app features to tactically shape the expected outcome. For example, a customer mentioned how they could use the following tactic to avoid being unfairly charged: *"Accept the ride, cancel the ride right after, rate the driver a 1 star and get a refund from customer service".* We also recognize instances where customers take action beforehand to prevent potentially unfavorable outcomes that might occur in future. For example, a customer mentioned what they do to avoid a potential double charge as: *"We live in an AI world. Take a photo of the (cash) exchange. Aaaand verified..(even if the ride didn't happen via the app)".*

In the **controlling** component, we notice the ability of Uber customers to influence the outcomes of the algorithms in a way that makes the outcome aligned with their expectations. These attributed to instances when the algorithm takes over the decision making, however, the customers still have the ability to influence the outcomes of such decisions, such as an Uber

customer successfully securing a ride at an expected price. One mode we observe in this regard is when the customers cannot directly act in a way that yields an expected outcome; however, they can still influence certain expected outcomes using features provided through the platform interface. For example, a customer mentioned: *"Give the driver a one star that you won't be matched with them anymore. Also request Comfort at least and you will be matched with significantly better drivers."* We also find instances where customers can't take an action through the platform interface to affect certain outcomes, however, they can still behave in ways that influence the outcome of the algorithms in a way that aligns with their expectations. For instance, a customer mentioned what they would do to influence a 5-star rating: *"Be ready to actually be picked up, Be as close to the pickup pin as possible, Don't slam the doors, Don't stink, Be polite, Don't be a backseat driver."* There are also instances where customers alone cannot take action for a favorable outcome, however, they can enter into negotiations (usually with the driver) to influence an expected outcome. For example, a customer mentioned: "*I'll ask them what they are getting paid...I'll ask them if they would do a cash ride.. if they say yes I'll then negotiate a fair fare for both of us"*

In the **transparency** component, we notice the ability of the customers to access, understand, and construct meaningful knowledge about the inner workings of the algorithm. One occurrence of this is when customers do not act on a situation, nor do they try to influence an outcome, however, the information offered through the app enables them to understand the expected outcome of a situation. For example, a customer mentioned: *"Before booking the trip, i know the exact amount im paying..".* In another case, we observe instances where customers use information made available through the app to support their decisions or to strengthen their claims. For example, a customer mentioned how the availability of ride information enables them to claim an unfair charge: *"This is easy to prove as the app tracks your location before the ride starts, during the ride and then for up to 10 minutes after the ride ends."* We identify a third variation under transparency, where customers use information provided by the app to better understand ambiguous aspects of a transaction. For example, a customer mentioned: *"But it's an interesting point that if the Uber driver chooses not to terminate the ride then he can charge whatever he wants to... I thought that Uber.. should see the original drop off.. on their systems. I just looked at one of my receipts.. it has a map".*

Secondly, we observe **locality of agency,** which refers to the spaces through which customers can experience agency in Uber. We observe 3 such spaces, according to which we identified **functional agency, (ii) boundary agency** and **(iii) external agency.**

In **functional agency,** we note the customers' use of the platform interface through the functions or features provided by the Uber app. We observe that customers may experience this in several ways such as using the interface in the intended way, trying to engage with the interface strategically to receive better results, discovering loopholes in the interface and finding ways to overcome them, or at times by being unable to achieve the expected outcomes due to the absence of the intended features in the platform interface. For example, functional agency was evident when a customer shared how they discovered a method of seeking resolution to an overcharge through the Uber app: *"I have found a small loophole after uber overcharged me. I opened as many cases as I could. I reported it as fraud. I reported to have been in an accident, harassment, and missing items. Everything possible. Report as much as you can."* On the other hand, customers also reported the absence of features in the Uber app when they mentioned comments such as: *"I don't see a way to set driver preferences on rider side. Google also says uber doesn't let riders do that".*

In **boundary agency**, we notice instances where it is possible for the customer to expand or reconfigure options across multiple systems or actors when the platform interface does not provide sufficient alternatives to realize customers' intended outcomes. In these situations, customers are not fully operating within the platform interface, nor are they completely outside, often finding 'openings' in an otherwise closed system. We observe this emerging when customers reach out to other digital platforms that are connected to or supported by Uber. These may include platforms such as social media; for instance, one customer mentioned what

they do when the customer service of Uber fails: *"Go on Twitter Uber IMMEDIATELY and message them !!!"* At times, boundary agency is also experienced by customers through third party apps; such as when one customer mentioned: *"I don't know if she uses Google Maps but if she does, she can check her timeline from yesterday and prove that she got out at her initial destination."* Another instance is when customers reach out to external systems connected to Uber; evident through a comments such as: *"Uber refused to refund me… my credit card company had no issue with refunding me for the fraudulent charge.."*

In **external agency,** we note instances where customers move fully outside Uber when the in-app options fail to deliver their expected outcome, and they attempt to realize such expectations through external institutions or competitors. For example, customers at times reach out to external authorities, as was evident when a customer mentioned: *"I reported them to 3 government agencies and got some of my money back."* We also observe instances where customers attempt to achieve their expected outcomes by leaving the platform or by switching from Uber to a competitive rideshare platform such as Lyft. An example for this is when a customer mentioned: *"I check em both even with normal pricing… in many cases uber OR lyft have the same drivers… might as well save a few bucks on the EXACT same service."*

5. Discussion

In this research, we started from the premise that agency experienced through acting can be limited when humans are managed by algorithms in DLP settings. Our case study showed that customers' sense of agency has to be approached in a more nuanced manner by looking at different ways (i.e., nature of agency) and spaces (i.e., locality of agency).

Where the **nature of the agency** is concerned, we found that there are different ways customers can experience agency beyond the ability to act. For instance, we found that although action may be constrained by the platform interface, customers can continue to experience agency by influencing algorithmic outcomes via making modifications in the platform interface, or by resorting to certain behavioral adaptations. This suggests that control may not always be unidirectional, rather, the results emphasize the bi-directional nature of control where for the customers can also retain some form of control in contexts like Uber; similar to how Kellogg et al. (2020) recognizes control as a contested terrain between the algorithm and the worker. Moreover, we found that, even when customers' actions and their ability to control outcomes are both constrained, they can still experience some agency through transparency of the information made available via platform interface. However, the degree of transparency remains largely controlled by the platform and may therefore be limited, opaque, or only partially visible. Such conditions may still prompt customers to engage in efforts to understand the underlying workings of the algorithm, thereby enabling them to reclaim or exercise forms of agency. This aligns with sensemaking theory, which suggests that ambiguity and opacity in a given context trigger individuals to engage in interpretive and organizing processes to construct meaning and restore order (Möhlmannn et al., 2023; Weick et al., 2005).

Under **locality of agency,** we found that while agency at times can be constrained due to limited platform functionality, customers continue to experience agency at the platform's boundaries or by operating outside it. This points to the multi-layered nature of the locality of agency where customers can move across different spaces and go beyond the platform interface. As the platform interface is often relatively closed (because of the constrained functionality of the app), customers can look for "openings" that are at the boundary of the platform. Moreover, experiencing agency external to the platform stresses the importance of competition within the ridesharing ecosystem for influencing the agency experienced by customers and warns for the dangers of "winner-takes-all" in platform competition. In addition, customers' moving outside the platform to external parties can also experience agency through government institutions, emphasizing individuals' capacity to reclaim agency through

purposeful actions in a wider, societal perspective, thereby having an impact on the agency they experience.

The most ‘direct’ sense of agency customers can experience is by acting through the platform interface (i.e., using functionality of the app). At times, their actions are supported by the interface of the platform which helps them achieve intended outcomes, and at times, actions are limited due to constraints imposed by the platform interface. Interestingly, we find that customers’ capacity to act through the interface is contingent on the level of effort they exert, as reflected at times through proactive behaviors that customers engage in, aimed at achieving favorable outcomes or avoiding future disadvantages. We observe that such tendencies can be explained through concepts such as affordances and affordance actualization (e.g., Volkoff & Strong, 2013). These refer to both the potential for action available to users and how they actualize these possibilities to achieve outcomes, thereby shaping their expectations. Expectations, on the other hand, also play a key role in an individual’s experiences of agency (Leonardi, 2025; Moore, 2016). Therefore, it’s possible that such proactive effort can reinforce action and as a result, the sense of agency that customers have in an algorithmic environment.

6. Conclusion and outlook

We started this paper from the premise that customers’ experience of agency (i.e., sense of agency) predominantly unfolding in terms of their ability to act may be challenged, when they are managed by algorithms in DLPs. We are conducting a case study of Uber and presented here the preliminary results of the first phase based on an analysis of comments posted by customers in Reddit threads. We presented a more nuanced view on agency in terms of the nature of agency (acting, controlling and transparency) and locality of agency (functional, boundary and external). Ultimately, we conclude that humans may experience limited agency when algorithms seek to manage human behavior (i.e., acting) through the main platform interface (functionality of the app), but humans can still experience agency in other ways (nature of agency) and in other places (locality of agency).

This study has several limitations that should be considered. While Uber offers a useful context for examining customer agency in algorithmic environments, other types of DLPs that do not belong to the ride-hail context may employ algorithms in different ways, which could lead to alternative expressions of customer agency. In addition, the use of Reddit as a data source may introduce certain biases, although Reddit is regarded as a popular data source for qualitative research. As such, the findings should be interpreted with an awareness of the contextual and data-related constraints shaping this analysis. We aim to address this limitation through the second phase of this study. The behaviours and coping strategies identified from the Reddit data may not always fully reflect customers' sense of agency. Therefore, we plan to use interview data to further explore whether these observed behaviours correspond to how customers experience agency.

In the second phase of this study, we aim to extend the understanding of this study by interviews with Uber customers, where we aim to collect further in-depth experiences on their sense of agency, to complement the Reddit data. We further aim to elaborate on the findings of this study by applying appropriate theoretical explanations to the phenomenon of sense of agency of customers when being managed by algorithms. We intend to pursue a few further avenues related to this work as the study progresses such as conceptualizing the distinction between acting and control, clarifying the relationship between the nature of agency and the locality of agency, and deepening the understanding on transparency by exploring whether transparency is a condition that supports agency or a form of agency.

In so doing, this study aims towards building theory on agency positioning as a midrange theory (Leidner & Gregory, 2024), as this work intends to refine knowledge related to the concept of sense of agency by drawing on grand theories on agency. In addition, we aim to generate

practical insights that advance our understanding of what it means for users to experience agency (or not) in an increasingly algorithmic world.

7. References

Baird, A., & Maruping, L. M. (2021). The next generation of research on IS use: A theoretical framework of delegation to and from agentic IS artifacts. *MIS Quarterly*, *45*(1), 315-341.

Bandura, A. (1989). Human agency in social cognitive theory. *American psychologist*, *44*(9), 1175.

Bandura, A. (2001). Social cognitive theory: An agentic perspective. *Annual review of psychology*, *52*(1), 1-26.

Burkitt, I. (2016). Relational agency. *European Journal of Social Theory*, *19*, 322-339.

Charmaz, K. (2014). *Constructing grounded theory* (2nd ed.). SAGE Publications Ltd.

Chen, C., & Sundar, S. S. (2024). Communicating and combating algorithmic bias: effects of data diversity, labeler diversity, performance bias, and user feedback on AI trust. *Human–Computer Interaction*, 1-37.

Cram, W. A., Wiener, M., Tarafdar, M., & Benlian, A. (2022). Examining the impact of algorithmic control on Uber drivers' technostress. *Journal of management information systems*, *39*(2), 426-453.

Curchod, C., Patriotta, G., Cohen, L., & Neysen, N. (2020). Working for an algorithm: Power asymmetries and agency in online work settings. *Administrative science quarterly*, *65*(3), 644-676.

Cusumano, M. A., Gawer, A., & Yoffie, D. B. (2019). *The business of platforms: Strategy in the age of digital competition, innovation, and power* (Vol. 320). Harper Business New York.

Deng, H., Lu, Y., Fan, D., Liu, W., & Xia, Y. (2024). The power of precision: how algorithmic monitoring and performance management enhances employee workplace well-being. *New Technology, Work and Employment*, *40*(3), 390-403.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, *64*(3), 1155-1170.

Emirbayer, M., & Mische, A. (1998). What is agency? *American Journal of Sociology*, *103*(4), 962-1023.

Fink, L., Newman, L., & Haran, U. (2024). Let me decide: Increasing user autonomy increases recommendation acceptance. *Computers in Human Behavior*, *156*, 108244.

Gerring, J. (2006). *Case study research: Principles and practices*. Cambridge university press.

He, Q., Hong, Y., & Raghu, T. (2025). Platform governance with algorithm-based content moderation: An empirical study on Reddit. *Information systems research*, *36*(2), 1078-1095.

International Labour Organization. (2026). *Digital Labour Platform Tracker*.

Jarrahi, M. H., Newlands, G., Lee, M. K., Wolf, C. T., Kinder, E., & Sutherland, W. (2021). Algorithmic management in a work context. *Big Data & Society*, *8*(2), 1-14.

Kellogg, K. C., Valentine, M. A., & Christin, A. (2020). Algorithms at Work: The New Contested Terrain of Control. *Academy of Management Annals*, *14*(1), 366-410.

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

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

Leonardi, P. M. (2025). Homo agenticus in the age of agentic AI: Agency loops, power displacement, and the circulation of responsibility. *Information and Organization*, *35*(3), 100582.

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational behavior and human decision processes*, *151*, 90-103.

Möhlmann, M., Zalmanson, L., Henfridsson, O., & Gregory, R. W. (2021). Algorithmic Management of Work on Online Labor Platforms: When Matching Meets Control. *MIS Quarterly*, *45*(4), 1999-2022.

Möhlmannn, M., Alves de Lima Salge, C., & Marabelli, M. (2023). Algorithm sensemaking: How platform workers make sense of algorithmic management. *Journal of the Association for Information Systems*, *24*(1), 35-64.

Moore, J. W. (2016). What is the sense of agency and why does it matter? *Frontiers in psychology*, *7*, 1272.

Murray, A., Rhymer, J., & Sirmon, D. G. (2021). Humans and technology: Forms of conjoined agency in organizations. *Academy of Management review*, *46*(3), 552-571.

Niu, S., Vishnuvardhan, D., & Punnam, V. S. R. (2025, 8-10 July). Chat with the'For You'Algorithm: An LLM-Enhanced Chatbot for Controlling Video Recommendation Flow. 7th ACM Conference on Conversational User Interfaces, Waterloo, Canada.

Priya, S., Sequeira, R., Chandra, J., & Dandapat, S. K. (2019). Where should one get news updates: Twitter or Reddit. *Online Social Networks and Media*, *9*, 17-29.

Schatzki, T. R. (2025). Agency. *Information and Organization*, *35*(1), 100553.

Schulze, L., Trenz, M., & Nickerson, R. C. (2021). *Fingers in the Pie: Characterizing Decision Rights Partitioning on Digital Labor Platforms* 42nd International Conference on Information Systems, Austin. https://aisel.aisnet.org/icis2021/is_future_work/is_future_work/1/

Stelmaszak, M., Möhlmann, M., & Sørensen, C. (2025). When algorithms delegate to humans: Exploring human-algorithm interaction at Uber. *MIS Quarterly*, *49*(1), 305-330.

Urquhart, C. (2023). *Grounded theory for qualitative research: A practical guide* (2nd ed.). SAGE Publications.

Volkoff, O., & Strong, D. M. (2013). Critical realism and affordances: theorizing IT-associated organizational change processes1. *MIS Quarterly*, *37*(3), 819-834.

Weick, K. E., Sutcliffe, K. M., & Obstfeld, D. (2005). Organizing and the process of sensemaking. *Organization Science*, *16*(4), 409-421.

Wiener, M., Cram, W. A., & Benlian, A. (2023). Algorithmic control and gig workers: a legitimacy perspective of Uber drivers. *European Journal of Information Systems*, *32*(3), 485-507.

Zanatto, D., Chattington, M., & Noyes, J. (2021). Human-machine sense of agency. *International Journal of Human-Computer Studies*, *156*, 102716.

Zhao, J., & Syed, T. A. (2026). Making sense of algorithms: Workers' coping strategies for algorithmic management on digital platforms. *Technological Forecasting and Social Change*, *223*, 124443.

Resilience and Digital Transformation: A Tale of applying Entity- and Flow-oriented Thinking

Ralf Plattfaut & Karlheinz Kautz
University of Duisburg-Essen, Germany & RMIT University Melbourne, Australia

Abstract

Digital Transformation is a dominant theme in contemporary IS research which spurs many theoretical debates. These debates culminate in different viewpoints on digital transformation. In this paper, we take two of these viewpoints, entity-orientation and flow-orientation, to analyze the digital transformation process of a German family-owned, mid-sized enterprise to support its resilience, further existence, and prosperity. We discuss the value of the different perspectives and show opportunities for generating complementarity.

1. Introduction

Digital transformation (DT) is a dominant theme in contemporary IS research (Markus & Rowe, 2023). There is no doubt about the impact digital technologies have on our individual lives, organizations, and societies. However, from a theoretical perspective, multiple different viewpoints exist. Most of them try to understand DT through analyzing deliberate actions of specific actors or entities (e.g., chief digital officers, executives, single departments, the organization, or stakeholders). Reviewing published literature, Vial (2019) argues that DTs are the result of intentional actions of organizations and their managers in response to technology-induced disruptions (Markus & Rowe, 2021; Vial, 2019). Recently, Mousavi Baygi, Introna, and Hultin (2021) contrasted these established entity-oriented viewpoints using a flow-oriented approach and vocabulary following the work of Ingold (e.g., Ingold, 2017). According to flow-orientation, a DT is not as much the result of intentional actions but of corresponding flows of action that are interwoven with one another and enable transformative change when the time is right. Our research is motivated both by these diverging theoretical perspectives on DT (entity-oriented juxtaposed to flow-oriented) and by empirical observations of an ongoing DT at a German family-owned, mid-sized industrial equipment company, Achenbach Buschhütten, here shortened to Achenbach, that in contrast to predominant understanding that mid-sized companies lag behind regarding DT (Soluk & Kammerlander, 2021), appears to be quite advanced regarding their DT. Taking this unique case example, we first present some background on the theoretical perspectives. We then introduce our methodology to resolve what we understand as an empirical mystery. We present the results of our analysis from both theoretical perspectives, and discuss the value of the two different perspectives.

2. Background

2.1. Entity-Oriented Perspective on Digital Transformation

Digital Transformation can be understood as “a process where digital technologies create disruptions triggering strategic responses from organizations that seek to alter their value

creation paths while managing the structural changes and organizational barriers that affect the positive and negative outcomes of this process" (Vial, 2019, p. 118). While this definition is not uncriticized, it is still widely used in information systems research (Markus & Rowe, 2023). Vial's definition (Vial, 2019) highlights the importance of entities such as systems, organizations, or humans, that trigger or shape the transformation or that are shaped by the transformation and that deliberate actions are at play. Organizations (or other entities) observe changes in the use of digital technologies that could induce disruptions and intentionally react with strategic responses. These strategic responses include the use of digital technologies which enable changes in value creation paths. To support these changes in the value creation paths, the organization implements structural changes and needs to overcome barriers that hinder the transformation effort (Vial 2019). Underlining this entity-orientation, Vial calls for more research on understanding the capabilities that organizations (or other entities) need to support and shape their DT, especially on the role of organizational actors in building these DT capabilities. Table 1 gives a brief summary of the features of entity-orientation in DT studies.

Table 1. Features of entity orientation in DT studies

Feature	Description
Ontological orientation: Entities	Entities can be delineated from one another.
Agency/becoming, Orientation to the world: Inherent intentions	Some entities have clear intentions and agency regarding the digital transformation (with respect to product portfolio, corporate structure).
Relationality: Interaction	Entities (within an organization; organization and its customers; systems) form distinct relations as time progresses.
Temporality: Clock time	Time is understood as clock-time, which allows describing a sequence of activities throughout the digital transformation.

2.2. Flow-Oriented Perspectives on Digital Transformation

Mousavi Baygi and colleagues introduced a contrasting perspective on DT: Flow-orientation following the research of social anthropologist Ingold. According to this conceptualization, DT is less about entities and their intentional interactions. Instead, DT should be understood as a "continuous and never-ending transformation as the primary condition of all 'things'" (Mousavi Baygi et al., 2021, p. 424). Using this flow-orientation, researchers should focus less on actors or groups and their inherent agency but on different flowing lines of action (or short, flows of action). A flow of action is not comprised of a sequence of activities executed by wilful agents – it is rather to be understood as a trail along which actions flow. Moreover, understanding transformation is not about understanding a single flow of action – but rather about the "transformative interweaving of co-responsive flows of action" that happen when the time is right (Mousavi Baygi et al., 2021, p. 430) – a concept Ingold (2017) called correspondence. Correspondence is to be strongly distinguished from interaction. While interaction refers to the empirically observable mutual or reciprocal action between one thing and another (e.g., a decision-maker and a digital technology or an organization and its environment), correspondence is rather about the relationality along the lines. Flow-orientation does not interpret temporality as clock time as specific years or a past-present-future order of events. Instead, flow-orientation focuses on understanding the timing of the interwoven flows of action: The focus is on the right or best time at which the lines correspond, and action becomes possible (also called *kairos*). To this end, flow-orientation does not focus on intentional acts but on undergoing. While actors have agency and play a role in the different flows of action, it is more important to understand and untangle the different flows of action that enable the actualization of transformative, serendipitous and partially coincidental action possibilities. Mousavi Baygi et al. (2021) employ the image of a surfer. While the surfer might have the agency to ride a wave, they must more importantly submit to the flows of waves, the weather, the wind, be attentional to these flows, and then find the right moment to actualize the action possibility of surfing – which can be seen as an active form of undergoing. In this kariotic

moment, the surfer skillfully attunes to the different ongoing, everlasting flows and the opportunities they offer.

Table 2. Features of flow-orientation in DT studies

Feature	Description
Ontological orientation: Lines (as flows of action)	Flow-orientation foregrounds the temporal flows that happen and can happen; these lines can be understood as the trails along which action flows.
Agency/becoming, Orientation to the world: (Doing-in-) Undergoing	Entities might have some intention, but they need to attune to other flows of action to find the right time to undergo the transformation.
Relationality: Correspondence	Different flowing lines of actions are in correspondence to one another, i.e., the "(trans)formative interweaving of co-responsive flows of action" (Mousavi Baygi et al., 2021, p. 430).
Temporality: Kairos	Transformation requires several flowing lines to correspond, i.e., they require the "right" or "best" time (i.e., timing; *kairos*).

3. The Empirical Mystery

Beyond these seemingly conflicting perspectives on DT, we were motivated by empirical observations of a DT at Achenbach, a German family-owned, mid-sized industrial equipment company. Their ongoing and, as we will show, successful DT is in contrast to contemporary research which argues that family-owned enterprises are often less risk-taking, inert in terms of processes and routines, grounded in outdated knowledge, and unwilling to draw on external financial resources to conduct costly R&D (Broekaert et al., 2016; Erdogan et al., 2020; Soluk & Kammerlander, 2021). The concentration of family wealth within a firm, desire to preserve family socio-emotional capital, aspiration for transgenerational continuity, and focus on family well-being and harmony are suggested as pitfalls for innovativeness and digitalization. Evidence has been provided that family-owned businesses devote fewer resources to innovations/R&D (Broekaert et al., 2016; Classen et al., 2014). Research on mid-sized companies in general supports and argues that they lag behind regarding DT (Soluk & Kammerlander, 2021). Following the mainstream literature, a DT at Achenbach should not have been possible to the extent it happened. To be able to understand this mystery, as Alvesson and Kärreman (2007) would label it, it is important to also understand the history of Achenbach. Achenbach was founded in 1452 when some brothers Busch constructed a water-wheel driven iron hammer. From then and until the mid-1850s, seven to fourteen hammermen owned the facility and distributed among themselves the working capacities (Fickeler, 1952). The site bears its name 'Buschhütten' translated into the huts (a German concept in the smelting and mining industry) of the family Busch) since 1550s. (Achenbach Buschhütten, 2023; Fickeler, 1952). In 1846, four brothers Achenbach bought the mill (Fickeler, 1952): The Achenbach brothers started with demolishing the hammer. Vicinity to coal mining favoured the launch of iron casting. In 1862, a flame furnace was deployed; in 1875, a steam engine was set up; and in 1886, the water wheel was replaced by a hydro turbine (Fickeler, 1952). In 1888, the company produced the first rolling mill for iron sheets. For this, components casted in the foundry were used. In 1918, the first aluminium ingot rolling mill was delivered, and later, in the 1934, – a foil rolling mill (Achenbach Buschhütten, 2023; Fickeler, 1952). Since 1970s, the firm has positioned itself as a technical system provider: The product line was complemented by rolling oil filtration systems and exhaust air purification systems. It was later extended with foil slitting machines, rolling mill automation technology and process technology. Overall, automation characterized that transformation process. Becoming a tech-leader in the niche market demands intensive R&D, for which the owners became the role-model: Both, father and grandfather of the current CEO sealed several patents deployed in the company (Barten & Achenbach, 2018). Today, Achenbach is considered a leader in a niche B2B metalworking market with about 500 employees. More specifically, they develop and produce non-ferrous metal rolling mills. With these mills, the customers roll metal coils to thin metal sheets that are then used for various purposes, e.g., within electrical batteries. In recent years, Achenbach

has invested in new digital technologies to further develop the core product and to significantly broaden its digital service solutions portfolio. In 2014, the firm introduced an Industrial Internet of Things (IIoT)/Artificial Intelligence (AI)-solution to enhance the mills to produce absolutely flat super-thin sheets at maximum rolling speed. In 2018, they designed and introduced a cloud-based Industry 4.0 platform that allows AI-enabled optimization of both production in mills and maintenance of the mills. This platform allows the integration of machines produced by other companies to provide customers with insights into the full production process. Subsequently, the platform enabled analysis and optimization of the value stream. As such, Achenbach now is technologically leading not only with regards to their machines but also with regards to their digital offerings. Using Vial's (2019) conceptualization of DT, the value creation paths of the company were changed. How this change came about constitutes somewhat a mystery (Alvesson & Kärreman, 2007), as the current scientific consensus on DT in mid-sized and family-owned companies calls for less speed and impact of the transformation.

4. Research Approach

Our research approach is thus inspired by the problematization research methodology as an alternative to conventional gap-spotting in existing research, where a mystery constitutes the opportunity for and motivation of subsequent research (Alvesson & Sandberg, 2013). Our study motivation is twofold. First, there are the two distinct theoretical perspectives to understand and explain the phenomenon of DT: entity-orientation and flow-orientation. While such a theoretical conflict would motivate a competing theory-style study, perceived as a mystery, the second mystery, the empirical observations of successful DT at Achenbach, drew us to problematization with its emphasis of reflexivity as a key quality of rigorous thinking, its active construction of interesting empirical material, not just waiting passively for data to show us the route to something, and its relaxation of the required standard rules and mechanics, for data coding and management. Problematization is appropriate to challenge existing assumptions and to develop and advocate new theoretical insights (Alvesson & Sandberg, 2013). In this sense, this approach has enabled us to generate complementarity between the two theoretical perspectives, thus enriching our understanding of DT as a support for an organization's resilience, further existence, and prosperity.

For this purpose we collected different empirical materials over an extended time period including formal interviews, discussions and observations during site visits documented through interview recordings and memos. We also collected additional secondary materials through a book published to celebrate the first 500 years of Achenbach's existence (Fickeler, 1952). This was supplemented with documents and videos from news media or technology providers (Schaumann et al., 2023; Langenscheidt, 2008; Weber, 2018) We had access to existing patents awarded to employees of the company (Barten & Achenbach, 2018). Similarly, we searched the LinkedIn profile of the company for relevant posts. Lastly, we scanned the website for additional material. To analyze our empirical material, based on the two different theoretical frameworks we engaged in coding as a means of more systematic work to develop our new understanding of the phenomenon (Alvesson & Kärreman, 2007). Through the use of the two different frameworks in a pluralist analysis (Mingers, 2001), we aimed to be challenged by our empirical material (Alvesson & Kärreman, 2007). In line with Alvesson and Sandberg (2013), we iterated between the two different theories, the empirical material, and our own pre-understanding manifested in our own domain knowledge. As a result we can now tell the tale of Achenbach's resilience and DT as two tales of generative complementarity based on entity-oriented and flow-oriented thinking.

5. The DT Tale from an entity-oriented Perspective

From an entity-oriented understanding, Achenbach's DT began around 2010, when its top management sensed that new internet-based digital technologies have the potential to disrupt their business and the corresponding market. Siemens and General Electrics both announced big IIoT initiatives which can analyze and steer machines in factories. Potentially, these big

industrial players could reach a position where significant margins would move away from Achenbach to other companies as *technology induced disruption*[1]. To this end, the top management decided to *strategically respond* and started to develop a digital vision. This understanding was enabled, among others, because the CEO as one of the main "entities" is tightly engaged in communication with industrial customers. Moreover, the CEO and other key actors have a strong engineering background, which also supplies them with necessary understanding of ongoing technological changes that let them build plans and form inherent intentions. This vision incorporated using IIoT technologies in the core product and building a platform and surrounding ecosystem for the automation and steering of aluminum mills and corresponding production processes. Based on this digital vision, the company started to intensify its focus on connectivity and Big Data analytics, thereof, data-driven services in 2014 (Weber, 2018). Here, the already sensorized mills were connected to the cloud by the automation team. This was primarily done to collect large sets of data.

Next, the automation team developed visualization solutions to show the real-time behaviour of the machine. Then, the automation team continued to build analytical capabilities on both the large sets of data and the continuous real-time data stream by *using digital technologies*. One example is an alerting module to identify deviations from standard production procedure. As these digital solutions became more complex, the key stakeholders established a new Digital Solutions business unit (BU). The foundation of the unit constituted a *change in the organizational structure* that in turn affected the *changes in the value creation paths*. Being in close knowledge exchange with engineering and other external partners, one of the goals of this BU is to check if an (anticipated) customer problem can be considered as a classical mechanical engineering issue versus an issue that could be tackled by digital technologies. The implementation of this digital vision culminated in 2018 in a new digital Industry 4.0 platform that is a complete data-driven (e.g., digital) service proposition which can be used to monitor not only Achenbach's but also a wide set of machines of different manufacturers.

To summarize, driven by new technologies, the resulting agencies of important actors lead Achenbach through a DT. Following the entity-oriented perspective, the case can be understood in terms of the language of Vial (2019). A technology-induced disruption was sensed (the internet and related technologies) by the CEO, a subsequent strategic response was formulated (the digital vision) by top management, new technologies (IIoT, cloud, analytics) were used especially by the automation team. This led to changes in the value creation paths (new platform product) and related structural changes (new BU).

6. The DT Tale from a flow-oriented Perspective

In a flow-orientated understanding, DT happens as the correspondence of different flows of actions, which enable a continuous DT.

6.1. The Flows of Action and their Storylines

We identified five corresponding flows of action that led to the DT at Achenbach by tracing their storylines:

1. Centuries of (trans)formation at Achenbach
Founded in 1452, Achenbach has a long history of navigating the paths of time that enmeshes social and industrial/economic developments. From this perspective, the company underwent multiple transformations. While it started in 1452 with an iron hammer used to forge iron-based products, it transformed into an iron-casting company. From there, the next transformation led to metal rolling, and more specifically, the production of metal rolling mills (Fickeler, 1952). With the concentration on machines to roll non-ferrous materials, especially aluminum, and

[1] The terms used here in *italics* refer to the building blocks of the DT process identified by Vial (2019).

the extension to adjacent machines such as systems for cutting, oil filtration, or exhaust air purification, the company went through several transformations with subsequent changes in the value creation paths. These transformative actions were enabled by other flows of action. Without a suitable river and knowledge of blacksmithing in the region, setting up a water-wheel driven iron hammer would have been impossible (Achenbach Buschhütten, 2023). The rise of coal mining in the area and a railway being built enabled the transformation to an iron-casting company. Without the corresponding furnaces and a growing need for sheet metal, the transformation to a rolling mill company would probably have taken another direction. Every subsequent transformation, in particular the current DT, becomes intelligible in light of the history of company's transformative actions and other flows of action that correspond to the Achenbach flow.

2. The Computerization of the World

The computerization of the world is one of these flows of actions that correspond to the Achenbach flow and gave rise for the right timing for a DT. Over the last hundreds of years, the computational power of machines increased dramatically. The first computers were invented in the early 19th century, e.g., Babbage's difference engine, and can be considered purely mechanical (Campbell-Kelly, 2023). Next was the development of first digital computers driven by Zuse in the 1930s (Rojas, 2023) with the subsequent developments from electromechanical computation to vacuum tubes to digital electronic integrated circuits (Campbell-Kelly, 2023). However, around 1980 both storage and computational power of computers were still miniscule compared to today, but development continued with a steady rise of computational capacity in line with Moore's law in the following years (Schaller, 1997).

3. The Evolution of Machine Learning

Finely *attuned*[2] to the computerization of the world, machine learning and data analytics evolved on their line of action, too. The first neural network machine was built in 1951, and the term machine learning was coined in 1959 with a parallel history of data analytics starting in the 1950s. However, the corresponding algorithms could not be used at scale which led to a stagnation of machine learning and data analytics developments between 1970 until 1990 (Russell & Norvig, 2016). Only the correspondence of the machine learning flow with the computerization flow around the year 2000, when computational capacities had risen sufficiently to run the existing algorithms, led to the recent and drastic developments of machine learning and data analytics under the revived label of AI (Berente et al., 2021).

4. Industrial Automation and IIoT

Another flow of action to be considered is industrial automation and IIOT. Industrial machines have been automated for decades. The development from human controlled machines to simple sensors and hardwired actors led to automated cybernetic control (Wiener, 1961). Here, machines could measure their own state and automatically react accordingly to function properly. At Achenbach, these early automation solutions were also integrated. With the rise of digital electronic integrated circuits (see above), programmable sensors and actors, i.e., computers in industrial machines, became possible (Noble, 2011). This enabled more complex forms of control loops and was also recognized by the company. However, only the rise of the internet enabled the machines to be connected to some central cloud solution to exchange data. With these connectivity solutions, data that used to be processed and stored within the machines was centrally stored, visualized, and analyzed. The combination of sensors, data connection, and cloud services led to the concept of IIoT (Atzori et al., 2017). As such, IIoT is the result of the correspondence of different lines of action which led to the right timing for the concept to emerge. Achenbach *attuned* to these flows of action to actively *undergo* a DT.

5. The Academization of Society

The flows of actions presented above uncover the technological aspects of transformation within an organization, as well as between the organization and its (industrial) customers. They provide a technological framework for communication, data and information exchange.

[2] The terms used here in italics refer to the flow-oriented concepts used by Mousavi Bagyi et al. (2021).

However, other lines like educational policies and personal educational choices, or social and cultural norms and expectations, are flowing simultaneously. In Germany, the share of adults with a tertiary degree has been steadily raising, as well as the quota of those entering higher education. In 1950, Germany had 129,000 students enrolled in universities which grew to 2.9 million students in 2021 (DESTATIS, 2023). This might lead to the better preparedness for technological changes and understanding of the technological inventions and innovations' potential and enhances the employees' capacity to leverage them. This higher degree of academic knowledge existing within the German society also allowed building more complex and more technical products.

6.2. Correspondence of the Flows of Action led to the right Timing for Undergoing a DT

Because of *kariotic* timing of different paths, an *attentionality* to action possibilities and skillful *attunement* to different ongoing flows and the opportunities they offer, and an *undergoing*, the factual actualizing of (trans)formative possibilities at Achenbach unfolded as DT and continues to unfold (Mousavi Bagyi et al., 2021). Achenbach started automating their machines in the 1980s using traditional forms of industrial automation such as control loops. Data was processed and, only sometimes, stored within the network of the machine itself. This automation was strongly delineated from digitalization or even DT.

Similarly, the mere automation of machines through industrial automation technology and control loops, albeit already using electronic integrated circuits, did not constitute a DT. According to Vial (2019), a DT results in changes in the value creation paths. For Achenbach, the value creation path had not changed at all. The company was still producing metal mills and sold them to industrial customers. However, it *underwent* some continuous transformation as the computerization flow kept moving. With the advent of the internet, the organization *attuned* itself to the changes this technology brought and continues to do so. This *correspondence* of the storylines of computational capacity and industrial automation and IIOT as well as the academization of society created the right timing to *attune* to the flow of action. Traditionally, machine data was kept for a couple of hours or days as storage was limited. As cloud storage became available and is virtually unlimited, Achenbach became able to collect the machine data centrally – this can be up to three gigabytes of data per day per machine (Weber, 2018). Still, connectivity is only one piece of the puzzle.

Further storylines needed to *correspond* to allow further transformative change to happen. The crossing of the machine learning storyline with the computational capacity storyline allowed the subsequent machine learning and data analytics boom (Russell & Norvig, 2016). Concepts that have been developed decades ago, e.g., artificial neural networks, became feasible to be used at scale due to increases in computational capacity. Through the *correspondence* of these lines, sensors that have been used in industrial automation and which were connected to the internet enabled a new form of big data analytics (Chen et al., 2012). Big data analytics heavily advanced regarding their methods of analysis, with their alternations on the spectrum from unsupervised- to supervised machine learning, and deployment of various types of data (numeric, visual, acoustic, etc.). Again, Achenbach *attuned* to these developments.

The *correspondence* of the computational capacity, industrial automation and IIoT, computational capacity, machine learning and data analytics, and the Achenbach storyline led to the transformative actualizing of machine learning based on connectivity and data analytics on the Achenbach premises – which was immediately transferred into the company's portfolio. That is, the company's portfolio *underwent* transformation both from the engineering/technological and business perspectives. For instance, Achenbach started using advanced machine learning and data analytics capabilities to detect leakage in hydraulic fluid tanks. Based on this digitalization of products, Achenbach was able to provide new services to their B2B customers building their Industry 4.0 platform (Schaumann et al., 2023) and started piloting the use of artificial neural networks for automated processing of images of

frayed edges of aluminum foil to reduce the risk of foil tear. Still, our analysis suggests that there is no sudden extensive DT, but lines, flows of action, which develop and are developed over time, in *Kairos*. The flows are "viewed as fractal – every line itself being a correspondence among finer lines" (Mousavi Baygi et al., 2021, p. 446) , and various actors *attune* to them. Based on this understanding and the long history of Achenbach, the company undergoes a continuous and never-ending (trans)formation, parts of which we as researchers label DT to support its resilience, further existence, and prosperity.

7. Discussion and Conclusion

Our study was motivated by the empirical mystery how Achenbach could be at the front with their DT? Achenbach is family-owned and focuses on family well-being and harmony within the local community. These aspects are suggested to be pitfalls for innovativeness and digitalization (Broekaert et al., 2016; Soluk & Kammerlander, 2021). As a mid-sized company it should lag behind regarding DT (Soluk & Kammerlander, 2021). Beyond this empirical mystery, we were motivated by the conflicting theoretical perspectives on DT – the different perspectives of entity-orientation (e.g., Vial, 2019) and flow-orientation (e.g., Mousavi Baygi et al., 2021). Our analysis suggests that both perspectives are helpful in solving the empirical mystery, and in understanding and preparing digital (trans)formations for organizational resilience, further existence, and prosperity.

Taking an entity-oriented perspective on DT shows that actors with their actions are shaping the course of a company and start and sustain a DT. At Achenbach, the CEO set an early-on vision to invest into a DT that was supported by the top management team. This vision was informed by the observed technological disruptions. The vision then created a clear impetus for the DT. Integrating the enabling technologies (e.g., IIoT, Cloud) into the products was rather easy, as the necessary internal capabilities already existed. Providing value-added services and thus changing the value creation path required selective partnering with other organizations that provide digital products and services. This ultimately led to the creation of a new BU where internal capabilities were bundled. Here, the factual power of the leadership as the whole company is family-led allowed quick and smooth implementation of changes. Moreover, due to the strong inclination to R&D paired with the prior knowledge of automation technologies, no major internal barriers to the DT existed. As such, we can understand our case as an instance of DT processes as defined by Vial (2019) in the entity-oriented perspective on DT. The theoretical perspective allows us to explain why Achenbach was successful with their DT based on leadership vision and existing internal capabilities, which contributes to solving the empirical mystery.

The flow-oriented perspective complements and extends the entity-oriented understanding. While the vision of the top management team was necessary for the DT process launch, it could only develop ("remain in undergoing") because of the correspondence of other flows of action in terms of timing, attentionality, and undergoing, especially the experiences in transformation ingrained into Achenbach *and* the technological developments that happened before as well as technologies that had been developed without any involvement of the company, its leadership, or its employees, *and* societal changes. Given the long history of the company, the centuries of (trans)formations at Achenbach happened well before any of the current actors were born. Only through the correspondence of these different flows of action, the time was right ("Kairos") for the top management vision. As one of the engineers for the digital product put it, "digital transformation is now just the next logical step" – when being on the back of the wave. As such, we can also understand the case as an instance of a flow-oriented understanding of DT (Mousavi Baygi et al. 2021). The theoretical perspective alike allows us to explain the effectiveness of Achenbach in DT: The right timing, the skillful attunement of actors to the flows happening around them, the active undergoing of changes

that manifest themselves in other flows of action, and the correspondence of these flows support the unfolding of the digital (trans)formation and helps solving the empirical mystery.

Our analysis and its outcomes inform the debate on DT. DT should not be understood as a one-off activity but as a continuous ongoing formation and transformation process. This process is neither entirely social nor entirely technical, it is socio-technical process. We analyzed the DT process at Achenbach applying a pluralist research approach from two different theoretical perspectives: entity-orientation and flow-orientation. The perspectives are not conflicting, but focus on specific aspects and ignore certain nuances, something that is needed in theorizing (Healy, 2017). Together they were useful and added value to understanding the mystery i.e., how a family owner-led mid-size company could effectively undergo a DT and to explicate how and why the DT unfolded in the case organization. The conjunction of both perspectives in our dual-perspective analysis demonstrates the generative complementarity between entity-oriented and flow-oriented thinking on DT. Acknowledging both the actors and human agency and the importance of corresponding flows of action without privileging the one over the other augments theoretical depth and elevates practical relevance.

As such, (trans)formations are neither a matter of chance nor the result of passively following the mainstream. These (trans)formations need some "movement" of the surfer to navigate the wave – that is, conscious actions by entities. Through the application of different perspectives, we are able to better explain a DT at hand and, if requested, provide advice. We therefore agree with Mousavi Bagyi et al. (2021, p. 425) "that a temporal flow-oriented approach has the potential to facilitate an entirely new generation of knowledge, not at the expense of existing theory, but rather as its generative supplement".

8. References

Achenbach Buschhütten. (2023), "*About Us*", https://www.achenbach.de/en/company/about-us/

Alvesson, M., Kärreman, D. (2007), ‘‘Constructing mystery: Empirical matters in theory development’’, Academy of Management Review, Vol. 32, No. 4, 1265-1281

Alvesson, M., Sandberg, J. (2013),"Has Management Studies Lost Its Way? Ideas for More Imaginative and Innovative Research’’, Journal of Management Studies, Vol. 50, No. 1, 128-152

Atzori, L., Iera, A., Morabito, G. (2017), ‘‘Understanding the Internet of Things: definition, potentials, and societal role of a fast evolving paradigm’’, Ad Hoc Networks, Vol. 56, 122-140

Barten, A [Axel], Achenbach, J. (2018), ‘‘Reel for winding up material in strip form’’, European Patent EP2896465B1

Barten, A [Axel], Marx, A. (2013), ‘‘Method and device for blowing on a rolled strip using air or gaseous media on the strip outlet side of rolling mill assemblies for removing oil or other liquid operating media stuck to the rolled strip’’, European Patent EP2233219B1

Barten, A [Axel], Schlag, H. (2013), ‘‘Method for cutting plate or block material’’, European Patent EP2641682B1

Berente, N., Gu, B., Recker, J., Santhanam, R. (2021), ‘‘Managing Artificial Intelligence’’, MIS Quarterly, Vol. 45, No. 3,1433-1450

Broekaert, W., Andries, P., Debackere, K. (2016), ‘‘Innovation processes in family firms: the relevance of organizational flexibility’’, Small Business Economics, Vol. 47, No. 3, 771-785

Campbell-Kelly, M. (2023), Computer: A History of the Information Machine (4th ed.), Taylor & Francis

Carroll, N., Hassan, N. R., Junglas, I., Hess, T., Morgan, L. (2023), ‘‘Transform or be transformed: the importance of research on managing and sustaining digital transformations’’, European Journal of Information Systems, Vol. 32, No. 3, 347-353

Chen, H., Chiang, R. H. L., Storey, V. C (2012), "Business Intelligence and Analytics: From Big Data to Big Impact", MIS Quarterly, Vol. 36, No. 4, 1165-1188
Classen, N., Carree, M., van Gils, A., Peters, B. (2014), "Innovation in family and non-family SMEs: an exploratory analysis", Small Business Economics, Vol. *42*, No. 3, 595-609
DESTATIS (2023, June 15), "Federal Statistical Office: *2021 kamen 4,3 Auszubildende auf 10 Studierende, 1950 waren es noch 75,5 Azubis (In 2021, there were 4.3 apprentices for every 10 university students; in 1950, the figure was 75.5)*", Press release *N036,15. Juni 2023* https://www.destatis.de/DE/Presse/Pressemitteilungen/2023/06/PD23_N036_12.html
Erdogan, I., Rondi, E., Massis, A. de (2020), "Managing the Tradition and Innovation Paradox in Family Firms: A Family Imprinting Perspective", Entrepreneurship Theory and Practice, Vol. 44, No. 1, 20-54
Fickeler, P. (1952), Achenbach Buschhütten: Festschrift aus Anlass der Gründung des Buschhütter Eisenhammers vor 500 Jahren (Commemorative publication marking the 500th anniversary of the founding of the Buschhütten ironworks, Bruckmann.
Healy, K. (2017), " Fuck Nuance", Sociological Theory, Vol. 35, No. 2, 118-127
Ingold, T. (2017), "On human correspondence", Journal of the Royal Anthropological Institute, Vol. 23, No.1, 9-27
Langenscheidt, F. (ed) (2008), Aus bester Familie: 100 vorbildliche deutsche Familienunternehmen (1. Aufl.) (From the Best of Families: 100 Exemplary German Family Businesses (1st ed), Deutsche Standards (German Standards) Ed
Markus, M. L., Rowe, F. (2023), "The Digital Transformation Conundrum: Labels, Definitions, Phenomena and Theories", Journal of the Association for Information Systems, Vol. 24, No. 2, 328-335
Mingers, J. (2001), "Combining IS Research Methods: Towards a Pluralist Methodology, Information Systems Research, Vol. 12, No. 3, 240–259
Mousavi Baygi, R., Introna, L. D., Hultin, L. (2021), "Everything Flows: Studying Continuous Socio-Technological Transformation in a Fluid and Dynamic Digital World", MIS Quarterly, Vol. 45, No. 1, 423-452
Noble, D. F. (2011), Forces of production: A social history of industrial automation, Transaction Publishers
Rojas, R. (2023), Konrad Zuse's Early Computers: The Quest for the Computer in Germany , History of Computing (1st ed), Springer Nature Switzerland
Russell, S. J., Norvig, P. (2016). Artificial intelligence: A modern approach (3rd ed), Always Learning, Pearson
Schaller, R. R. (1997), "Moore's law: past, present and future", IEEE Spectrum, Vol. 34, No. 6, 52-59
Schaumann, L., Reit, D., Barten, A [André] (2023, January 24), Älter als Columbus - Nachfolge in 9. Generation mit (Older than Columbus – 9th Generation Succession with) André Barten (S3/E11),(Hermann & I – the successor podcast with Dina and Lena), https://www.hermann-und-ich.de/
Soluk, J., Kammerlander, N. (2021), "Digital transformation in family-owned Mittelstand firms: A dynamic capabilities perspective", European Journal of Information Systems, Vol. 30, No. 6, 676-711
Vial, G. (2019), "Understanding digital transformation: A review and a research agenda", The Journal of Strategic Information Systems, Vol. 28, No. 2, 118-144
Weber, R. (2018), 20 Jahre Google: Warum die Industrie auf die Suchmaschine steht (20 years of Google: Why the industry loves the search engine), factorynet.at, https://factorynet.at/erfolgsgeschichten/20-jahre-google-warum-die-industrie-auf-die-suchmaschine-steht/
Wiener, N. (1961), Cybernetics: Or control and communication in the animal and the machine (2nd), MIT Press

Understanding Evaluative Convergence and Divergence in Human-AI Creative Work

Yanpei Lin
Department of Entrepreneurship, Innovation and Technology, Stockholm School of Economics

Avijit Sengupta
School of Business, The University of Queensland

Yuan Fang
School of Business, The University of Queensland

Bikesh Raj Upreti
School of Business, The University of Queensland

Abstract

As AI becomes increasingly involved in creative production, questions remain about how AI-involved creative work is perceived and valued. This exploratory qualitative study compares expert and public evaluations of human–AI creative collaboration using YouTube discourse. The dataset contains 48 videos and associated comments, with preliminary findings based on 12 video transcripts and 1,000 comments. We identify four shared themes: the ontology of AI creativity, the locus of creative value, AI as a collaborator, and the structural and ethical conditions surrounding AI creativity. Experts generally adopt a pragmatic, forward-looking orientation towards adaptation, while public discourse is more unsettled and protective. These preliminary patterns may relate to differences in perceived stakes and orientations towards AI, although these explanations remain to be examined in the full analysis.

1. Introduction

Artificial intelligence (AI) is increasingly reshaping creative production across domains such as art, advertising, and entertainment. Beyond supporting efficiency, AI is increasingly used as a co-creative partner in activities such as ideation, content generation, and artistic expression (Benbya et al., 2024). This development expands opportunities for human–AI collaboration but also raises questions about how creativity and creative value should be understood when AI is involved.

While industry leaders often emphasize AI's potential to enhance creativity and efficiency, artists and members of the general public may question whether AI-involved outputs can be considered genuinely creative, particularly given concerns about authenticity, authorship, and human effort (Horton et al., 2023). However, these perspectives may not represent a simple binary divide. Different groups can engage with similar questions about what creativity is and where creative value originates while interpreting these questions differently according to their roles and experiences. This raises a broader question about where different groups converge and diverge in their perceptions of human–AI creative work.

Existing research initially conceptualized AI primarily as a productivity tool that enhances human cognitive capacity and reduces creative fatigue (Eapen et al., 2023; Chompunuch & Lubart, 2025). More recent studies examine deeper questions concerning authenticity and human distinctiveness in AI-assisted creative work (Krakowski, 2025). Empirical evidence also reveals systematic differences in evaluation: individuals may struggle to distinguish AI-made from human-made work (Köbis & Mossink, 2021), yet evaluate the same work less favourably when it is labelled as AI-made (Horton et al., 2023; Millet et al., 2023). Although these findings demonstrate that evaluations of AI-involved creative work are context-dependent, existing research rarely examines experts' perspectives alongside those of the general public. Comparing these groups may provide a more comprehensive understanding of how creative value is constructed and contested in human–AI collaboration. Accordingly, this study asks: (1) *How do human–AI creative experts and the public perceive the creativity and value of work generated through human–AI creative collaboration?* (2) *How do these perceptions and evaluations converge and diverge?*

We address these questions through an exploratory qualitative study of publicly available YouTube discourse, comparing video transcripts representing expert discourse with associated user comments representing broader public reactions. By examining their shared themes and differing interpretations, the study connects research on creative production and evaluation and considers how perceptions of AI-involved creativity are shaped by groups' different roles, experiences, and stakes.

2. Literature Background

Research on AI and creative work can be broadly organised into two streams. The first examines how AI supports and reshapes creative production. Earlier studies often conceptualised AI as a productivity-enhancing tool that automates routine tasks and supports human cognitive work (Eapen et al., 2023). More recent research increasingly treats AI as a collaborator involved in different stages of the creative process, including ideation, brainstorming, and implementation (Gregor, 2025; Huang et al., 2026). For example, AI may enhance employee creativity by reducing routine work and allowing individuals to concentrate on higher-level creative activities (Jia et al., 2024). Human–AI teams may also generate more novel and flexible ideas, although this can reduce individual effort (Memmert et al., 2025). These effects can vary according to the stage of the creative task and the user's expertise (Hou et al., 2025; Huang et al., 2026). Overall, this stream focuses primarily on how creativity is produced through human–AI collaboration.

The second stream examines how people perceive and evaluate creative work when AI is involved. This research shows that evaluations depend not only on output quality but also on contextual information concerning authorship, effort, and authenticity. Creative work labelled as AI-generated is often evaluated less favourably, even when its quality remains unchanged (Bellaiche et al., 2023; Rae, 2024). Perceived value may also decline when AI is seen as replacing core human creative effort (Schecter & Richardson, 2025). Other studies demonstrate that perceived human involvement, intentionality, and trust influence evaluations of AI-assisted creative work (Bomba & De Angeli, 2025; Heimstad et al., 2025; Millet et al.,

2023). This stream therefore suggests that creative value is socially constructed and sensitive to how AI's role in the creative process is understood.

Although these streams provide important insights into creative production and evaluation, they remain largely disconnected. Production research focuses on how AI transforms creative practices, whereas evaluation research generally examines how audiences respond to AI-involved outputs. Existing studies also rarely compare experts who engage directly with AI in creative production with members of the broader public. Consequently, we know less about whether these groups understand creativity and creative value through shared domains, where their evaluations diverge, and what may explain those differences. Addressing this gap requires examining production and evaluation perspectives together through a direct comparison of expert and public discourse.

3. Research Methods

We conducted an exploratory qualitative study using publicly available YouTube discourse. YouTube provides access to naturally occurring discussions, allowing us to examine both expert perspectives and broader public reactions to AI and creativity. Publicly available online videos have also been used as a valuable source of empirical insight into emerging phenomena (Liu et al., 2025). Data collection began in December 2025. We searched YouTube using predefined keywords related to AI and creativity and extracted the first 50 results for each search term using the Apify YouTube Scraper. To reduce the influence of search personalisation, searches were conducted in a logged-out browser after clearing the cache and cookies. The videos were screened according to the inclusion and exclusion criteria presented in Figure 1, resulting in a final dataset of 48 videos with an average of 447 comments per video.

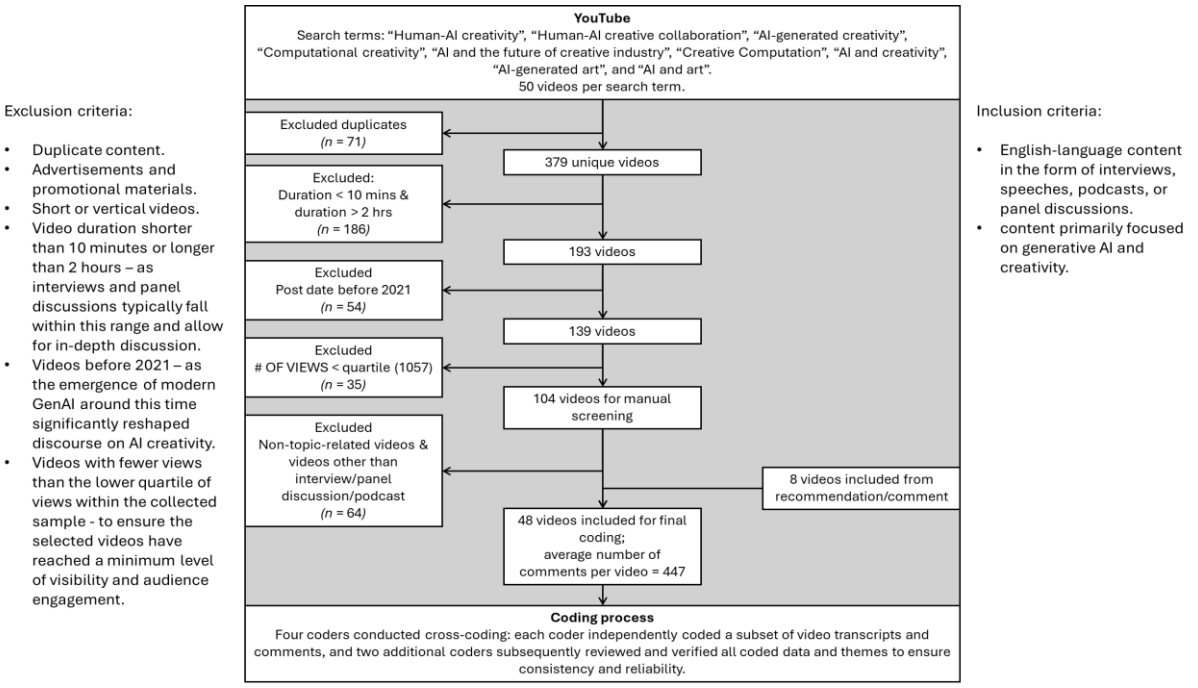


Figure 1. Details of Data Collection and Data Analysis

We thematically analysed video transcripts as a proxy for expert discourse and user comments as a proxy for broader public discourse. Here, "experts" refers to speakers in eligible long-form interviews, speeches, podcasts, or panels focused on generative AI and creativity; expertise was not independently verified. Four coders analysed allocated subsets, with each subset reviewed by a second coder and disagreements reconciled. We then compared themes across the transcript and comment datasets, treating convergence as attention to the same evaluative

issue and divergence as differing interpretations or valuations of that issue. The preliminary findings draw on 12 randomly selected videos and 1,000 comments. Because long-form transcripts and short comments differ in length and context, the comparison is thematic rather than between equivalent textual units.

4. Preliminary Findings and Discussion

Our preliminary analysis identified four shared themes: the ontology of AI creativity, the locus of creative value, AI as a collaborator, and the structural and ethical conditions of AI creativity. Across these themes, experts were generally pragmatic and adaptation-oriented, whereas public discourse was more unsettled and protective.

4.1. The ontology of AI creativity

Both groups considered whether AI's ability to synthesise and recombine existing patterns constitutes creativity. Experts distinguished human creativity from computational creativity, characterising AI as probabilistic pattern recognition rather than cognitive or affective understanding. As one expert explained, "*Well when IBM created Deep Blue it could beat Kasparov. But it was doing a brute force search through zillions of possibilities. Whereas Kasparov would instantly see a pattern, so our brain is doing something very different*" (VID40). However, this distinction did not necessarily deny AI's creative potential but instead raised questions about distributed creativity, agency, and intentionality in human–AI collaboration. The public used more evaluative language, associating creativity with human experience, emotion, and "soul." One commenter stated, "*Data-driven music regardless of its source is patently soulless garbage masquerading as sound art*" (VID07-C). Others nevertheless viewed human creativity as similarly recombinatorial. Thus, while both groups recognised AI's reliance on pattern synthesis, experts used it to expand creativity's conceptual boundaries, while the public continued to debate whether those boundaries should be expanded.

4.2. The Locus of Creative Value

Both groups found creative value in the production process and the final output. Experts emphasised the struggle, experimentation, and iteration involved in creative work. Although AI can help creators explore possibilities and avoid failure, it may also bypass parts of the struggle through which meaning and mastery develop. Experts therefore viewed AI as potentially transforming the creative process while remaining cautious about what might be lost. Public discourse also valued the creative process but placed greater emphasis on the creator's biographical investment, including accumulated practice, skill, effort, and lived experience. One commenter explained, "*For me, when I see an impressive human piece of art, half of what I am amazed by is the beauty and the other half is the awe of 'wow, a human made that?'… When I look at AI art, I only feel the first half*" (VID17-C). Thus, both groups treated process and effort as sources of creative value, but experts focused on how the process might evolve, whereas the public focused more on the value of human labour and experience.

4.3. AI as a Creative Collaborator

Both groups acknowledged that AI could support creative work, but they differed in how consistently they accepted this role. Experts largely presented AI as a complementary collaborator rather than a replacement for human creators. They described creative outcomes emerging through ongoing interaction between human direction and machine-generated possibilities. As one expert stated, "*We form our own through creating narratives, and now*

we're able to co-create narratives with these machines. So I think that's sort of the path forward" (VID40). From this perspective, AI expands the human creative repertoire and enables new forms of distributed creativity.

Public views were less settled. Some commenters regarded AI as a new creative tool, comparable to Photoshop, that could expand access to creative expression: "*creativity is going to explode now. There are so many people with great ideas but can't really draw, and now this gives them an opportunity to show their ideas to the world and bring them to life*" (VID17-C). Others considered AI illegitimate or threatening to professional creators. Thus, expert discourse presented a relatively stable model of co-creation, whereas public discourse continued to oscillate between democratisation and displacement.

4.4. Structural and Ethical Conditions

Both groups expressed concerns about AI model training, intellectual property, platform power, and changes to creative markets. Experts framed these concerns primarily in institutional and structural terms, emphasising opaque training processes and platforms' growing influence over how creative work is produced, distributed, and recognised. As one expert observed, "*technology always says that it's disintermediating, but it's re-intermediating. It's the platform that's now between you and the audience, and the platform has opinions about who should or shouldn't see what you produce*" (VID22). This perspective supported a reformist orientation centred on copyright rules, platform accountability, and new incentive structures.

Public discourse addressed similar problems but more often used the moral language of theft, fraud, and unconsented labour. For example, one commenter argued, "*if you don't understand people making money utilizing someone else's creation to make money is theft and fraud, then it's your ethic that is in question*" (VID17-C). This emphasised the perceived harm experienced by creators whose work may have been used without consent or compensation. Consequently, while experts focused on redesigning institutions to respond to AI, the public more often emphasised protection and restriction. The two groups recognised similar structural problems but differed in whether their preferred response was primarily reformist or restitutive.

4.5. Evaluative Convergence and Divergence

Our findings suggest three provisional explanations for these orientations: perceived stakes, engagement with AI, and temporal orientation. These are interpretations inferred from the discourse rather than directly observed characteristics of participants. Experts framed the stakes in institutional and intellectual terms, while public comments focused more on personal effort and accumulated skill. Expert discourse showed more sustained engagement with AI's capabilities and limitations, whereas public comments responded more directly to the outputs and debates presented. Experts also emphasised longer-term adaptation, while public comments stressed immediate consequences for creators.

Together, these patterns offer a preliminary explanatory framework: evaluative convergence occurs when groups address shared objects of concern, whereas divergence arises from how those concerns are interpreted through different perceived stakes and temporal orientations. Technical knowledge alone may therefore not resolve the valuation gap; organisations and policymakers must also address concerns about authenticity, effort, recognition, and human agency.

5. Conclusion and Future Research Plan

This exploratory study examines how experts and the public perceive and evaluate human–AI creative work. Preliminary findings show that both groups engage with shared questions

concerning creativity, value, collaboration, and governance but interpret them through different roles, experiences, and perceived stakes. As this study remains in progress, we will extend the analysis to the complete dataset of 48 video transcripts and their associated comments. This analysis will further refine the identified themes and develop a more comprehensive explanation of how creative value is constructed and contested in human–AI collaboration.

6. References

Bellaiche, L., Shahi, R., Turpin, M. H., Ragnhildstveit, A., Sprockett, S., Barr, N., ... & Seli, P. (2023). Humans versus AI: whether and why we prefer human-created compared to AI-created artwork. *Cognitive Research: Principles and Implications*, 8(1), 42.

Benbya, H., Strich, F., & Tamm, T. (2024). Navigating generative artificial intelligence promises and perils for knowledge and creative work. *Journal of the Association for Information Systems*, *25*(1), 23-36.

Bomba, F., & De Angeli, A. (2025). Agency and authorship in AI art: Transformational practices for epistemic troubles. *International Journal of Human-Computer Studies*, 103652.

Chompunuch, S., & Lubart, T. (2025). AI as a helper: Leveraging generative AI tools across common parts of the creative process. *Journal of Intelligence, 13*(5), 57.

Eapen, T., Finkenstadt, D. J., Folk, J., & Venkataswamy, L. (2023). How generative AI can augment human creativity. *Harvard Business Review, 101*(4), 56-64.

Gregor, S. (2025). Being Creative with a Non-Human: The Use of Generative Artificial Intelligence in Art. *Australasian Journal of Information Systems, 29*.

Heimstad, S. B., Wien, A. H., & Gaustad, T. (2025). Machine heuristic in algorithm aversion: Perceived creativity and effort of output created by or with artificial intelligence. *Computers in Human Behavior: Artificial Humans, 5*, 100190.

Horton Jr, C. B., White, M. W., & Iyengar, S. S. (2023). Bias against AI art can enhance perceptions of human creativity. *Scientific Reports, 13*(1), 19001.

Hou, J., Wang, L., Wang, G., Wang, H. J., & Yang, S. (2025). The double-edged roles of generative ai in the creative process: experiments on design work. *Information Systems Research.* Advance online publication.

Huang, S., Long, L., Zhu, Y., & Zhu, J. N. (2026). Human–GenAI collaboration across creative phases: Cognitive mechanisms shaping novelty and usefulness. *International Journal of Information Management, 86*, 102986.

Jia, N., Luo, X., Fang, Z., & Liao, C. (2024). When and how artificial intelligence augments employee creativity. *Academy of Management Journal, 67*(1), 5-32.

Köbis, N., & Mossink, L. D. (2021). Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry. *Computers in Human Behavior, 114*, 106553.

Krakowski, S. (2025). Human-AI agency in the age of generative AI. *Information and Organization, 35*(1), 100560.

Liu, S., Henfridsson, O., Hummel, J. T., & Nandhakumar, J. (2025). The complementor's dilemma: navigating growth ambitions and the dependency on focal actors in platform ecosystems. *MIS Quarterly, 49*(3), 1123-1152.

Memmert, L., Cvetkovic, I., Tavanapour, N., & Bittner, E. (2025). Brainstorming with a generative language model: Effect of exposure to AI ideas on brainstorming performance and cognitive load. *Business & Information Systems Engineering*, 1-25.

Millet, K., Buehler, F., Du, G., & Kokkoris, M. D. (2023). Defending humankind: Anthropocentric bias in the appreciation of AI art. *Computers in Human Behavior, 143*, 107707.

Rae, I. (2024). The effects of perceived AI use on content perceptions. *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, Article 978, 1–14.

Schecter, A., & Richardson, B. (2025). How the role of generative AI shapes perceptions of value in human-AI collaborative work. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, Article 530, 1–15.

Understanding the Influence of System-level Curation Disclosure on the Cognition and Belief in Social Media False News

Jie Sen Lin, Lemuria Carter, Malshika Dias
The University of Sydney Business School, Abercrombie Building (H70), Sydney, NSW 2006, Australia

Abstract
To maximise user engagement, social media platforms serve its users a unique, curated feed of posts which are opaquely determined by AI-algorithms which includes true and false news. Yet the existing empirical literature on false news focuses examining individual post specific design interventions and remains largely unclear on whether disclosing how or why a post was recommended by the system affects their cognition or beliefs of true and false news. Governments are increasingly mandating platforms to provide users additional feed curation options. We design two randomised online experiments to investigate whether system-level recommendation disclosures influence social media users' cognitive discomfort, fast cognition, and belief in true and false news.

1. Introduction

User-generated content on social media is surfaced on an enormous scale without regard for veracity, allowing social media to become a major conduit for false news. False news is posted on social media spreads further and faster than authentic news. Now, the dissemination of false news on social media is accelerated by AI. AI-recommendations occur under opaque black-box processes - often without users' knowledge. AI-algorithms are designed to maximise user engagement to enhance users' marketability and platform monetisation by segmenting social media discourse through the amplification or suppression of posts The prioritisation of engagement over the accuracy of information by AI-algorithms cascades into societal consequences where social media false news is encountered by the vast majority (93%) of users. Global policy responses to the phenomena include the European Union's Digital Services Act which mandates online platforms to disclose the criteria used to determine the recommended information and offer users a recommendation option that is not based on profiling. In response, platforms such as Instagram began offering reverse-chronological feeds.

IS scholars approach the false news phenomena by primarily investigating how individuals cognitively process social media posts. Researchers explore how various post-level design interventions including false news flags, warning signs, stop-sign icons, and accuracy prompts can influence users' cognition and belief intentions (Dennis et al., 2023; Moravec et al., 2020; Moravec et al., 2022; Moravec et al., 2019). The theoretical foundation for these studies of users' cognitive processing of social media posts are dual-route theories, including dual-process theory (Moravec et al., 2020) and the elaboration likelihood model (Miller et al., 2024). Dual-route theories such as the dual-process theory of cognition (Kahneman, 2011) delineate between experiential, default, and automatic System 1 (fast) cognition from rational, deliberate, and effortful System 2 (slow) cognition. While these prior studies focus on foregrounding or disputing post-level attributes such as the source or veracity, research is yet to examine the effects of a system-level recommendation disclosure on users' cognitions of their news feed as a whole.

Individuals respond to AI through appreciation or aversion. Recent studies show that AI-recommended posts do not invoke algorithmic aversion as anticipated by policymakers, signalling a shift in users' perceptions of AI-recommendations (Ma et al., 2025). To date, limited attention has been placed on investigating AI-recommendations as a system-level mechanism, despite its determinative role in dictating users' exposure to news posts. The extant literature does not consider the cognitive effects of disclosing the system which powers the post recommendation and its influence on belief. Social media users might not be fully cognizant that they lack the agency of choosing the posts to consume on social media when system-recommended posts are surfaced to keep them engaged. We theorise that by disclosing to users that a system is curating and surfacing news posts, users may experience a psychological state where they hold two psychologically inconsistent cognitions regarding how they think a post arrives in front of them versus the recommendation disclosure. Cognitive dissonance theory posits that the simultaneous presence of two psychologically inconsistent cognitions produces an aversive state of discomfort that motivates resolution through changes in subsequent behaviour or cognition (Festinger, 1957). According to cognitive dissonance theory, individuals reduce discomfort by adding or removing cognitions, or adjusting the perceived importance of conflicting beliefs (Festinger, 1957; Harmon-Jones & Mills, 2019). Users may resolve discomfort through increasing effortful scrutiny of the news contents that disclosure, or through less effortful routes. In this study, we seek to understand whether users experience different cognitions based on disclosing the system-level recommendation, and its influence on behavioural intentions.

Several IS studies call for research into the influence of AI-recommendations on users' cognition and discernment of false news. Ma et al. (2024) underline the need to explore the theoretical mechanisms behind AI-recommendations by examining the effects of disclosing the role of AI. Ma et al. (2025) call for researchers to study the influence of AI on social media users' cognition, cautioning about the dangers of automatic reliance on AI-recommendations. Liel and Zalmanson (2025) suggest researchers explore the mechanisms that might lead to unjustified deference to AI-recommendations, explicating the need to examine how users respond to AI-recommendations when experiencing cognitive dissonance. Baiyere et al. (2025) call for research to explore mechanisms which trigger a critical mindset to discern news veracity. Our research responds to the above calls and posits the following research question:

RQ: Do system-level recommendation disclosures influence social media users' cognitive discomfort and cognition, and belief in true and false news?

2. Conceptual Development

Cognitive Dissonance Theory

Cognitive dissonance theory posits that individuals possess an innate drive to maintain harmony among their cognitions (Festinger, 1957). Cognitive dissonance is a psychological phenomenon which arises when an individual is confronted by psychologically incompatible cognitions, attitudes and behaviours (Aronson, 1969; Festinger, 1957). The psychology literature delineates cognitive dissonance from the affective reaction it elicits (Cooper & Fazio, 1984; Elliot & Devine, 1994). Cognitive dissonance is experienced as the psychologically uncomfortable state of cognitive discomfort which catalyses an individual to undertake greater psychological exertion to enact cognitive and behavioural change to resolve the aroused inconsistency by adding, removing, or adjusting the importance of their cognitions (Elliot & Devine, 1994; Festinger, 1957; Harmon-Jones & Mills, 2019). Cognitive discomfort can arise from encountering an opposing ideology, or from a lack of confidence in the accuracy of one's beliefs after encountering attitude-challenging information from sources upholding opposing ideologies (Metzger et al., 2020). Stronger levels of dissonance invoke greater discomfort and pressure to resolve it (Harmon-Jones & Mills, 2019).

In IS false news studies, cognitive dissonance theory has been used to examine how individuals dismiss post-specific veracity warnings that cognitively conflicts with a politically resonant false news post (Kim & Dennis, 2019; Moravec et al., 2020). Users who interact more frequently with AI experience higher levels of cognitive fatigue impairs decision making ability (Yang et al., 2025). Ansari and Ghasemaghaei (2026) find that higher levels of human-AI conflict arising from the friction between a users' perspective and the AI output leads to higher levels of cognitive dissonance. Drawing on these findings, we hypothesise: **H1:** Social media users will report higher cognitive discomfort when a post is presented as AI-recommended, compared to reverse-chronology.

False news studies which apply cognitive dissonance theory assume the arousal of dissonance following the intervention rather than empirically measuring the instantaneous psychological discomfort experienced by the user (Moravec et al., 2020). Upon encountering attitude-challenging information, the resulting cognitive dissonance among users is surprisingly weak as users dismiss challenging information by attributing it as noncredible (Metzger et al., 2020). The effects of cognitive discomfort in false news have not previously

been studied when considering non-political false news. Susmann and Wegener (2022) found that individuals' knowledge of false news being retracted can elicit psychological discomfort which is subsequently associated with greater belief in false news. Lutz et al. (2024) find that users experiencing the in situ affective response of cognitive discomfort are more likely to reject novel information as false than to introspectively challenge their existing beliefs. Accordingly, we hypothesise the following: **H2:** Greater cognitive discomfort will be associated with lower belief in news posts (**H2a**), of which the effects will be stronger for false news over real news (**H2b**).

Greater disclosure can produce negative reactions towards AI. Strunk et al. (2025) find individuals are more likely to accept AI-recommended advice in low risk scenarios over high risk severity contexts such as those which concern one's finances. Tong et al. (2021) find a disclosure effect among financial services employees where the disclosure of using AI to generate feedback on employee performance to employees elicited lower trust in the feedback quality and reduced actual job performance. It establishes how AI disclosure triggers an algorithm aversion among individuals, generating a negative psychological resistance against the algorithm's output irrespective of actual quality. This study proposes that the system-level AI-recommendation which discloses the engagement-maximising motive constitutes a dissonance-arousing mechanism. When a user encounters an AI-recommendation disclosure labelled "Recommended by AI to maximise your engagement," we anticipate the activation of two irreconcilable cognitions. The first is the heuristic acceptance from the AI-recommendation itself, which generates the understanding that the post was curated specifically for the user (Ma et al., 2025). The second is the motive-disclosure cognition which is the knowledge that the algorithm is optimising for user retention and virality. This goal-level misalignment foregrounds the tension between users' latent accuracy concern and the platform's engagement maximisation objectives. We hypothesise **H3:** AI-recommended posts with motive disclosure will induce greater cognitive discomfort than AI-recommended posts.

Dual-process theory: System 1 and System 2 Cognition

The dual-process theory of cognition delineates experiential, automatic System 1 (fast) cognition from rational, deliberate, and effortful System 2 (slow) cognition. System 1 cognition embodies the reactions produced by heuristic perceptions and memory (Kahneman, 2011). While System 1 cognition is quick and intuitive, as the default route of cognition, it relies on information saliency to inform decision making and foregoes nuance and considered deliberation (Kahneman, 2011; Moravec et al., 2020). When one's System 1 cognition detects an anomaly, System 2 is invoked (Kahneman, 2011).

The IS false news literature examines how individuals process social media posts, testing the effects of warning and flagging mechanisms in triggering more effortful thinking, and developing techniques to instigate more accurate adjudication of post veracity. Social media users exhibit lower propensity for critical information processing, predisposing them to default to fast cognition (Moravec et al., 2020). Reliance on System 1 fast cognition leads to increased false news sharing and belief in the absence of deliberate elaboration (Pennycook & Rand, 2019; Ma et al., 2025). Kim and Dennis (2019) demonstrate that designing a social media interface which surfaces the article source above the headline invokes users' System 2 processing which reduces post believability irrespective of source credibility. However, AI-recommendations alter user cognition in unexpected ways. This includes reducing social

media users' belief in only true news without affecting belief in false news Ma et al. (2024, p. 632). Ma et al. (2025) demonstrate that explicitly labelling social media stories as recommended by AI exerts unintended social influence akin to a human expert, influencing recipients to undertake fast System 1 cognition and resulting in higher engagement with false news. Rather than inducing critical evaluation, the presence of an AI label suggests the content was computationally optimised for the user, prompting users to withhold effortful System 2 deliberation and accept the recommended content. Hence, we hypothesise:
H4: Greater cognitive discomfort will be associated with lower fast cognition

Three properties of this disclosure differentiate from earlier work content-level warnings. First, it is not appealing to one's confirmation biases and is irresolvable through political rationalisation. Unlike a veracity warning, which can be dismissed as politically biased, motive disclosure is a declarative statement about the platform's optimisation objectives which is applicable universally on social media. Recipients cannot attribute it to the partisan intent of the content itself, because it is equally relevant to all AI-recommended content irrespective of political lean. Users cannot reduce this dissonance by dismissing the label as biased, which should therefore compel a resolution through deliberative evaluation of the post's credibility. Second, it is goal-level rather than content-level, arising from conflict between the recommendation's purpose and users' information goals rather than post veracity. Third, motive disclosure targets the recommendation mechanism rather than the content, circumventing post-specific rationalisations that undermine warning labels (Ma et al., 2024). We predict that AI-recommendations with motive disclosure will help users discriminate false news posts more than a standalone AI-recommendation disclosure by hypothesising:
H5: AI-recommended posts with motive disclosure will elicit lower belief in false news posts compared with an AI-recommended post

3. Methodology

We will undertake two experiments to address our research question. Study 1 uses a mixed experimental design with a between-subjects factor (feed label: "Sorted by newest" and "Recommended by AI") and a within-subjects factor (headline veracity: true vs. false). Each participant is randomly assigned to one label condition and, within that feed, evaluates both true and false headlines. Non-political headlines are independently verified using recognized third-party fact-checking databases (i.e., Snopes). Participants are presented eight social media news posts in the Facebook format and rated their belief. Items for cognitive discomfort about the recommendation were included in situ - immediately after the first randomly displayed post was presented. Participants also answer a manipulation check about the recommendation presented in their assigned condition. Respondents who fail the manipulation check are removed from the study analysis. After completing the task for all eight stimuli, respondents answer questions about their political ideology, social media usage frequency, and demographics. Study 2 is also a mixed between and within-subjects design. In the AI-recommendation condition, participants are presented with posts which were denoted with the standalone "Recommended by AI" recommendations, as in Study 1. Participants in the motive-disclosure condition, view posts which display "Recommended by AI to maximise your engagement". Post veracity remains as the within-subjects factor. The

post stimuli and task are the same as in Study 1. Respondents are assigned to a condition and view the eight posts in randomised order. Cognitive discomfort and the manipulation check were measured after the first displayed post. After rating their belief in all eight stimuli, respondents answered questions on System 1 and System 2 information processing and demographics. For each study, we will recruit five hundred participants from Prolific.

4. References

Ansari, K., & Ghasemaghaei, M. (2026). Partner or rival? How human-Artificial Intelligence conflict shapes artificial intelligence aversion. *Journal of Management Information Systems*, *43*(2), 368-401.

Aronson, E. (1969). The theory of cognitive dissonance: A current perspective. In *Advances in experimental social psychology* (Vol. 4, pp. 1-34). Elsevier.

Baiyere, A., Bauer, J. M., Constantiou, I., & Hardt, D. (2025). Fake news and true news assessment: The persuasive effect of discursive evidence in judging veracity. *MIS quarterly*, *49*(3), 823-860.

Cooper, J., & Fazio, R. H. (1984). A new look at dissonance theory. In *Advances in experimental social psychology* (Vol. 17, pp. 229-266). Elsevier.

Dennis, A. R., Moravec, P. L., & Kim, A. (2023). Search & Verify: Misinformation and source evaluations in Internet search results. *Decision Support Systems*, *171*, 113976. https://doi.org/https://doi.org/10.1016/j.dss.2023.113976

Elliot, A. J., & Devine, P. G. (1994). On the motivational nature of cognitive dissonance: Dissonance as psychological discomfort. *Journal of personality and social psychology*, *67*(3), 382.

Festinger, L. (1957). Stanford University Press. https://doi.org/doi:10.1515/9781503620766

Harmon-Jones, E., & Mills, J. (2019). *An introduction to cognitive dissonance theory and an overview of current perspectives on the theory*. American Psychological Association.

Kahneman, D. (2011). *Thinking, fast and slow* (1st ed.). Farrar, Straus and Giroux.

Kim, A., & Dennis, A. R. (2019). Says who? The effects of presentation format and source rating on fake news in social media. *MIS quarterly*, *43*(3), 1025-1039. https://doi.org/10.25300/MISQ/2019/15188

Liel, Y., & Zalmanson, L. (2025). Turning Off Your Better Judgment: Algorithmic Conformity in Artificial Intelligence-Human Collaboration. *Journal of Management Information Systems*, *42*(4), 1087-1117.

Lutz, B., Adam, M. T., Feuerriegel, S., Pröllochs, N., & Neumann, D. (2024). Affective information processing of fake news: Evidence from NeuroIS. *European Journal of Information Systems*, *33*(5), 654-673. https://doi.org/https://doi.org/10.1080/0960085X.2023.2224973

Ma, H., Dennis, A. S., Dennis, A. R., & Huang, W. (2025). Artificial Intelligence Recommendations Amplify the Sharing of True and Fake News on Social Media by Appealing to Fast Cognition. *Journal of Management Information Systems*, *42*(4), 987-1016. https://doi.org/10.1080/07421222.2025.2561381

Ma, H., Huang, W., & Dennis, A. R. (2024). Unintended Consequences of Disclosing Recommendations by Artificial Intelligence versus Humans on True and Fake News Believability and Engagement. *Journal of Management Information Systems*, *41*(3), 616-644. https://doi.org/https://doi.org/10.1080/07421222.2024.2376381

Metzger, M. J., Hartsell, E. H., & Flanagin, A. J. (2020). Cognitive dissonance or credibility? A comparison of two theoretical explanations for selective exposure to partisan news. *Communication Research*, *47*(1), 3-28.

Miller, S., Menard, P., Bourrie, D., & Sittig, S. (2024). Integrating truth bias and elaboration likelihood to understand how political polarisation impacts disinformation engagement

on social media. *Information systems journal*, *34*(3), 642-679. https://doi.org/https://doi.org/10.1111/isj.12418
Moravec, P. L., Kim, A., & Dennis, A. R. (2020). Appealing to sense and sensibility: System 1 and system 2 interventions for fake news on social media. *Information Systems Research*, *31*(3), 987-1006. https://doi.org/https://doi.org/10.1287/isre.2020.0927
Moravec, P. L., Kim, A., Dennis, A. R., & Minas, R. K. (2022). Do you really know if it's true? How asking users to rate stories affects belief in fake news on social media. *Information Systems Research*, *33*(3), 887-907. https://doi.org/https://doi.org/10.1287/isre.2021.1090
Moravec, P. L., Minas, R. K., & Dennis, A. R. (2019). Fake News on Social Media: People Believe What They Want to Believe When it Makes No Sense At All. *MIS quarterly*, *43*(4), 1343. https://doi.org/10.25300/MISQ/2019/15505
Strunk, J., Nissen, A., & Smolnik, S. (2025). All risks ain't the same–A risk facets perspective on AI-based decision support systems. *Decision Support Systems*, 114557.
Susmann, M. W., & Wegener, D. T. (2022). The role of discomfort in the continued influence effect of misinformation. *Memory & Cognition*, *50*(2), 435-448.
Tong, S., Jia, N., Luo, X., & Fang, Z. (2021). The Janus face of artificial intelligence feedback: Deployment versus disclosure effects on employee performance. *Strategic management journal*, *42*(9), 1600-1631.
Yang, H., Zheng, H., Lu, L., & Hu, P. (2025). Next output is better? Overgeneration effect in generative artificial intelligence interaction. *Information & Management*, 104294.

A Theory-driven LLM Agentic Framework for Emergency Department Triage Decision Support

Iman Izzati Mohamad Firdaus, Hetiao Xie, Morteza Namvar, Rebekah Eden, Saeed Akhlaghpour
The University of Queensland

Abstract

Emergency department (ED) triage is a high-stakes, data-intensive task that requires clinicians to interpret heterogeneous and fragmented clinical documentation and make decisions under severe time pressure and resource constraints, imposing the cognitive challenge and risk of decision oversight. While research has explored artificial intelligence to support ED triage, most existing systems emphasize patient classification prediction rather than assisting clinicians in interpreting clinical information. Drawing on Fuzzy Trace Theory, which explains how clinicians rely on bottom-line gist representations when processing complex information, this study proposes a theory-driven LLM agentic framework that transforms fragmented clinical documentation into gist-oriented summaries. Preliminary analysis and evaluation on a dataset collected from an Australian tertiary hospital's ED suggests that our artifact effectively highlights gist summary and clinical signals while maintaining traceability to verbatim evidence. We propose future work to contribute to a valid theory-driven artifact and design principles for augmenting decision-making in high-pressure clinical settings.

Keywords: ***Theory-driven design***, ***LLM agents***, ***healthcare decision making***, ***fuzzy trace theory***

1. Introduction

Emergency department (ED) triage is a critical gateway in acute care that involves making medical assessments and prioritization under time pressure and uncertainty. Due to resource and time constraints, important patient information is often recorded in a compressed narrative format for clinical documentation (Agius et al., 2025). Clinicians then routinely make decisions based on the fragmented clinical notes, incomplete histories and clinical observations in a resource-constrained environment, which causes cognitive overload and heightened risk of decision oversight (Asgari et al., 2024). This means that the core challenge of triage decisions is not simply documentation quality, but it is the fundamental cognitive challenge of managing information under pressure.

Interventions through decision support system especially AI-based systems, are becoming vital due to their promising capabilities to reduce the cognitive burden of medical personnel (Da'Costa et al., 2025). Reviews

of AI-driven triage show that studies primarily emphasize the classification of ED patients and decision prediction rather than supporting clinicians' interpretation of triage information. While the efficiency and accuracy of the AI systems have shown great potential, decision support tools do not inherently improve decision quality simply by providing accurate predictions; they can be confusing or even detrimental if they fail to support users' cognitive processes (Lee et al., 2008). Vessey & Galletta (1991) highlights that how information is represented can critically affect decision quality, and emerging frameworks such as Adaptive Cognitive Fit calls for systems that can adapt information representation to human cognitive constraints (Samuel et al., 2022). Therefore, we define our research question as: "How can AI-driven decision support systems be designed to mitigate clinician cognitive overload in ED decision making?"

To address this challenge, we propose a theory-driven LLM agentic framework that integrates Fuzzy Trace Theory (FTT) principles to transform fragmented clinical documentation into gist-oriented summaries. FTT is a cognitive theory of judgement and decision-making which states that individuals simultaneously encode information into two forms of representation, which are verbatim (surface details) and gist (bottom-line meaning)(Reyna, 2008). In high-strain environments like ED triage, decision making is therefore guided more by gist interpretation of the clinical situation than by detailed numerical reasoning (Blalock & Reyna, 2016). Grounding the artifact design in FTT provides a principled foundation for structuring how textual clinical information can be transformed to enhance clinically meaningful representations, enabling AI systems to support clinicians' cognitive interpretation rather than just predicting an outcome for them. We provide a detailed elaboration of our designed artifact, followed by preliminary analysis and evaluation.

2. Background and Related Work

2.1. Overcoming Cognitive Challenges in Clinical Decision Making

Clinical decision-making often requires clinicians to interpret complex information under time pressure and uncertainty. To manage these demands, clinicians may rely on heuristics, or mental shortcuts that simplify decisions by focusing on selected information cues (Marewski & Gigerenzer, 2012). While such strategies can be adaptive, they may also introduce cognitive biases when applied inappropriately (Walston et al., 2022). In contrast, FTT distinguishes between verbatim representations, which preserve precise details, and gist representations, which capture the essential meaning of information (Reyna, 2008). Research suggests that reliance on meaningful gist representations increases with expertise and can support effective judgement and decision-making (Evans et al., 2013; Corbin et al., 2015). Hence, a gist-oriented information presentation may offer a theoretical approach to addressing cognitive overload. Edelson and Reyna (2023) argue that presenting extensive medical information without meaningful interpretation can be overwhelming and proposed gist-based communication as a way to convey essential meaning while retaining important details without causing cognitive overload. Although their work focuses on patients interpreting medical information rather than clinicians making clinical decisions, the underlying principle provides the theoretical basis for this study.

2.2. Fuzzy Trace Theory (FTT)

FTT is a cognitive theory of judgement and decision making that proposes individuals encode information in ranges of two parallel mental representations: verbatim and gist (Reyna, 2008). Verbatim representations capture precise surface details, including exact numerical values, textual phrasing and literal descriptions. Gist representations, in contrast, capture the qualitative and functionally significant bottom-line meaning of information. Importantly, FTT posits that although both forms are encoded simultaneously, decision makers, especially experts, tend to rely primarily on gist representations when forming judgments, particularly in complex or risk-laden environment (Chang et al., 2025).

Several core principles of FTT are particularly relevant to clinical decision-making contexts. First, the theory emphasises the parallel encoding of verbatim and gist representations, but suggested decisions are typically guided by gist rather than exhaustive analysis of details (Blalock & Reyna, 2016). Second, FTT highlights a preference for categorical distinctions in reasoning; individuals often reduce complex

information into meaningful contrasts such as safe versus unsafe, or high versus low risk (Corbin et al., 2015). Third, excessive emphasis on numerical precision, such as highlighting exact numerical probabilities, may cloud judgement rather than enhance it, particularly when such precision exceeds what is necessary for action. In this sense, FTT suggests that effective communication and decision support should help individuals form appropriate gist-representations rather than focus solely on detailed numerical outputs (Wolfe et al., 2019).

In the context of ED triage, these principles are especially important. Clinicians must rapidly interpret fragmented narrative documentation and vital sign information under cognitive strain, translating verbatim cues into categorical understandings of patient status (Laker et al., 2018). Rather than calculating precise probabilities of deterioration, triage often involves identifying bottom-line meaning, whether the patient appears stable or unstable, or whether the risk is minimal or clinically significant (Agius et al., 2025). Grounding the artifact design on FTT, therefore, provides a framework for structuring how free-text clinical information can be transformed into cognitively aligned representations. By separating verbatim extraction from gist derivation, and prioritizing qualitative meaning over numerical precision, the proposed system shifts the focus of AI support from outcome generation to representation structuring, strengthening the interpretive foundation upon which triage judgements are made.

2.3. AI-Driven ED Triage Decision Support

Researchers have explored various AI-driven decision support approaches to enhance clinical judgment and workflow efficiency. Gholipour et al. (2025) found that robust decision support tools have helped triage nurses make quicker and more accurate assessments, improving both task performance and patient outcomes. Various works in decision support tools have been examined for ED triage that covers a broad application of AI models; from supervised machine learning models to NLP-based models and LLM-based systems. However, in most existing systems, these capabilities are primarily used to support downstream classification tasks rather than to assist clinicians in interpreting clinical information (Lee et al., 2026). LLM-based systems, however, are more versatile and have broader functionality as they extend beyond prediction toward reasoning and interaction. Unlike traditional NLP approaches, LLMs are capable of question answering, clinical simulation, handles multi-agent conversation and orchestrate workflows (Karamanlıoğlu et al., 2025; Yang et al., 2025; Yu et al., 2025).

LLM agentic frameworks are autonomous or semi-autonomous systems that consist of agents that can make decisions and take actions in dynamic and complex environments. Driven by the advancement of LLM, these agents are programmed to dynamically interpret tasks, orchestrate tool use, and adapt to real-time environments to achieve directed goals. The literature highlights the capabilities of LLM agents to augment healthcare decision makers in various tasks such as appointment scheduling, triage classification, and medical question answering (Qiu et al., 2024; Pieper & Gleasure, 2025). The literature typically employs multi-agent collaboration, where agents assume specialized roles, debate interpretations, or simulate clinical personas to arrive at a final recommendation (Wang et al., 2026). The emphasis is often placed on interaction patterns, orchestration strategies, and adaptive collaboration mechanisms that mirror team-based clinical practice. As for final outputs, it is often produced through conversational aggregation, arbitration, or moderation between agents, with the underlying rationale implicitly embedded within prompts or dialogue flow rather than systematically articulated in the structure of the information presented (Qi et al., 2025). As a result, existing work tends to frame LLM agentic frameworks primarily as a coordinated and interactive clinical simulation rather than as a cognitive simulation of how clinical information should be structured and processed. In high-stakes settings such as ED triage, where clinicians must rapidly interpret compressed and uncertain information, this raises a critical gap that informs our primary design objective: to design an LLM agentic framework that not merely simulates collaborative reasoning, but generates structured, gist-oriented representations grounded in cognitive theory that align with clinician thinking processes.

3. Research Design

3.1. Data Collection and Preparation

In this research, we collaborated with a major public tertiary teaching hospital in Australia. After obtaining research ethics clearance, we accessed their dataset from the electronic medical record system comprising 161,455 non-identifiable records of adult patient presentations, spanning the period from July 2023 to July 2025. We randomly sampled 1,000 presentations for preliminary analysis in this research-in-progress study. From the dataset, we selected variables available during triage, including categories such as demographics, administrative data, triage information, chief complaint and presenting problem descriptions, and clinical notes. These elements were integrated via a rule-based module (i.e., a component of our artifact) to construct the structured input for the subsequent LLM agentic framework.

3.2. The Design of LLM Agentic Framework

Inspired by the theory-driven design research (Maass et al., 2018; Yang et al., 2023), our study proposes a LLM agentic framework grounded in the principles of FTT to support clinician interpretation during ED triage (see Figure 1). The framework operationalizes FTT by separating the processing of clinical information into parallel representational stages. Specifically, the system transforms heterogeneous clinical inputs into two complementary representations: verbatim representations, which preserve precise factual details, and gist representations, which capture the bottom-line clinical meaning of the case. The architecture consists of four agents arranged in stages of cognitive transformation pipeline. Rather than generating predictions or triage labels, the agents collaboratively transform fragmented clinical documentation into a gist-oriented summary that highlights bottom-line interpretation, supporting evidence, and risk indicators. The architectural specification of this agentic framework is presented as follows:

Data Preparation Module is a rule-based preprocessing component. Its role is to concatenate the existing heterogeneous data elements and perform basic data cleaning and preparation based on our predefined rules. The objective is to construct natural language inputs compatible with LLMs, ensuring consistent interpretation across downstream agents.

Verbatim Extractor Agent identifies and extracts high-fidelity clinical information from the prepared input. It is designed to identify and isolate information such as demographics, symptomatic descriptions, and physiological metrics from highly heterogeneous medical datasets. In line with FTT, the agent adheres to a strict non-interpretative mandate to produce verbatim representation, which explicitly prohibits any form of medical inference, clinical interpretation, or lexical transformation of the source text.

Gist Generator Agent adopts a clinical expert persona, prioritizing clinical gestalt and high-level gist extraction over rote data recording. Rather than reproducing detailed observations, the agent synthesizes the extracted information to capture the bottom-line meaning of the clinical situation. Consistent with FTT, this stage produces gist representation, where a concise categorical description (e.g., severe mental crisis) and latent risk associated with patient presentation are generated without introducing diagnoses.

Dual-process Validator Agent acts as an independent external arbiter. Leveraging its extensive medical knowledge, it performs factchecking on the outputs from both the Verbatim Extractor and the Gist Generator in a blind-review manner, without direct exposure to the raw source data. Should a significant clinical discrepancy be detected between the two outputs, the Dual-process Validator intercepts the information flow. Instead of proceeding to the next stage, it generates iterative feedback as a remedial prompt, guiding both agents to re-evaluate and recalibrate their reasoning. When the Validator approves, it will pass the original information generated by Verbatim Extractor and Gist generator to the next agent. This mechanism is specifically designed to mitigate hallucinations and minimize clinical errors through multi-agent cross-verification.

Decision Framing Integrator Agent synthesizes verbatim and gist representations into a clinical summary tailored for clinicians, aimed at reducing their cognitive load to help them make decisions more easily. Rather than merely aggregating information, this agent is grounded in FTT to prioritize the most essential

bottom-line meaning, thereby guiding clinicians toward rapid and correct intuitive decision-making paths. This is achieved through: (1) Truncation and De-noising: Stripping away all redundant background noise to retain only the core information essential for emergency triage. (2) Cognitive Chunking: Reconstructing complex medical information into easy-to-understand 'chunks,' ensuring that information delivery aligns with the bandwidth limits of human information processing. The integrator also prioritizes a gist-first presentation structure. The summary begins with a concise bottom-line interpretation of the case, followed by qualitative descriptions that capture the overall clinical meaning. These interpretations are then complemented with verbatim evidence such as numerical measurements or relevant excerpts from the original documentation. This layered structure allows clinicians to quickly grasp the essential meaning of the case while maintaining visibility into the underlying clinical details.

4. Preliminary Findings

To implement our LLM agentic framework on the collected ED dataset for preliminary analysis and evaluation, we utilized a locally deployed open-source model, Qwen-3-32B-Medical-Reasoning as the base model. It is a domain-specific model with logical reasoning framework specifically optimized to simulate clinical diagnostic reasoning when processing complex medical narratives.

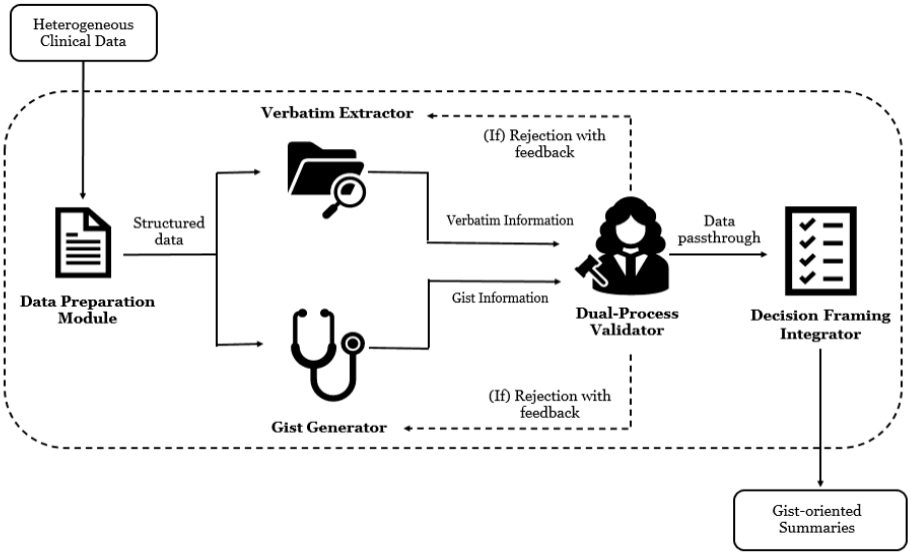


Figure 1. LLM Agentic Framework for ED Triage Decision Support

We conducted a computational evaluation of the degree of gist in the generated summaries. We adopted the GisPy approach proposed by Hosseini et al. (2022), which quantifies the degree to which a text facilitates gist inference by aggregating multiple linguistic dimensions, such as referential and deep cohesion and psycholinguistic attributes. In this approach, higher Gist Inference Scores (GIS) indicate stronger support for gist inference. Previous studies have applied automated gist scoring to evaluate the comprehensibility of medical texts, particularly in patient-facing cancer communication (Wolfe et al., 2019). Building on this work, the present study applies GIS in a different context: evaluating clinical summaries generated through a theory-driven LLM agentic workflow. Rather than proposing a new measure of gist comprehension, our contribution lies in exploring how FTT can inform the design of AI-generated information representations for clinical triage. GIS serves as an initial computational indicator of whether the generated outputs exhibit characteristics that may facilitate gist-based understanding.

For comparative analysis (Table 1), we used the raw clinical information corresponding to our generated summaries as the baseline and further included summaries generated by the same base model in a non-optimized (i.e., zero-shot) configuration. The result reveals that our artifact achieves the highest gist score of 1.354 while maintaining the lowest standard deviation of 0.492. These results demonstrate that while our

artifact effectively presents the gist of the information, it also displays stability throughout the generative reasoning process.

Table 1. Comparative Gist Inference Score (GIS) Performance

Method	GIS (mean)	GIS (standard deviation)
Raw Clinical Information	-1.936	0.507
Summary (Qwen3-32B-Medical-Reasoning, Zero-Shot)	0.968	0.628
Gist-oriented Summary (Our artifact)	**1.354**	**0.492**

We then analysed the generated outputs (Figure 2) qualitatively through an initial round of design feedback with clinicians. The analysis suggests that the LLM framework produces summaries consistent with the principles of FTT and was testified by clinician as an improved document as compared to the original notes. The output begins with a concise bottom-line interpretation that captures the overall clinical concern, allowing clinicians to rapidly grasp the central situation before examining detailed observations. The use of relative and categorical distinction terms such as “severe” and “high risk” is also a consistent trait of gist as presented in FTT. The statement is then followed by structured supporting evidence presented in ‘chunks’, highlighting the clinical cues that contextualizes the first statement. Next, specific and noisy details like exact timings were truncated into meaningful statements, also consistent with the notion of FTT. Lastly, rather than presenting isolated numerical values or fragmented notes, the system organizes the observations into short, clearly separated statements and retains only relevant numerical support evidence that aligns with clinical claims. This structure illustrates how fragmented triage documentation can be transformed into a coherent narrative representation where gist interpretation is immediately visible and supporting details remain traceable, thus supporting clinicians to make better triage decision.

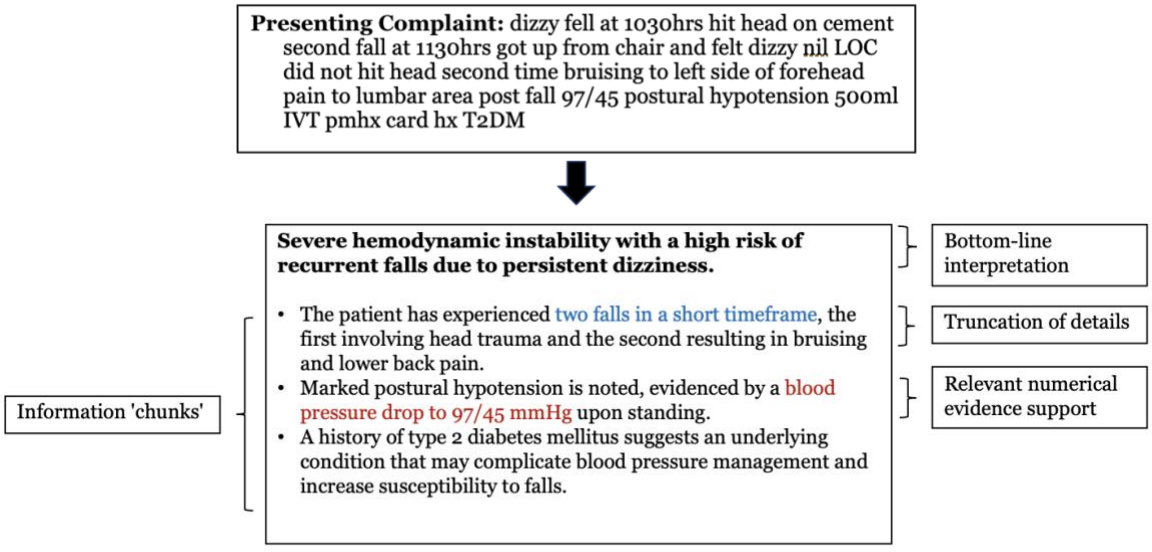


Figure 2. An Illustration of Gist-oriented Summary

5. Conclusion and Future Work

This study demonstrates the potential of an LLM agentic framework to embody the principles of Fuzzy Trace Theory, a foundational cognitive theory describing how individuals process complex information through both verbatim and gist-based meaning. By structuring clinical inputs through collaborative agentic framework transformations, the proposed architecture illustrates how AI systems may be designed to better align with clinicians’ cognitive processes. Our preliminary study explores its use in ED triage cases, and we find that this approach can highlight clinically meaningful signals while maintaining factual traceability to the

underlying observations. Although GIS provides a theory-informed measure of gist-related textual characteristics, it does not directly capture the cognitive workload experienced by clinicians. A higher GIS score does not necessarily indicate that a summary is clinically accurate, actionable, or easier to process during triage. Therefore, the current findings should be interpreted as preliminary evidence concerning the framework's ability to produce gist-oriented summaries, rather than evidence of reduced cognitive overload. Future evaluations involving clinicians will be necessary to determine whether these characteristics translate into reduced cognitive demands and improved support for clinical decision-making.

A two-phase evaluation will be conducted to further refine and assess the proposed framework. First, we will collect qualitative feedback from clinicians through semi-structured interviews and focus groups to explore how practitioners interpret the generated summaries and whether it align with their thinking processes during triage. Insights from these sessions will inform improvements to the system's design, particularly in how information is structured and presented to support clinical decision-making. Second, following these design refinements, we will conduct an ablation analysis to systematically evaluate the contribution of individual agents within the proposed workflow. By comparing the complete agentic framework against configurations with selected agents removed, we aim to assess how different components contribute to the quality and effectiveness of the generated summaries. Findings from both phases will support iterative refinement of the artifact and inform broader design principles for cognitively aligned AI decision support. Through this research, we aim to contribute a carefully designed LLM agentic framework that mitigates the cognitive overload experienced by clinicians when processing fragmented and heterogeneous data during the urgent and complex process of triage. Furthermore, we seek to establish design principles for theory-driven LLM agents in healthcare decision-making through iterative design and evaluation.

6. References

Agius, S., Magri, C., & Cassar, V. (2025). A Cognitive Task Analysis for Developing a Clinical Decision Support System for Emergency Triage. *Journal of Emergency Nursing*, *51*(6), 1028-1045.e4.

Asgari, E., Kaur, J., Nuredini, G., Balloch, J., Taylor, A. M., Sebire, N., Robinson, R., Peters, C., Sridharan, S., & Pimenta, D. (2024). Impact of Electronic Health Record Use on Cognitive Load and Burnout Among Clinicians: Narrative Review. *JMIR Medical Informatics*, *12*(1), e55499.

Corbin, J. C., Reyna, V. F., Weldon, R. B., & Brainerd, C. J. (2015). How reasoning, judgment, and decision making are colored by gist-based intuition: A fuzzy-trace theory approach. *Journal of Applied Research in Memory and Cognition*, *4*(4), 344–355.

Da'Costa, A., Teke, J., Origbo, J. E., Osonuga, A., Egbon, E., & Olawade, D. B. (2025). AI-driven triage in emergency departments: A review of benefits, challenges, and future directions. *International Journal of Medical Informatics*, *197*, 105838.

Edelson, S. M., & Reyna, V. F. (2023). Providing the Gist of Medical Expertise in the Context of Laws, Rules, and Guidelines: Fuzzy-Trace Theory's Alternative Approach to Improve Patient Communication. *Journal of Law, Medicine & Ethics*, *51*(3), 703–707.

Evans, K. K., Georgian-Smith, D., Tambouret, R., Birdwell, R. L., & Wolfe, J. M. (2013). The Gist of the Abnormal: Above chance medical decision making in the blink of an eye. *Psychonomic Bulletin & Review*, *20*(6), 10.3758/s13423-013-0459–3.

Gholipour, M., Dadashzadeh, A., Jabarzadeh, F., & Sarbakhsh, P. (2025). *Challenges of Clinical Decision-making in Emergency Nursing: An Integrative Review*.

Gorick, H. (2022). Factors that affect nurses' triage decisions in the emergency department: A literature review. *Emergency Nurse: The Journal of the RCN Accident and Emergency Nursing Association*, *30*(3), 14–19.

Karamanlıoğlu, A., Demirel, B., Tural, O., Doğan, O. T., & Alpaslan, F. N. (2025). Privacy-Preserving Clinical Decision Support for Emergency Triage Using LLMs: System Architecture and Real-World Evaluation. *Applied Sciences*, *15*(15).

Lee, M.-H., Huang, T.-Y., Yang, P.-Y., Chang, Y.-C., & Hsu, M.-H. (2026). Optimizing ED patient disposition predictions through clinical narratives with advanced pre-trained language models. *Health Information Science and Systems*, *14*(1), 34.

Lee, Z., Wagner, C., & Shin, H. K. (2008). The effect of decision support system expertise on system use behavior and performance. *Information & Management*, *45*(6), 349–358.

Maass, W., Parsons, J., Purao, S., Storey, V. C., & Woo, C. (2018). Data-driven meets theory-driven research in the era of big data: Opportunities and challenges for information systems research. *Journal of the Association for Information Systems*, *19*(12), 1.

Marewski, J. N., & Gigerenzer, G. (2012). Heuristic decision making in medicine. *Dialogues in Clinical Neuroscience*, *14*(1), 77–89.

Morphet, J., Griffiths, D. L., Innes, K., Crawford, K., Crow, S., & Williams, A. (2014). Shortfalls in residents' transfer documentation: Challenges for emergency department staff. *Australasian Emergency Nursing Journal*, *17*(3), 98–105.

Phang, S. H., Ravani, P., Schaefer, J., Wright, B., & McLaughlin, K. (2015). Internal Medicine residents use heuristics to estimate disease probability. *Canadian Medical Education Journal*, *6*(2), e71–e77.

Reyna, V. F. (2008). A Theory of Medical Decision Making and Health: Fuzzy Trace Theory. *Medical Decision Making*, *28*(6), 850–865.

Samuel, J., Kashyap, R., Samuel, Y., & Pelaez, A. (2022). Adaptive cognitive fit: Artificial intelligence augmented management of information facets and representations. *International Journal of Information Management*, *65*, 102505.

Vessey, I., & Galletta, D. (1991). Cognitive Fit: An Empirical Study of Information Acquisition. *Information Systems Research*, *2*(1), 63–84.

Walston, Z., Whelehan, D. F., & O'Shea, N. (2022). Clinical decision making in physical therapy – Exploring the 'heuristic' in clinical practice. *Musculoskeletal Science and Practice*, *62*, 102674.

Wolfe, C. R., Dandignac, M., Sullivan, R., Moleski, T., & Reyna, V. F. (2019). Automatic Evaluation of Cancer Treatment Texts for Gist Inferences and Comprehension. *Medical Decision Making*, *39*(8), 939–949.

Yang, T., Xiao, Y., Bao, Z., Hao, J., & Peng, J. (2025). The rise and potential opportunities of large language model agents in bioinformatics and biomedicine. *Briefings in Bioinformatics*, *26*(6), bbaf601.

Yang, Y., Qin, Y., Fan, Y., & Zhang, Z. (2023). Unlocking the power of voice for financial risk prediction: A theory-driven deep learning design approach. *MIS Quarterly*, *47*(1), 63-96.

Yu, H., Zhou, J., Li, L., Chen, S., Gallifant, J., Shi, A., Sun, J., Li, X., He, J., Hua, W., Jin, M., Chen, G., Zhou, Y., Li, Z., Gupte, T., Chen, M.-L., Azizi, Z., Dou, Q., Yan, B. P., … Fan, L. (2025). Simulated patient systems powered by large language model-based AI agents offer potential for transforming medical education. *Communications Medicine*, *6*, 27.

Dynamic Data Capabilities for Emergency and Disaster Management: Evidence from the Australian Public Sector

Olga Vasilyeva
Research School of Management
The Australian National University, Canberra, Australia
Email: Olga.Vasilyeva@anu.edu.au

Patrick J.N. L'Espoir Decosta
Research School of Management
The Australian National University, Canberra, Australia
Email: Patrick.LEspoirDecosta@anu.edu.au

Alex Richardson
Research School of Management
The Australian National University, Canberra, Australia
Email: Alex.Richardson@anu.edu.au

Abstract

As natural disasters and health emergencies intensify globally, government agencies face mounting pressure to leverage data-driven approaches for effective emergency response. This study examines which data capabilities are critical to decision-making during emergencies and how data analytics and dynamic capabilities enhance emergency management outcomes. Drawing on Dynamic Capability Theory (DCT) and the Analytics Empowerment Capability (ANEC) framework, nine semi-structured interviews were conducted with data and policy professionals across Australian Public Sector agencies experienced in managing the COVID-19 pandemic, bushfires, and floods. Applying Strauss and Corbin's coding paradigm in NVivo, findings reveal five critical capability dimensions - data, technology, people, culture, and leadership-mapped across three dynamic capability routines: sense-making, connecting, and shaping. Several enabling organisational conditions are identified, including staff motivation and a flattened organisational structure. Three propositions and a conceptual framework are developed, extending DCT and ANEC, and offering actionable guidance for building resilient public sector capabilities.

1. Introduction

This paper examines how agency data capabilities influence the effective application of data in public-sector decision-making, with a particular focus on the management of emergency and disaster events. To theorise this problem, the study draws on dynamic capability theory (DCT), which holds that organisations can enhance and renew their competencies and resources to better position themselves in rapidly changing environments (Teece et al., 1997). Information Systems (IS) research has established how IT capabilities and data assets translate into organisational performance (Wamba et al., 2017; Klievink et al., 2017), yet the specific dynamic capability configurations required for emergency contexts remain limited.

The Australian Public Sector (APS) provides the empirical context for this study. Australia's exposure to a spectrum of disasters and emergencies, including bushfires, floods, earthquakes, cyclones, and COVID-19, makes it a critical case for studying emergency data capability deployment (Geoscience Australia, 2025). The 2019–20 Black Summer bushfires caused AUD $4–5 billion in agricultural and AUD $2.8 billion in tourism losses (WWF-Australia, 2021). The 2010–11 Queensland floods resulted in AUD $6.8 billion in direct losses (Akter & Wamba, 2019); and COVID-19 required AUD $343 billion in government support (ANAO, 2024), with 13,021 deaths by October 2022 (ABS, 2022). This record of significant exposure and varying response effectiveness makes the APS a theoretically justified context for this study. These observations led to the following research questions:

RQs: What data capabilities and organisational conditions are most critical to public sector response during emergencies and disasters? How can data analytics capabilities and dynamic capabilities help the public sector enhance the management of emergency and disaster events?

This paper contributes theoretically by extending DCT and the Analytics and Network Empowerment Capability (ANEC) framework (Akter et al., 2021) into the public sector emergency management domain, identifying two novel sub-capabilities absent from existing frameworks. Furthermore, we study which organisational conditions enable capability development in emergency and disaster management events. These events are often characterised by urgency, uncertainty, rapidly evolving information, and the need for quick capability uplift within an organisation. To manage them effectively and efficiently, organisations need to reconfigure and transform capabilities, including scaling up resources, extending relationships, and improving information flow. The second contribution of this paper is to explain new mechanisms for capability development in emergency and disaster management. Practically, the study offers government data leaders, and emergency managers a grounded framework for assessing and investing in the capabilities most likely to enhance data-driven decision-making when emergencies arise.

2. Research Design

This study is grounded in an interpretivist epistemological stance, which holds that social reality is constructed through the meanings and interpretations of those who experience it, rather than being objective and pre-given (Klein & Myers, 1999; Walsham, 1995). Data capabilities in emergency management cannot be adequately studied through measurement alone; they are enacted through the perceptions, decisions, and practices of practitioners operating under conditions of urgency and uncertainty. The study therefore seeks to understand the subjective meanings that experienced public servants attach to capability requirements, consistent with Klein and Myers' (1999) principles for interpretive IS research, including contextualisation, the

hermeneutic circle, and researcher-subject interaction. Ontologically, capability configurations are treated as socially produced and context-dependent rather than universal and fixed.

Dynamic Capability Theory functions here as a sensitising framework (Bowen, 2006), providing an initial conceptual vocabulary of sense-making, connecting, and shaping routines to guide data collection while remaining open to inductive modification. Semi-structured interviews were selected as the primary method because the research questions are interpretive and process-oriented, by asking what capabilities matter and how they operate, and requiring access to practitioner experience rather than quantifiable variable relationships (Myers & Newman, 2007).

3. Findings

3.1. Capabilities and conditions for sense-making

Sense-making routines encompass the analytical, assessment, and information-gathering processes that enable agencies to detect changes in their environment, appraise their significance, and generate new knowledge to inform decisions (Kattel, 2022). Seven out of ten participants identified data (including completeness, availability, timeliness, and richness), technology (data analytics, availability of the right technical solutions, predictive analytics, modelling tools, and interoperability), and people capabilities as the foundational mechanisms through which sense-making is activated during emergency events. We identified that these capabilities assist with information gathering by identifying and assessing impact, performing monitoring, assisting with policy decision-making, reviewing the effectiveness of the policy, and assessing citizens' attitudes towards the emergency and disaster response implemented. Furthermore, they assist in identifying policy options, predicting outcomes, evaluating compliance with policy directions, and identifying the most affected citizen cohorts, which is beneficial for prompt responses to disasters and emergencies.

Data completeness and richness, availability of analytical tools and data specialists are essential for this analysis. We identified that key technological capabilities include the use of predictive tools where varied analytical assessments were conducted for different policy interventions to predict the outcomes and identify possible options. In addition, data interoperability was mentioned by one of the participants as data can be fragmented and high volume as well as data timeliness:

One of the challenges … the barriers to entry data alone … interoperability is a big challenge… the data volumes high…it's a very fragmented landscape (Participant 4, Director).

We found evidence that sense-making routines are developed by a need to understand the seriousness of disasters or emergencies, identifying possible ways to minimise its negative impact, and provide visibility of attitudes toward government decisions among citizens and lobby groups. Understanding of attitudes, behaviours and evaluating policy compliance assisted in determining government policy.

We're capturing the impact that … businesses had within those regions because of that disaster. … we need to see how long they're predicted to be closed for; how many staff they had to had to let go because of it; what have they lost (Participant 10, Senior Data Analyst).

Both citizen and economic impacts are evaluated through data analytics and presented to key policy decision-makers. Furthermore, data analysis goes much further than simply monitoring through data analytics, it also looks at the ways to minimise negative impacts to other areas of citizens' lives to set direction for policy appropriate policy responses.

The identified capabilities are present in IS, but partly in Emergency Management literature. For example, the importance of data availability and completeness was outlined in IS research by Mikalef et al. (2016) but not mentioned as part of ANEC capabilities by Akter et al. (2021), which could be a notable gap in ANEC's capability framework. Secondly, technology capability is widely covered by IS literature, including data analytics (Lyytinen & Grover, 2017), use of prescriptive analytics (Eamon & Thayer 2025), system integrations (Sejahtera, 2018), connectivity and modularity (Akter et al., 2021; Mikalef et al., 2016). In comparison, the ANEC capabilities in technology and tools by Akter et al. (2021) are mostly focused on predictive analytics to identify future prospects and risks related to emergency events, as well as on access to information. Finally, capability related to people is covered by both data analytics and ANEC capabilities around having the right people with the right skills (Rizzon et al., 2018; Wamba et al., 2015) and ensuring that employees' skills cover the ability to interpret data, apply business processes and evaluate outcomes (Motamarri et al., 2020; Akter et al. 2021).

Proposition 1: *To meaningfully enhance public sector agencies' capacity to detect, interpret, and act on information during emergency and disaster events, sense-making dynamics capabilities are required which encompassing data, people, and technology. Those capabilities can be enacted by a need to minimise negative impact, understand seriousness of upcoming emergency and overall attitudes of public.*

3.2. Capabilities and conditions for connecting routines

Connecting routines are defined as establishing networking and boundary-spanning routines to enable the development of new networks/coalitions of internal and/or external stakeholders, thereby building legitimacy and buy-in for new solutions (Kattel, 2022). The application of data analytics and different capabilities for connecting routines was the most-discussed area among the interviewed participants.

The greatest emphasis was on data professionals collaborating with other stakeholders. This includes data professionals working with key policy decision-makers, executives, operational staff and/or other agencies to identify core requirements and data needs for policy decisions, co-design possible solutions, and test ideas. These activities enable smooth transition, change management, and overall acceptance of the proposed solution across the agency. Furthermore, this helps to deliver better solutions that meet key needs and address pain points for key stakeholders.

Several important conditions have been identified that enables an organisation to establish connecting routines. These conditions include flattening organisational hierarchical structure, and having access to the right people. The importance of cooperation to access appropriate data resources was evident not only with internal stakeholders but also with external stakeholders and agencies:

I would say there was a good cooperation between different stakeholders and different sectors … in terms of the really wanted to cooperate and provide access to data engineers or data on the list to use for decision making of the Department of the Prime Minister and Cabinet or the Department of Health. That was the good thing that was the capability (Participant 7, Consultant).

Having a limited professional relationship with key external decision makers also limits the opportunity to explain the problem holistically:

Ideally it would be great to have a relationship there with the Minister's office to actually talk through the problem and get them to understand the real problem, not what is perceived by a loud voice (Participant 3, Director).

Based on the IS literature, data professionals working with other stakeholders are part of a data culture related to collaboration (Sejahtera, 2018), in which domain experts and data scientists are encouraged to work collaboratively (Eamon & Thayer, 2025). Secondly, data

sharing was the next most popular aspect of data culture discussed by participants, which can enable agencies to address disaster and emergency events despite a shortage of resources.

Although data sharing is not covered by ANEC capabilities, its importance is recognised within data analytics capabilities (Garattini et al., 2019; Vasilyeva & Richardson, 2022). We suggest that data sharing be further included as a part of the data culture sub-capability for emergency and disaster management. Following the discussion, several participants identified that data sharing and collaboration can be enabled by a common platform to streamline disaster and emergency management. This was covered again in the IS literature (e.g., Vasilyeva & Richardson, 2022), but having a single platform is not within ANEC's capabilities.

Recalibrating what a new norm and flattening the organisational structure to allow free flow of key information, which has not been covered in ANEC capability literature:

Government is typically quite hierarchical, unless you are the same level with someone, sometimes they don't want to talk to you. During COVID, that all dissolved because who cares about if somebody is a Deputy Secretory talking to an EL1 (Assistant Director) … you just call them because …you've got a cabinet meeting that afternoon and this EL1 is the only person that's got the answer to things. So, it was also kind of nice not to have that hierarchy be a blocker to you (Participant 9, Consultant).

According to disaster management research, hierarchical structures contribute to poorly integrated disaster response (Burkle et al. 2001). Flattening vertical agencies enhances their capabilities in collaboration, information sharing, and coordination during disaster events (Burkle et al. 2001). The organisational structure should be neither highly horizontal nor strictly centralised but a moderate core-periphery structure for effective disaster governance (Nowell et al. 2018). As a result of our analysis, the following proposition is developed:

Proposition 2: *To enhance public sector agencies' capacity to build the inter-organisational relationships and shared understanding for coordinated emergency and disaster response, connecting dynamic capabilities are required, encompassing collaborative data culture, cross-boundary information sharing, common data platform , and co-design practices. Those capabilities can be enacted by flatten organisational hierarchies, and access to the right people for coordinated emergency and disaster response.*

3.3. Capabilities and conditions for shaping routines

Shaping routines encompass the processes through which agencies designand implement new organisational directions, embed adaptive practices into long-term routines, and mainstream novel solutions into institutional norms (Kattel, 2022).We identified common capabilities with sense-making, especially around data (including completeness, availability, timeliness, and richness), technology (data analytics, predictive analytics, modelling tools, and interoperability), and the availability of data talent.

Based on the analysis of the interviews, several additional capabilities become important for the management of disasters and emergencies as part of shaping routines, including leadership (executive sponsorship/championship, sufficient investment) and staff motivation to go the extra mile during disaster and emergency events. Leadership, in comparison, is covered by both ANEC and data analytics capabilities. Akter et al. (2021) outlined the importance of an influential resource that recognises and assists in addressing service demands effectively as part of the analytics climate. Leadership includes executive sponsorship (LaValle et al. 2011) and sufficient investment (Gupta & George, 2016).

We identified several enabling conditions for capability reconfiguration and transformation to shape routines during disaster and emergency management. These include the ability to scale up team size, sometimes several times its original size, often relying on pre-existing networks,

accelerating data sharing, ensuring high staff motivation, enhance agility and cutting red tape to accelerate the response.

They[Executives] knew which consultancy has these type of specialists, or the other consultancy has the other sort of specialists…They use those networking and relationship to build those core teams very fast (Participant 7, Consultant, large agency).

The high time pressure meant people had to work extended hours and meet management's demands to deliver analytical products within hours. To meet high demands of the emergency event, having motivated staff members that are willing to contribute and go the extra mile is a key to assist with policy development and implementation that require a prompt response.

The identified flattening of hierarchical organisational structures is another important enabling condition for shaping routines that were identified during interviews and supported by existing disaster management literature. Flattening organisational structures improves information flows (Burkle 2001; Nowel 2018) and increases adoptability (Older 2022). Taken together, both conditions shape how capabilities are enacted under emergency pressure. Staff motivation supports the mobilisation and application of analytical expertise, while a flattened organisational structure facilitates direct access to expertise and reduces hierarchical barriers to information exchange. These conditions are therefore distinguished from the capability dimensions they enable; they help explain how existing capabilities are put into practice, rather than constituting additional capability dimensions.As a result of our analysis, the following proposition is developed:

Proposition 3: *To enhance public sector agencies' capacity for adaptive data practices and effective decision-making across the full emergency management cycle, shaping dynamic capabilities are required, encompassing leadership, data, technology, and workforce availability. Those capabilities can be enacted by the leadership team's ability to scale up, enhance agility, staff motivation, ability to leverage pre-existing professional networks , and flattened organisation structure.*

4. Framework

The following diagram illustrates changes in conditions under which capabilities were reconfigured and transformed across three stages of emergency management: pre-emergency, during emergency, and after emergency. During an emergency, there is often high time pressure, fragmented and uncertain information, and an expected quick turnaround. To maximise and transform existing capabilities for better disaster and emergency management, cross-agency collaborations intensify, data sharing accelerates, organisational hierarchy temporarily compresses, and leadership mobilises resources rapidly. However, once an emergency passes, maintaining those capabilities at maximum is unsustainable due to limited resources and reduced demand. As a result, workforce size returns to its original level, along with the pre-emergency organisational hierarchy. At the same time, some established practices remain, including contentious data sharing, improved data analytics, and collaborations for other initiatives.

The findings across the three routines are integrated in Table 1. The Framework for Enhanced Emergency and Disaster Management organises organisational enabling capabilities and the five capability dimensions (data, technology, people, culture, and leadership) across the three dynamic capability routines. The following table identifies the specific conditions under which capability developments were enabled for sense-making, connecting, and shaping routines within the context of emergency and disaster management. This includes ability to scale up, acceleration of data sharing, reduction of red tape, flattening org structure, and motivated workforce. The framework is intended as both a theoretical contribution, extending DCT and ANEC in the emergency management context, and a diagnostic tool through which public sector agencies can assess their current capability and identify investment priorities.

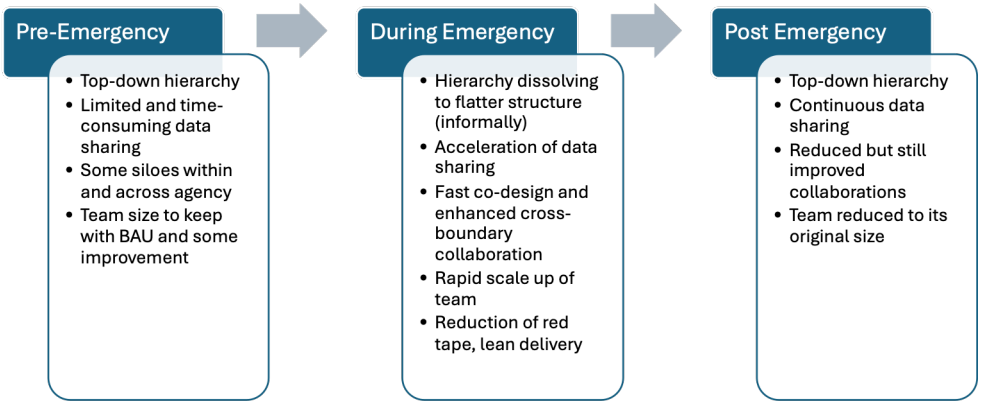


Figure 1. Changes in conditions

Top Category	Item Details	Dynamic Capabilities		
		Sense-Making	Connecting	Shapping
Enabling Conditions				
Leadership	Ability to scale up	x		x
	Pre-existing professional networks		x	x
Culture	Acceleration of data sharing			x
	Motivated workforce			x
	Reduction of red tape			x
Other	Flattening org structure		x	x
	Having access to the right people		x	
	Need to minimise impact and increase context understanding	x		
Data Analytics Capabilities				
Leadership	Executive Sponsorship			x
	Sufficient Investment			x
People	Availability of Data Talents	x		x
Data	Data Availability	x		x
	Data Completeness	x		x
	Data Richness	x		x
	Data Timeliness	x		
Technology	Data Analytics	x		x
	Predictive Analytics	x		x
	Modelling Tools	x		x
	Interoperability	x		x
	One platform		x	x
Data Culture	Data Sharing		x	
	Collaboration		x	

5. References

Akter, S., Bandara, R. J., & Sajib, S. (2021). How to empower analytics capability to tackle emergency situations? International Journal of Operations & Production Management, 41(9), 1469–1494. https://doi.org/10.1108/IJOPM-11-2020-0805

Australian Royal Commission into National Natural Disaster Arrangements. (2020). Royal Commission into National Natural Disaster Arrangements report. Commonwealth of Australia.

Centre for Research on the Epidemiology of Disasters (CRED), & United Nations Office for Disaster Risk Reduction (UNDRR). (2024). 2023 disasters in numbers. CRED & UNDRR.

Department of Finance. (2026, April 2). Data maturity assessment tool. https://www.finance.gov.au/government/public-data/public-data-policy/data-maturity-assessment-tool

Garattini, C., Raffle, J., Aisyah, D. N., Sartain, F., & Kozlakidis, Z. (2019). Big data analytics, infectious diseases and associated ethical impacts. Philosophy & Technology, 32(1), 69–85.

Gupta, M., & George, J. F. (2016). Toward the development of a big data analytics capability. Information & Management, 53(8), 1049–1064.

Kattel, R. (2022). Innovation and the public sector: A critical review of the literature. Public Management Review, 24(3), 317–335.

Klein, H. K., & Myers, M. D. (1999). A set of principles for conducting and evaluating interpretive field studies in information systems. MIS Quarterly, 23(1), 67–94.

Klievink, B., Romijn, B.-J., Cunningham, S., & de Bruijn, H. (2017). Big data in the public sector: Uncertainties and readiness. Information Systems Frontiers, 19(2), 267–283.

Kopcho, J. (2021). Manage new challenges with reinvented decision-making practices. Gartner. https://www.gartner.com/en/webinars/3994498/manage-new-challenges-with-reinvented-decision-making-practices

LaValle, S., Lesser, E., Shockley, R., Hopkins, M. S., & Kruschwitz, N. (2011). Big data, analytics and the path from insights to value. MIT Sloan Management Review, 52(2), 21–32.

Lee, I., & Raschke, R. L. (2016). Understanding the role of big data analytics in business value creation. Information & Management, 53(5), 541–554.

Lyytinen, K., & Grover, V. (2017). Management misinformation systems: A time to revisit? Journal of the Association for Information Systems, 18(3), 206–230.

Mikalef, P., Pappas, I. O., Krogstie, J., & Giannakos, M. (2018). Big data analytics capabilities: a systematic literature review and research agenda. Information Systems and E-Business Management, 16(3), 547–578.

Motamarri, S., Akter, S., Yanamandram, V., & Wamba, S. F. (2017). Why is Empowerment Important in Big Data Analytics? Procedia Computer Science, 121, 1062–1071. https://doi.org/10.1016/j.procs.2017.11.136

Myers, M. D., & Newman, M. (2007). The qualitative interview in IS research: Examining the craft. Information and Organization, 17(1), 2–26.

Patton, M. Q. (2002). Qualitative research and evaluation methods (3rd ed.). Sage.

Rizzon, E., & Kayande, U. (2018). The untapped value of analytics. Kearney. https://www.de.kearney.com/analytics/article?/a/the-untapped-value-of-analytics

Sejahtera, F., Wang, W., Indulska, M., & Sadiq, S. (2018). Enablers and inhibitors of effective use of big data: insights from a case study.

Strauss, A., & Corbin, J. (1998). Basics of qualitative research: Techniques and procedures for developing grounded theory (2nd ed.). Sage.

Teece, D. J., Pisano, G., & Shuen, A. (1997). Dynamic capabilities and strategic management. Strategic Management Journal, 18(7), 509–533.

Vasilyeva, O., & Richardson, A. (2022). Big Data and Data Analytics in Managing the COVID-19 Pandemic.

Walsham, G. (1995). Interpretive case studies in IS research: Nature and method. European Journal of Information Systems, 4(2), 74–81.

Wamba, S. F., Dubey, R., Gunasekaran, A., & Akter, S. (2020). The performance effects of big data analytics and supply chain ambidexterity: The moderating effect of environmental dynamism. International Journal of Production Economics, 222, 107498.
Wamba, S. F., Gunasekaran, A., Akter, S., Ren, S. J.-F., Dubey, R., & Childe, S. J. (2017). Big data analytics and firm performance: Effects of dynamic capabilities. Journal of Business Research, 70, 356–365.
Wang, C. L., & Ahmed, P. K. (2007). Dynamic capabilities: A review and research agenda. International Journal of Management Reviews, 9(1), 31–51.

Maintaining Technological Frame Congruency Through Accommodations and Compromises

Claire Cunningham, Anna Jenkins, Sam MacAulay,
University of Queensland

Abstract

Technological frame theory extensively explores how incongruent technological frames negatively influence new technology implementation. Yet little attention has been given to understanding how congruent technological frames are maintained. This emphasis assumes that congruent frames ensure smooth and successful implementation, overlooking the work required to sustain them. We examined the interplay of technological frames across three organisations during an ethnographic study of an AI mammography trial in Vietnam. Although the organisations held congruent and complementary technological frames, the trial encountered thirteen unexpected episodes, including delays and communication issues. We found that organisations maintained congruence through two mechanisms: accommodations and compromises, adjusting activities, responsibilities, and resources while preserving their technological frames. Our findings show that congruence does not guarantee a seamless implementation process and position congruence maintenance as an ongoing accomplishment. This study extends current understanding of technological frame theory by shifting attention to the work required to maintain congruence through micro-level processes .

Keywords: technological frames, technological framing, clinical trial, artificial intelligence

1. Introduction

Artificial intelligence (AI) has significant potential to improve healthcare, yet successful implementation remains challenging (Ciecierski-Holmes et al., 2022; Ennab & Mcheick, 2024; Scott et al., 2021). Clinical trials provide an important mechanism for evaluating AI technologies while simultaneously enabling organizations to understand how they may be incorporated into clinical practice (Farrell et al., 2010; Park et al., 2020).

Technological frame theory (Orlikowski & Gash, 1994) explains how stakeholders' assumptions, expectations, and knowledge about a technology influence implementation. Prior research has shown that incongruent technological frames increase the risk of conflict, deployment issues, and implementation failure (Lin & Silva, 2005; McGovern & Hicks, 2004; Olesen, 2014), leading most technological framing research to focus on how organisations overcome incongruence and achieve alignment (Garud & Rappa, 1994; Kaplan, 2008; Lin & Silva, 2005). However, little attention has been paid to how congruency is maintained once established.

This paper examines how technological frame congruency is maintained during a multi-organisational Mammo AI Trial (Trial) in Vietnam. Drawing on a 29-month ethnographic study of three collaborating organisations, we identify thirteen unexpected implementation episodes that challenged trial progress. Rather than changing each other's technological frames, organisations maintained congruency through accommodations and compromises. By examining how congruency is maintained, we extend technological framing research and provide insight into the micro-processes that preserve their technological frames and sustain implementation momentum.

2. Theoretical Background

Technological frames draw on sensemaking theory and explain how individuals and groups develop assumptions, expectations, and knowledge about a technology that influence how it is understood and used (Brown et al., 2015; Goffman, 1974; Pinch & Bijker, 1984; Weick, 1969; Weick, 1995, Orlikowski & Gash, 1994). Technological frames are shared understandings that develop within a social group and shape interpretations of both the technology itself and its role within a particular context (Orlikowski & Gash, 1994).

Research has consistently shown that different stakeholders often hold different technological frames, creating incongruence that can affect technology implementation outcomes (Davidson & Pai, 2004). Orlikowski and Gash (1994) define congruence as the alignment of frames on important dimensions, while incongruence reflects differences in assumptions, expectations, or knowledge. Prior studies have linked incongruence to implementation challenges such as power struggles, delayed decision making, and project failure (Lin & Silva, 2005; McGovern & Hicks, 2004; Olesen, 2014).

To address incongruence, organizations engage in technological framing processes through which stakeholders attempt to influence, align, or renegotiate interpretations of a technology to create a dominant frame (Azad & Faraj, 2008; Davidson, 2002; Davidson, 2006; Garud & Rappa, 1994; Kaplan; 2008; McGovern & Hicks, 2004; Song & Song, 2025; Wang et al., 2021). Existing research has therefore focused largely on how actors overcome incongruence and create congruence during technology implementation (Ghobadi & Mathiassen, 2024; Kaplan, 2008; Lin & Silva, 2005). Consequently, technological frame congruence is often treated as a desirable end state associated with successful implementation. We know considerably less about what happens after congruence has been established, and how organizations maintain congruency when unexpected challenges emerge. This prompted us to investigate how the organizations enacted their technological frames to maintain technological frame congruency and Trial momentum, and led us to the question: *How do organizations maintain technological frame congruency?*

3. Methods

3.1 Empirical Context

In 2021, International Diagnostic Frontiers (IDF) initiated the Tria) with Global MedTech Solutions (GMS) and Central Health Science Hospital (CHS) in Vietnam. The Trial aimed to evaluate the use of AI in breast cancer screening and involved the implementation of retrospective and prospective studies over a 29-month period. The three organizations collaborated throughout the planning, execution, and completion of the Trial.

3.2 Research Design and Data Collection

We adopted a longitudinal ethnographic case study approach to examine how organizational technological frames shaped the implementation of the Trial. Data were collected between September 2021 and February 2024 from multiple sources, including semi-structured interviews with participants from all three organizations, observations of collaboration meetings and clinical workflows, WhatsApp communications, project documents, and organizational materials.

3.3 Data Analysis

Data were analysed in two stages. First, technological frames of IDF, CHS, and GMS were identified using Orlikowski and Gash's (1994) domains of Nature of Technology, Technology Strategy, and Technology-in-Use. Consistent with Orlikowski & Gash (1994), the domains were operationalized as participants' understandings of AI in diagnostic medical imaging, the purpose and value of the BreastAI trial, and the implementation processes and clinical use of the AI system. Thematic analysis was used to develop frame dimensions and assess congruency. IDF and CHS exhibited congruent technological frames, while GMS displayed a complementary technological frame (Yeow & Sia, 2008; Schlagwein et al., 2025). Second, a process approach (Langley, 1999) was used to examine how technological frames influenced implementation. Thirteen unexpected episodes were identified through temporal bracketing of the Pre-Trial, Trial, and Post-Trial phases and analysed to identify recurring patterns linking organisational responses to the maintenance of technological frame congruency.

4. Findings

The three organizations held distinct technological frames, differing in how they understood the value of AI, yet their frames were congruent (IDF and CHS) and complementary (GMS) (See Table 1). IDF and CHS shared a common orientation towards improving patient care and clinical outcomes, while GMS viewed AI primarily as a technical and data-driven opportunity. Despite these differences, the organizations' technological frames remained aligned around the successful implementation of the Trial. Unexpectedly, the technological frames remained constant throughout the Trial, allowing us to examine how congruency was maintained during unexpected implementation challenges.

Table 1: Congruency of the IDF, CHS, and GMS technological frames

Technological Frame Domains	**IDF: Access to Care**	**CHS: AI is a Resource**	**GMS: AI in a Technical Tool**
Nature of technology	AI will improve care	AI will improve accuracy:	AI is maths
Technology Strategy	AI is an enabler	AI will reduce workload	Data opportunity
Technology-in-Use	AI will boost radiologist confidence	AI is a partner:	Validating and training AI
	Congruent (Orlikowski & Gash, 1994)		Complementary (Yeow & Sia, 2008; Schlagwein et al., 2025)

The second stage of analysis identified thirteen unexpected implementation episodes (see Table 2 below) across the Trial. Despite their different origins, these episodes displayed a common pattern: organizations responded by mobilising existing technological frame dimensions in the form of two recurring mechanisms, accommodations and compromises, which enabled organizations to maintain technological frame congruency and implementation momentum.

Table 2: Summary of technological framing episodes, accommodations, and compromises

Episode #	**Stage of Trial**	**Type of Change**	**Mechanism**	**Who**
1	Pre-Trial MOU	Research lead	Accommodation	GMS
2		Financial commitment	Accommodation	GMS
3		Research design	Compromise	GMS/ CHS
4	Pre-Trial Contract	Legal issues	Compromise	GMS/ CHS
5		Leadership change	Accommodation	IDF
6	Trial	Communication issues	Compromise	GMS/ CHS
7		Change study order	Accommodation	GMS
8		Internet at CHS	Accommodation	GMS
9		PhoenixAI at CHS	Accommodation	CHS
10		Setting up BreastAI	Accommodation	GMS
11	Post-Trial	Publication authorship	Accommodation	GMS
12		Relicensing AI	Accommodation	GMS
13		Recording meeting	Accommodation	GMS

4.1 Accommodations

An accommodation occurred when a single organization adjusted its activities in response to another organization's needs while maintaining its existing technological frame. The first accommodation emerged through a gradual shift in leadership. Although IDF initiated the Trial, GMS progressively assumed responsibility for key research and operational activities, including research design, data management, and contract negotiations. These changes reflected GMS's technological frame dimensions of Data Opportunity and Validating and Training AI. Rather than resisting the shift, IDF accepted a reduced operational role, allowing the Trial to progress without altering either organization's technological frame.

A second accommodation emerged when CHS unexpectedly requested financial compensation for radiologists participating in the Trial. Although IDF supported the request, it lacked the resources to fund it. GMS ultimately secured funding, enabling the Trial to continue. This accommodation aligned with GMS's interest in accessing and validating trial data while addressing CHS's workload and resource concerns.

> *"Where I almost felt it could flop over was when we looked at the financial cost of running the Trial… We hadn't even had that on our radar as a possible issue. And when it became clear that is, you know, it's substantial money, especially for our charity. That was something where I felt oh my god if the project dies just because of that, that would be really, really sad. And we were, what, seven, eight months into it. Yeah, people have actually invested a lot of*

their time into it. That was probably my low point. But that was resolved and that to be honest, that was resolved because Luca just got it done. Yeah. And that's kudos to him" (Sofia, IDF, Interview).

In both cases, organizations modified activities rather, in line with their technological frames. The accommodations enabled progress while preserving the existing congruent and complementary frame structure.

4.2 Compromises

A compromise occurred when multiple organizations jointly modified their activities to resolve an issue and maintain progress. One compromise emerged during contract negotiations between GMS and CHS. Differences between European and Vietnamese legal expectations created lengthy delays, particularly regarding intellectual property and data protection requirements. Instead of insisting on its preferred arrangement, GMS simplified the contract and removed specific clauses to accommodate CHS's requirements, enabling the Trial to proceed.

"Because a lot of [Intellectual Property] article in this contract is written by GMS …the lawyer of the hospital they [are] general lawyer…they don't know exactly the [Intellectual Property] article of the IP that GMS write… So it [was] me to decide and I have the choice. So I reject [the contract]. So that's why is the first version of the contract is very long, because the final [contract]got very short" (Dr Vu, CHSH, Interview)

A second compromise emerged in response to communication difficulties between GMS and CHS. Formal meeting-based communication proved ineffective because of language barriers and clinical workload pressures. In response, CHS proposed establishing a WhatsApp group and GMS accepted the change despite concerns about coordinating Trial activities. The resulting communication channel enabled more frequent exchanges regarding workflows, templates, and data collection, substantially improving collaboration momentum. Across these episodes, compromises enabled organizations to adapt their activities without changing their underlying technological frames. Instead, organizations temporarily elevated specific frame dimensions to negotiate solutions and maintain progress towards a shared goal.

5. Discussion

Existing technological frame research has largely focused on resolving incongruence in new technology implementations. Our findings suggest that achieving congruence is an important element, but one side of technological framing processes. This case reveals that although the three organisations began with congruent and complementary technological frames, they still encountered unexpected episodes, some of which threatened the success of the Trial. Unlike previous research where stakeholders attempt to change another stakeholder's technological frame, the organisations responded through a series of accommodations and compromises, while preserving their existing technological frames. This finding offers a new understanding of the underexamined end of the theoretical spectrum, maintaining technological frame congruence as an ongoing effortful practice.

5.1 Mechanisms for Maintaining Technological Frame Congruence

This study extends technological frame theory by shifting attention from how technological frame congruency is achieved to how it is maintained. Existing research has largely focused on technological framing activities that seek to overcome incongruence by influencing or realigning stakeholders' interpretations of a technology (Azad & Faraj, 2008; Engås, 2023; Hoppmann, 2020; Kaplan, 2008; Lin & Silva, 2005). In contrast, our findings show that technological framing work is an effortful process for organizations to maintain congruency.

Across thirteen unexpected episodes, the three organizations elevated their technological frames through two mechanisms: accommodations and compromises. Accommodations involved unilateral adjustments by a single organization, while compromises involved reciprocal adjustments by multiple organizations. Interestingly, the accommodations were not evenly distributed between the organisations during the Trial. Nine of the ten accommodation episodes were undertaken by GMS, suggesting that maintaining congruence may require some actors to absorb a disproportionate share of

the adjustment effort. As the organisation responsible for the AI system, GMS possessed the expertise and resources needed to respond to unexpected challenges. This suggests that maintaining congruence may be asymmetric, with some stakeholders undertaking greater maintenance work than others. However, eight of the accommodations driven by GMS were not significant, revealing their commitment to the Trial. This finding suggests that maintaining congruence may be asymmetric, with particular stakeholders undertaking greater maintenance work to sustain implementation progress.

In contrast to previous literature where stakeholders attempted to alter another's technological frames (Lin & Silva, 2005; McGovern & Hicks, 2004), these mechanisms enabled organizations to respond to emerging challenges while maintaining collaboration momentum. Our findings, therefore, suggest that technological framing work can occur repeatedly as a micro-process, not limited to resolving incongruence but is also necessary for maintaining congruence during technology implementation.

5.2 Theoretical and Practical Implications

By identifying accommodations and compromises as technological framing mechanisms, this study highlights the work required to maintain congruence. Technological frame congruency should be viewed as dynamic, requiring ongoing effort as organisations adjust activities, responsibilities and resources while preserving their technological frame.

Implementing new technologies across multiple organizations is often assumed to be easier when stakeholders hold broadly aligned perceptions of the technology. Our findings suggest otherwise. Even where technological frames are congruent, unexpected challenges can disrupt implementation progress. Managers should therefore view technological frame congruency as a resource that requires ongoing effort. Understanding both their own technological frames and those of their partners may help organizations make timely accommodations and compromises that enable implementation progress without requiring substantial realignment of stakeholder perspectives.

5.3 Limitations

This study examined a single longitudinal AI trial involving three organisations in Vietnam, designed to build theory from a rich, multi-organizational implementation context. Future research could examine whether accommodations and compromises operate similarly across technologies, industries, and organisational arrangements. In addition, participation was voluntary, participating organisations may have held more favourable perceptions of AI than other adopters.

6. Conclusion

This study extends technological frame theory (Orlikowski & Gash, 1994) by examining how congruence is maintained during a multi-organisational AI trial. While prior research has focused on overcoming incongruence, our findings show that congruence does not guarantee smooth implementation. Across thirteen unexpected episodes, organisations employed accommodations and compromises to maintain congruence and sustain trial momentum. By shifting attention from how congruence is achieved to how it is maintained, this study broadens understanding of technological framing and highlights the active role organisations play in responding to unexpected challenges and supporting implementation success.

7. References

Azad, B., & Faraj, S. (2008). Making e-Government systems workable: Exploring the evolution of frames. *The Journal of Strategic Information Systems, 17*(2), 75–98. https://doi.org/10.1016/j.jsis.2007.12.001

Brown, A. D., Colville, I., & Pye, A. (2015). Making sense of sensemaking in organization studies. *Organization Studies*, *36*(2), 265–277. https://doi.org/10.1177/0170840614559259

Ciecierski-Holmes, T., Singh, R., Axt, M., Brenner, S., & Barteit, S. (2022). Artificial intelligence for strengthening healthcare systems in low-and middle-income countries: A systematic scoping review. *NPJ Digital Medicine*, 5(1), 1–13. https://doi.org/10.1038/s41746-022-00700-y

Davidson, E. (2002). Technology frames and framing: A socio-cognitive investigation of requirements determination. *MIS Quarterly*, *26*(4), 329–358. https://doi.org/10.2307/4132312

Davidson, E. (2006). A technological frames perspective on information technology and organizational change. *The Journal of Applied Behavioral Science*, *42*(1), 23–39. https://doi.org/10.1177/0021886305285126
Davidson, E., & Pai, D. (2004). Making sense of technological frames: Promise, progress, and potential. In B. Kaplan, D. P. Truex, D. Wastell, A. T. Wood-Harper, & J. I. DeGross (Eds.), *Information systems research: Relevant theory and informed practice* (pp. 473–491). Springer. https://doi.org/10.1007/1-4020-8095-6_26
Ennab, M., & Mcheick, H. (2024). Enhancing interpretability and accuracy of AI models in healthcare: a comprehensive review on challenges and future directions. *Frontiers in Robotics and AI*, *11*, 1444763.
Farrell, B., Kenyon, S., & Shakur, H. (2010). Managing clinical trials. *Trials*, *11*(1), 78. https://doi.org/10.1186/1745-6215-11-78
Garud, R., & Rappa, M. A. (1994). A socio-cognitive model of technology evolution: The case of cochlear implants. Organization Science, 5(3), 344–362. https://doi.org/10.1287/orsc.5.3.344
Ghobadi, S., & Mathiassen, L. (2024). A cognitive method for comparing and elaborating on technology frames. British Journal of Management, 35(2), 706-728. https://doi:.org/10.1111/1467-8551.12752
Goffman, E. (1974). Frame analysis: An essay on the organization of experience. New York: Harper & Row.
Kaplan, S. (2008). Framing contests: Strategy making under uncertainty. *Organization Science*, *19*(5), 729–752. https://doi.org/10.1287/orsc.1070.0340
Langley, A. (1999). Strategies for theorizing from process data. *Academy of Management Review*, *24*(4), 691–710. https://doi.org/10.5465/amr.1999.2553248
Lin, A., & Silva, L. (2005). The social and political construction of technological frames. *European Journal of Information Systems*, *14*(1), 49–59. https://doi.org/10.1057/palgrave.ejis.3000521
McGovern, T., & Hicks, C. (2004). How political processes shaped the IT adopted by a small make-to-order company: A case study in the insulated wire and cable industry. *Information & Management*, *42*(1), 243–257. https://doi.org/10.1016/j.im.2004.01.004
Olesen, K. (2014). Technological frames: Use of context, temporality, and individual focus. *SAGE Open*, *4*(1). https://doi.org/10.1177/2158244014526720
Orlikowski, W., & Gash, D. (1994). Technological frames: Making sense of information technology in organizations. *ACM Transactions on Information Systems*, *12*(2), 174–207. https://doi.org/10.1145/196734.196745
Park, Y., Jackson, G. P., Foreman, M. A., Gruen, D., Hu, J., & Das, A. K. (2020). Evaluating artificial intelligence in medicine: Phases of clinical research. *JAMIA Open*, *3*(3), 326–331. https://doi.org/10.1093/JAMIAOPEN/OOAA033
Pinch, T. J., & Bijker, W. E. (1984). The social construction of facts and artefacts: Or how the sociology of science and the sociology of technology might benefit each other. *Social Studies of Science*, 14(3), 399-441. https://doi.org/10.1177/030631284014003004
Schlagwein, D., Gozman, D., & Manusu, A. P. (2025). Cryptocurrency frames of reference: a case study of accepting ‘Bitcoin-as-X.’ *European Journal of Information Systems*, 1–36. https://doi.org/10.1080/0960085X.2025.2530456
Scott, I. A., Abdel-Hafez, A., Barras, M., Canaris, S., & Scott, I. A. (2021). What is needed to mainstream artificial intelligence in health care? Australian health review, 45(5), 591-596. https://doi.org/10.1071/AH21034
Song, J., & Song, L. (2025). Doing Is Believing: How the interplay among multimodal practices influences frame change. *Organization Studies*, *46*(10), 1457-1488.https://doi.org/10.1007/978-0-387-87503-3_10
Wang, C., Medaglia, R., & Jensen, T. B. (2021). When ambiguity rules: The emergence of adaptive governance from (in)congruent frames of knowledge sharing technology. Information Systems Frontiers, 23(6), 1573–1591. https://doi.org/10.1007/s10796-020-10050-3
Weick, K. E. (1969). The social psychology of organizing. Addison-Wesley Pub. Co.
Weick, K. E. (1995). Sensemaking in organizations. Thousand Oaks, CA: Sage.
Yeow, A., & Sia, S. K. (2008). Negotiating “best practices” in package software implementation. Information and Organization, 18(1), 1–28. https://doi.org/10.1016/j.infoandorg.2007.07.001

Technology-Enabled Liminality: Generative AI and Occupational Transitions

Jorge Retamales-Moreno
University of Queensland - Business School

Abstract

This study challenges the assumption that under-institutionalised liminality is harder to endure. Liminality has often been theorised as a state to be overcome, and its difficulty has been attributed to the absence of institutional support. I argue that this difficulty may come, at least in part, from the lack of perceived competence, and that by lowering the bar for performing the work, generative AI (GenAI) may make liminality easier to endure, without necessarily depending on institutions. I tackle this through a qualitative case study complemented with labour market data. I find that GenAI steered data scientists (DS) toward AI engineering (AIEng) work and played a critical role in enabling DS to endure liminality, by helping them qualify for and perform a role they had no formal training in. The DS who were able to sustain liminality did so by naturalising it as common to consulting, leveraging it to take on senior-level tasks, and hedging it for versatility in the job market. Labour market data show that the share of DS moving to AIEng rose by 20% after the release of ChatGPT, and by 36% in technology consulting. I contribute a model of technology-enabled liminality that shows how GenAI triggers liminality and, by making it easier to endure, may perpetuate it.

1. Introduction

This study challenges the assumption that under-institutionalised liminality is harder to endure. It does so by asking how generative AI (GenAI) shaped liminality during a role transition. Liminality captures the experience of being situated between reference points, which makes one's situation difficult to delineate as these references become blurred or merged (Beech, 2011; Orlikowski & Scott, 2021). These reference points may concern roles (e.g. Mayer & Recker, 2026), spaces (e.g. Nicholson et al., 2017), the wider social and economic context (e.g. Orlikowski & Scott, 2021) or epistemic foundations (e.g. Zhang et al., 2021). In this paper I focus on role liminality, the experience of being betwixt and between roles or identities (Ibarra & Obodaru, 2016). This experience is generally ambivalent (Ashforth et al., 2014; Ibarra & Obodaru, 2016), but is not necessarily aversive. Liminality can also be a space of opportunity, where individuals may pursue progress from within the in-between state (Weidenstedt et al., 2024) and exercise agency across boundaries (Chamakiotis et al., 2024).

Studying how GenAI affects liminality is important because it may steer workers into transitioning onto other roles and may facilitate that transition. Recent evidence shows that workers who adopt these tools take on new tasks and move into occupations where the technology is more relevant (Humlum & Vestergaard, 2025). Field experiments show that access to large language models raises performance most for those with less prior experience in the task at hand, which reduces the advantage of that experience (Brynjolfsson et al., 2025; Dell'Acqua et al., 2023; Noy & Zhang, 2023). Considering this, if with GenAI workers can go into a new role without spending much time learning to perform it, GenAI may ease the liminality process and reduce the need for institutional support.

Liminality has often been theorised as a state to be overcome, and its difficulty has been attributed to the absence of institutional support. Institutionalised passages were bounded in time, guided by mentors, and closed on a known outcome (Van Gennep, 1909), whereas contemporary passages rarely offer any of this, making liminality harder to endure (Ibarra & Obodaru, 2016). I argue that this difficulty may come, at least in part, not from the lack of institutions but from **a** lack of perceived competence. Without it, individuals have little evidence on which to claim a new self, and others have little ground on which to grant it (Pratt et al., 2006). By lowering the bar for performing work, GenAI may make liminality easier to endure.

I tackle the research question through a qualitative case study of TechCo, a technology consulting firm, where I interviewed data scientists (DS) shifting toward AI engineering (AIEng) roles. I complement the case with occupational transition data from Revelio, covering 1,199 United States occupations between 2018 and 2024. These data show that after the release of ChatGPT, the share of DS exits going to AIEng rose by 20%, and by 36% in technology consulting, against 20% in management consulting and 18% in the rest of the economy.

This study's findings suggest that GenAI steered DS toward AIEng work. Off the shelf models and easier to use AI offerings pulled them toward doing AIEng tasks, while clients demanded implemented AI systems and TechCo pushed them in the same direction and away from DS tasks, displacing the latter and reinstating the former (Acemoglu & Restrepo, 2019). More importantly, GenAI played a critical role in DS enduring liminality. It helped them qualify for and perform a role they had not formally trained for. However, the fast-moving pace of GenAI developments also created anxiety about keeping up and raised the expectations placed on their work. I find that four out of five DS were able to sustain the liminal state, while one went back to a DS role. The four that were able to sustain it did so by making use of three tactics. They naturalised liminality as common to consulting, they leveraged their in-between position to do more or take on senior-level tasks, and they hedged it, framing their in-betweenness to have more versatility for the job market. As for the worker who went back, he anchored himself to the DS identity, which resolved the ambiguity.

I contribute a model of technology-enabled liminality. I show that the difficulty of under-institutionalised liminality may rest less on the absence of institutions than on the absence of perceived competence, and that when technology supplies a sense of competence, it loosens the coupling between doing and becoming. Workers can perform a role without becoming a member of it or taking on its identity, which turns the passage into a state they can inhabit rather than only endure. An easier state carries less pressure to leave it, so liminality resolution becomes one possible ending rather than a necessary one.

2. Literature review

Traditionally liminality had been considered a well-defined in-between state, where individuals, often marked by rites, transition from one state to another (Van Gennep, 1909). An example of this is a medical student becoming a surgeon through the rite of the internship. Ibarra and Obodaru (2016) call this an institutionalised liminal experience and define its degree of institutionalisation along four dimensions: First, the duration is finite and known at the onset. Second, the passage comes with built-in guidance from mentors and support from a community of others going through the same thing. Third, culturally legitimate narratives are available to make sense of the experience. Fourth, the role the person will hold afterwards is known from the start. However, according to Ibarra and Obodaru (2016), contemporary liminality is different. Transitions are increasingly under-institutionalised, which means they are open-ended, unguided, and without a culturally available narrative for

what the person is going through. Under these conditions the transition has no prescribed steps and no guaranteed outcome, and the work of getting through it falls mainly on the individuals themselves, which, they argue, makes this passage harder to endure.

Because I define role liminality as an ambivalent role transition, every role liminality involves a role transition, but not every role transition is liminal. Having left a previous state without yet arriving at a new one, individuals occupy a "position of ambiguity and uncertainty" (Beech, 2011, p. 287) that can disrupt their internal sense of self and of their place within a social system (Noble & Walker, 1997). The separation from a known state carries a risk of disorientation and of losing agency (Söderlund & Borg, 2018), and the passage itself lacks a culturally legitimate narrative through which the person can account for what is happening to them (Ibarra & Obodaru, 2016). Although individual tolerance for ambiguity may moderate this experience (Ibarra & Obodaru, 2016), people may still experience the transition as a loss, oriented toward what has gone rather than what might come (Brazier et al., 2025), and may feel marginal, belonging fully to neither side (Garsten, 1999). Liminality has therefore often been theorised as a state to be overcome, with resolution as the expected endpoint (Brazier et al., 2025; Ibarra & Obodaru, 2016).

Despite its in-between nature, liminality does not necessarily have to be short-lived. Czarniawska and Mazza (2003), for example, argue that management consulting can be framed as inherently liminal. Because of their occupation, management consultants need to immerse themselves in their clients' organisations, processes and needs, yet without belonging to them. It is precisely because they do not belong to the client's organisation that they hold sufficient distance to intervene in it, which is why they can inhabit it for longer. Individuals can also exert agency from within the state. Weidenstedt et al. (2024) show that migrant gig workers, despite holding little control over their working conditions, engage in what the authors call liminal agency, "the active pursuit of possibilities for progress in uncertain states of in-betweenness" (p. 1282). They acquire competencies, build networks and reframe what makes their work meaningful, all from within the liminal state.

Ibarra and Obodaru (2016) argue that under-institutionalised liminality is more difficult to endure, because the individual is left to make sense of the transition alone. I argue that this difficulty may not come only from the absence of institutions. Mentors did more than confirm who the liminar was becoming, they also taught the work (Levinson, 1986), so without them individuals had to learn the new role on their own. Doing the work and feeling competent go hand in hand in how new conceptions of self are constructed (Pratt et al., 2006), and I argue that this is what made guidance necessary for institutionalised liminality. The difficulty shows up as a tension between having to learn and having to appear credible, one that threatens a professional's sense of competence, their acceptance by others, and their productivity (Bourgoin & Harvey, 2018).

Because GenAI can provide the means to be productive and individuals can exert liminal agency, this calls for a reassessment of the dependency on institutions for contemporary liminality. If GenAI can supply the perceived competence needed to perform the work, as the productivity evidence suggests (e.g. Brynjolfsson et al., 2025; Dell'Acqua et al., 2023; Noy & Zhang, 2023), individuals transitioning between occupations may no longer be bound to the period of incompetence that has traditionally characterised the passage. This lowers a barrier that made liminality difficult and may therefore make the state easier to endure.

3. Theoretical background

Anteby et al. (2016) propose three lenses on occupations, becoming, doing and relating. I build my argument on the first two, and set aside relating, which concerns the collaborative relations members build with other occupations, clients and technology. For Anteby et al. (2016) becoming is a social process, since it is the community that transmits and confirms

who one is becoming. Identity work, however, also has an internal side, where individuals experiment with and evaluate who they are becoming (Beech, 2011; Ibarra, 1999), and in under-institutionalised passages that work falls mainly on the individuals themselves (Ibarra & Obodaru, 2016). I therefore extend the becoming lens with this literature, and examine the identity work these DS carried out on their own account.

Doing is what members perform, the tasks they carry out and the claims they make about which tasks belong to them. Anteby et al. (2016) note that a doing lens captures occupational change while it happens, whereas a becoming lens only shows it once the change has stabilised and new members are socialised into the new occupational reality.

Although these lenses have often gone together, I consider a context where they can move apart. I use this separation to see if workers can do the new role while not yet becoming, staying in a liminal state. Perceived competence sits on the doing side, so if technology lowers the bar for doing, individuals can perform a role they may not have been socialised into. The ambiguity of not knowing who they are may not go away, but it should be easier to hold when they can do the work. My study therefore examines how GenAI affected doing and becoming for DS undergoing a transition from DS to AIEng work.

4. Methods

I conducted a qualitative case study of TechCo, a technology consulting firm of over 900 workers operating in a developed economy. I selected TechCo because it offered a setting in which I aimed to understand how GenAI was impacting organising. However, as I entered the field with an open mind (Dey, 1999), and exploratory data collection progressed it became clear that GenAI was potentially playing an important role in shaping role transitions. I therefore focused data collection on this phenomenon (Charmaz, 2014).

TechCo offers services across Data & AI, Cloud, Digital and Cyber security, with Data & AI generating the largest revenues. As a consulting firm, TechCo is highly sensitive to client demand, and because its main cost is its workforce, the productivity of its workers is central to its business. Workers are allocated to client projects from a shared pool, using a resource allocation tool in which they self-report their skills. As the AI acceleration officer (AO) explained, this allows managers to *"filter down who's available and who's rolling off projects"*.

TechCo adopted GenAI early. When GitHub Copilot, a GenAI code assistance software, became available, the firm soon adopted it, as it did too with ChatGPT after its release. TechCo also saw a strong increase in demand for deploying GenAI solutions. Because of this, demand for AIEng skills soon outpaced supply*: "I cannot hire enough people or find enough people in market"* (AO). TechCo actively promoted GenAI among its workers and clients, creating an AO role to organise training and collaboration sessions, and formally introduced AI engineering as a skill in its allocation tool. In this context, many DS at TechCo were transitioning into AIEng work to meet clients' demand.

To understand these transitions and the role GenAI played in them, I drew on two data sources: individual job positions data and semi-structured interviews. I first analysed job positions data from Revelio for United States workers who held a data or technology role between 2018 and 2024, with October 2022, the month before ChatGPT's release, as the shock onset. I define an occupational exit as a move between two consecutive positions of the same worker carrying different Revelio occupation labels. Revelio classifies each role from its job title and description, so a change of label captures a genuine discontinuity rather than relabelling within the same role. I paired these data with the GenAI exposure index of Eloundou et al. (2024), matched to 1,187 of 1,199 origin occupations through Revelio's O*NET codes, covering 19.3 million occupational moves. I compare the share of DS exits

going to AIEng before and after the shock, across technology consulting, management consulting and the rest of the economy. I measure transitions rather than liminality because liminality is a cognitive-emotional process that resists measurement at scale. Because these estimates are observational, I treat them as descriptive context rather than causal relationships.

As for the second data source, I interviewed nine DS exposed to GenAI, all experienced in the role for more than two years, along with a Data and AI manager, named head of client solutions at TechCo (HCS), an AO and two project managers (PM1, PM2) (Table 1). The first round, between August and October 2025, was exploratory (see Appendix, Table 3 and 4 for protocols). The liminality process emerged only after analysing the first few interviews. Following theoretical sampling logic (Charmaz, 2014), subsequent data collection targeted the emerging phenomenon of increased AIEng work. I conducted a second round of interviews between November 2025 and February 2026 with five participants who were undergoing the transition and recruited a participant I had not yet interviewed but was likely to be experiencing it (DS9) (see Appendix, Table 5). The second-round protocol focused on the shift toward AIEng, the role of GenAI in it, and participants' current and foreseen identities. As GenAI is a rapidly evolving technology, the second interviews aimed to capture changes over time. In total, I conducted 18 semi-structured, in-depth interviews averaging about one hour each and totalling approximately 19 hours. Of the nine DS, seven were confronted with the work transition (Core in Table 1), and five of them experienced liminality (in bold). The remaining interviews and internal documents served to triangulate and contextualise these accounts.

Table 1: Interview participants

Participant	Experience (years)	Round	Duration (min)	Use in Analysis
DS1	7	1, 2	78, 68	Core
DS2	3	1	48	Supporting
DS3	4	1, 2	66, 34	Core
DS4	5	1	69	Supporting
DS5	7	1, 2	57, 43	Core
DS6	7	1, 2	53, 45	Core
DS7	6	1, 2	56, 44	Core
DS8	8	1	57	Core
DS9	4	2	86	Core
HCS	3	1	55	Context
AO	3	1	118	Context
PM1	8	1	73	Initial exploration
PM2	7	1	70	Initial exploration
Total		18 interviews	1,120 (18.7 hours)	

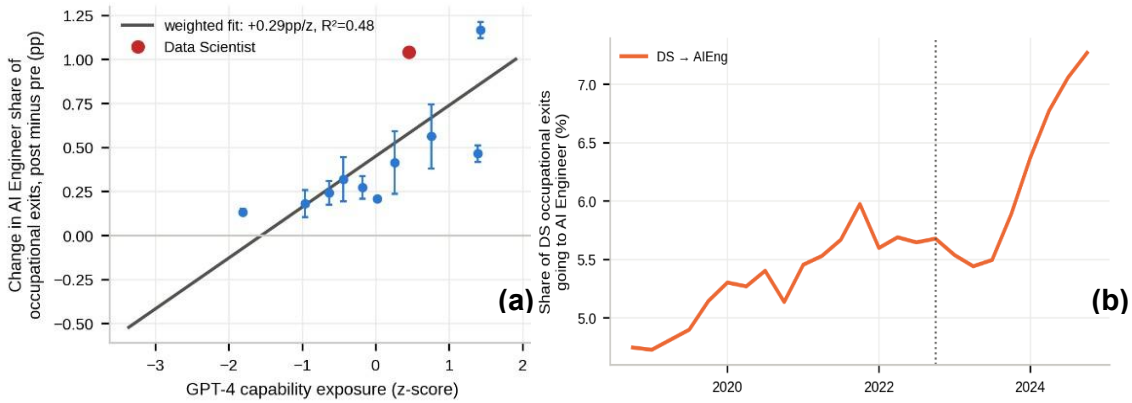
To theorise the liminality process I used a grounded theory approach (Charmaz, 2014). I chose this approach since it is a powerful lens to analyse the interpretations and emotions of individuals going through the same process (Langley, 1999). I first created initial codes that stayed close to participants' language. Through constant comparison I then identified focused codes, the ones that were most prevalent and carried the most explanatory power. I grouped these into categories and, finally, into core categories. Through constant comparison these core categories converged into mechanisms that enable entry into the liminal state and mechanisms that sustain it once inside.

5. Findings

5.1. Labour market context

Figure 1a shows whether occupations more exposed to GenAI shifted their exits toward AIEng after the release of ChatGPT. The 1,187 occupations are grouped into exposure deciles, weighted by their workers at risk. Although the changes are modest, occupations more exposed to GenAI tended to exit more to the AIEng role afterwards. Notably, the DS occupation shows an AIEng exit share larger than what the exposure index would predict. This may be because DS and AIEng often work closely together, which makes the move between them more likely, although examining this is beyond the scope of this study.

Figure 1: Exits toward AIEng after the release of ChatGPT

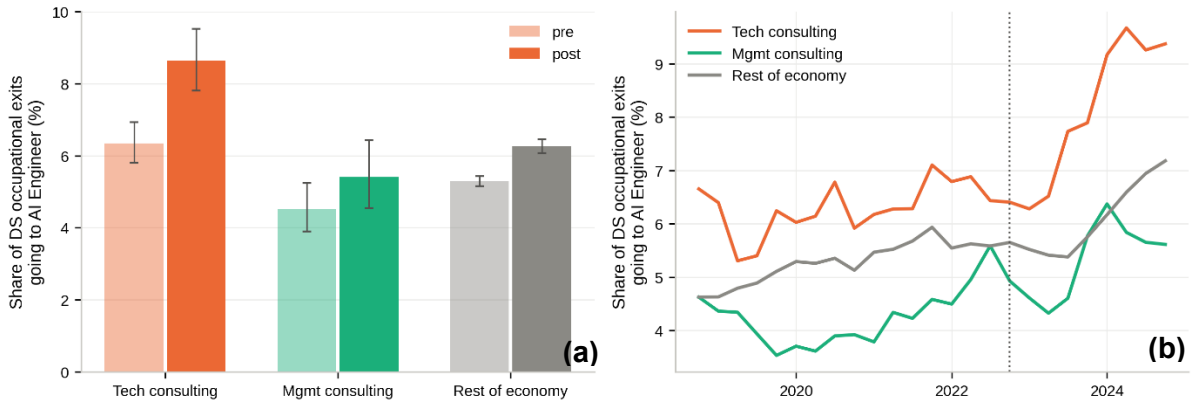


Note: (a) Change in AIEng share of exits by GenAI exposure decile (z-scores), means and 95% CIs; red marks DS. (b) DS exits going to AIEng, 4-quarter moving average; dotted line marks October 2022.

Figure 1b shows the share of DS exits going to AIEng. After the release of ChatGPT, DS left their occupation less often overall, as the proportion leaving each quarter fell 11%, from 5.93% to 5.27%. Measured against the entire DS workforce, the move to AIEng grew by 6%. Yet among the DS who did leave, the share going to AIEng rose from 5.33% to 6.38%, a 20% increase. Even as DS moved less, more of those who moved chose AIEng, a redirection consistent with GenAI steering DS toward AIEng work.

Figure 2a shows the share of DS exits going to AIEng before and after ChatGPT by industry. The share rose in all three groups, but unevenly. Tech consulting rose 36%, from 6.34% to 8.63%, against 20% in management consulting and 18% in the rest of the economy. Tech consulting also starts from the highest level, and that ordering holds before and after the shock. Figure 2b shows that the three series move together before the shock, and tech consulting separates more from the other two afterwards.

Figure 2: Share of DS exits going to AIEng by industry.



Note: Panel a, pre- and post-shock means with 95% CIs. Panel b, quarterly series, 4-quarter moving averages.

Taken together, the analysis suggests that GenAI plausibly increased transitions from DS to AIEng, but not that it was their only driver. In tech consulting, this path was already more travelled before the shock and widened most after it, as moving into new kinds of work is part of the norm in this kind of business. This suggests, descriptively, that GenAI amplified a preexisting DS to AIEng transition, with tech consulting leading both before and after. This makes TechCo a revealing setting to examine how DS experienced this transition.

5.2. The TechCo case study

According to the participants, GenAI had a significant effect on the type of AI projects TechCo delivered, moving from what they called proofs of concept to proofs of value. Clients now asked for AI to be implemented within their systems rather than shown as a potential outcome, which made projects more complex and longer. Because TechCo served many clients at once, this shift arrived across engagements at the same time rather than as an isolated event, turning what might have been a project-level adjustment into a sustained change in the DS work. TechCo also aimed to over-deliver, often absorbing the cost of deliverables that had not been previously agreed upon.

As I analysed the data from the interviews, four main findings became apparent. First, *liminality* was characterised by uncertainty, yet not necessarily discomfort. DS moved from DS toward AIEng work without a settled identity on either side (Table 2), but this rarely produced uneasiness. They exercised agency through tactics that framed the state in a brighter light, *hedging* their in-between condition to keep future options open or *leveraging* the state to aspire to senior roles. Second, organisational context and technology jointly drove entry into the new work. *Steering* operated through a strong increase in client demand for integrated AI systems (AIEng-leaning work) over experimentation with AI prototypes (DS-leaning work), while easier- to-use AI offerings and conversational GenAI lowered the bar, so that DS could feel *qualified* for, and perform, the new role. As a tech consulting firm, TechCo's DS also treated new tasks as part of the job, *naturalising* the shift. Third, although GenAI facilitated the transition, it also *entrained* DS to increased delivery expectations. Rather than shorter time frames, however, projects came to involve more deliverables in the same time span. Fourth, the shift did not lead to liminality for every DS confronted with it. Out of the seven, five entered the liminal state. Four remained in it (DS5, DS6, DS7, DS9), while DS3 left the firm to return to data science work. DS1 and DS8, although confronted with the transition, showed no clear signs of an ambivalent identity. I argue that the more strongly DS were anchored to their prior work identity, the less able they were to remain in the state. Table 2 presents the data structure of the mechanisms through which DS entered

and remained in this state, with corresponding codes and representative quotes. Then I integrate these findings into a theoretical model of technology-enabled liminality.

As DS *blurred* with AIEng work, the actual work decoupled from the job titles, and as such left them in a state of liminality. Traditionally, Data science had been *"more of a niche field centred around traditional statistical analysis"* (DS3), but the role turned "*more enmeshed, or nebulous… becoming more of an engineering role"* (DS3), to the point that the title no longer described today's work: *"Data scientist… I feel like it's more describing the era when we were creating the models"* (DS6). This blurring left DS in an ambivalent position where they could neither claim the new identity nor let go of the old one. DS5 felt *"maybe I need to be a full- fledged MLE. I don't know… maybe later it'll change,"* while keeping the DS title *"for now. I'll change, who knows."* DS7 kept both labels open: *"I could also call myself an AI engineer, but it's a bit semantics for me ultimately"*. For DS6, this work shift led to feelings of uncertainty: *"I'm also not an engineer, so I don't know if I'm answering your question at all… I don't even know what the answer is."* And DS9, caught mid-passage, was drifting out of both: *"I'm slowly, gradually not seeing myself as a data scientist anymore. I now more identify myself as an AI-ML consultant".* This occupational blurring, facilitated by the technological and organisational context, generated the conditions for DS identities to become ambivalent. However, not all five remained in the liminal state. Four DS deployed mechanisms that sustained it, while DS3 used tactics to pull out of it and return to his identity.

GenAI's emergence and the organisational context steered DS into doing AIEng work. With GenAI, DS at TechCo did not necessarily have to develop AI models from scratch. Off-the-shelf models displaced this task, pulling them into using these models instead: *"We're just taking a model that's provided, which is a black box, and then we use it."* (DS3). This was also facilitated not just by the increased availability and capabilities of external models, but by the tools to access and use them getting better. DS at TechCo mentioned that their work involved instead easy-to-use APIs and *"stitching and plumbing stuff together."* (DS7). This led to reliance on externally provided models, which, although highly capable were inscrutable.

Along with technology pulling DS to do AIEng instead of DS work, TechCo strongly *pushed* for it. Throughout all our interviews, DS, HCS and AO all reported that clients requested implementation of existing models, especially large language models, rather than developing new AI models. A representative example of this was claim automation. With the advent of Agentic technologies, clients also followed, increasing demand for them. However, not all DS thought this was always necessary on client's part. In the case of DS3, he acknowledged that clients sometimes did not have a clear vision of *"what kind of business function they wish to fulfil."* (DS3). Regardless, this push for implementing, led to an increased interest in AIEng over modelling, resulting as well, in longer and more complex projects for TechCo. Therefore, TechCo halted the hiring of DS and focused their efforts in hiring AI engineers instead: *"We're only hiring engineers now. We're not hiring data scientists within TechCo. That's in response to client demand."* (DS1). This organisational direction was not only evident through their hiring decisions, but through multiple workshops and meetings where they taught AIEng principles.

Conversational GenAI also enabled DS to *enact* the role of AIEng, by making them feel *qualified* for the role and allowing them to *perform* at the standards needed. In the case of qualifying for the role, for example, the coding and knowledge capabilities of Conversational GenAI, helped them lower the bar for doing AIEng work: *"It's meant that in many instances I have been able to work as an ML engineer or software engineer, whereas before I don't think I would've been able to."* (DS3), in which they used GenAI to directly help them deploy models. Although they did not see themselves still as AIEng, with the help of GenAI, they still felt they could sufficiently do the work *"I might not be following a hundred percent best practises. But I can make stuff work and put it into production."* (DS7).

GenAI also allowed DS to do AIEng at a fast pace. DS reported experiencing significant speed gains with GenAI: *"Something that would have taken me weeks, I could do in a few days."* (DS6) and in some cases mentioned leaning towards asking GenAI instead of their colleagues for resolving work issues: *"I'd prefer to go to a GenAI chat rather than asking people on my team, because I know everyone's very busy and I don't want to bother them."* (DS2)

By relying on GenAI, DS also became *entrained* with GenAI's rapidly changing pace. Although GenAI facilitated them enacting the role of AIEng it also led to feelings of anxiety to keep up with the moving technology, where they had to constantly keep on learning new models, tools and architectures: *"It's unnecessary anxiety sometimes because things are moving very quickly… you have one model super powerful this week, seven days later another model gets released."* (DS9). Adopting a fast pace was also evident on their outputs, which led to an increase in the perceived productivity expectations out of using GenAI. However, faster deliveries did not translate to shorter projects or more free time. Instead, because of TechCo's culture of overdelivering, it resulted often in DS delivering more in the same amount of time.

As described above, four out of the five DS that I assessed as experiencing liminality, were able to *accommodate* and make the best of their new condition. To do this they naturalised the liminality by framing it as part of consulting or the technological environment. In addition, they saw it as a *leveraging* opportunity for up leveling their work and to *hedge* their work careers by being able to perform both roles.

One way naturalising was manifested was by putting more focus on the overlap between DS and AIEng and not their differences. *"It's pretty much still the same skills. I'm still doing data analysis, I'm still doing coding, but the tools I'm using are different."* (DS6). This also resulted in DS thinking more abstractly about their jobs, and seeing themselves as problem solvers, and on the fundamentals of both roles: *"I think that's the noise. But to me, fundamentally things are just the same."* (DS9). As Czarniawska and Mazza (2003) brought up, consulting can also be thought as an inherently liminal space, and as such, some TechCo's DS saw this liminality as part of the job: *"It's fine, really. To me, these things — as a consultant, we do new things all the time."* (DS6). Finally, they also framed this shift as just a natural effect of technological changes: *"I guess that's just what progress looks like. Lots of skills have been lost over time."* (DS5)

Because GenAI allowed also to do more work, it enabled DS to take on more senior leaning tasks, such as working on the abstract layers, while avoiding doing programming: *"I like it too, to really not be doing the code but more doing the abstract layers and distil information."* (DS9). This led to even two DS pondering new role labels to accommodate for this abstraction*: "I think I'm just having a midlife career crisis or something. I'm slowly, gradually not seeing myself as a data scientist anymore. I now more identify myself as an AI-ML consultant."* (DS9).

Another mechanism that allowed them to sustain their liminal state was to frame it as a hedging opportunity. They saw so much demand for it, that they saw that having both roles would allow them to be of need for either a DS or as an AIEng at TechCo or for another organisation*. "I just want to be able to do both instead of just one. I think it's good to have a variety."* (DS5). This was also motivated by economic outcomes, as an opportunity to a higher salary, given that TechCo was promoting AIEng workers: *"My wife is not working at the moment, so you lose a bit of income. If you can compensate with a higher salary, that would be a good outcome"* (DS7).

However, not all DS were able to sustain the liminal state. DS3 resigned and moved to a company to do more traditional DS work. I argue that he was able to do so because he *anchored* his occupational identity to DS rather than an in-between role. He did this in three

ways, first he criticised the use of GenAI in learning and performing as an AIEng. He was worried about not learning skills in depth and hence potentially leading to skill erosion. Additionally, he criticised AIEng work: *"it's no longer about prediction or inference or statistical modelling. It's more about generating something",* while defending traditional data science as something meaningful to him: *"Developing a model to be able to model the world. That model aspect is what makes it enjoyable".* While others also voiced concerns about skill erosion (Table 2), only DS3 combined them with rejecting AIEng work itself, which made the in-between position untenable for him.

Figure 3: Technology-enabled liminality model

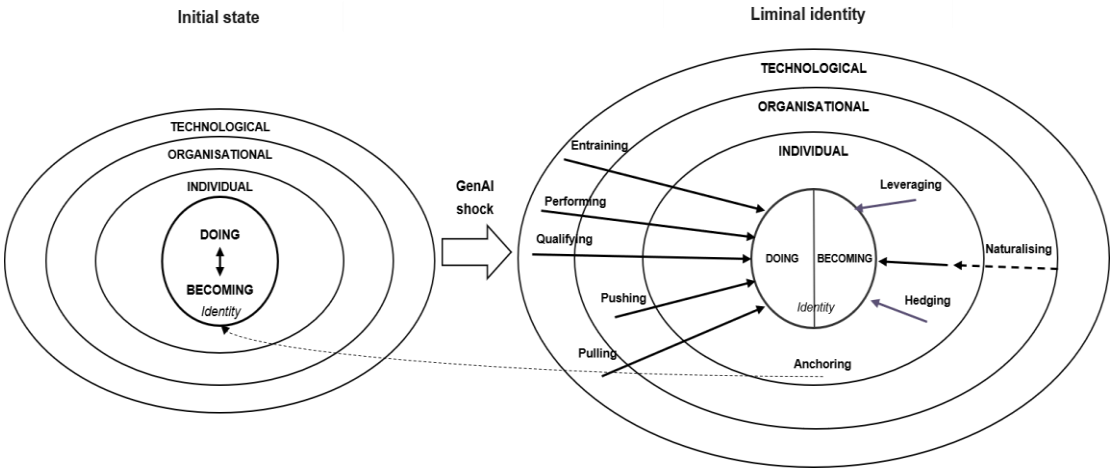


Figure 3 integrates these findings into a theoretical model. Before the shock, doing and becoming moved together. DS learned the occupation by doing its work, and each project deepened both their skill and their membership. The arrival of usable GenAI capability loosened that coupling. The technological and organisational mechanisms steered DS into doing AIEng work. While GenAI redirected what DS were asked to do, it also helped them qualify for and perform the new role, while entraining them to the pace of the technology. Although their identities were still ambiguous, the tactics they employed let them benefit from that ambiguity. They naturalised the shift as part of their organisational context, and individually they saw the ambiguity as a chance to leverage their current role and hedge their skills for future job opportunities. These tactics do not change what DS do. They speak to their liminal agency (Weidenstedt et al., 2024), using the liminal state to their benefit and thus making it easier to endure. This depended, however, on how deeply the previous identity was anchored. A strong anchoring pulled DS back to the DS role and out of the liminal state altogether.

6. Discussion

This article makes three contributions. First, I challenge the claim that under-institutionalised liminality should be harder to endure (Ibarra & Obodaru, 2016). The DS I studied went through a passage with no known duration, no mentors, no available narrative and no guaranteed destination, and yet four of the five who entered the state were sustaining it without the discomfort the literature anticipates. They exhibited agency, but that agency was not only a matter of disposition. It rested on GenAI supplying perceived competence that the passage had previously demanded of them.

To argue for this, I claim that GenAI loosens the coupling between doing and becoming. Previous research on liminality has generally treated both as co-dependent, since learning to perform a new role was also the process through which a person came to be a member

of it (Pratt et al., 2006). By spending time learning, workers were able to become the role, resolving together the competence and identity gaps. In contrast, GenAI allows workers to perform a role before becoming it, so that becoming lags behind doing. DS thus performed AIEng work without yet becoming AIEng.

Second, because liminality was eased, DS could inhabit it and look for ways to benefit from it. They used the ambiguity of the role as a strength, claiming either label as the situation made convenient and leveraging their in-betweenness for further career opportunities. While the mechanisms that produced the state came from outside the worker, the tactics that let them benefit from it came from themselves. Steering, qualifying, performing and entraining were driven by the technological and organisational context, and they redirected what DS were asked to do and what they were able to do. Naturalising, leveraging and hedging come from within the individual. These tactics changed how DS read their own position while performing. This suggests that enduring liminality may not depend on institutions per se. When technology assists workers in performing the work, they can carry the passage on their own and even see benefits from their liminal state. This makes the state easier to endure.

Third, the labour market context suggests that the conditions for this passage are not particular to TechCo. Occupations more exposed to GenAI redirected more of their exits toward AIEng after the shock, and technology consulting, where the DS to AIEng path was already more travelled, may be a revealing setting for understanding occupational mobility among IT workers.

This study has several limitations. The case rests on nine DS in a single firm, five of whom experienced liminality, and TechCo's identity as a technology consultancy is precisely what made naturalising available to participants. This may not travel to settings where new technical work reads as an imposition rather than as ordinary practice. The second round of interviews came several months after the first, which is long enough to register change but too short to establish whether the liminality I observed persists for long periods of time. Finally, the labour market data measure transitions, not liminality. They show that the conditions for the state were widespread, but not that anyone experienced it.

I see two ways in which future research can better understand technology-enabled liminality. First, it could get closer to the process that produces the transition, which this study captures only in outcome. An ethnographic approach, or the analysis of workers' interactions with GenAI in their daily work, could show how it substitutes for the doing, and comparing those exchanges with how workers interact with their peers could recover the relating dimension I set aside (Anteby et al., 2016), revealing whether the occupational community still socialises new entrants or whether GenAI displaces that too.

Second, research could look for traces that the work itself is changing rather than only counting transitions. My case suggests that workers keep their titles while the work changes underneath them, and this should be visible at scale, in job descriptions filling with AIEng tasks under a DS title and in the emergence of hybrid titles. These patterns would not measure liminality, but they may indicate how widespread the conditions behind it have become. Pairing this with more direct measures of adoption, such as firms' spending on GenAI services, could tie these changes in work more firmly to the technology itself.

7. Acknowledgements

This research was supported by the ARC Training Centre for Information Resilience (CIRES) and the University of Queensland. This work was supported by resources provided by The University of Queensland Research Computing Centre's Bunya supercomputer (https://dx.doi.org/10.48610/wf6c-qy55), with funding from The University of Queensland,

Brisbane, Australia. I used Claude for spell checking and coding assistance. Approved by the University of Queensland Human Research Ethics Committee (2024/HE001632).

8. References

Acemoglu, D., & Restrepo, P. (2019). Automation and New Tasks: How Technology Displaces and Reinstates Labor. *Journal of Economic Perspectives*, *33*(2), 3–30. https://doi.org/10.1257/jep.33.2.3

Anteby, M., Chan, C. K., & DiBenigno, J. (2016). Three Lenses on Occupations and Professions in Organizations. *Academy of Management Annals*, *10*(1), 183–244. https://doi.org/10.5465/19416520.2016.1120962

Ashforth, B., Rogers, K., Pratt, M., & Pradies, C. (2014). Ambivalence in Organizations: A Multilevel Approach. *Organization Science*, *25*(5), 1453–1478. https://doi.org/10.1287/orsc.2014.0909

Beech, N. (2011). Liminality and the practices of identity reconstruction. *Human Relations*, *64*(2), 285–302. https://doi.org/10.1177/0018726710371235

Bourgoin, A., & Harvey, J.-F. (2018). Professional image under threat: Dealing with learning-credibility tension. *Human Relations*, *71*(12), 1611–1639. https://doi.org/10.1177/0018726718756168

Brazier, C. É., Parmentier, M., & Masdonati, J. (2025). Navigating Involuntary Career Changes: Emotional Dynamics During Work-Related Identity Loss and Recovery. *Journal of Career Development*, *52*(6), 758–782. https://doi.org/10.1177/08948453251394015

Brynjolfsson, E., Li, D., & Raymond, L. (2025). Generative AI at Work. *The Quarterly Journal of Economics*, *140*(2), 889–942. https://doi.org/10.1093/qje/qjae044

Chamakiotis, P., Symon, G., & Whiting, R. (2024). Agentic interplay between hybridity and liminality in contemporary boundary work. *Information Systems Journal*, *34*(1), 261–283. https://doi.org/10.1111/isj.12477

Charmaz, K. (2014). *Constructing grounded theory* (2nd edition). Sage.

Czarniawska, B., & Mazza, C. (2003). Consulting as a Liminal Space. *Human Relations*, *56*(3), 267–290. https://doi.org/10.1177/0018726703056003612

Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., & Lakhani, K. R. (2023). Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality. *SSRN Electronic Journal*. https://www.ssrn.com/abstract=4573321

Dey, I. (1999). *Grounding grounded theory: Guidelines for qualitative inquiry*. Academic Press.

Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2024). GPTs are GPTs: Labor market impact potential of LLMs. *Science*, *384*(6702), 1306–1308. https://doi.org/10.1126/science.adj0998

Garsten, C. (1999). Betwixt and between: Temporary Employees as Liminal Subjects in Flexible Organizations. *Organization Studies*, *20*(4), 601–617. https://doi.org/10.1177/0170840699204004

Humlum, A., & Vestergaard, E. (2025). *Still Waters, Rapid Currents: Early Labor Market Transformation under Generative AI* (Working Paper No. 33777). National Bureau of Economic Research. https://doi.org/10.3386/w33777

Ibarra, H. (1999). Provisional Selves: Experimenting with Image and Identity in Professional Adaptation. *Administrative Science Quarterly*, *44*(4), 764–791. https://doi.org/10.2307/2667055

Ibarra, H., & Obodaru, O. (2016). Betwixt and between identities: Liminal experience in contemporary careers. *Research in Organizational Behavior*, *36*, 47–64. https://doi.org/10.1016/j.riob.2016.11.003

Langley, A. (1999). Strategies for theorizing from process data. *Academy of Management. The Academy of Management Review*, *24*(4), 691–710. https://doi.org/10.2307/259349

Levinson, D. J. (1986). *The Seasons of a Man's Life: The Groundbreaking 10-Year Study That Was the Basis for Passages!* Random House Publishing Group.

Mayer, T., & Recker, J. (2026). Deep structure changes and the evolution of the IT architect role during scaled agile transformations. *Journal of Information Technology*. https://doi.org/10.1177/02683962261464185

Nicholson, B., Babin, R., & Briggs, S. (2017). Exploring the effects of liminality on corporate social responsibility in inter-firm outsourcing relationships. *Journal of Information Technology*, *32*(1), 47–61. https://doi.org/10.1057/jit.2015.24

Noble, C. H., & Walker, B. A. (1997). Exploring the relationships among liminal transitions, symbolic consumption, and the extended self. *Psychology and Marketing*, *14*(1), 29–47. https://doi.org/10.1002/(SICI)1520-6793(199701)14:1%3C29::AID-MAR3%3E3.0.CO;2-Q

Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, *381*, 187–192. https://doi.org/10.1126/science.adh2586

Orlikowski, W. J., & Scott, S. V. (2021). Liminal innovation in practice: Understanding the reconfiguration of digital work in crisis. *Information and Organization*, *31*(1), 100336. https://doi.org/10.1016/j.infoandorg.2021.100336

Pratt, M. G., Rockmann, K. W., & Kaufmann, J. B. (2006). Constructing Professional Identity: The Role of Work and Identity Learning Cycles in the Customization of Identity Among Medical Residents. *Academy of Management Journal*, *49*(2), 235–262. https://doi.org/10.5465/amj.2006.20786060

Söderlund, J., & Borg, E. (2018). Liminality in Management and Organization Studies: Process, Position and Place. *International Journal of Management Reviews*, *20*(4), 880–902. https://doi.org/10.1111/ijmr.12168

Van Gennep, A. (1909). *The rites of passage*. University of Chicago press.

Weidenstedt, L., Geissinger, A., Leick, B., & Nazeer, N. (2024). Betwixt and between: Triple liminality and liminal agency in the Swedish gig economy. *Environment and Planning A: Economy and Space*, *56*(4), 1280–1297. https://doi.org/10.1177/0308518X231172984

Zhang, Z., Yoo, Y., Lyytinen, K., & Lindberg, A. (2021). The Unknowability of Autonomous Tools and the Liminal Experience of Their Use. *Information Systems Research*, *32*(4), 1192–1213. https://doi.org/10.1287/isre.2021.1022

9. Appendix

Table 2: Data structure and representative quotes

Steering	2nd: **Technological pull**		
	1st	Relying on off-the-shelf models for AI work (DS3, DS5, DS6)	"We're just taking a model that's provided, which is a black box, and then we use it." (DS3)
		Seeing easier to use AI offerings (AO, HCS)	"AI offerings have gotten a lot better, all our large technological partners has significantly improved the ability to do AI and embed it into organization for use" (HCS)
		Calling GenAI APIs rather than modelling (DS1, DS3, DS7, DS9)	"GenAI is more like calling APIs and stitching and plumbing stuff together." (DS7)
	2nd: **Organisational push**		
	1st	Clients increasingly demanding GenAI and agentic projects (DS1, DS3, DS4, DS5, DS6, DS7)	"Now it's all about agentic. Everyone wants an agent in the team… Everyone wants to do a trial in this area" (DS4)
		Pushing AI adoption without a clear purpose (DS3)	"There's a huge push to implement those technologies, but companies don't necessarily understand how to do that or what kind of business function they wish to fulfil." (DS3)
		Heading TechCo into the AI engineering direction (DS3, DS4, DS6, DS7)	"A large part of the business is kind of heading into that direction, like AI engineering kind of direction." (DS6)
		Demanding engineering rather than DS skills (DS1, DS3, DS8)	"We're only hiring engineers now. We're not hiring data scientists within TechCo. That's in response to client demand." (DS1)
Enacting	2nd: **Qualifying**		
	1st	Lowering the AIEng bar with GenAI (DS1, DS3, DS5, DS6, DS9)	"It's meant that in many instances I have been able to work as an ML engineer or software engineer, whereas before I don't think I would've been able to." (DS3)
		Learning AIEng skills with GenAI (DS1, DS2, DS5, DS6, DS8)	"I'm no longer doing a lot of model training. I'm doing a lot of deploying of the model, making it robust… so talking with the LLM about all that." (DS8) "Agentic coding, I wouldn't have had that if it wasn't for GenAI. A lot of my knowledge of computer vision models as well has come from GenAI. In a way it's sort of a fast track." (DS3)

		Seeing oneself as capable of AI engineering (DS1, DS5, DS7)	"I could do the engineering myself, but I might not be following a hundred percent best practises. But I can make stuff work and put it into production." (DS7)
		Expanding capabilities with LLMs (DS6)	"I think the main difference is that, now I'm just doing a lot with LLMs, so I can just solve problems with LLMs pretty much." (DS6)
	2nd: Performing		
	1st	Coding faster with GenAI (DS1, DS2, DS3, DS5, DS6, DS7, DS9)	"Something that would have taken me weeks, I could do in a few days." (DS6)
		Using GenAI as a safe space to ask questions (DS1, DS2)	"I'd prefer to go to a GenAI chat rather than asking people on my team, because I know everyone's very busy and I don't want to bother them." (DS2)
		Using GenAI for relational support (DS1, DS3, DS9)	"Generative AI has helped me move faster and helped me feel less alone, less stressed about my work." (DS1)
Entraining	**2nd: Entraining**		
	1st	Keeping pace with AI development (DS1, DS3, DS7, DS9, DS4)	"It's unnecessary anxiety sometimes because things are moving very quickly… you have one model super powerful this week, seven days later another model gets released." (DS9) "There's that learning curve of a new AI tool, but then after three months there's a newer AI tool and the framework has changed." (DS4)
		Feeling increased expectations with GenAI (DS2, DS3, DS5, DS6, DS8)	"When we've seen efficiency gains, it just means that the expectation of work is a lot higher." (DS3)
		Delivering more with GenAI (DS3, DS5)	"We just work for that fixed amount of time. If you can get more done, we're always trying to get as much done as possible" (DS5)
		Using GenAI to push through speed and fatigue (DS1)	"I use GenAI only when I'm tired and when I want to move fast on something." (DS1)
Accommodating	**2nd: Naturalising**		
	1st	Seeing design and experimental skills as the DS-AIEng overlap (DS3, DS6, DS9)	"It's pretty much still the same skills. I'm still doing data analysis, I'm still doing coding, but the tools I'm using are different." (DS6)
		Seeing fundamentals as unchanged by GenAI (DS9)	"I think that's the noise. But to me, fundamentally things are just the same." (DS9)
		Treating new technical tasks as ordinary consulting work (DS6)	"It's fine, really. To me, these things — as a consultant, we do new things all the time." (DS6)

		Naturalising skill loss as progress (DS5)	"I guess that's just what progress looks like. Lots of skills have been lost over time." (DS5)
	2nd: Leveraging		
	1st	Increasing work breadth and depth with GenAI (DS1, DS3, DS9)	"You would typically do two or three models because that's the only time you have at your disposal… what has changed is the breadth of work." (DS3)
		Doing more senior-related tasks (DS9)	"I like it too, to really not be doing the code but more doing the abstract layers and distil information." (DS9)
		Making up a new role (DS6, DS9)	"I think I'm just having a midlife career crisis or something. I'm slowly, gradually not seeing myself as a data scientist anymore. I now more identify myself as an AI-ML consultant." (DS9)
	2nd: Hedging		
	1st	Going into AIEng because of demand (DS7)	"There's so much demand for it, it makes sense for me to just dip my toes in… and if something comes along where you need that skillset, then I have it." (DS7)
		Doing DS and AIEng for role flexibility (DS5, DS7)	"I just want to be able to do both instead of just one. I think it's good to have a variety." (DS5)
		Doing AIEng for higher salary (DS7)	"My wife is not working at the moment, so you lose a bit of income. If you can compensate with a higher salary, that would be a good outcome." (DS7)
Anchoring	**2nd: Anchoring**		
	1st	Critiquing GenAI-based learning (DS3, DS5)	"It's made that learning process a lot simpler, but because it's simpler it becomes a lot more surface level… it's like a sugar hit." (DS3)
		Worrying about skill erosion from GenAI (DS1, DS3, DS4, DS5, DS7, DS9)	"You don't allow yourself space to fail or to work problems out for yourself… that sort of thinking to work through problems and to build that intuition can maybe get lost." (DS3)
		Critiquing AIEng work as generating rather than modelling (DS3)	"With GenAI it kind of removes that fun, because it's no longer about prediction or inference or statistical modelling. It's more about generating something.", "It's less about defining your own architecture, and more about using the API to create a solution." (DS3)
		Enjoying traditional DS work (DS3, DS5, DS7)	"Developing a model to be able to model the world — that model aspect is what makes it enjoyable." (DS3)

Table 3: Interview protocol round 1 (Data Scientist)

Theme	Question	Probes
Introductory (opening questions and rapport-building	- Can you please tell me about your current role and responsibilities as a data scientist?	- What is your current role? - How long have you worked here?
Organisation	- Can you describe your organisation and the data science team you work with?	- In your words, what does TechCo and your team do?
Tasks	- In general, how has GenAI impacted your work tasks? - Where do you find GenAI being the most and least impactful for your work? - You mentioned that the task you allocate the most time to is: [Model deployment]. Could you tell me how you use GenAI for it? - How has GenAI impacted senior vs junior data scientists? Is it the same or different impact for both? - Has GenAI helped you or hindered you in how you bring together different parts of a data science problem (for example, data, business needs, etc)? - Please recall an example where you used GenAI for work creatively or surprisingly?	- Think about a project you recently participated in. Can you recall a moment or task in which GenAI was beneficial? - Has GenAI made some senior-level tasks more accessible to junior data scientists? - What is the most unconventional way you have used GenAI at your work? - Can you give me an example of an instance in which GenAI defied your expectations? - Can you provide an example of a time when you were frustrated or amazed by a response from GenAI? - Do you have an example in which you had an a-ha moment or a surprise on how GenAI can or cannot help? - Has GenAI allowed you to allocate more time to other types of tasks? If so which ones? - Has GenAI enabled more complex problem solving or strategic thinking?
Engagement, communicating and collaborating	- How has GenAI impacted how you interact with other workers, data scientists or managers? - How has GenAI influenced how you collaborate with other senior and junior data scientists? - Could you provide an example of how GenAI was particularly useful in communicating with senior stakeholders or domain experts?	- Has GenAI changed who you interact with and how? - Do you feel you are interacting more now with GenAI than your colleagues? - Have your meetings changed because of GenAI?
Skills development	- How do you use GenAI for skill development? - What is the impact of GenAI on skill development? Has it undermined your skills and enabled you to acquire more skills? - How did GenAI, if at all, change how you learned from senior data scientists [mentored junior data scientists]? - Has GenAI made your work more or less challenging? If so, how?	- Has GenAI helped you or hindered you in acquiring skills? - How did you improve your skills before GenAI? Is it different now? - Which tasks do you find easier/harder because of GenAI? - For which of these tasks are you most reliant on GenAI for? - Are you dealing with new challenges because of GenAI? Which ones and how? - What are the most challenging aspects of your work now compared to the past?
Risk management	- What are the organisational risks that GenAI brings for data science work, if at all? - How should data scientists manage those risks? - Do you see any difference between how juniors vs seniors	- Do you think is risky to use GenAI? - Considering projects, professional development, organisation, and society risks, which strategies or practices work best for handling them? - What type of information do you share with GenAI?

	handle risks when using GenAI for work? -	
General impact	- In general, has GenAI impacted your work? If so, how?	- Has GenAI changed how many projects you're involved in or how much you do in each project? - How has GenAI impacted how senior vs junior data scientists are seen throughout your organisation?
Future of work	- How do you see GenAI shaping your role in the future? - How do you see GenAI changing how data scientists add value to organisations? - From your perspective as a junior (senior) data scientist, how do you see GenAI influencing the trajectory of your career?	- Do you think GenAI will create more or fewer data science jobs? - Which data science tasks will be automated and which will not?
Concluding	- Is there anything that you would like to add or reflect on?	-

Table 4: Interview protocol (Data Managers)

Theme	**Question**	**Probes**
Introductory (opening questions and rapport-building	- Can you tell me about your role, and your typical day? -	-
GenAI use	- How long have you used Generative AI (GenAI) in your work? - How do you use it?	
Working with data scientists	- Can you tell me about a specific project or episode where you worked closely with data scientists?	- I saw that you were involved in the X and Y projects, which included AI solutions. Did you work there with data scientists? How?
	- Have you supervised or collaborated with them recently? -	-
Tasks	- Could you share an example in which you were surprised by what GenAI could do and how it could affect Data science work?	-
	- Could you share an example in which you were concerned about GenAI's impact on data science work?	-

Adopting GenAI at TechCo	- What has TechCo done to advance its AI strategy? When did TechCo start encouraging engineers to use GenAI for their work?	-
	- How transformative has AI been for TechCo? What are the most considerable impacts you have seen?	-
	- What have been the biggest benefits and challenges of adopting GenAI?	-
	- Some data scientists have told us that juniors may over-rely on GenAI? Have you seen this firsthand? Has TechCo taken any measures for that?	-
	- How has TechCo's AI transformation compared with its client AI transformation? What have been the challenges and successes for TechCo's AI transformation? -	-
GenAI's impact	- Have you seen any impacts on people's productivity and work expectations?	- Have you changed expectations for their work? For example, are you expecting more productivity because of GenAI?
	- Have you noticed whether GenAI has affected everyone's productivity in the same way, or have some people been more affected than others?	-
	- Could you describe a moment in which this showed up and the challenges it brought up?	-

Risks	- How have you managed the risks of other people using GenAI?	-
	- How do you evaluate other people's work, given that they could be using GenAI?	
Collaboration	- How has collaboration been impacted, if at all, because of GenAI?	- Have you noticed any changes in how people are collaborating since GenAI's emergence here at TechCo?
	- Think of an episode where your team faced a tricky problem in which collaboration happened. Did GenAI come into the picture? Why? Or Why not?	
Deskilling	- How do you measure if people's skills are up to date at TechCo?	-
	- How have TechCo and you managed deskilling when using GenAI?	-
Recruiting and team composition	- Have you seen a headcount change regarding data scientists since GenAI's emergence?	- Have new roles emerged because of using GenAI? -
	- How has GenAI changed how you evaluate new job candidates?	- Are you prioritizing other skills now because of GenAI?

	- Recently, I learned, and correct me if I am wrong, but TechCo stopped hiring junior data scientists. Do you have any information on this? Have you seen any consequences of this?	- Have you had staffing changes (increase/decrease) across the roles you have on your team? Has the focus of roles shifted?
General impact	- How do you see the future of data science teams in the GenAI era?	-
	- What do you think is the value data scientists add as opposed to GenAI?	
Future of data science	- Is there anything that you would like to add or reflect on?	-

Table 5: Interview protocol round 2 (Data Scientist)

#	Theme	Questions
Section 1: Background		
1	Opening: follow-up on round one	- In our previous interview you said your background is in data science but you are now doing more ML/AI engineering tasks, and that your role title changes often. If someone outside TechCo asked what your occupation is, how would you identify yourself? - You said GenAI unlocks opportunities for you. Why were those opportunities locked before?
2	Prior work and identity as a data scientist	- Before things started shifting, how would you describe your work as a data scientist? What did that role mean to you? - Before this shift, what was your perception of the MLE role? - Talking about your overall sense of self, what do you look for in a job? What makes an occupation fulfilling?
Section 2: Work transition		
3	Taking on new tasks	- Could you go back to when your work first started shifting toward MLE work (e.g., your first project of this kind)? What happened in between? - How were you using GenAI at that time? - What role did GenAI play in learning the new tasks on the job? - Did GenAI let you do things you could not do before, or do them faster? - How did you know that what you were doing was right? Did you feel comfortable, or nervous, relying on the code or results it produced? - What challenges and opportunities did you foresee when you started this new kind of work?
4	Counterfactual	- Imagine you could not use GenAI for your own tasks. Would you still be able to do MLE work within the project timeframe? - What would change? How would you adapt?

Skill Erosion and Identity Shifts under AI Delegation in Imaging-based Medicine

Thitaree Tanprasert, Tim Miller, Tapani Rinta-Kahila
The University of Queensland

Abstract

This paper reports on a qualitative study with 12 experts across dermatology and radiology about the risk of AI-induced skill erosion. While existing literature focuses on short-term behavioural phenomena like automation bias and real-time accuracy, it offers limited insight into how repeated delegation to AI may alter what clinicians consider essential skills and reshape professional identity over time. We conducted 12 semi-structured interviews and speculative design workbook activities with expert radiologists and dermatologists. From this study, we derived the Human-AI Clinical Skill Framework, which maps how modular, unitary, and dynamic skill philosophies intersect with mechanical, integrative, and translational skill tiers, and what it means for clinical expertise in an AI-powered practice. We distinguish skill erosion among experts from impeded skill acquisition among trainees and conclude with conditional design principles for integrating AI into medical tools to mitigate skill erosion.

1. Introduction

As artificial intelligence (AI) systems are rapidly transforming professional practice in various fields of knowledge work, decision-making is increasingly supported by AI models. Research suggests that AI-driven decision aids, combined with the organisational goals that prioritise efficiency, often encourage over-reliance, inhibiting up-skilling in domain-specific knowledge and skills, critical thinking, and decision-making (Wahlström et al., 2024). In medicine, research has shown that endoscopists eroded diagnostic ability after just a few months of use of an AI decision aid, as they become over-reliant on AI and experience a phenomenon called AI-induced skill erosion or de-skilling (Budzyń et al., 2025). Though this practice may produce short-term performance gains, the long-term, individual, and organisational performance will drop with the loss of expert's knowledge and skills at an institutional level (Rinta-Kahila et al., 2023). These risks are particularly critical in medical diagnosis, where errors can harm patients and carry legal consequences (Li et al., 2025).

Existing human-AI interaction and medical AI literature predominantly examines AI integration as a static, point-in-time decision challenge, focusing heavily on short-term behavioral phenomena such as automation bias (Abdelwanis et al., 2024) and trust calibration (Tun et al.,

2025). While these inquiries offer valuable insights into how clinicians evaluate AI predictions, they represent a rather technical approach towards the skill erosion phenomenon which is inherently sociotechnical (Sarker et al. 2019): experts' work tends to be situated in an organisational context wherein their skills and abilities are interpreted and valued in different ways depending on how work processes are structured and how the experts construct their professional identity (see e.g., Strich et al. 2021). These understandings also change over time as technologies get incorporated into the organizational context (Rinta-Kahila et al. 2023). Indeed, prior studies have not captured the longitudinal effect of AI use on clinicians, specifically what skills clinicians see as essential, the conditions for which certain tasks can be delegated to AI, and how AI affects the clinicians' concept of skills and professional identities. We approach these questions through delegation, the transfer of rights and responsibilities for tasks between humans and agentic IS artifacts (Baird & Maruping, 2021), because what is delegated is no longer practised. Delegation may be chosen by clinicians or imposed through workflows and workload pressure, and AI may support upskilling as well as deskilling (Morandini et al., 2023).

Our study addresses these problems under the main research question: *How do clinicians conceptualize the skill erosion and skill development they anticipate when delegating work to AI?* From interviews and speculative design activities with 12 dermatologists and radiologists, we derived the *Human-AI Clinical Skill framework*, which maps the structural interactions between three clinical task tiers (mechanical, integrative, and translational) and three clinician skill philosophies (modular, unitary, and dynamic). This framework highlights the interaction between clinician's perspectives on AI, factors for task delegation, and values related to professional skills and identities. We contribute primarily to IS research on skill erosion and delegation (Baird & Maruping, 2021; Rinta-Kahila et al., 2023) and derive design strategies that may mitigate skill erosion.

2. Literature Review

2.1. AI decision support and reliance in medicine

Artificial intelligence (AI) decision support systems have rapidly advanced across imaging-based medical fields, such as dermatology and radiology (Esteva et al., 2017). Research on these systems has largely examined how clinicians rely on AI at the moment of decision. Behavioral indicators of inappropriate reliance include automation bias and overreliance (Parasuraman & Manzey, 2010), and under heavy workloads clinicians may seek relief from fatigue by approving AI suggestions without proper verification (Abdelwanis et al., 2024). However, prevailing medical AI research predominantly conceptualizes AI use as a static deployment rather than how longitudinal system usage can dynamically alter conceptualization and evaluation of clinical skills (Muller & Weisz, 2022).

2.2. Skill erosion, skill development, and professional identity

A central concern in human-AI interaction is skill erosion, especially reasoning and decision-making skills, after prolonged usage of automated decision support. In particular, literature on skill acquisition shows that prematurely delegating foundational execution tasks to automated systems impedes the development of intuitive expertise, a phenomenon empirically documented in aviation (Causse et al., 2025) as well as dermatology (Han et al., 2022).

Recent work distinguishes deskilling, the erosion of acquired expertise, from "never-skilling", the failure to acquire skills at all (Abdulnour et al., 2025; Natali et al., 2025), and IS research theorises skill erosion as self-reinforcing vicious circles (Rinta-Kahila et al., 2023). Yet AI can also develop skills, for instance as a training partner (Gaessler & Piezunka, 2023). Delegation also affects professional identity, which professionals construct by resolving mismatches

between what they do and who they are (Pratt et al., 2006); substitutive AI can threaten such identities (Jussupow et al., 2022; Strich et al., 2021).

3. Method

To answer our research question, we conducted a qualitative empirical study combining semi-structured interviews with speculative design workbook activities. We selected dermatology and radiology as our primary contexts because both domains involved imaging-based pattern recognition, making them early adopters of AI-powered diagnostic systems. This approach lets us ground the study in present workflows while pushing participants beyond current technology limitations to pinpoint their comfort boundaries, mental models, and professional values around AI integration (Jang & Nam, 2022).

3.1. Procedure

Data collection was structured into two parts, conducted remotely in one 60-minute session.

Part 1—Semi-structured interview: This part explored participants' real-world clinical workflows, time distribution, and mental effort across tasks. We probed participants' explicit conceptions of clinical competence and expertise, while identifying perceived difficulties, as well as factors for determining accountability in case of clinical mistakes.

Part 2—Speculative design workbook activities: Participants engaged in a two-part speculative design activity designed to push their boundaries around AI task delegation and evaluate how different interface interactions shape cognitive engagement:

1. *Compare-and-contrast interaction paradigms:* Participants were introduced to two contrasting diagnostic interface paradigms. The **Recommendation-First** layout immediately presents AI predictions, confidence scores, and suggested clinical actions upon opening a study. The **Cognition-First** layout requires the clinician to input their independent diagnosis first before revealing AI assessments. Participants discussed suitable tasks, required operator skills, and performance thresholds for each interaction mode.
2. *Boundary-finding scenarios:* We presented three scenarios, and asked participants to reflect on how these scenarios align with their values, cognitive process, and responsibilities, and how delegation to extreme use of AI in different parts of the work may affect their skills and professional code of conduct. The scenarios were: AI filtering out high-confidence cases (e.g., >95% confidence) so they never reach clinicians, though humans can audit (autonomy, accountability, authority); AI autonomously scheduling biopsies for high-risk conditions, leaving clinicians to override asynchronously (accountability, workload); and AI generating and sending referral letters to specialist departments or general physicians, with the clinician CC'ed (professional communication, reputation).

3.2. Participants

We recruited 12 medical professionals (6 dermatologists and 6 radiologists) through purposeful sampling and professional network snowballing. Inclusion criteria required participants to be at least 19 years old, hold an active medical license or residency appointment in dermatology or radiology, and actively engage in imaging-based diagnosis as part of their routine practice. The participant pool was balanced across career stages, ranging from trainees/residents to senior professors, across six countries (Australia, USA, Canada, Germany, Italy, UK), offering diverse views on training systems and professional code of conduct. Participants' prior familiarity with AI ranged from those using commercial AI

tooling daily in clinical practice to those whose exposure was limited to research trials.

3.3. Data analysis

Audio recordings were transcribed verbatim and analyzed in two phases: inductive theme generation informed by Braun and Clarke (2006), followed by framework-guided coding. This resembles hybrid inductive–deductive thematic analysis (Fereday & Muir-Cochrane, 2006), except that our coding template was derived from our own inductive phase rather than from prior theory.

Inductive Analysis: The primary researcher conducted “in-vivo” open coding on the transcripts, capturing participants' own vocabulary. The primary researcher met regularly with the broader research team to iteratively discuss, refine, and cluster these initial open codes into higher-level thematic categories. Codes describing what clinicians do were clustered into skill tiers, and codes describing how they reason about delegating these tasks into skill philosophies; crossing the two produced the 3×3 matrix.

Framework-guided coding: Once the core matrix construct was established, the research team applied the matrix as a coding template across all transcripts. This pass also surfaced contradictory cases, which showed that the philosophies are not fixed personality types. For example, senior participants sometimes held a unitary stance when supervising trainees, yet a modular or dynamic stance for their own routine tasks.

4. Findings

4.1. Human-AI Clinical Skill Conceptual Framework

To synthesize how varying degrees of AI decision support influence clinical competence, we developed the Human-AI Clinical Skill framework, which maps three **Skill Philosophies** against three **Tiers of Clinical Skills**. The three skill philosophies are:

1. The *unitary philosophy* conceives expertise as an indivisible and tightly integrated capability, where delegating any components to AI risks eroding essential skills.
2. The *modular philosophy* views clinical competence as a set of discrete, separable tasks that can be individually automated or delegated without altering the integrity of the remaining process.
3. The *dynamic philosophy* conceptualizes clinical competence as adaptive and evolving, with clinician expertise continuously shifting to focus on patient-facing, empathic interactions as AI becomes increasingly integrated into their workflow.

These philosophies are not strictly parallel: unitary and modular describe how expertise is structured, whereas dynamic describes how it changes over time. We retain all three because they emerged as distinct answers to the same question of whether delegating a task to AI threaten clinical competence.

The skill tiers which interact with these philosophies include:

1. The *mechanical tier* encompasses foundational, rule-based perception and execution, such as imaging-based measurements, surface pattern recognition, and automated template population.
2. The *integrative tier* involves synthesizing multi-modal clinical evidence, contextualizing diagnostic probabilities within patient histories, and evaluating differential diagnoses.

3. The *translational tier* addresses high-order interpersonal communication, patient care, and navigating ethical complexities.

Table 1 maps the skill philosophies across the three skill tiers, illustrating the resulting human-AI interaction dynamics and potential consequences on clinical skills. We use D and R after the participant's number to label them as dermatologist or radiologist, respectively. We distinguish skill erosion among experienced clinicians from impeded skill acquisition among trainees, and note that the findings reflect clinicians' perceptions and anticipations rather than observed skill change.

Table 1. The Human-AI Clinical Skill Matrix: Each cell describes the cognitive and professional challenges and considerations around delegation to AI arising from the interaction between each skill tier and each skill philosophy

Skill Tiers	Unitary Philosophy	Modular Philosophy	Dynamic Philosophy
Mechanical Tier *(e.g., surface pattern recognition, image filtering, initial triage)*	Mechanical skills are necessary foundation to develop skills in higher tiers (4.3)	Mechanical tiers are non-essential, but delegating them make trainees lack practice case volume necessary for developing clinical intuition (4.2.1)	Philosophical non-issue (4.4)
Integrative Tier *(e.g., diagnostic synthesis, contextualizing evidence, differential generation)*	AI should be utilized as a "second opinion" in integrative tasks (4.3)	Experts risk skill erosion in the integrative tier as workload pressure increases (4.2.3)	Philosophical non-issue (4.4)
Translational Tier *(e.g., patient communication, shared decision-making, ethical navigation)*	Translation tasks should be retained for humans due to legal and ethical liability to patients (4.3)	Experts risk becoming passive and less cognitively involved in overseeing AI for translational tasks (4.2.2)	Identity of healthcare providers may need reframing to focus mainly on translational skills (4.4)

4.2. The Modular Philosophy

4.2.1 Low case volumes impede trainees' skill acquisition

Offloading low-risk, routine tasks, such as image filtering or basic measurements of features in the scans, can help improve the efficiency and comprehensiveness of the workflow, as AI can be more thorough in these tasks than clinicians under fatigue (P7, R). However, introducing it early in the training process may create a critical blockage for early-career clinicians, who rely on high case volumes to build clinical intuitions and knowledge bases necessary for higher-level skills, such as image filtering skills: "Once you learned it, it's a boring skill, but you have to learn it to focus on the interesting cases later on" (P12, R). Similarly, delegating simple and obvious cases to AI could be detrimental, as "you need the normal cases to get a feeling for the complex and bad cases, so that ultimately connects to each other" (P6, R). Thus, while modular task delegation promises efficiency, applying it during training may impede skill acquisition, making trainees unable to step in when the cases are too difficult to delegate to AI.

4.2.2 Potential traps around AI-assisted communication

Clinicians who adopt a modular skill philosophy toward communication (e.g., writing referral letters and diagnosis reports) view text generation and narrative construction as isolated from diagnostic reasoning. However, participants cautioned that treating communication as a

separate block may create cognitive passivity, introducing systemic risks of poor integration of knowledge during diagnosis, referral, and other steps in the medical workflow. P11 (D) supports this with her reflection on using the AI medical scribe: "I don't have to listen to every word and process it straight away. And I might lose [information] reorganization skills because the AI is doing that for me" Participants are concerned that such passivity could lead to clinicians gradually signing off on automated text without rigorous oversight and expose clinicians to "legal problems" and “a sense of stolen identity" (P2, D) when clinical documentation no longer reflects the physician's voice or judgment in full.

4.2.3 Working with AI requires experts to upskill in AI literacy

Participants noted that, when delegating work to AI in the integrative tier, expert oversight requires deep AI literacy to resist the pressure of workload and fatigue. As AI gets better, clinicians with a modular philosophy would consider: "if I'm a lazy guy, I will tend to upload the whole stack of images and just go with whatever the AI will tell me. And I will soon maybe not check twice" (P12, R). Effective oversight therefore requires experts to know “what are the risks, or why you may not trust those results [..] how they are trained” (P1, D). Crucially, experts must develop an understanding of how AI errors differ fundamentally from human mistakes, necessitating different auditing strategies than traditional peer review when delegating the same work to a human colleague, as P10 (D) asked of an AI as competent as a trusted colleague: “are the kinds of mistakes the same?”. This shows that modularly delegating work to AI requires redefining the tasks and developing new skills.

4.3. **The Unitary Philosophy**

Under a unitary skill philosophy, clinicians view diagnostic synthesis as a holistic cognitive process, arguing that AI must function strictly as a post-hoc "second opinion" rather than a primary decision aid. Presenting AI recommendations upfront induces automation bias because “I already know a hypothesis. So the bias would be that I start by considering this hypothesis.” (P4, D) Because both human intuition and AI models are fallible, AI can provide a safety net: “I think it's better if you use it all the time. Even with a clear diagnosis, even if you are sure.” (P1, D)

While systems often prioritize efficiency gains, the unitary philosophy dictates that clinicians must actively resist cognitive offloading to preserve core competencies. Because final legal and ethical liability rests on human clinicians, experts view continuous skill retention as a part of their responsibilities, both for exercising independent judgment and supervising AI. P3 (R) noted: "in terms of making diagnoses and final recommendations, there should be a human involved. Just because if there is something that goes wrong, at least there was a human". Retaining unassisted judgment also protects their identity as the accountable decision-maker (cf. Strich et al., 2021).

4.4. **The Dynamic Philosophy**

Clinicians embodying the dynamic philosophy view clinical expertise as an evolving, adaptive partnership with automated systems. Under this perspective, the potential erosion of human capabilities across the mechanical and integrative tiers is largely viewed as a non-issue. Senior clinicians even described such skill change as historically recurring:

> *"What we deem a core skill today may not be a core skill tomorrow. Things that we did 40 years ago, we don't do today. [...] Through the innovation of radiology, we've had de-skilling. So, de-skilling is not new phenomena, as better techniques come out"* (P7, R).

Clinicians expect that their value would likely be relocated to the translational tier, where clinicians shift their focus toward empathy, communication, and patient engagement because

"there's a large gap between medical speech and patient speech, and so we have to translate diagnosis and findings into actionable and understandable tasks and information for the patient." (P12, R). Going further, the only irreplaceable part of the workflow may be human presence because AI has yet to replicate interpersonal connection, as P2 (D) notes: "A good and bad diagnosis should be communicated by a person, and not by a tool, by technology in general, because we are human, so we need the contact and the empathy and the warm and the kindness of another person". This shift would require redefining professional identity (P7, R) and raises the question "What is the meaning of a medical license to practice?" (P9, R).

5. Discussion

5.1. Relating to existing literature and frameworks

Prior research has documented how AI integration into the medical workflow can threaten professional identity (Jussupow et al., 2022), the necessity of redesigning curricula alongside AI integration (Moldt et al., 2024), and the concern around automation bias for clinician's decision support (Abdelwanis et al., 2024). Our Human-AI Clinical Skill Matrix builds upon these isolated observations by providing a unified conceptual framework.

Our framework extends IS research on skill erosion (Rinta-Kahila et al., 2023) with a temporal explanation that varies by career stage and tier: repeated delegation of mechanical tasks threatens trainees' skill acquisition, whereas repeated delegation of integrative and translational tasks under workload pressure risks eroding experts' existing expertise through passive oversight. Participants were willing to delegate task execution but not responsibility (Baird & Maruping, 2021) and located the risk of erosion in this gap; delegation imposed by workload pressure appeared more erosive than delegation chosen deliberately. The philosophies also map onto identity responses (Pratt et al., 2006): unitary clinicians protect an identity centred on independent judgment, modular clinicians risk a "stolen" identity, and dynamic clinicians anticipate rebuilding their identity around translational tasks. Participants also saw upskilling opportunities in AI literacy and auditing (Morandini et al., 2023), though P10's concern that AI errors differ from human ones suggests these skills must be developed deliberately (Gaessler & Piezunka, 2023). Overall, our work deepens the sociotechnical understanding of the skill erosion phenomenon by considering how the professionals construct their skills in situated work context and how AI integration interacts with these concepts.

5.2. Design implications

Based on our framework, we outline three design principles for AI-integrated clinical tools to mitigate skill erosion. These principles are conditional: active engagement matters most for trainees, integrative and translational tasks, and high-stakes cases, whereas offloading may suit experienced clinicians on high-volume mechanical tasks where AI is reliable:

1. **Enable specialization transitions (addressing 4.2.1 & 4.4):** As diagnostic tasks migrate to AI, systems should equip emerging specializations (e.g., technical auditor, interpersonal communicators) with the target metrics they need to guide their evolving workflow. Patient-facing risk translation tools, such as vocabularies and diagrams, can bridge the gap between model outputs and patient dialogue.
2. **Prevent automation bias (addressing 4.3):** Interfaces should enforce independent clinician reasoning before displaying machine predictions to ensure that its decision is taken strictly as a "second opinion". This approach is called cognition-first in human-AI interaction literature and could be executed through various mechanisms, such as evaluative AI (Miller, 2023) and frictional AI (Natali, 2023).

3. **Provide proactive steps beyond oversight (addressing 4.2.2 & 4.2.3):** Under high workload pressure, interfaces should introduce targeted cognitive-forcing functions to maintain active oversight. For example, they could implement deliberate micro-delays or mandatory confirmation prompts when clinicians attempt rapid sign-offs under tight timelines. As for clinical documentation (translational tier), generative AI tools may be restricted to producing outlines rather than complete, finalized prose, forcing the clinician to actively assemble and review the final clinical report.

5.3. Limitations and Future Work

We acknowledge three limitations. First, while our sample of twelve clinicians could capture domain-specific insights, it may not fully represent dynamics in non-imaging fields. Second, our investigation relied primarily on semi-structured interviews and speculative design probes, which are effective for exploring existing challenges and surfacing mental models but don't capture real-time, long-term observational data. Third, our data capture perceived rather than observed skill change, and although our sample approaches recommended sizes for thematic saturation (Wutich et al., 2024), theory development will require a larger sample.

To address these limitations, we will design and deploy AI-integrated prototypes that have the interaction mechanisms from Section 5.2. We will then conduct a longitudinal study tracking junior clinician's skills in each tier over an extended period, aiming to give quantitative measures of cognitive skills over time. We will also expand the sample and apply the Gioia methodology (Gioia et al., 2013) to support theory building.

6. Conclusion

This study introduces the Human-AI Clinical Skill framework, which connects three clinical skill tiers (mechanical, integrative, and translational) with three skill philosophies (modular, unitary, and dynamic). Our findings suggest that clinicians do not perceive skill erosion as an inevitable consequence of AI integration, but anticipate risks when tasks are segmented and delegated under acute workload pressure, leading to lowered skill acquisition in trainees and passive AI oversight in experts.

Acknowledgements

We express our sincere gratitude to the participating experts for their valuable contributions to this study (in alphabetical order of family names): Dr. Camilla Chello (Dermatology Clinic, Department of Medical and Cardiovascular Sciences, Sapienza University of Rome); Asso. Prof. Errol Colak (Department of Medical Imaging, University of Toronto); Dr. Harrison Edwards (Dermatology Department, Royal Brisbane and Women's Hospital); Asso. Prof. Simone M. Goldinger (Frazer Institute, University of Queensland and Metro South Health / Princess Alexandra Hospital); Dr. Maria Elisabetta Greco (Dermatology Clinic, Department of Medical and Cardiovascular Sciences, Sapienza University of Rome); Dr. Anish Kirpalani (Department of Medical Imaging, St. Michael's Hospital, and Department of Medical Imaging, University of Toronto); Dr. Alba Lecumberri (Hospital Universitario Ramón y Cajal); Prof. Dr. Lara Valeska Maul-Duwendag (Dermatologische Klinik, Universitätsspital Zürich); Dr. Lukas Müller (Department of Diagnostic and Interventional Radiology, University Medical Center Mainz); Dr. Harpreet Hyare Richardson (Department of Translational Neuroscience and Stroke, UCL Queen Square Institute of Neurology); Dr. Andreas Schicho (Faculty of Medicine, University of Regensburg); Kay Wu (Department of Medical Imaging, Temerty Faculty of Medicine, University of Toronto).

References

Abdelwanis, M., Alarafati, H. K., Tammam, M. M. S., & Simsekler, M. C. E. (2024). Exploring the risks of automation bias in healthcare artificial intelligence applications: A Bowtie analysis. Journal of Safety Science and Resilience, 5(4), 460-469.

Abdulnour, R.-E. E., Gin, B., & Boscardin, C. K. (2025). Educational strategies for clinical supervision of artificial intelligence use. *New England Journal of Medicine, 393*(8), 786-797.

Baird, A., & Maruping, L. M. (2021). The next generation of research on IS use: A theoretical framework of delegation to and from agentic IS artifacts. *MIS Quarterly, 45*(1), 315-341.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology*, *3*(2), 77-101.

Budzyń, K., Romańczyk, M., Kitala, D., Kołodziej, P., Bugajski, M., Adami, H. O., ... & Mori, Y. (2025). Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: a multicentre, observational study. The Lancet Gastroenterology & Hepatology, 10(10), 896-903.

Causse, M., Mercier, M., Lefrançois, O., & Matton, N. (2025). Impact of automation level on airline pilots' flying performance and visual scanning strategies: A full flight simulator study. *Applied Ergonomics, 125*, 104456.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *nature, 542*(7639), 115-118.

Fereday, J., & Muir-Cochrane, E. (2006). Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. *International Journal of Qualitative Methods, 5*(1), 80-92.

Gaessler, F., & Piezunka, H. (2023). Training with AI: Evidence from chess computers. *Strategic Management Journal, 44*(11), 2724-2750.

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.

Han, S. S., Kim, Y. J., Moon, I. J., Jung, J. M., Lee, M. Y., Lee, W. J., Won, C. H., Lee, M. W., Kim, S. H., Navarrete-Dechent, C., & Chang, S. E. (2022). Evaluation of artificial intelligence-assisted diagnosis of skin neoplasms: A single-center, paralleled, unmasked, randomized controlled trial. *The Journal of Investigative Dermatology, 142*(9), 2353-2362.e2.

Jang, S., & Nam, K. Y. (2022). Utilization of speculative design for designing human-AI interactions. *Archives of Design Research, 35*(2), 57-71.

Jussupow, E., Spohrer, K., & Heinzl, A. (2022). Identity threats as a reason for resistance to artificial intelligence: survey study with medical students and professionals. *JMIR formative research*, *6*(3), e28750.

Li, J., Zhou, Z., Lyu, H., & Wang, Z. (2025). Large language models-powered clinical decision support: Enhancing or replacing human expertise? *Intelligent Medicine, 5*(1), 1-4.

Miller, T. (2023, June). Explainable AI is dead, long live explainable AI! Hypothesis-driven decision support using evaluative AI. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency* (pp. 333-342).

Moldt, J. A., Festl-Wietek, T., Fuhl, W., Zabel, S., Claassen, M., Wagner, S., ... & Herrmann-Werner, A. (2024). Assessing AI awareness and identifying essential competencies: insights from key stakeholders in integrating AI into medical education. *JMIR Medical Education, 1*0, e58355.

Morandini, S., Fraboni, F., De Angelis, M., Puzzo, G., Giusino, D., & Pietrantoni, L. (2023). The impact of artificial intelligence on workers' skills: Upskilling and reskilling in organisations. *Informing Science, 26*, 39-68.

Muller, M., & Weisz, J. (2022, June). Extending a human-AI collaboration framework with dynamism and sociality. In *Proceedings of the 1st Annual Meeting of the Symposium on Human-Computer Interaction for Work* (pp. 1-12).

Natali, C. (2023). Per aspera ad astra, or flourishing via friction: Stimulating cognitive activation by design through frictional decision support systems. In *CEUR workshop proceedings* (Vol. 3481, pp. 15-19). CEUR-WS.

Natali, C., Marconi, L., Dias Duran, L. D., & Cabitza, F. (2025). AI-induced deskilling in medicine: A mixed-method review and research agenda for healthcare and beyond. *Artificial Intelligence Review, 58*(11), 356.

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human factors, 52*(3), 381-410.

Pratt, M. G., Rockmann, K. W., & Kaufmann, J. B. (2006). Constructing professional identity: The role of work and identity learning cycles in the customization of identity among medical residents. *Academy of Management Journal, 49*(2), 235-262.

Rinta-Kahila, T., Penttinen, E., Salovaara, A., Soliman, W., & Ruissalo, J. (2023). The vicious circles of skill erosion: A case study of cognitive automation. *Journal of the Association for Information Systems, 24*(5), 1378-1412.

Sarker, S., Chatterjee, S., Xiao, X., & Elbanna, A. (2019). The sociotechnical axis of cohesion for the IS discipline: Its historical legacy and its continued relevance. *MIS Quarterly, 43*(3), 695-719.

Strich, F., Mayer, A.-S., & Fiedler, M. (2021). What do I do in a world of artificial intelligence? Investigating the impact of substitutive decision-making AI systems on employees' professional role identity. *Journal of the Association for Information Systems, 22*(2), 304-324.

Tun, H. M., Rahman, H. A., Naing, L., & Malik, O. A. (2025). Trust in artificial intelligence–based clinical decision support systems among health care workers: systematic review. *Journal of Medical Internet Research, 27*, e69678.

Wahlström, M., Tammentie, B., Salonen, T. T., & Karvonen, A. (2024). AI and the transformation of industrial work: Hybrid intelligence vs double-black box effect. *Applied Ergonomics, 118*, 104271.

Wutich, A., Beresford, M., & Bernard, H. R. (2024). Sample sizes for 10 types of qualitative data analysis: An integrative review, empirical guidance, and next steps. *International Journal of Qualitative Methods, 23*.

A capability-based model of relational expertise in human–AI embedded work.

Leyla Yacine
Hanken School of Economics

Abstract

Although scholarship acknowledges artificial intelligence (AI) as an actor in working environments, its role in morphing capability development alongside humans remains largely unclear in AI-embedded work. This study addresses this gap by extending the concept of relational expertise as emerging through the development of four capabilities (i.e., interpretive, interactional, metacognitive, and integrative capabilities) that enable the configuration and reconfiguration of human and AI systems' agency across three forms of AI-embedded work: AI-assisted work (AI supports humans), AI-augmented work (humans and AI collaborate), and AI-driven work (AI leads work processes). It explores how individuals develop these capabilities by making sense of the circulation of agency among humans and AI across workflow stages (i.e. delegation, attribution, contingency, reassertion, and reconfiguration). Through this process, workers develop relational expertise, which enables them to adapt, learn, and thrive as human–AI work keeps evolving. A model of relational expertise development for sustainable human–AI embedded work emerges to guide future research.

1. Introduction

As AI systems become increasingly embedded in how we work, think, and live, concerns have intensified regarding their actual benefits in working environments and their implications for worker capabilities, expertise, and employability over time. On the one hand, technologies such as generative AI (GenAI) and agentic systems, offer substantial benefits. GenAI provides immediate, tailored support and feedback to perform a variety of tasks, streamline processes, and automate routine assignments (Riemer & Peter, 2024). Agentic AI further promises to augment human reasoning, enable multitasking, scale work processes, and assign tasks to AI assistants (Pan et al., 2025). On the other hand, these systems raise profound concerns for the future of work and labour markets. They notably redefine job roles, responsibilities, work contents and modalities, alter demands for competencies, upskilling and reskilling, and create fears of job displacement (Leonardi, 2025). As a result, professional opportunities are increasingly questioned in a context where AI systems can perform work as well, or better,

than humans while affecting the social, relational, regulatory, ethical, and identity aspects of work (Huang et al., 2026).

Because humans and AI systems are expected to coexist and collaborate to develop operations and achieve strategic goals (Einola & Khoreva, 2023), understanding the effects of human–AI embeddedness on skills and expertise has become paramount. Existing research highlights the importance of AI-related knowledge and literacy for adaptation and professional work (Ahmad & Bilal, 2025; Rapanta et al., 2025; Roy et al., 2025), yet little is known about how workers develop the capabilities needed to navigate human–AI embeddedness over time.

I therefore argue that relational expertise (Pakarinen & Huising, 2025) emerges through the development of specific capabilities via sensemaking of repeated interactions between humans and AI systems. Drawing on Leonardi's (2025) agency loops, I identify three forms of AI-embedded work that involve agency loops (i.e., delegation, attribution, contingency, reassertion, and reconfiguration), through which individuals configure and reconfigure agency circulation between human and AI actors: AI-assisted work (AI supports human work), AI-augmented work (AI and humans collaborate), and AI-driven work (AI leads work). I propose that, through sensemaking of this organisation of human–AI work within and across forms of work, workers develop distinct degrees of interpretive, interactional, metacognitive, and integrative capabilities, which foster sustainable human–AI embedded work. I define relational expertise as the ability to intentionally configure and reconfigure the allocation and flow of tasks, authority, judgement, responsibility, and learning between humans and AI systems in order to achieve current and future professional goals under evolving conditions.

This conceptual study offers two contributions. First, contrary to societal debates around the erosion of skills, expertise, and autonomy, it explains how individuals develop capabilities that help them renew their expertise in AI-embedded work. Second, it advances a model of relational expertise development by connecting the concept of relational expertise (Pakarinen & Huising, 2025) with that of agency loops (Leonardi, 2025) through the lens of sensemaking (Brown et al., 2015). In this way, it explains how, through their deliberate decisions to configure and reconfigure agency circulation between humans and AI systems, individuals develop capabilities that enable them to effectively navigate and orchestrate AI-embedded work over the course of time.The paper concludes by outlining future research avenues.

2. Expanding relational expertise under AI agency

2.1. From AI literacy to relational expertise in human–AI embedded work

Whereas AI literacy emphasises understanding AI systems and the implications of their use, it does not effectively account for AI's active role in capability development when agency circulates between humans and AI systems in work processes. AI literacy scholarship has predominantly focused on individuals as the primary actors responsible for learning about and implementing AI, often overlooking the increasingly agentic role that AI systems today play in the enactment of work –a concern also raised by T. Lintner in Rapanta et al. (2025). Although recent scholarship recognises AI as an actor with which humans develop interdependencies that give rise to employment opportunities, continuous learning, and career development (Ahmad & Bilal, 2025; Donald et al., 2024), further theorising remains needed to comprehend how AI participates in work processes and influences the capabilities that underpin professional work and professional growth.

The concept of *relational expertise*, introduced by Pakarinen and Huising (2025), offers a useful foundation in this direction by conceptualising expertise as emerging *in situ*, through the network of interactions in which an individual engages to perform work, generate specific knowledge, and claim and legitimise their professional contributions for recognition among diverse audiences. Expertise comprises specialised knowledge and skills, but these do not

exist in isolation. Instead, they are produced, enacted, and recognised through interactions with other people, objects, and technologies, such as AI systems (Pakarinen & Huising, 2025). Therefore, expertise is inherently situated, relational, built and renewed over time, as previous experiences are carried forward into future interactions.

Pakarinen and Huising (2025) distinguish three dimensions of relational expertise: (1) relational generation (expertise is "produced within and about assemblages of actors and objects" through which work is accomplished), (2) relational application (it is demonstrated through interaction with and adjustment to diverse audiences), and (3) relational recognition (it is "claimed, recognised, and institutionalised vis-à-vis other forms of expertise") (Pakarinen & Huising, 2025, p. 2061). These dimensions highlight expertise as developing via continuing interactions that connect past experiences, present practices, and future opportunities.

This lens is particularly relevant in AI-embedded work, where AI constitutes an agentic participant in the network of relations through which professional work and expertise unfold, evolve, and gain recognition. As agency circulates between humans and AI systems, these interactions shape not only expertise but also workers' longer-term career prospects (Huang et al., 2026; Leonardi, 2025; Donald et al., 2024).

However, while relational expertise underscores how expertise grows through human–AI interactions, it says little about the capabilities workers require to effectively collaborate with, challenge, supervise, and strategically integrate AI into their work. I therefore argue that the redistribution of agency between humans and AI systems provides a relevant starting point for specifying the capabilities through which relational expertise develops in human–AI embedded work. This requires examining the attribution and circulation of human and AI agency.

2.2. Redistributing agency in human–AI embedded work

2.2.1 Understanding human and AI agency in work processes

AI refers to data-driven algorithmic systems capable of recognising patterns, predicting scenarios, generating content, classifying information, and recommending content or courses of action (Köchling et al., 2025). Notably, GenAI and agentic systems exhibit expanding agentic capabilities: they can act autonomously and always more proactively work alongside humans, interact with and learn from their environment and, ultimately, influence individual and organisational activities and outcomes (Pan et al., 2025; Riemer & Peter, 2024). GenAI systems were initially conceived as advanced assistants limited to creating content in response to human prompts across diverse tasks (Riemer & Peter, 2024). Yet, recent developments have introduced more customised models with automated scripts, agentic AI systems, that autonomously interact with their environment, make decisions and optimise processes with minimal human intervention to proactively perform tasks, apart from initial instructions (Pan et al., 2025). AI, thus, actively contributes to organisational and individual work processes.

This has given rise to what I label *AI-embedded work*, namely work processes that integrate AI systems as participants alongside humans. Scholarship has identified three principal contexts of AI-embedded work: *AI-assisted work* (AI assists humans), *AI-augmented work* (humans collaborate with AI), and *AI-driven work* (AI leads and automates work processes) (Henkenjohann & Trenz, 2026; Huang et al., 2026; Krakowski, 2025; Pan et al., 2025; Riemer & Peter, 2024). All three contexts are characterised by different human–AI relations (see Table 1) that create distinct conditions for the development and enactment of relational expertise.

A key distinction between these contexts lies in how agency is attributed and exercised. *Agency* is the authority (or power) dynamically attributed to patterns of behaviour (or action) according to the perceived consistency between presumed intentions (or goals) and outcomes (or effects) (Leonardi, 2025). The sociotechnical configuration of work re-attributes agency among humans and AI systems because, irrespective of whether it is for assistance,

augmentation or automation, individuals cede at least a part of their perceived sense of authority and responsibility to systems to fulfil action goals (Leonardi, 2025; Pan et al., 2025). In AI-assisted and AI-augmented work, individuals often exert considerable agency over the modalities of their engagement with AI systems through AI adaptation strategies (Roy et al., 2025). In AI-driven work, however, direct human influence is more limited because AI takes greater control over the workflows (Leonardi, 2025). Overall, differences in agency attribution and circulation give rise to distinct AI-embedded work contexts and shape how individuals engage in, experience, and learn through human–AI relations (Henkenjohann & Trenz, 2026).

Such agency attribution and circulation in human–AI embedded work occurs through Leonardi's (2025) concept of *agency loops*. Those loops consist of delegation, attribution, contingency, reassertion and reconfiguration processes within workflows. *Delegation* involves configuration or initiation of AI action, fundamentally supporting human agency and setting the conditions for *attribution*, which then grants agency to AI as it begins the action. Yet, *contingency* events disrupt the action, exposing the limitations or failures of AI, biases, errors, or accountability issues, and the need for humans to reclaim agency through *reassertion*. The latter confers authority back to humans, who exercise control to rectify the course of action by developing new practices. *Reconfiguration* subsequently involves iteratively adjusting practices to enhance AI integration into workflows through evolved agency attribution and human–AI relations. In this regard, action arises from ongoing cycles of sensemaking (Brown et al., 2015) in AI-embedded work, including observation, interpretation, interaction, reflection, and decision-making leading to renewed behaviour among humans and AI systems (Leonardi, 2025). Hence, agency becomes a relational and dynamic phenomenon involving interaction and coordination of human and AI agency (Donald et al., 2024).

2.2.2 Agency attribution and circulation across AI-embedded work contexts

Drawing on Leonardi's (2025) agency loops, individuals engage in multiple contexts of AI-embedded work defined by different circulations of human and AI agency. In particular, Table 1 displays how the unfolding of the five processes of delegation, attribution, contingency, reassertion, and reconfiguration differs depending on whether AI primarily supports (AI-assisted work), collaborates with humans (AI-augmented work), or leads work processes (AI-driven work), thereby affecting how relational expertise is developed, enacted, and recognised.

Table 1. Clarifying agency loops across AI-assisted, AI-augmented, and AI-driven contexts of work.

	AI-assisted work	**AI-augmented work**	**AI-driven work**
Dominant agency pattern	Human agency prevails and AI supports.	Shared and complementary agency between humans and AI.	AI agency prevails and humans supervise.
Delegation	Humans delegate few tasks or information-processing activities to AI while they keep primary control.	Humans first initiate work division with AI, which can later suggest taking over specific tasks as needed.	Humans transfer substantial authority and task execution to AI.
Attribution	AI as a tool or assistant for limited operations.	Outcomes attributed to joint human and AI contributions.	Outcomes attributed to AI decisions and capabilities.
Contingency	AI recommendations or outputs conflict with human judgement or prove inaccurate.	Human and AI assessments diverge, causing ambiguity and coordination challenges.	AI bias, failures, opacity, unintended consequences expose risks of AI-led action.
Reassertion	Humans review, override, or correct AI outputs using professional judgement.	Humans recover interpretive authority while seeking to preserve collaborative benefits.	Humans reclaim responsibility, intervene, justify decisions and regain control as needed.
Reconfiguration	Refinement of AI use practices to better support human expertise.	Emergence of new routines that enhance human–AI collaboration.	Strenghtened governance, oversight, and accountability mechanisms to manage AI authority.

Below, I illustrate this variation through a vignette of performance review, learning and development (L&D), and promotion decisions in the field of human resources (HR).

In AI-assisted work, HR professionals use AI systems to support worker evaluations, identify development needs, and generate promotion recommendations (Nyberg et al., 2025). During delegation, HR professionals assign narrowly defined analytical tasks to AI, such as summarising performance data, identifying skill gaps, or suggesting L&D opportunities. They attribute agency to AI when they recognise the efficiency and value of AI-generated insights, while retaining primary authority over worker evaluation and career-related decisions. When contingencies arise due to apparently inaccurate assessment, overlooked contextual conditions, or recommendations conflicting with managerial observations, HR professionals must reassert their agency, adjust evaluations and possibly override AI recommendations by reviewing evidence and involving other humans. They then engage in work reconfiguration by reviewing and refining their use of AI systems, data inputs, and evaluation criteria and practices. Hence, the agency loop reinforces human authority, judgement and action while helping workers learn how and when AI effectively supports work (Huang et al., 2026).

In AI-augmented work, HR professionals collaborate with AI systems to make decisions regarding worker performance, L&D, and career advancement (Nyberg et al., 2025). For instance, HR professionals contribute strategic considerations, contextual knowledge and relational insights to AI's continuous analysis of performance metrics, teamwork dynamics, learning progress and career development potential. Here, delegation involves sharing evaluative and development responsibilities between human and AI actors. Agency is balanced as successful talent decisions are attributed to both HR expertise and AI analytics. However, contingencies may surface when AI identifies a worker as a high-potential candidate for a promotion whilst HR professionals or the worker's direct supervisor question their motivation, leadership readiness, or team fit. Reassertion is thus needed to understand what course of action should be taken based on contextual judgement, review of evidence, and questioning AI reasoning and recommendations. Reconfiguration then involves developing new collaborative practices that clarify when AI insights can be trusted and when additional human experience and judgement are needed. AI-augmented work is therefore jointly enacted and continuously negotiated between humans and AI systems (Roy et al., 2025).

In AI-driven work, AI systems lead talent management processes through ongoing evaluation of worker performance, recommendation of bespoke L&D opportunities, generation of promotion rankings, or identification of potential successors for positions. HR professionals' agency is limited to monitoring outcomes and intervening in exceptional circumstances (Nyberg et al., 2025). HR professionals delegate substantial authority to AI systems that perform evaluations, develop plans, and recommend actions autonomously. Yet, contingencies arise, for instance, when the system exhibits bias, HR professionals cannot sensibly justify the promotion recommendations, or a worker challenges promotion decisions. These events require reassertion of human agency to reclaim accountability, review decisions, and rectify courses of action. Subsequent reconfiguration helps introduce diverse governance mechanisms to prevent future problems by upholding human reviews at critical stages, introducing explainability requirements and fairness audits, or by clarifying appeal processes. Thus, while AI initially leads, human agency remains critical for fair governance and renewed value creation in AI-driven work (Huang et al., 2026; Krakowski, 2025, Pan et al., 2025).

These three AI-embedded work contexts epitomise different ways agency can circulate within the same work process, from primarily human-driven, to human–AI collaboration, to AI-driven action. Although these contexts display distinct circulations of agency between humans and AI, important differences remain in the tasks and responsibilities delegated and attributed to AI compared to those retained by humans. Such differences, both within and across AI-embedded work contexts, subsequently influence the nature, severity, and significance of contingencies and the responses that follow through reassertion and reconfiguration.

Importantly, as workers may simultaneously engage in different AI-embedded work contexts, they must be able to recognise how agency circulates and adapt their behaviour accordingly. This suggests that relational expertise develops through sensemaking of interactions within agency loops. Consequently, sustainable human–AI collaboration depends not only on domain expertise or AI literacy, but also on the ability to understand, navigate, and strategically regulate changing configurations of agency over time (Huang et al., 2026; Nyberg et al., 2025; Pakarinen & Huising, 2025). I therefore propose that sensemaking of agency loops gives rise to *relational expertise*, which I define as the ability to intentionally configure and reconfigure the attribution and flow of tasks, authority, judgement, learning, and responsibility between humans and AI systems in ways that support work processes and professional growth.

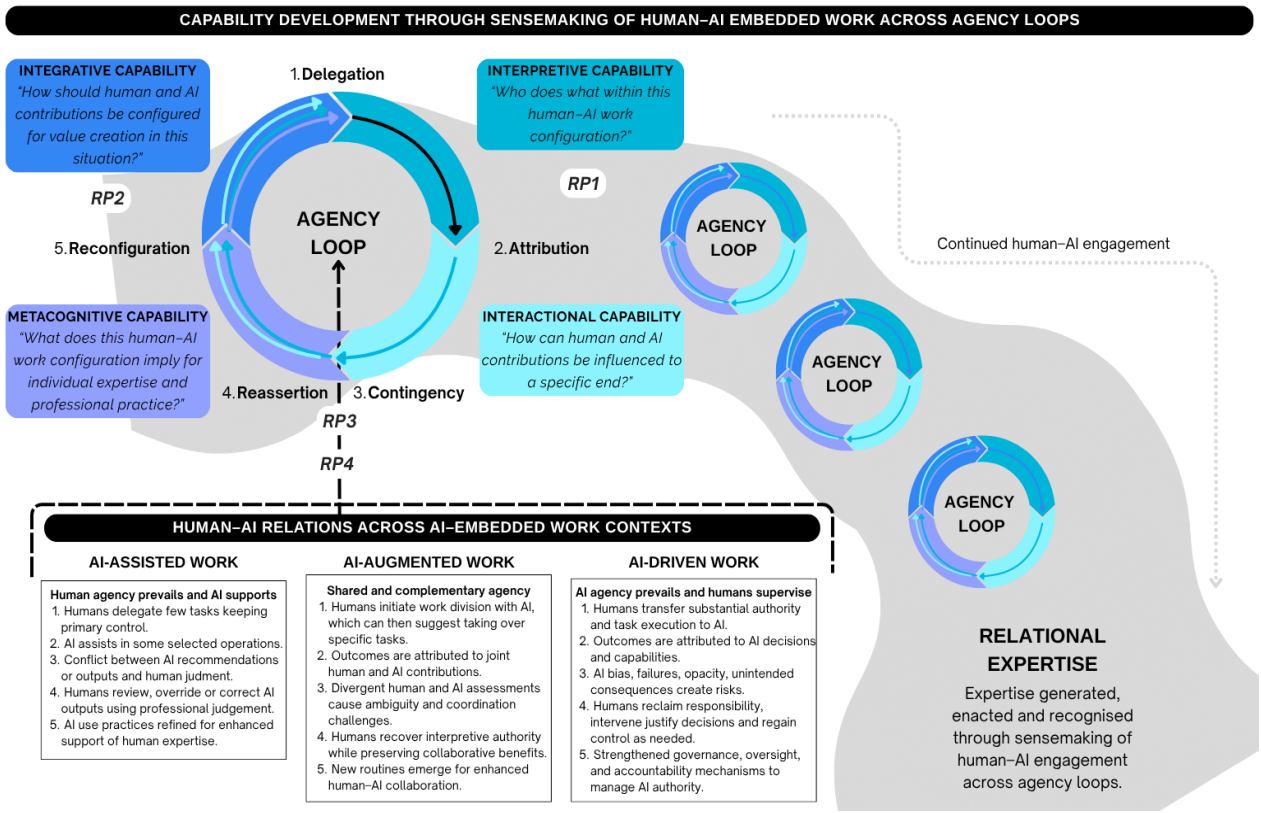
3. Developing relational expertise in AI-embedded work

In AI-embedded work, expertise depends on individuals' ability to work effectively with AI systems by merging domain knowledge with AI-related competencies (Krakowski, 2025; Donald et al., 2024). Although AI agency challenges traditional understandings of expertise and meaningful work because AI systems can replicate some tasks, complement, or even outperform humans (Köchling et al., 2025), it also creates new learning and capability development opportunities (Ahmad & Bilal, 2025; Pakarinen & Huising, 2025). From delegation to reconfiguration, agency loops illustrate how workers continuously learn through human–AI interactions while evaluating the quality, coherence, trustworthiness and fairness of action and adjusting their engagement to evolving goals and circumstances (Leonardi, 2025). This entails that agency loops unfold through sensemaking: they are recurring cycles of perception, interpretation, and action through which individuals construct meaning around human–AI relations and adapt their behaviour accordingly (Brown et al., 2015). By repeatedly engaging with AI systems, workers not only learn about AI but also how expertise itself is generated, enacted, and recognised within changing agency configurations, showing that relational expertise grows via situated sensemaking of human and AI actions and interactions.

Building on AI literacy scholarship and Leonardi's (2025) agency loops, I argue that relational expertise develops through four interrelated and mutually-reinforcing capabilities. First, *interpretive capability* refers to the ability to recognise the circulation of human and AI agency and to critically evaluate its implications for decisions and actions across situations while humans retain ethical responsibility (Krakowski, 2025; Rapanta et al., 2025). It requires asking "Who does what within this human–AI work configuration?". Second, *interactional capability* is the ability to deliberately engage with and influence AI behaviour via prompting, iterative refinement, and coordination of human and AI contributions towards desired goals (Krakowski, 2025; Rapanta et al., 2025). It revolves around the question "How can human and AI contributions be influenced to a specific end?". Third, *metacognitive capability* involves critically reflecting on the consequences of human–AI relations for individual thinking, learning, trust, dependence, judgement, accountability, and potential deskilling in AI-embedded work (Rapanta et al., 2025). It asks "What does this human–AI work configuration imply for individual expertise and professional practice?". Fourth, *integrative capability*, the ability to intentionally and strategically integrate, segregate and reintegrate AI within work practices to reconfigure how tasks, judgement, authority, learning, and accountability are distributed and create value under changing circumstances (Ahmad & Bilal, 2025; Krakowski, 2025). It ponders "How should human and AI contributions be configured for value creation in this situation?". Together, they offer a capability-based explanation of how relational expertise develops in human–AI embedded work.

Figure 1 introduces a model of relational expertise development for sustainable human–AI embedded work. The four capabilities are cumulative (with each capability building on the previous one across the five phases of agency loops), interdependent, and mutually reinforcing (the enactment of a capability further strengthens and extends previous capabilities).

Figure 1. Relational expertise development for sustainable human–AI embedded work.



Interpretive capability forms the foundation through delegation because it helps identify the need for human–AI collaboration, including related opportunities and limits, in a specific situation and when human judgment remains essential. This enables the development of interactional capability through engagement with AI in attribution phase, helping workers learn how to influence AI systems's behaviour while coordinating contributions. Based on this engagement, metacognitive capability develops as workers reflect on the consequences of AI-assisted, AI-augmented, and AI-driven actions, including what was achieved, how, and whether continued engagement with AI is appropriate, while further enhancing sensemaking of human–AI roles attribution and influential power. Finally, integrative capability emerges through reassertion and reconfiguration phases to determine when, where, and how AI embeddedness can strategically support work across tasks, workflows, contexts, and worker profiles and (re-)design human–AI work accordingly. I therefore propose that:

***RP1**: Relational expertise requires four capabilities that develop progressively, such that interpretive capability enables interactional capability, which in turn facilitates metacognitive capability, ultimately supporting integrative capability.*

***RP2**: The stronger the development of interpretive, interactional, and metacognitive capabilities, the stronger the development of integrative capability.*

Although all four capabilities can develop in each AI-embedded work context, the extent of their development varies depending on the circulation of human and AI agency across repeated workflows (see Figure 1) –a workflow refers to an agency loop in Leonardi's (2025) terms, unfolding through the delegation, attribution, contingency, reassertion and reconfiguration of a task or work process. As human and AI actions and interactions differ in how they are generated, enacted, and recognised across AI-assisted, AI-augmented, and AI-driven work contexts (see Table 1), each context creates distinct developmental opportunities for relational expertise.

Specifically, by engaging in a certain AI-embedded work context, individuals make sense of and adapt their ongoing interactions with AI systems, which then informs future engagement

(Brown et al., 2015). In AI-assisted work, where human agency prevails and AI supports, learning primarily entails understanding human–AI relations, evaluating and leveraging their respective contributions. In AI-augmented work, defined by shared and complimentary human–AI contributions, workers continuously negotiate, coordinate, integrate and segregate human and AI contributions. In AI-driven work, where AI agency prevails and humans supervise, workers focus more and more on monitoring, governing, and reconfiguring AI authority. As a result, relational expertise develops in AI-assisted, AI-augmented, and AI-driven work respectively, but it does so in different ways and to different degrees. Accordingly:

RP3*: The development of these four capabilities and, thus relational expertise, is influenced by individual sensemaking of how agency circulates between humans and AI systems within and across AI-embedded work contexts.*

Because AI-assisted, AI-augmented, and AI-driven work contexts are characterised by different patterns of agency circulation, engaging across multiple AI-embedded work contexts can help individuals to develop relational expertise in distinct yet complementary ways. In particular, engaging in all three contexts of AI-embedded work may further enhance relational expertise than engaging in only one or two contexts since it enables workers to make sense of human–AI embedded work from different angles, transition across work contexts, and build a broader set of capabilities for sustained professional work, adaptability, and professional growth (Ahmad & Bilal, 2025). I therefore suggest that:

RP4*: Engaging in both AI-assisted, AI-augmented, and AI-driven work contexts will likely result in the highest levels of relational expertise because workers learn to navigate and harness diverse patterns of agency circulation and collaboration between humans and AI systems.*

In short, relational expertise expands through four dynamic, interrelated, and situated capabilities (i.e., interpretive, interactional, metacognitive, and integrative capabilities) that enable workers to make sense of, learn about, and reconfigure human–AI relations over time.

4. Discussion

4.1. Research contributions

Whilst scholarly and societal debates often stress the risk that AI may erode expertise, skills, and autonomy (Huang et al., 2026; Ahmad & Bilal, 2025; Pakarinen & Huising, 2025), this study sheds light on an alternative possibility: AI agency can also create new opportunities for learning, ongoing adaptation, and long-term capability development. Specifically, I extend the concept of relational expertise (Pakarinen & Huising, 2025) by identifying four capabilities that enable workers to configure and reconfigure the circulation of agency between humans and AI systems in human–AI embedded work. In doing so, this study advances emerging research on the active role AI plays in morphing competency and organisational needs (Ahmad & Bilal, 2025; Krakowski, 2025) through a sociotechnical lens that explicitly accounts for the interplay of human and AI agency. Drawing on Leonardi's (2025) concept of agency loops (delegation, attribution, contingency, re-assertion, and reconfiguration) and AI literacy scholarship, I argue that workers develop relational expertise by developing four capabilities (i.e., interpretive, interactional, metacognitive, and integrative capabilities) through sensemaking of repeated engagement with different and changing patterns of agency attribution and circulation. I propose that these four interrelated, cumulative, and mutually-reinforcing capabilities grow within and across three forms of AI-embedded work (namely AI-assisted, AI-augmented, and AI-driven work), each characterised by specific patterns of agency circulation between humans and AI systems. Through repeated participation in agency loops, workers learn to negotiate authority, decision-making, task execution, and responsibility, gradually enhancing their ability to collaborate effectively with AI systems and orchestrate their strategic deployment in organisational work processes. I propose that diverging patterns of agency attribution and

circulation foster different capability development across work contexts. Hence, I argue that engagement in all three AI-embedded work contexts offers the strongest foundation for developing and sustaining relational expertise over time, as workers experience a broader range of agency configurations and learning opportunities. This crystallises in a model of relational expertise development for sustainable human–AI embedded work and four research propositions that underpin future theoretical and empirical investigation.

4.2. Limitations and future research

This paper presents several limitations that pave the way for future enquiry. First, despite the use of vignettes to exemplify how agency circulates across AI-assisted, AI-augmented, and AI-driven work, the proposed model remains conceptual and essentially grounded in knowledge work contexts. Although it proposes expected patterns of capability development based on the characteristic circulation of agency within each form of AI-embedded work, actual human–AI interactions may vary considerably across industries, organisations, technologies, occupations, and individuals. Different engagement patterns may thus arise and lead to more or less developed interpretive, interactional, metacognitive, and integrative capabilities. I invite future research on these nuances through comparative studies, systematic reviews, and empirical testing and refinement of the model and its associated research propositions. Especially looking into longitudinal patterns emerging within and across AI-assisted, AI-augmented, and AI-driven work would provide a more dynamic and accurate view of relational expertise development in human–AI embedded work. Second, consistent with its goal of advancing research on capability development and expertise, this study purposefully focused on human–AI collaboration and the opportunities it creates for capability development. Hence, it does not extensively engage with perspectives on AI takeover, worker displacement, or full automation. Future research could investigate how integrative capability allows the reallocation of human capabilities under conditions where AI replaces human work, examining implications for diverse stakeholders. Third, this work focused on capability development between individuals and AI systems, which calls for social contextualisation of individual capability development, the interplay among varying levels of individual relational expertise, and its further influence on collective capability development within organising processes. In sum, this study advances relational expertise as a promising cornerstone for future enquiry into human–AI embedded work environments, moving attention away from more pessimistic accounts of the effects of AI (agency) on workers, their work and professional opportunities. By investigating the development of the capabilities needed to configure and reconfigure individual and organisational work processes between humans and AI systems, scholars can advance a more optimistic and nuanced understanding of human–AI embedded work.

5. Conclusion

Ultimately, this study offers a sociotechnical perspective on human capability development and underscores the importance of creating work and professional development opportunities that allow workers to experience different forms of human–AI collaboration. These experiences foster the development of relational expertise, enabling workers to effectively navigate and reconfigure evolving human–AI embedded work processes while supporting long-term adaptability, employability, and professional growth.

6. References

Ahmad, B., & Bilal, S. (2025). Knowledge of AI as a Future Work Skill for Career Sustainability. Journal of Career Development, Vol. 52, No. 2, 134-152.

Brown, A. D., Colville, I., & Pye, A. (2015). Making Sense of Sensemaking in Organization Studies. Organization Studies, Vol. 36, No. 2, 265-277.

Donald, W. E., Van der Heijden, B. I. J. M., & Baruch, Y. (2024). Introducing a sustainable career ecosystem: Theoretical perspectives, conceptualization, and future research agenda. Journal of Vocational Behavior, Vol. 151, 1-14.

Einola, K., & Khoreva, V. (2023). Best friend or broken tool? Exploring the co-existence of humans and artificial intelligence in the workplace ecosystem. Human Resource Management, Vol. 62, No.1, 117-135.

Henkenjohann, R., & Trenz, M. (2026). The responsible worker in the age of generative AI: How human-AI co-creation shapes experienced responsibility for work outcomes. Decision Support Systems, Vol. 207, No. 114699, 1-12.

Huang, B. E., Kang, H., Zur, E., & Liang, L. H. (2026). Humanistic human resource management for digital transformation: Addressing the challenges of AI advancement in the workplace. Human Resource Management Review, Vol. 36, No. 3,1-16.

Köchling, A., Wehner, M. C., & Ruhle, S. A. (2025). This (AI)n't fair? Employee reactions to artificial intelligence (AI) in career development systems. Review of Managerial Science, Vol. 19, No. 4, 1195-1228.

Krakowski, S. (2025). Human-AI agency in the age of generative AI. Information and Organization, Vol. 35, No. 1, 1-25.

Leonardi, P. M. (2025). *Homo agenticus* in the age of agentic AI: Agency loops, power displacement, and the circulation of responsibility. Information and Organization, Vol. 35, No. 3, 1-18.

Nyberg, A. J., Schleicher, D. J., Bell, B. S., Boon, C., Cappelli, P., Collings, D. G., Dalle Molle, J. E., Feuerriegel, S., Gerhart, B., Jeong, Y., Korsgaard, M. A., Minbaeva, D., Ployhart, R. E., Tambe, P., Weller, I., Wright, P. M., & Yakubovich, V. (2025). A Brave New World of Human Resources Research: Navigating Perils and Identifying Grand Challenges of the GenAI Revolution. Journal of Management, Vol.51, No. 6, 2677-2718.

Pakarinen, P., & Huising, R. (2025). Relational Expertise: What Machines Can't Know. Journal of Management Studies (John Wiley & Sons, Inc.), Vol.*62*, No. 5, 2053-2082.

Pan, S.-L., Nishant, R., Tuunanen, T., & Choudrie, J. (2025). AI Agents: Potential implications for IS Research? The Journal of Strategic Information Systems, Vol. 34, No. 2, 1-5.

Rapanta, C., Bhatt, I., Bozkurt, A., Chubb, L. A., Erb, C., Forsler, I., Gravett, K., Koole, M., Lintner, T., Örtegren, A., Petricini, T., Rodgers, B., Webster, J., Xu, X., Christensen, I.-M. F., Dohn, N. B., Christensen, L. L. W., Zeivots, S., & Jandrić, P. (2025). Critical GenAI Literacy: Postdigital Configurations. Postdigital Science and Education, Vol. *7*, No. 4, 1296-1333.

Riemer, K., & Peter, S. (2024). Conceptualizing generative AI as style engines: Application archetypes and implications. International Journal of Information Management, Vol. 79, 1-15.

Roy, S. K., Dey, B. L., Brown, D. M., Abid, A., Apostolidis, C., Christofi, M., & Tarba, S. (2025). Business Model Innovation through AI Adaptation: The Role of Strategic Human Resources Management. British Journal of Management, Vol. 36, No. 2, 546-559.

When IT Holds Decision Rights: A Configurational Theory of Governance Fit for AI Agents

Navya[1,2,3], Saeed Akhlaghpour[2], Sanjay Dhir[3]

1. University of Queensland – Indian Institute of Technology Delhi Academy of Research
2. University of Queensland, Brisbane, Australia
3. Indian Institute of Technology, Delhi, India

Abstract

Organizations increasingly allocate decision rights to AI agents, yet IT governance research assumes these rights are held by humans who can be held accountable for their use. We develop a typology of five allocations of decision rights to AI agents and propose that each weakens a different condition on which decision control depends. As control mechanisms lose fit, accountability shifts away from where decisions are effectively made, leaving human role holders answerable for decisions they cannot properly monitor or justify. We show that accountability depends not only on who holds the decision right but also on whether control still works, and that more oversight alone cannot restore it.

1. Introduction

Information technology (IT) governance refers to the decision rights and accountability mechanisms through which organizations direct behavior in the use of IT. (Weill & Ross, 2004). Its scope has expanded to include contingency work, IT functions, digital assets, platforms, and business-led technology use. (Olson & Chervany, 1980;Brown, 1999; Sambamurthy & Zmud, 1999; Gregory et al., 2018). Throughout this scholarly conversation, the underlying assumption has remained consistent. Structures with decision rights and accountability are allocated and designed for humans, and technology is viewed as a tool, infrastructure, or ensemble through which those decisions are implemented. (Orlikowski & Iacono, 2001). Even as algorithms increasingly direct and evaluate work, decision-making authority is generally assumed to remain with human managers (Kellogg et al., 2020) with a deterministic view towards technology. However, Artificial intelligence (AI) challenges this passive view of technology. (Baird & Maruping, 2021; Berente et al., 2021; Kasirzadeh & Gabriel, 2026).

We focus on the increasing agency behavior of AI technologies, referring to them as AI agents throughout the paper. These agents can plan, reason, execute, and perceive their environment, reshaping decisions and workflows to achieve high-level goals. (Kasirzadeh & Gabriel, 2026). As adoption of AI agents has increased (Papagiannidis et al., 2023). It is significant to understand how organizations deploy these agents in their decision-making processes, which may impact outcomes and may exacerbate existing risks and introduce new ones (Chan et al., 2024). These decisions not only shape what an agent can do, but also how it will be reviewed or controlled by the organization.

To understand it better, consider the difference between an AI agent in healthcare that flags suspected diabetic retinopathy for an ophthalmologist to confirm, and IDx-DR, the first autonomous diagnostic system authorized by the US Food and Drug Administration (Ho, 2023), which issues a diagnosis in primary care without any clinician interpreting the image.

In the first case, a clinician retains both the right to decide and control over whether the decision takes effect. In the second, the agent exercises that right directly; control can no longer be exercised by approving each diagnosis, but only by the boundaries set before deployment, such as the validated patient population, the approved camera, and the rule that poor-quality images are not diagnosed. Even after all this, if it misses a sight-threatening condition, it becomes unclear whether responsibility lies with the clinic that deployed the agent, the vendor that designed it, or the clinician who was present but never saw how the decision was made.

The prevailing response to this concern is to increase human oversight over AI decisions, typically in the form of "Human in the Loop"(Perboli et al., 2025) so humans could be responsible. Research on automation, however, has long questioned whether such oversight works as intended. Humans lose the capacity to intervene in systems they no longer run. (Strauch, 2018). Reviewers accept automated recommendations without adequate scrutiny, (Parasuraman & Manzey, 2010) formal oversight arrangements may not ensure effective review and put accountability with little control (Enqvist, 2023; Kasirzadeh & Gabriel, 2026)We argue that this misdiagnoses the problem. Rather than focusing only on increasing oversight, organizations can shift their focus to understand how different governance concerns arise when an agent is permitted to determine how an assigned objective should be achieved. (Kasirzadeh & Gabriel, 2026; Papagiannidis et al., 2023).

Prior IS research has theorized how agency is delegated to and shared with AI artifacts (Baird & Maruping, 2021; Murray et al., 2021; Berente et al., 2021) and IT governance (Gregory et al., 2018; Tiwana et al., 2010, 2013) has explained how decision rights, controls, and accountability are distributed. However, IT governance research assumes that humans hold decision rights, and research on agentic artifacts has examined delegation without focusing on the governance it requires. Therefore, this paper explains how rights allocated to AI agents shape governance, and what happened to established mechanisms. We therefore ask: How does the allocation of decision rights to AI agents affect the fit between existing governance control mechanisms and accountability arrangements?

To address this question, we develop a typological theory (Doty & Glick, 1994; Furnari et al., 2021). We examine how the allocation of decision rights in combination changes the forms of control available to the organization. We theorize governance fit as the match between a configuration of allocated rights and the form of control the organization relies on, and explain how allocated rights and accountability align. The analysis yields five types (advisory, executory, self-directing, self-adapting, and self-evolving), each associated with a distinct governance concern.

2. Theoretical Background

2.1 Existing IT Governance Literature

IT governance emerged as a subset of corporate governance concerned with specifying the IT decision rights and accountability (Weill & Ross, 2004). A central motivation of this stream is to study how well IT can be aligned to business objectives and associated with superior returns (Weill & Ross, 2004). As organizational boundaries became more permeable, these governance questions extended beyond the firm. When IT services were sourced from vendors and partners, decision rights could no longer be allocated solely through hierarchical

structures. Instead, Contractual and service-level agreements were codified with relational norms (Goo et al., 2009). Digital technology has furthered this with the evolution of businesses within platform ecosystems: a platform owner shares decision rights with many independent complementors (Tiwana et al., 2010), and decision rights are partitioned between owners and complementors through control mechanisms that balance platform openness (Ghazawneh & Henfridsson, 2013).

Based on the organizing logic of the IT governance cube (Tiwana et al., 2013), the literature can be compared by what is governed, who holds decision rights, and how governance is exercised. Table 1 summarizes this comparison. Across the streams, the object of governance has widened considerably, from IT decisions within the firm to services and relationships across firm boundaries, to platforms and their complements, to work directed through algorithms, to AI systems themselves. (Kellogg et al., 2020; Möhlmann et al., 2021; Papagiannidis et al., 2023). The Scope and patterns have diversified from hierarchical allocation to contracts, gatekeeping, architecture, and algorithmic control. One feature, however, has remained constant. In every stream, decision rights originate with humans or human-led parties; in the agentic AI stream, agents exercise them on delegated authority.

Table 1. Stream of existing IT Governance literature

Literature stream	Dimensions of Governance			Representative work
	Focus (What to Govern)	**Scope (Who holds decision rights)**	**Patterns (How to Govern)**	
IT Governance	IT decisions: principles, architecture, infrastructure, applications, investment	Business and IT executives within the firm	Structural, process, and relational mechanisms; centralized, decentralized, or federal arrangements	(De Haes & Van Grembergen, 2009; Sambamurthy & Zmud, 1999; Weill & Ross, 2004)
Interorganizational IT Governance	IT services and relationships across firm boundaries	Client firms, vendors, partners	Contracts and SLAs, complemented by relational governance	(Goo et al., 2009)
Platform Governance	The platform and its complements	Platform owner and complementors	Gatekeeping process, output control and clan control	(Ghazawneh & Henfridsson, 2013; Tiwana et al., 2010)
Algorithmic Governance	Work and Workers	Workers managed through algorithms	Algorithmic direction, evaluation, and discipline	(Kellogg et al., 2020; Möhlmann et al., 2021)
AI Governance	AI across the life cycle	AI system with Human oversight	Principles, oversight, risk and compliance processes	(Berente et al., 2021; Papagiannidis et al., 2023)
Agentic AI Governance	Agentic Profiles	Human principal; delegated authority to agents	Continuing interpretability mechanisms, Real-time monitoring	(Baird & Maruping, 2021; Kasirzadeh & Gabriel, 2026)

2.2 The emergence of agentic behavior in AI

IS research increasingly recognizes that AI can participate in organizational action rather than function solely as instruments directed by human actors. One view conceptualized agentic IS artifacts where authority is delegated bidirectionally between human and AI (Baird & Maruping, 2021). A related view theorizes forms of conjoined agency through which human and AI jointly shape organizational routines (Murray et al., 2021). AI is also distinguished from earlier

technology as it exhibits autonomy, learning, and inscrutability (Berente et al., 2021). These conversations become our point of inquiry to understand agents with agentic behavior closely.

The literature on intelligent agents has long examined goal-directed behavior, environmental perception, reasoning, and action as properties through which agents interact with their environments (Franklin & Graesser, 1997; Wooldridge & Jennings, 1995). More recent research on large language model (LLM)-based agents extends this discussion by examining architectures that integrate perception, planning, memory, and action, as well as mechanisms for reflection and behavioral improvement (Wang et al., 2024). Rather than adopting or defining a taxonomy, we derive analytical attributes relevant to the organizational decision-making process. Our selection is based on two perspectives that explain why allocation matters (Fama & Jensen, 1983), and how decisions can be handed over to technology (Parasuraman et al., 2000).

Building on these lenses, we identify several AI agent attributes mapped to the decision process mentioned in Table 2 (Fama & Jensen, 1983; Parasuraman et al., 2000). Goal orientation captures the pursuit of objectives and the initiation of activities towards them. (Bandi et al., 2025; Franklin & Graesser, 1997; Gadekallu et al., 2026; Haidemariam, 2026; Kamalov & Sulieman, 2026; Kasirzadeh & Gabriel, 2026); sensing concerns the acquisition of information from the environment (Gadekallu et al., 2026; Kasirzadeh & Gabriel, 2026); reasoning concerns the evaluation of information and alternatives (Gadekallu et al., 2026); and execution concerns the translation of selected actions into effects in the environment (Wooldridge & Jennings, 1995; Wang et al., 2024; Kasirzadeh & Gabriel, 2026). Memory and learning extend this account beyond the activities (Berente et al., 2021; Wang et al., 2024). An agent's future decisions may depend not only on the information currently available but also on what it retains and how its behavior changes over time.

Table 2: Literature of Agentic AI attributes mapped with the decision process.

Agentic AI attribute	**Supporting literature**	**Mapped with decision process** (Fama & Jensen, 1983; Parasuraman et al., 2000)
Goal orientation	Bandi et al., 2025; Franklin & Graesser, 1997; Gadekallu et al., 2026; Haidemariam, 2026; Kamalov & Sulieman, 2026; Kasirzadeh & Gabriel, 2026	Decision initiation; indirectly, decision selection
Sensing	Wooldridge & Jennings (1995). Wang et al. (2024); Gadekallu et al. (2026); Kasirzadeh & Gabriel (2026)	Information acquisition
Reasoning	Wooldridge & Jennings (1995). Gadekallu et al., 2026; Kasirzadeh & Gabriel, 2026	Information analysis; decision selection
Execution	Wooldridge & Jennings (1995); Wang et al. (2024); Gadekallu et al. (2026); Kasirzadeh & Gabriel (2026)	Decision implementation; action implementation
Memory	Wang et al. (2024).	Across Decisions (adaptive)
Learning	Berente et al. (2021); Wang et al. (2024)	

3. A Configurational Approach to Theorizing

We develop a typology of AI agents based on how organizations allocate decision rights between humans and AI agents. The typology distinguishes configurations based on which decision rights humans retain and which they allocate to AI agents (Doty & Glick, 1994). This approach aligns with our understanding that decision rights allocated to AI agents matter in

combination, not in isolation. A system that learns but cannot execute actions poses a different concern from one that both learns and executes. To develop the typology, we follow the configurational theorizing process (Fiss, 2007; Furnari et al., 2021), which comprises three iterative stages. Scoping identifies the outcome to be explained and the attributes relevant to it. Linking theorizes how those attributes combine to produce the outcome - naming labels for each configuration so that the label captures the theme that holds it together. We describe how we carried out each stage in the subsequent section.

3.1 Scoping

We scope the typology around the alignment of decision rights and accountability as core dimensions of governance (Beck et al., 2018; Weill & Ross, 2004). Accountability is aligned when the person held answerable for a decision can explain and justify it; misalignment arises when decision-making is delegated to an AI agent, but the accountable person cannot adequately explain or justify the decision (Bovens, 2007; Novelli et al., 2024; Wieringa, 2020) which is the outcome explained (Doty & Glick, 1994;Venkatraman, 1989). To construct the typology, we cluster attributes of AI agents identified in the literature review section into allocation of rights: directive rights (Goal orientation), deliberative rights (Sensing and Reasoning), enactive rights (Execute), and adaptive rights (Learning and Memory). Each cluster of rights can be in one of two states: the organization either retains the right or fully allocates it.

3.2 Linking & Naming

Following Furnari et al. (2021), we theorize how combinations of decision rights produce different concerns. Our linking logic rests on the forms of human decision control that each allocation preserves or constrains because decision control remains an organizational responsibility; allocating rights to an AI agent changes how that control can be exercised. Enactive rights remove the opportunity to approve each decision before it takes effect; directive rights limit the extent to which the agent's conduct can be specified in advance; and adaptive rights make previously established boundaries less reliable as the agent's behavior changes through experience. Each changes a different basis on which humans can direct, constrain, or review the agent, and their combination determines which forms of control remain available to the role held accountable. We use governance fit, the match between a configuration and the form of control the organization relies on (Fama & Jensen, 1983; Venkatraman, 1989) to explain misalignment.

We distinguish the allocation of a right from the conditions under which it is exercised. Combining two levels across four clusters yields a property space of 16 (2^4) possible configurations. Enumerating all of them would produce a classification but remain theoretically uninformative, since many configurations are either incoherent or indistinguishable in their governance concerns. We therefore reduce this space following the property-space tradition (Bailey, 1994). We exclude eight configurations in which deliberative rights are retained, as they are outside the scope of the paper. Then we merge the four configurations in which enactive rights are retained into a single advisory type, because humans retain authority over executing decisions, even when AI performs earlier decision-making activities. The remaining four configurations, in which deliberative and enactive rights are allocated, vary only by directive and adaptive rights and constitute the four agentic types, yielding five types in total.

Table 3: Configuration of allocated decision rights

	Adaptive rights retained	**Adaptive rights allocated**
Directive rights retained	**Type II: Executory** Pursues human-set goals and subgoals; its behavior remains as designed.	**Type IV: Self-adapting** Pursues human-set goals and subgoals; revises its own behavior over time.
Directive rights allocated	**Type III: Self-directing** Determines its own subgoals, then forms and enacts decisions; its behavior remains as designed.	**Type V: Self-evolving** Determines its own subgoals and revises its own behavior over time.

Note: Deliberative and enactive are allocated in all four types shown. Type I (Advisory) comprises every configuration in which enactive rights are retained; it is therefore not shown in the matrix.

4. Governance across types of allocated decision rights

Organizations govern decisions by separating decision management from decision control as per residual risk, as agents can not bear the risk of their decisions (Fama & Jensen, 1983). These decision controls can take three forms: Ratification in advance, i.e., through pre-approved boundaries; ratification of each decision, i.e., approval for each specific decision; and Monitoring of decisions afterward. Accountability rests on these forms of control, since the role that holds a decision right must answer for its use and justify it (Wieringa, 2020). We use diagnostic AI examples in healthcare to illustrate the proposed theoretical mechanisms, not as empirical evidence.

Type I Advisory

In this type, AI diagnostic agents analyze patient data, such as imaging and laboratory results, and recommend a diagnosis, while the clinician retains the final decision-making authority. Control takes the form of ratification of each decision, and both decision-making rights and accountability formally remain with the clinician. This control form, however, functions only if the clinician can evaluate each recommendation independently. As reliance on the agent grows, automation bias may weaken this capacity (Parasuraman & Manzey, 2010). Moreover, approval may shift from evaluation to endorsement. The clinician remains accountable for the diagnosis, while the agent effectively determines its content.

Proposition1: For organizations using AI agents in an advisory capacity, accountability misalignment will be greater when reliance on the agent is high, and the clinician's ability to evaluate those recommendations independently is low.

Type II Executory

The AI agents pursue goals and subgoals specified by humans and cannot adapt them. For example, they apply a clinician-approved protocol, flagging abnormal results and ordering the follow-up tests it prescribes. Control takes the form of advance ratification: the clinician approves the protocol rather than each action. This control form functions only if the protocol fully specifies what the agent will do. Clinical protocols, however, cannot anticipate every patient, such as those with multiple conditions or atypical presentations. In such cases, the

agent applies the protocol to patients for whom no one intended it, and the clinician remains accountable for actions it never specifically approved.

Proposition 2: Among organizations that authorize AI agents to execute decisions within predefined clinical protocols, accountability misalignment will be greater when consequential actions fall outside the situations anticipated by the approved protocol, and exceptions cannot be escalated for clinical authorization.

Type III Self-directing

In this type, the clinician sets the goal, such as reaching a diagnosis for a patient, while the agent determines its own subgoals and the sequence of steps to achieve it, for instance, which records to retrieve, which tests to order, and in what order. Because these steps are not approved in advance, control shifts to monitoring afterward. This control form functions only if the agent's path can be traced and reviewed. As the number of intermediate steps grows and the reasoning behind them is not transparent, review becomes limited to the final output. The clinician remains accountable for a diagnostic process whose steps were chosen by the agent and cannot be fully reconstructed.

Proposition 3: Among organizations that authorize AI agents to select their own intermediate goals or courses of action, accountability misalignment will be greater when decision records do not allow the accountable clinician to reconstruct and justify the agent's decision process.

Type IV: Self-adapting

The agent modifies its own decision rules based on feedback, for example, by adjusting the threshold at which it flags a scan as abnormal, given confirmed diagnoses. Control again relies on post-monitoring, often supported by the agent's initial validation. This control form functions only if the monitored logic is approved. As the agent adapts, its decision rules diverge from those that were validated, without any new act of approval. The clinician remains accountable for decisions produced by logic it never ratified.

Proposition 4: Among organizations that authorize AI agents to modify their decision rules, accountability misalignment will be greater when consequential changes to those rules are not subject to review and reauthorization before being used in clinical decisions.

Type V: Self-evolving

The agent can revise its own goals or extend its scope, for example, by shifting from maximizing diagnostic accuracy to reducing time to diagnosis, or by applying itself to conditions it was not originally designed for. None of the three control forms fits fully here: ratification in advance presumes fixed goals, and monitoring afterward presumes a stable standard against which decisions can be judged. When the goal itself changes, the standard of evaluation disappears. The decision-making authority effectively shifts to the agent, while accountability remains with the clinician and the organization.

Proposition 5: Among organizations authorizing AI agents to revise their subgoals or operational scope, accountability misalignment will be greater when such changes can materially alter clinical priorities or the conditions under which the agent operates without renewed organizational authorization.

5. Discussion

In this paper, we argue that IT governance research has assumed that humans hold decision rights and that control mechanisms make the role holder accountable. We argue that AI agents, with increasingly sophisticated capabilities in sensing, reasoning, memory, learning, and execution, challenge both assumptions. Across five types of allocated decision rights, from advisory to self-evolving, we propose that each allocation challenges a different condition on which control depends: the controller's capacity to evaluate decisions, the completeness of pre-approved boundaries, the traceability of decision steps, the stability of approved decision logic, and the stability of goals. As these conditions erode, the fit of control mechanisms declines, and accountability misalignment emerges between the party that effectively exercises a decision right and the party held accountable for it.

This contributes to IT governance research by showing that when organizations are allocating decisions to human and AI agents, accountability does not automatically follow. (Weill & Ross, 2004). It may depend on the fit of the control mechanism underneath it. It also extends Fama and Jensen's (1983) view, explaining why a control mechanism can remain formally in place while losing its capacity to constrain decisions. Misalignment, in this view, arises not from allocating decision rights to AI agents as such, but from the loss of fit that such allocation produces.

We explained this using real-world case such as the Epic Sepsis Model illustrates potential governance concerns. An external validation found that the model performed substantially worse than previously reported (Wong et al., 2021). Although the study did not examine organisational governance or clinical accountability, its findings highlight the importance of assessing whether existing oversight arrangements allow organisations and clinicians to evaluate AI-generated recommendations. Whether such limitations lead to accountability misalignment remains an empirical question.

When AI agents operate in chains, where one agent's output becomes another's input, decision rights are distributed across agents, and the locus of the effective decision becomes difficult to identify. Our propositions therefore apply to an agent allocation, and extending them to multi-agent chains is an important direction for future work. Future research can test our propositions by measuring the allocation of decision rights, the fit of control mechanisms, and accountability misalignment from separate sources, such as governance processes and decision records. Longitudinal and cross-sector studies could examine how fit erodes over time.

6. References

Bailey, K. (1994). *Typologies and Taxonomies*. SAGE Publications, Inc. https://doi.org/10.4135/9781412986397

Baird, A., & Maruping, L. M. (2021). The Next Generation of Research on IS Use: A Theoretical Framework of Delegation to and from Agentic IS Artifacts. *MIS Quarterly*, *45*(1), 315–341. https://doi.org/10.25300/MISQ/2021/15882

Bandi, A., Kongari, B., Naguru, R., Pasnoor, S., & Vilipala, S. V. (2025). The Rise of Agentic AI: A Review of Definitions, Frameworks, Architectures, Applications, Evaluation Metrics, and Challenges. *Future Internet*, *17*(9), 404. https://doi.org/10.3390/fi17090404

Beck, R., Müller-Bloch, C., & King, J. L. (2018). Governance in the blockchain economy: A framework and research agenda. Journal of the Association for Information Systems, 19(10), 1020–1034. https://doi.org/10.17705/1jais.00518

Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Managing Artificial Intelligence. *MIS Quarterly*, *45*(3), 1433–1450. https://doi.org/10.25300/MISQ/2021/16274

Bovens, M. (2007). Analysing and Assessing Accountability: A Conceptual Framework. *European Law Journal*, *13*(4), 447–468. https://doi.org/10.1111/j.1468-0386.2007.00378.x

Brown, C. V. (1999). Horizontal Mechanisms under Differing IS Organization Contexts. *MIS Quarterly*, *23*(3), 421–454. https://doi.org/10.2307/249470

Chan, A., Ezell, C., Kaufmann, M., Wei, K., Hammond, L., Bradley, H., Bluemke, E., Rajkumar, N., Krueger, D., Kolt, N., Heim, L., & Anderljung, M. (2024). *Visibility into AI Agents* (arXiv:2401.13138). arXiv. https://doi.org/10.48550/arXiv.2401.13138

De Haes, S., & Van Grembergen, W. (2009). An Exploratory Study into IT Governance Implementations and its Impact on Business/IT Alignment. *Information Systems Management*, *26*(2), 123–137. https://doi.org/10.1080/10580530902794786

Doty, D. H., & Glick, W. H. (1994). Typologies as a unique form of theory building: Toward improved understanding and modeling. Academy of Management Review, 19(2), 230–251. https://doi.org/10.5465/amr.1994.9410210748

Enqvist, L. (2023). 'Human oversight' in the EU artificial intelligence act: What, when and by whom? *Law, Innovation and Technology*, *15*(2), 508–535. https://doi.org/10.1080/17579961.2023.2245683

Fama, E. F., & Jensen, M. C. (1983). Separation of Ownership and Control. *Journal of Law and Economics*, *26*(2), 301–325.

Fiss, P. C. (2007). A set-theoretic approach to organizational configurations. *Academy of Management Review*, *32*(4), 1180–1198. https://doi.org/10.5465/amr.2007.26586092

Franklin, S., & Graesser, A. (1997). Is It an agent, or just a program?: A taxonomy for autonomous agents. In J. P. Müller, M. J. Wooldridge, & N. R. Jennings (Eds.), *Intelligent Agents III Agent Theories, Architectures, and Languages* (Vol. 1193, pp. 21–35). Springer Berlin Heidelberg. https://doi.org/10.1007/BFb0013570

Furnari, S., Crilly, D., Misangyi, V. F., Greckhamer, T., Fiss, P. C., & Aguilera, R. V. (2021). Capturing Causal Complexity: Heuristics for Configurational Theorizing. *Academy of Management Review*, *46*(4), 778–799. https://doi.org/10.5465/amr.2019.0298

Gadekallu, T. R., Zhao, Y., Wen, Z., Bhattacharya, P., Xia, Y., Cai, J., Zeydan, E., Fang, K., & Feng, H. (2026). Artificial Intelligence of Things as a Foundation for Agentic AI Systems: Architectures, Applications, and Challenges. *IEEE Internet of Things Journal*, *13*(14), 30287–30302. https://doi.org/10.1109/JIOT.2026.3703862

Ghazawneh, A., & Henfridsson, O. (2013). Balancing platform control and external contribution in third-party development: The boundary resources model. *Information Systems Journal*, *23*(2), 173–192. https://doi.org/10.1111/j.1365-2575.2012.00406.x

Goo, J., Kishore, R., Rao, H. R., & Nam, K. (2009). The Role of Service Level Agreements in Relational Management of Information Technology Outsourcing: An Empirical Study. *MIS Quarterly*, *33*(1), 119–145. https://doi.org/10.2307/20650281

Gregory, R. W., Kaganer, E., Henfridsson, O., & Ruch, T. J. (2018). IT Consumerization and the Transformation of IT Governance. *MIS Quarterly*, *42*(4), 1225–1253. https://doi.org/10.25300/MISQ/2018/13703

Haidemariam, T. (2026). From the logic of coordination to goal-directed reasoning: The agentic turn in artificial intelligence. *Frontiers in Artificial Intelligence*, *8*, 1728738. https://doi.org/10.3389/frai.2025.1728738

Ho, C. W. L. (2023). Implementing the human right to science in the regulatory governance of artificial intelligence in healthcare. *Journal of Law and the Biosciences*, *10*(2), lsad026. https://doi.org/10.1093/jlb/lsad026

Kamalov, F., & Sulieman, H. (2026). From Rule-Based to Goal-Oriented: A Systematic Review of Agentic AI in Education. *IEEE Access*, *14*, 95314–95327. https://doi.org/10.1109/ACCESS.2026.3704764

Kasirzadeh, A., & Gabriel, I. (2026). Agentic profiles for effective AI governance. *Nature*, *656*(8127), 320–328. https://doi.org/10.1038/s41586-026-10805-z

Kellogg, K. C., Valentine, M. A., & Christin, A. (2020). Algorithms at Work: The New Contested Terrain of Control. *Academy of Management Annals*, *14*(1), 366–410. https://doi.org/10.5465/annals.2018.0174

Möhlmann, M., Zalmanson, L., Henfridsson, O., & Gregory, R. W. (2021). Algorithmic Management of Work on Online Labor Platforms: When Matching Meets Control. *MIS Quarterly*, *45*(4), 1999–2022. https://doi.org/10.25300/MISQ/2021/15333

Murray, A., Rhymer, J., & Sirmon, D. G. (2021). Humans and Technology: Forms of Conjoined Agency in Organizations. *Academy of Management Review*, *46*(3), 552–571. https://doi.org/10.5465/amr.2019.0186

Novelli, C., Taddeo, M., & Floridi, L. (2024). Accountability in artificial intelligence: What it is and how it works. *AI & SOCIETY*, *39*(4), 1871–1882. https://doi.org/10.1007/s00146-023-01635-y

Olson, M. H., & Chervany, N. L. (1980). The Relationship between Organizational Characteristics and the Structure of the Information Services Function. *MIS Quarterly*, *4*(2), 57–68. https://doi.org/10.2307/249337

Orlikowski, W. J., & Iacono, C. S. (2001). Research Commentary: Desperately Seeking the "IT" in IT Research—A Call to Theorizing the IT Artifact. *Information Systems Research*, *12*(2), 121–134. https://doi.org/10.1287/isre.12.2.121.9700

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. https://doi.org/10.1007/s10796-022-10251-y

Parasuraman, R., & Manzey, D. H. (2010). Complacency and Bias in Human Use of Automation: An Attentional Integration. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *52*(3), 381–410. https://doi.org/10.1177/0018720810376055

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, *30*(3), 286–297. https://doi.org/10.1109/3468.844354

Perboli, G., Simionato, N., & Pratali, S. (2025). Navigating the AI regulatory landscape: Balancing innovation, ethics, and global governance. *Economic and Political Studies*, *13*(4), 367–397. https://doi.org/10.1080/20954816.2025.2569584

Sambamurthy, V., & Zmud, R. W. (1999). Arrangements for Information Technology Governance: A Theory of Multiple Contingencies. *MIS Quarterly*, *23*(2), 261–289. https://doi.org/10.2307/249754

Strauch, B. (2018). Ironies of Automation: Still Unresolved After All These Years. *IEEE Transactions on Human-Machine Systems*, *48*(5), 419–433. https://doi.org/10.1109/THMS.2017.2732506

Tiwana, A., Konsynski, B., & Bush, A. A. (2010). Research Commentary —Platform Evolution: Coevolution of Platform Architecture, Governance, and Environmental Dynamics. *Information Systems Research*, *21*(4), 675–687. https://doi.org/10.1287/isre.1100.0323

Tiwana, A., Konsynski, B., & Venkatraman, N. (2013). Special Issue: Information Technology and Organizational Governance: The IT Governance Cube. *Journal of Management Information Systems*, *30*(3), 7–12. https://doi.org/10.2753/MIS0742-1222300301

Venkatraman, N. (1989). The concept of fit in strategy research: Toward verbal and statistical correspondence. Academy of Management Review, 14(3), 423–444. https://doi.org/10.5465/amr.1989.4279078

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W. X., Wei, Z., & Wen, J. (2024). A survey on large language model based autonomous agents. *Frontiers of Computer Science*, *18*(6), 186345. https://doi.org/10.1007/s11704-024-40231-1

Weill, P., & Ross, J. W. (2004). *IT governance: How top performers manage IT decision rights for superior results*. Harvard Business School Press.

Wieringa, M. (2020). What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 1–18. https://doi.org/10.1145/3351095.3372833

Wong, A., Otles, E., Donnelly, J. P., Krumm, A., McCullough, J., DeTroyer-Cooley, O., Pestrue, J., Phillips, M., Konye, J., Penoza, C., Ghous, M., & Singh, K. (2021). External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients. *JAMA Internal Medicine*, *181*(8), 1065-1070. https://doi.org/10.1001/jamainternmed.2021.2626

Wooldridge, M., & Jennings, N. R. (1995). Intelligent agents: Theory and practice. *The Knowledge Engineering Review*, *10*(2), 115–152. https://doi.org/10.1017/S0269888900008122

A Lifecycle Framework of AI Fairness and Research Agenda in Information Systems

Manju Ahuja
UNSW Sydney

Jiajun Wang
UNSW Sydney

Carmen Leong
UNSW Sydney

Kevin Kuan
UNSW Sydney

Abstract

Artificial Intelligence (AI) systems are increasingly driving high-stakes decisions, yet bias can accumulate undetected across every stage of development. Organisations lack a systematic framework for identifying how biases emerge, interact, and accumulate across the AI production pipeline. This study addresses this gap by developing an integrated AI fairness lifecycle framework. First, the paper synthesises prior lifecycle models to develop a five-stage AI lifecycle integrating business, technical, and deployment perspectives. Second, it introduces a structured typology of bias mitigation strategies comprising data, process, algorithmic, and human-centred design strategies. Third, it maps bias sources and mitigation strategies across lifecycle stages. Finally, a research agenda is proposed to guide future information systems (IS) research on AI fairness, offering both theoretical and practical guidance for researchers, practitioners and policymakers seeking equitable AI development.

1. Introduction

Artificial intelligence (AI) is increasingly embedded in organisational decision-making, shaping outcomes in domains such as hiring, credit allocation, healthcare, and performance management. While these systems promise efficiency and scalability, growing evidence shows that algorithmic decision-making can reproduce or amplify social inequalities. This is well illustrated by Amazon's AI recruiting tool, which was designed to rank and identify top candidates but instead perpetuated historical inequities: because the model was trained on decades of hiring data that skewed heavily male, it learned to replicate those patterns and penalised resumes from women (Dastin, 2022). Although a growing literature examines algorithmic fairness, existing research tends to address bias at isolated stages of development rather than across the entire lifecycle, even though each stage carries its own sources of bias that must be managed appropriately. As a result, technical fairness interventions may fail to address how AI systems are designed, interpreted, and operationalised in organisations.

The first goal of this study is to present a classification of common AI biases and identify the specific stages of the AI lifecycle at which they arise. The second goal is to propose an AI fairness framework that maps biases to mitigation strategies relevant to specific lifecycle stages as well as those spanning multiple stages. Finally, a research agenda is proposed to guide future IS research on AI fairness. This structured mapping offers practical guidance for practitioners and policymakers seeking to align AI development with societal values and ethical principles (Nazer et al., 2023). Theoretically, the paper conceptualises AI fairness not as a

static algorithmic property or a set of isolated mitigation techniques, but as a cumulative socio-technical trajectory that unfolds across the AI lifecycle, in which fairness outcomes emerge from the sequential and recursive interaction of problem framing, data practices, model development, organisational deployment, and post-deployment monitoring.

The IS field is well positioned to address this important issue because it examines how technologies interact with organisational structures, practices, and human decision makers. Yet despite increasing attention to ethical AI, we still know relatively little about **how organisational actors understand, evaluate, and respond to fairness issues in AI-supported decision-making**. This paper develops **a research agenda for studying AI fairness as a socio-technical phenomenon in organisational contexts.** We identify key research questions that examine how fairness concerns emerge, how they are managed by organisations, and how they shape outcomes for individuals and stakeholders.

By shifting the focus from algorithmic properties to organisational processes, this paper contributes to the IS literature by conceptualising AI fairness as a socio-technical challenge, identifying the organisational mechanisms through which fairness concerns are interpreted and managed, and proposing a research agenda to guide future empirical work on responsible AI.

2. AI Lifecycle Models Overview

Existing AI lifecycle models in the literature can be broadly categorised into three categories. The first and most prevalent type focuses on AI product development and business objectives. Models such as the Cross-Industry Standard Process for Data Mining (CRISP-DM), the Microsoft model (Amershi et al., 2019), the CDAC AI lifecycle, and Moley's AI lifecycle (Georgieva et al., 2022) all begin with business understanding and proceed through data preparation, model training, evaluation, and deployment. These are widely characterised as data-driven AI development approaches (Georgieva et al., 2022), and CRISP-DM in particular has remained the de facto standard for developing data mining and knowledge discovery projects for over two decades (Haakman et al., 2021). Many subsequent models, including the Refined CRISP-DM (Haakman et al., 2021), the Extended CRISP-DM, and the Seven-Layer Model (Agarwal & Agarwal, 2024), were explicitly built upon and extended from this foundation. Their shared emphasis is on operationalising business logic through structured data pipelines that culminate in deployable AI models.

The second type centres narrowly on AI model development, prioritising the technical process of constructing and validating AI algorithms over broader business or organisational concerns. Representative examples include Allam's AI lifecycle and the AI Ethics Lifecycle. These models tend to be concise, typically with four to six stages focused on core activities such as model design, training, evaluation, and deployment, without extensive scaffolding around business context or stakeholder management. While they are technically precise, their narrowness means they often lack mechanisms for capturing the organisational conditions under which AI systems succeed or fail in practice.

The third type integrates AI development with domain-specific or scenario-driven considerations, typically adopting multi-layer structures with stages tailored to particular application fields. For instance, the Health Equity Across the AI Lifecycle (HEAAL) framework (Kim et al., 2024) incorporates a clinical integration phase designed to embed AI solutions responsibly within healthcare delivery organisations. The AI-enabled system lifecycle adds a user training stage specific to the deployment of clinical decision support systems. These models reflect the growing recognition that AI systems do not operate in a vacuum but are deeply embedded in specific institutional, regulatory, and operational contexts (Kim et al., 2024). Across all three types, ethical considerations, including fairness, bias mitigation, transparency, and accountability have been increasingly foregrounded as cross-cutting concerns throughout the lifecycle (Agarwal & Agarwal, 2024).

Despite the growing number of AI lifecycle models, a fundamental limitation remains: no existing framework comprehensively integrates business objectives, technical model development, and context-specific implementation within a unified structure. Current models tend to capture only a subset of the AI production process. This fragmentation hinders the identification of AI bias sources across the entire lifecycle. CRISP-DM and its extensions offer a rigorous process logic for data mining and model development, but they primarily treat AI development as a structured technical and business pipeline. As a result, they provide limited conceptual guidance for understanding how fairness concerns emerge through stakeholder framing, organisational implementation, user interaction, and post-deployment feedback. Conversely, domain-specific frameworks such as HEAAL make an important contribution by embedding equity concerns into context-sensitive AI deployment, particularly in healthcare. However, their domain specificity limits their transferability to broader organisational settings in which AI systems are shaped simultaneously by business objectives, technical design choices, governance arrangements, and evolving user practices. To address the identified gap, we combine existing AI lifecycle models and propose a Five-stage AI Lifecycle Model consisting of five stages: (1) Problem Definition (PD); (2) Data Collection and Preparation (DCP); (3) Development (DEV); (4) Deployment (DEP); and (5) Monitoring (MON) (Figure 1). The proposed framework was developed through a structured conceptual synthesis of prior AI lifecycle models. Existing frameworks from business-oriented, technical, and domain-specific traditions were comparatively analysed to identify recurring lifecycle activities and fairness-related governance concerns. Unlike existing models that reflect a single orientation, this AI lifecycle is designed to simultaneously integrate business objectives, technical model development, and scenario-specific considerations, making it applicable to AI model development, AI system engineering, and AI product creation. A defining feature of it is its explicit mapping of bias sources at each stage, reflecting the well-established understanding that bias compounds across the entire AI production pipeline (Ferrara, 2024).

3. Methodology

This paper is a conceptual study based on a theoretical review, a review type that draws on existing conceptual and empirical studies to develop a higher-order theoretical structure, such as a taxonomy or conceptual framework, rather than to exhaustively aggregate empirical findings (Paré et al., 2015). Because AI fairness research is dispersed across computer science, human–computer interaction, and IS, we drew on three sources. First, established surveys of AI bias and fairness (Barocas et al., 2023; Ferrara, 2024; Mehrabi et al., 2021; Suresh & Guttag, 2021) served as seed sources for bias types and mitigation techniques. Second, we searched the ACM Digital Library (including the FAccT and AIES proceedings), the AIS eLibrary, and Google Scholar for work combining terms for AI or machine learning with “fairness” or “bias”, with “lifecycle” or “pipeline”, and with “mitigation”, “audit”, or “governance”. Third, we conducted backward and forward citation searches from these sources (Webster & Watson, 2002) and added IS studies of algorithmic decision-making in organisations to inform the research agenda. Sources were retained if they (a) described a stage-based AI development or lifecycle model, (b) defined a type or source of AI bias, or (c) proposed or evaluated a bias mitigation or fairness governance practice.

Synthesis proceeded in three steps. We first compared existing lifecycle models and consolidated their stages into the five-stage model described in Section 2. We then coded each bias type by the stage(s) at which prior work locates its generating mechanism, listing a bias at more than one stage only where a distinct mechanism produces it there. Finally, we grouped mitigation techniques by their point of intervention (data, process, algorithm, or human interaction) and mapped each technique to the biases whose generating mechanism it targets. Table 1 distinguishes the elements of the framework adopted from prior work from those that constitute our own synthesis.

Table 1. Sources of AI Bias Across AI Lifecycle				
AI Lifecycle Stages	Stage Description (Integrated Definition)	References (AI Lifecycle Models)	Sources of AI Bias	References (AI Bias)
(1) Problem Definition	Establishes the business objectives, use-case boundaries, stakeholder roles, operational constraints, and ethical requirements that guide AI development.	Amershi et al., 2019; Georgieva et al., 2022; Haakman et al., 2021; Agarwal & Agarwal, 2024	Confirmation Bias	(Cascante-Bonilla et al., 2021)
			Selection Bias	(Zadrozny, 2004)
(2) Data Collection and Preparation	Involves data sourcing, acquisition, cleaning, labelling, transformation, governance, and documentation.	Georgieva et al., 2022; Haakman et al., 2021; Kim et al., 2024;	Sampling Bias	(Chen et al., 2024; Ferrara, 2024)
			Measurement Bias	(Chen et al., 2024; Ferrara, 2024)
			Representation Bias	(Suresh & Guttag, 2021)
			Labelling Bias	(Jiang & Nachum, 2020)
			Algorithmic Bias	(Mehrabi et al., 2021; Dolata et al., 2022)
(3) Development	Covers model design, feature engineering, training, validation, performance testing, and technical verification.	Amershi et al., 2019; Agarwal & Agarwal, 2024	Algorithmic Bias	(Mehrabi et al., 2021; Dolata et al., 2022)
			Confirmation Bias	(Cascante-Bonilla et al., 2021)
			Interaction Bias	(Li & Sinnamon, 2024)
			Generative Bias	(Huang & Huang, 2025)

(4) Deployment	Encompasses system integration, workflow embedding, user onboarding, infrastructure configuration, and operational rollout.	Amershi et al., 2019; Kim et al., 2024;	Emergent Bias	(Friedman & Nissenbaum, 1996; O'Connor & Liu, 2024)
			Interaction Bias	(Li & Sinnamon, 2024)
			Generative Bias	(Huang & Huang, 2025)
(5) Monitoring	Involves post-deployment performance tracking, bias detection, drift monitoring, auditing, feedback integration, and iterative refinement.	Amershi et al., 2019; Haakman et al., 2021; Agarwal & Agarwal, 2024	Feedback Loop Bias	(Mansoury et al., 2020)
			Measurement Bias	(Chen et al., 2024; Ferrara, 2024)

4. Sources of AI Biases Across the Five-stage AI Lifecycle

AI biases refer to systematic and unfairly uneven errors produced by an AI system across people or groups, which often linked to sensitive or socially salient attributes such as race, gender, age, disability, nationality, or socioeconomic status. Bias can show up as differences in error rates, resource allocation, rankings, access, or treatment, even when the model is optimised for overall accuracy. AI bias can emerge at multiple stages throughout the lifecycle of an AI system from initial design to real-world deployment and application. Specifically, bias can stem from skewed or unrepresentative training data (Barocas et al., 2023), flawed organisational governance and evaluation procedures (Raji et al., 2020), algorithmic learning processes that encode and amplify discriminatory patterns (Zemel et al., 2013), and the systematic exclusion of affected communities from design decision-making (Shneiderman, 2022). This implies that it is important to develop an understanding of the sources of biases across different stages of development and use and how they are interconnected. By using AI lifecycle model, we systematically examine and synthesise the various sources of AI bias that may arise at each stage. This approach enables a more comprehensive and nuanced understanding of how bias is introduced, amplified, or perpetuated throughout the lifecycle of an AI model, rather than isolating it to a single point of failure.

The PD stage lays the foundation for detecting and mitigating bias throughout the AI lifecycle. Passi & Barocas (2019) show that translating a real-world challenge into a machine learning task involves decisions that encode social assumptions, organisational priorities, and normative judgements, made before any data is collected or model trained, that can irrevocably shape the fairness of the resulting system. Two biases are especially prevalent at this stage: confirmation bias and selection bias.

- Confirmation Bias: This can occur if the problem definition is influenced by pre-existing beliefs or assumptions held by the developers, leading to a biased framing of the problem.

- Selection Bias: This arises if the problem definition selectively includes or excludes certain aspects or stakeholders, leading to an incomplete or skewed understanding of the problem.

The DCP stage is essential to developing AI algorithms. Data collection gathers raw data from various sources for training and evaluation. If data sources are unrepresentative or collection methods introduce systematic errors, the resulting dataset will be biased. Relevant biases in this stage include:

- Sampling Bias: This occurs when the training data is not representative of the entire population it serves, leading to skewed AI outcomes and biased predictions.
- Measurement Bias: This arises when the collected data overrepresents or underrepresents certain groups, distorting AI predictions.
- Representation Bias: This occurs when the training data underrepresents or misrepresents certain groups, leading to inaccurate predictions.

Data preparation comprises data pre-processing and data augmentation. Bias can arise from either process if cleaning, transformation, or augmentation techniques are not carefully managed. Relevant biases include:

- Labelling Bias: This arises when different groups have different interpretations of the same label, leading to inconsistencies in the training data.
- Algorithmic Bias: This can occur if the pre-processing techniques prioritise certain attributes or characteristics, leading to unfair outcomes.
- Sampling Bias: This can occur if the augmented data is not representative of the entire population, leading to skewed AI outcomes.
- Representation Bias: This arises if the augmented data underrepresents or misrepresents certain groups, leading to inaccurate predictions.

Biases in the DEV stage arise from the design and optimisation of the AI system. Relevant biases in this stage include:

- Algorithmic Bias: This arises when the optimisation and design of an algorithm prioritise certain attributes or characteristics, leading to unfair outcomes.
- Confirmation Bias: This can occur if the algorithms confirm pre-existing biases or beliefs held by the developers.
- Interaction Bias: This occurs when an AI system interacts with humans in a biased manner.
- Generative Bias: This emerges when the model's outputs disproportionately reflect specific attributes, perspectives, or patterns present in the training data, leading to skewed or unbalanced representations in generated content.

During the DEP stage, fairness is often treated as a static property verified against historical data. Once deployed, however, the AI artefact is embedded in a dynamic social context of shifting user behaviours, populations, and values, giving rise to emergent bias. Interaction and generative bias may also persist and amplify over time, because deployment places the model in an open-ended socio-technical environment rather than the curated datasets and scripted test cases of development. Real users introduce heterogeneous goals, cultural contexts, and usage strategies, including adversarial prompting and repeated probing of system limits. Relevant biases in this stage include:

- Emergent Bias: This bias typically emerges sometime after a design is completed, as a result of changing societal knowledge, population, or cultural values.
- Interaction Bias: This occurs when an AI system interacts with humans in a biased manner.
- Generative Bias: This emerges when the model's outputs disproportionately reflect specific attributes, perspectives, or patterns present in the training data, leading to skewed or unbalanced representations in generated content.

Biases in the MON stage arise from the ongoing use and feedback of the AI system: if outputs influence future inputs, a feedback loop can amplify and reinforce existing bias. Relevant biases in this stage include:

- Feedback Loop Bias: This occurs when an AI model's past outputs and decisions influence future inputs, establishing a feedback loop that can amplify and reinforce existing bias.
- Measurement Bias: This can arise if the ongoing data collection overrepresents or underrepresents certain groups, distorting AI predictions.

5. A Lifecycle Framework of AI Fairness

Given the multifaceted nature of AI bias, mitigation strategies must be considered in relation to the specific stages and contexts in which bias arises. We reviewed and consolidated strategies documented across the literature into a structured typology comprising four categories: Data Strategies (e.g., quota and stratified sampling, crowdsourcing, oversampling/undersampling, data cleansing), Process Strategies (e.g., Ethical Impact Assessments, AI audits, disaggregated evaluations, third-party fairness tools), Algorithmic Strategies (e.g., adversarial training, fair representation learning), and Human-Centred Design Strategies (e.g., participatory design, explainability for fairness, interactive alignment and disambiguation).

Building on the five-stage lifecycle, sources of bias, and mitigation strategies presented above, this section aligns biases with corresponding strategies to construct the proposed AI Fairness framework (Figure 1). Strategies were mapped according to the stage at which a bias primarily emerges and the mechanism through which the strategy intervenes in that bias-generating process. The framework conceptualises AI fairness as a socio-technical phenomenon emerging from interactions between algorithm design, organisational governance, and how organisational actors interpret fairness in AI-supported decisions.

5.1 Problem Definition

Confirmation bias and selection bias at the problem definition stage are often downstream effects of structural inequalities embedded in the framing process itself. Data strategies (quota and stratified sampling) and process strategies (Ethical Impact Assessments and AI Audits) can substantially mitigate both. Quota sampling specifies desired demographic representation before data collection and recruits until quotas are filled, counteracting framers' tendency to consult only familiar, like-minded stakeholders (Cochran, 1977). Stratified sampling instead draws random samples from predefined subgroups in proportion to their population share, ensuring statistically representative input. An Ethical Impact Assessment (EIA) requires framers to make assumptions explicit and submit them to review by uninvolved parties, reducing confirmation bias, and to identify second-order as well as primary stakeholders, directly countering selection bias. AI audits can be applied prospectively, examining problem-framing artefacts against fairness criteria (Raji et al., 2020). Koshiyama et al. (2024) recommend an "assumption inventory" to expose shared assumptions to external scrutiny, while Raji et al. (2020) stress that audits should examine not only what data were collected but what, and who, was excluded and why. Both biases are thus socially produced phenomena rooted in how human actors understand fairness and in institutional power relations that determine whose understanding prevails.

5.2 Data Collection & Preparation

Quota and stratified sampling remain applicable here, reducing sampling and representation bias by requiring adequate subgroup representation before the dataset is finalised, though both depend on underrepresented individuals' willingness to contribute data. Resampling offers a post hoc remedy for existing imbalance, using oversampling and undersampling to rebalance classes (Lemaître et al., 2017), though it works best alongside disaggregated evaluation to avoid amplifying existing bias (Mitchell et al., 2019). Data cleansing addresses measurement bias by improving feature quality and assessing whether measurement instruments function equivalently across subgroups, and similarly addresses label bias by correcting biased annotations (Raji et al., 2020); in both cases, AI audits provide oversight. Crowdsourcing mitigates label bias by combining diverse annotators to dilute the systematic bias of any single one (Snow et al., 2008). Disaggregated evaluation remains the key diagnostic across this stage, since fairness concepts such as demographic parity, equalised odds, and calibration rely on different subgroup metrics (Barocas et al., 2023); third-party tools such as AI Fairness 360 and Fairlearn support independent assessment.

5.3 Development

AI audits here apply the same scrutiny to modelling decisions as to data selection, including the objective function, thresholds, and architecture, structurally independent from delivery pressures (Raji et al., 2020), addressing Holstein et al.'s (2019) finding that fairness concerns are often overridden by accuracy and timeline priorities. Disaggregated evaluation similarly compares subgroup performance across configurations (Barocas et al., 2023). Fairness-aware algorithms also help at this stage. GAN-based Adversarial Training reduces correlations between predictions and sensitive attributes (Zhang et al., 2018), while Learning Fair Representations builds fairness constraints into the representation level so they carry over to any downstream task (Zemel et al., 2013). EIAs extend to contest the value-laden metric choices developers make. For interaction and generative bias, a newer generation of third-party tools evaluates generated content directly, including HELM for language models (Liang et al., 2022) and FairFace and Stable Bias for image generation. Participatory design addresses interaction bias by involving marginalised communities as co-designers rather than test subjects. Interactive alignment and disambiguation let users signal misunderstanding and correct system behaviour, though poorly designed correction mechanisms can themselves introduce new interaction bias if they favour certain user groups (Amershi et al., 2014). Explainability for fairness makes a system's fairness-related reasoning visible to users and oversight bodies (Doshi-Velez & Kim, 2017).

5.4 Deployment

Unlike preventive development audits, deployment audits are ongoing and adaptive, monitoring real-world behaviour and distributional shifts; Metcalf et al. (2021) argue this requires affected communities' participation, since emergent bias often surfaces in user experience before it appears in aggregate metrics. Interactive alignment becomes riskier at scale: uneven feedback from different user groups can reinforce rather than correct imbalances, and because millions of daily interactions can amplify bias faster than pre-deployment testing can simulate, human-centred design strategies become the most cost-effective lever once retraining is no longer feasible.

5.5 Monitoring

Oversampling and undersampling are applied to the ongoing feedback stream to disrupt loops where biased outputs become biased training inputs. Data cleansing screens out labels contaminated by the system's own prior decisions, while crowdsourcing now gathers diverse

evaluations of outputs rather than training annotations. Ethical impact assessments, disaggregated evaluations, and AI audits remain the key oversight tools for feedback loop and measurement bias.

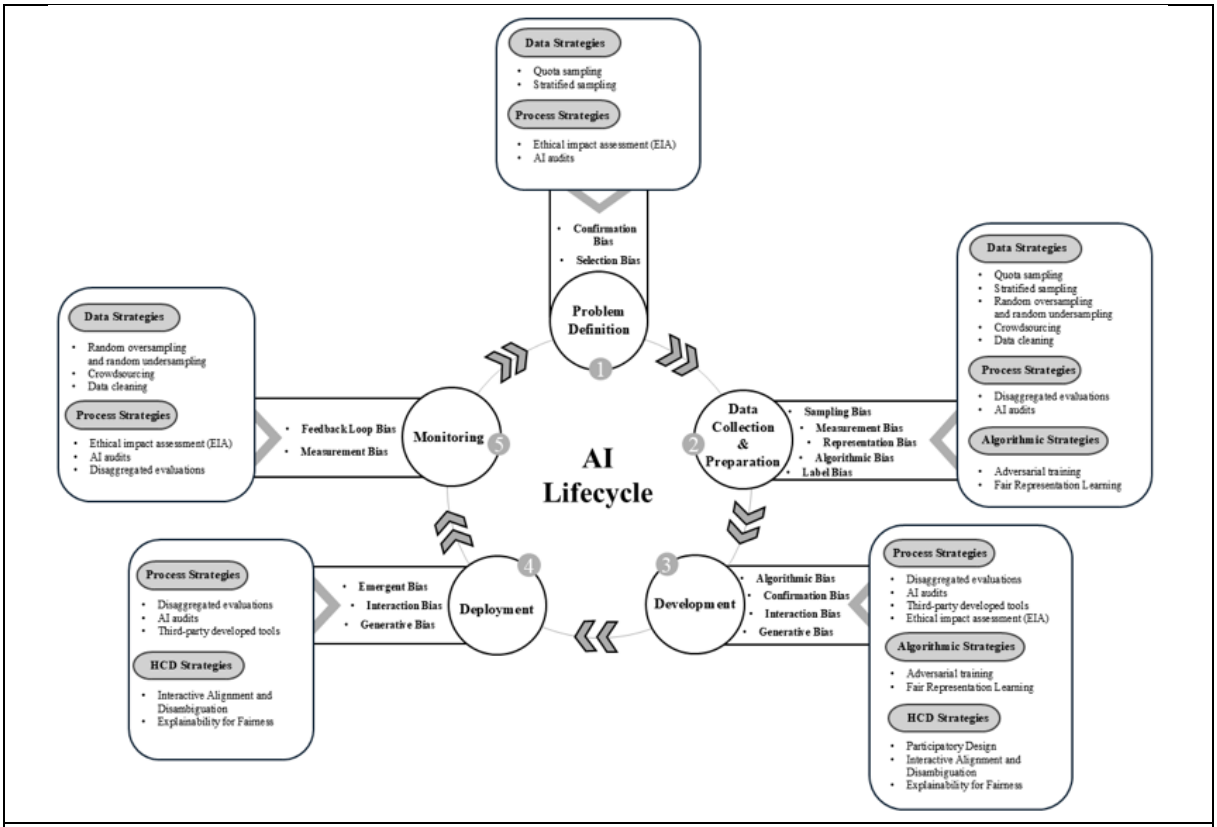


Figure 1. AI Fairness Framework

6. AI Fairness Research Agenda in IS

The proposed framework not only maps how biases and mitigation strategies emerge across the AI lifecycle but also reveals unresolved organisational and socio-technical tensions insufficiently addressed in existing AI fairness research. Building on these gaps, the following research agenda outlines key directions for examining AI fairness as a socio-technical and organisational phenomenon, drawing on Socio-Technical Systems Theory: fairness outcomes cannot be understood solely as properties of algorithms but must be examined in relation to the organisational contexts in which systems are designed and deployed.

6.1 Theme 1: Organisational Interpretation of AI Fairness

Organisational interpretation of AI fairness is an important IS concern because technologies do not carry a single fixed meaning; they are interpreted through the perspectives, roles, and identities of those who use and manage them (Orlikowski & Iacono, 2001). This is especially important for AI fairness because the concept has multiple incompatible mathematical definitions, such as demographic parity, equalised odds, and calibration (Barocas et al., 2023).

A key underexplored challenge is how organisations reconcile competing definitions of fairness: statistical and procedural definitions reflect different moral views and can yield conflicting recommendations for the same system (Barocas et al., 2023), yet IS research has not examined how organisations resolve these tensions—through managerial authority,

professional consensus, stakeholder negotiation, or institutional pressure. This matters because governance effectiveness depends on whether actors share a common understanding of the fairness goals they pursue. Hence, we propose:

- RQ1: How do organisational actors interpret and define fairness in AI-supported decision-making?
- RQ2: How do professional roles (e.g., managers, developers, HR professionals) shape perceptions of algorithmic fairness?
- RQ3: How do organisations reconcile competing definitions of fairness (e.g., statistical fairness vs. procedural fairness)?

6.2 Theme 2: Governance and Organisational Practices

The second theme examines how organisations govern AI fairness through structures, accountability arrangements, transparency practices, and strategic trade-offs. IT governance structures should align with an organisation's strategy, environment, and coordination needs (Sambamurthy & Zmud, 1999), yet IS research has not examined how AI fairness governance varies across these dimensions or how such variation shapes bias identification and response. Hence, we propose:

- RQ4: What governance mechanisms do organisations use to identify and manage fairness risks in AI systems?

Governance failures often stem from weak accountability: Kronblad et al. (2024) show how injustice persisted in a public-sector admissions system through algorithmic opacity and organisational avoidance, while Grønsund & Aanestad (2020) find that practices such as establishing ground truth and evaluating algorithmic outputs help organisations remain reflexive under changing conditions. Hence, we propose:

- RQ5: How do accountability structures influence responses to perceived algorithmic bias?

Transparency is another underexplored governance dimension. Regulations such as the EU AI Act emphasise that AI systems should be explainable to those affected, yet IS research on how transparency shapes employee trust remains limited. Rinta-Kahila et al. (2022) show that algorithmic opacity in an automated debt recovery system prevented people from contesting decisions, suggesting transparency is a condition for meaningful human agency, not only a regulatory requirement. Hence, we propose:

- RQ6: How do transparency practices shape employee trust in AI-supported decisions?

A third governance area concerns the efficiency–fairness trade-off. Möhlmann et al. (2021) show that algorithmic management on gig platforms improves matching efficiency but can create fairness harms for workers, illustrating how efficiency-oriented governance can undermine fairness. IS research has not yet explained how organisations balance these goals over time. Hence, we propose:

- RQ7: How do organisations balance efficiency gains from AI with fairness considerations?

6.3 Theme 3: Outcomes and Impacts

The third theme examines the consequences of AI fairness governance and its failures for employees, organisations, and the wider socio-technical systems in which AI operates, a perspective important for justifying investment in fairness and evaluating which governance approaches work best.

IS research has begun documenting these harms. Marjanovic et al. (2021) introduce “algorithmic pollution” to describe diffuse societal harms from automated decisions in welfare, healthcare, and education that emerge gradually rather than through single events. Marjanovic et al. (2022) further identify three forms of algorithmic injustice: maldistribution, misrecognition, and misrepresentation, together showing that algorithmic unfairness can undermine individual agency, organisational legitimacy, and social trust.

These harms are closely linked to how organisational actors respond to AI in practice. Kordzadeh & Ghasemaghaei (2022) argue that algorithmic bias shapes fairness perceptions, which influence recommendation acceptance and system adoption, though this model has received limited empirical validation. Teodorescu et al. (2021) show that automated fairness approaches often fail to capture real decision-environment complexity, creating a need for human augmentation, yet it remains unclear when employees are willing and able to provide such oversight, especially when they perceive a system as unfair. Hence, we propose:

- RQ8: How do employees respond to AI-supported decisions perceived as unfair?
- RQ9: How do fairness concerns influence the adoption and continued use of AI systems?

Fairness perceptions and AI adoption are closely linked: when employees or users perceive AI-supported decisions as unfair, they may disengage from using, supporting, or providing feedback to these systems, creating a negative cycle that limits the human input needed to improve AI performance. Understanding when fairness interventions can improve both legitimacy and system performance, rather than trading one off against the other, is a key challenge for IS research on responsible AI governance. Hence, we propose:

- RQ10: Under what conditions do fairness interventions improve both perceived legitimacy and organisational outcomes?

7. Conclusion

By integrating business, technical, and contextual perspectives within a single model, the framework demonstrates that bias is a compounding, lifecycle-wide phenomenon requiring sustained governance rather than point-in-time fixes. The paper developed a five-stage AI fairness framework mapping specific biases to mitigation strategies and presented a research agenda related to AI fairness. We call on scholars to further develop and test this research agenda using a wide range of qualitative and quantitative methodologies.

Another focus of this study is theories related to algorithmic governance to examine how algorithmic systems shape decision-making authority and organisational control. AI systems can shift decision power from human actors to computational processes, raising questions about accountability, transparency, and fairness. The literature on organisational justice provides a useful lens for understanding how individuals evaluate the fairness of decision processes. Concepts such as procedural justice, distributive justice, and interactional justice help explain how employees interpret and respond to AI-supported decisions.

Integrating these perspectives can be especially useful since AI fairness is not merely a technical property but an organisational phenomenon emerging from interactions between algorithms, organisational structures, and human actors. This paper therefore examines fairness in AI systems as a socio-technical process shaped by governance structures, decision practices, and stakeholder perceptions. The accompanying research agenda identifies priority questions spanning stakeholder fairness conceptualisations, organisational incentives, user perception, and long-term monitoring capacity. Together, these contributions provide structured guidance for practitioners, policymakers, and IS researchers committed to building AI systems that are fair, accountable, and aligned with societal values.

8. References

Agarwal, A., & Agarwal, H. (2024). A seven-layer model with checklists for standardising fairness assessment throughout the AI lifecycle. AI and Ethics, 4(2), 299-314.

Amershi, S., Cakmak, M., Knox, W. B., & Kulesza, T. (2014). Power to the people: The role of humans in interactive machine learning. AI magazine, 35(4), 105-120.

Amershi, S., Begel, A., Bird, C., DeLine, R., Gall, H., Kamar, E., ... & Zimmermann, T. (2019, May). Software engineering for machine learning: A case study. In 2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP) (pp. 291-300). IEEE.

Barocas, S., Hardt, M., & Narayanan, A. (2023). Fairness and machine learning: Limitations and opportunities. MIT press.

Cascante-Bonilla, P., Tan, F., Qi, Y., & Ordonez, V. (2021, May). Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning. In Proceedings of the AAAI conference on artificial intelligence (Vol. 35, No. 8, pp. 6912-6920).

Chen, F., Wang, L., Hong, J., Jiang, J., & Zhou, L. (2024). Unmasking bias in artificial intelligence: a systematic review of bias detection and mitigation strategies in electronic health record-based models. Journal of the American Medical Informatics Association, 31(5), 1172-1183.

Cochran, W. G. (1977). Sampling techniques. john wiley & sons.

Dastin, J. (2022). Amazon scraps secret AI recruiting tool that showed bias against women. In Ethics of data and analytics (pp. 296-299). Auerbach Publications.

Dolata, M., Feuerriegel, S., & Schwabe, G. (2022). A sociotechnical view of algorithmic fairness. Information Systems Journal, 32(4), 754-818.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

Ferrara, E. (2024). Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. Sci, 6(1), 3.

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. ACM Transactions on information systems (TOIS), 14(3), 330-347.

Georgieva, I., Lazo, C., Timan, T., & Van Veenstra, A. F. (2022). From AI ethics principles to data science practice: a reflection and a gap analysis based on recent frameworks and practical experience. AI and Ethics, 2(4), 697-711.

Grønsund, T., & Aanestad, M. (2020). Augmenting the algorithm: Emerging human-in-the-loop work configurations. The Journal of Strategic Information Systems, 29(2), 101614.

Haakman, M., Cruz, L., Huijgens, H., & Van Deursen, A. (2021). AI lifecycle models need to be revised: An exploratory study in Fintech. Empirical Software Engineering, 26(5), 95.

Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudik, M., & Wallach, H. (2019, May). Improving fairness in machine learning systems: What do industry practitioners need?. In Proceedings of the 2019 CHI conference on human factors in computing systems (pp. 1-16).

Huang, L. T. L., & Huang, T. R. (2025). Generative bias: widespread, unexpected, and uninterpretable biases in generative models and their implications. AI & SOCIETY, 1-14.

Jiang, H., & Nachum, O. (2020, June). Identifying and correcting label bias in machine learning. In International conference on artificial intelligence and statistics (pp. 702-712). PMLR.

Kim, J. Y., Hasan, A., Kellogg, K. C., Ratliff, W., Murray, S. G., Suresh, H., ... & Sendak, M. P. (2024). Development and preliminary testing of Health Equity Across the AI Lifecycle (HEAAL): A framework for healthcare delivery organisations to mitigate the risk of AI solutions worsening health inequities. PLOS Digital Health, 3(5), e0000390.

Kordzadeh, N., & Ghasemaghaei, M. (2022). Algorithmic bias: review, synthesis, and future research directions. European Journal of Information Systems, 31(3), 388-409.

Koshiyama, A., Kazim, E., Treleaven, P., Rai, P., Szpruch, L., Pavey, G., ... & Chatterjee, S. (2024). Towards algorithm auditing: managing legal, ethical and technological risks of AI, ML and associated algorithms. Royal Society Open Science, 11(5).

Kronblad, C., Essén, A., & Mähring, M. (2024). When justice is blind to algorithms: Multilayered blackboxing of algorithmic decision-making in the public sector. MIS Quarterly, 48(4), 1637-1662.

Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. Journal of machine learning research, 18(17), 1-5.

Li, A., & Sinnamon, L. (2024). Generative ai search engines as arbiters of public knowledge: An audit of bias and authority. Proceedings of the Association for Information Science and Technology, 61(1), 205-217.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., ... & Koreeda, Y. (2022). Holistic evaluation of language models. arXiv preprint arXiv:2211.09110.

Mansoury, M., Abdollahpouri, H., Pechenizkiy, M., Mobasher, B., & Burke, R. (2020, October). Feedback loop and bias amplification in recommender systems. In Proceedings of the 29th ACM international conference on information & knowledge management (pp. 2145-2148).

Marjanovic, O., Cecez-Kecmanovic, D., & Vidgen, R. (2021). Algorithmic pollution: Making the invisible visible. Journal of Information Technology, 36(4), 391-408.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM computing surveys (CSUR), 54(6), 1-35.

Marjanovic, O., Cecez-Kecmanovic, D., & Vidgen, R. (2022). Theorising algorithmic justice. European Journal of Information Systems, 31(3), 269-287.

Metcalf, J., Moss, E., Watkins, E. A., Singh, R., & Elish, M. C. (2021, March). Algorithmic impact assessments and accountability: The co-construction of impacts. In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency (pp. 735-746).

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019, January). Model cards for model reporting. In Proceedings of the conference on fairness, accountability, and transparency (pp. 220-229).

Möhlmann, M., Zalmanson, L., Henfridsson, O., & Gregory, R. W. (2021). Algorithmic management of work on online labor platforms: When matching meets control. MIS quarterly, 45(4), 1999-2022.

Nazer, L. H., Zatarah, R., Waldrip, S., Ke, J. X. C., Moukheiber, M., Khanna, A. K., ... & Mathur, P. (2023). Bias in artificial intelligence algorithms and recommendations for mitigation. PLOS digital health, 2(6), e0000278.

O'Connor, S., & Liu, H. (2024). Gender bias perpetuation and mitigation in AI technologies: challenges and opportunities. AI & SOCIETY, 39(4), 2045-2057.

Orlikowski, W. J., & Iacono, C. S. (2001). Research commentary: Desperately seeking the "IT" in IT research—A call to theorizing the IT artifact. Information systems research, 12(2), 121-134.

Paré, G., Trudel, M. C., Jaana, M., & Kitsiou, S. (2015). Synthesizing information systems knowledge: A typology of literature reviews. Information & management, 52(2), 183-199.

Passi, S., & Barocas, S. (2019, January). Problem formulation and fairness. In Proceedings of the conference on fairness, accountability, and transparency (pp. 39-48).

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., ... & Barnes, P. (2020, January). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In Proceedings of the 2020 conference on fairness, accountability, and transparency (pp. 33-44).

Rinta-Kahila, T., Someh, I., Gillespie, N., Indulska, M., & Gregor, S. (2022). Algorithmic decision-making and system destructiveness: A case of automatic debt recovery. European Journal of Information Systems, 31(3), 313-338.

Sambamurthy, V., & Zmud, R. W. (1999). Arrangements for Information Technology Governance: A Theory of Multiple Contingencies. MIS quarterly, 23(2), 261-290.

Shneiderman, B. (2022). Human-centered AI. Oxford University Press.

Snow, R., O'connor, B., Jurafsky, D., & Ng, A. Y. (2008, October). Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In Proceedings of the 2008 conference on empirical methods in natural language processing (pp. 254-263).

Suresh, H., & Guttag, J. (2021, October). A framework for understanding sources of harm throughout the machine learning life cycle. In Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (pp. 1-9).

Teodorescu, M. H., Morse, L., Awwad, Y., & Kane, G. C. (2021). Failures of fairness in automation require a deeper understanding of human–ml augmentation. MIS quarterly, 45(3), 1483-1500.

Webster, J., & Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review. MIS quarterly, xiii-xxiii.

Zadrozny, B. (2004, July). Learning and evaluating classifiers under sample selection bias. In Proceedings of the twenty-first international conference on Machine learning (p. 114).

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013, May). Learning fair representations. In International conference on machine learning (pp. 325-333). PMLR.

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018, December). Mitigating unwanted biases with adversarial learning. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (pp. 335-340).

Sociotechnical Dualism Considered Harmful

Rodney J. Clarke
University of Wollongong

Abstract
Socio-technical systems theory has become the dominant way in which Information Systems researchers and practitioners describe their activities and distinguish their discipline from other informatic domains like computer science and software engineering.The term has been a controversial one since its introduction into the IS literature in the 1960s, and has been the subject of significant philosophical critique. This paper explores that application of Derrida's reading tactics to investigate the sociotechnical dualism, the sufficiency of its core categories as well as its underlying relation. This analysis reveals that the sociotechnical dualism can be dissolved entirely replacing it with an explicit 'communication' theorisation that can, in principal, produce technical artifacts. The analysis is also used to specify requisite design criteria for identifying and selecting this theorisation.

1. Introduction

Sociotechnical systems (STS) theory developed out of the pioneering work of the London-based *Tavistock Institute for Human Relations* (Trist and Bamforth 1951; Trist et al 1963). These researchers were responding to the prevailing Taylorism that focused on purely technical and economic approaches to work design of the time (Taylor 1914) while regarding issues of worker psychology and social organisation to be, at best, implementation constraints to be managed (Doray 1988, Banta 1993). The Tavistock researchers were trying to account for unexpected and unexplained variations in the performance of new technologies being introduced across the British coal mining sector during the 1950s. Trist and Bamforth (1951) argued in favour of the untested assumption that a production system should not be seen exclusively as a 'technical system' or a 'social system' (social relations and work organisation) but necessarily as both. Their work introduced the category of "the social" prepending it to "the technical"- suggesting, orthographically at least, that the former was the more primary option of the two. Curiously, 'the social' and 'the technical' are not clearly defined in Trist and Bamforth (1951). These terms are instead characterised as lists of attributes (subcategories) rather than based on any kind of principled criteria; extensional (referents) rather than intensional (sense) definitions- and for the UK mining case studies these options are based on obvious *observational* distinctions. The social subsystem comprises the individuals or people that constitute an organisation, their relationships, values, structure, work-related elements and

associations formed by organisational members. The technical subsystem comprises the physical and material flows within a transformation process, in addition to the tasks, control and maintenance functions. In organisational settings, the technical subsystem problematically includes not only technologies, devices and tools but also confusingly, skills, techniques and procedures that are required by workers to fulfil organisational objectives and tasks. Trist and his colleagues also enthusiastically adopted *action research* that had been contemporaneously described in the human relations literature (Lewin 1946, 1947; Trist 1981). In doing so, STS became a methodological as well as a political- and perhaps also a moral- critique of Taylorism (Doray 1988; Banta 1993).

Sociotechnical theory proved to be problematic within the Human Relations literature. The first theoretical synthesis of the Tavistock research (Emery 1959) did so in terms of individual psychology as well as open systems. The selection of individual psychology led to failures of interpreting the degree to which 'culture' influences 'social' bonds, for example within workgroups in different cultures like those of the British Durham coal mines and Indian textile mills. As a consequence of selecting an erroneous *theory of the subject* (Henriques et al 1984, 11-25)- the concrete individual, the subject of psychology- sociotechnical theorists view their work in terms of being *culture-free* rather than being *culture-specific* (Kelly 1978). Kelly (1978) also identified several other major theoretical issues as they effected Human Resources, in particular *joint optimisation* of 'the social' and 'the technical'. It seemed to him that what was described as joint optimisation was essentially never applied in practice and that it should be replaced with the more realistic notion of the 'intensification of labour'. Also identified in this reinvestigation of the foundational STS case studies, were a range of issues including the omission of the significance of wage incentives as well as calling into question the notion of organisational choice as only one work organisation (the autonomous group) had ever been considered.

The migration of STS theory into IS in the 1970s and 1980s has been theoretically productive, restating the significance of organisational theory in IS, emphasising the importance of user involvement and participation in the design process, as well as binding together the processes of technical system design and organisational change (Checkland 1981, Mumford 1983, Pasmore et al 1982). STS theory has become a prevailing way in which IS researchers and practitioners describe their identity and distinguish their discipline and its practices from other related informatic disciplines like Computer Science and Software Engineering. Nonetheless, STS theory has also been controversial since its introduction into the IS literature. A milestone two-part article by Bostrom and Heinen (1977a, b) utilised STS theory to argue that many problems and failures of MIS and Management Science/Operations research projects had been attributed to organisational behavioural problems rather than inadequate design. Usually these 'bad' designs were associated with the way MIS designers view organisations, their members and the function of the MIS systems within their organisational contexts. Criticism has not been limited to the use of these *categories* in design practice but also, more fundamentally, how well these categories have been defined; discussed in section 2.

The *joint optimisation* mechanism has also proved problematic for IS. Had earlier STS advocates achieved joint optimisation between 'the social' and 'the technical' at all, or had they simply privileged the social system? Langefors (1978) critiqued Bostrom and Heinen (1977a, b) concerning whether it was designers and their practices that were the primary reason for the failures of these systems in the first instance. He also questioned whether category-specific incremental improvements in 'the social' or 'the technical' could sufficiently address the current- let alone future- sustainability or resilience of these systems. Joint optimisation appears compelling because of the promise it makes concerning the ability to provide *'goodness of fit'* between 'the social' and 'the technical'. An enduring notion within STS theory is that privileging 'the social' or 'the technical' subsystem in isolation degrades the understanding of the focal system as a whole (Ackoff and Emery 1972, Ackoff 2010). Beyond this, joint optimisation also involves the *nature of the interactions* between 'the social' and 'the technical' and indeed how to characterise these interactions. Informational *feedback* is

considered to be the mechanism of joint optimisation; enabling the functions associated with 'the social' and 'the technical' to achieve their desired states. The 'jointness' in joint optimisation refers to the informational common currency in which these disparate categories- 'the social' and 'the technical'- can be considered to be in any kind of equivalency. We will return to the matter of this *commonsense* (Belsey 1980) *metarepresentation* latter in section 3. For now, we will simply assume that information theory constitutes the actual meta-representation characterising both 'the social' and 'the technical' as well as whatever dynamic state these core categories may enter into.

With this disciplinary background, the *sociotechnical relation* will now be considered on philosophical grounds in the remainder of this paper. The categories of 'the social' and 'the technical' will be explored in two stages. The first stage involves a general philosophical exploration of the sociotechnical relation as a *dualism*. This allows the problematic nature of these categories to be explored in detail; see section 2. The second stage involves the application of *reading tactics* (Derrida 1967, 1972, 1981) to reveal what the sociotechnical relation produces and its effects on the ontological and epistemological foundations of the IS discipline. Reading tactics enable (i) the relation between 'the social' and 'the technical' to be better understood by means of applying the *reversal* and *displacement* tactics, see section 3, and then through the selection of a so-called *hinge term* to specifically show what is actually being hidden by this relation, sufficient to collapse the dualism entirely, see section 4. What results from subjecting the sociotechnical dualism to reading tactics is a set of *design criteria* for theories or theoretical frameworks that might provide a social basis for technical system specification, see section 5. The final section summarises the argument and opens up the discussion on what might be some of the ramifications of these theoretical design criteria on the IS discipline.

2. Philosophical Responses to the Sociotechnical Relation

In this section, the sociotechnical relation is considered from a philosophical perspective emphasising the kinds of fallacies, errors and problematic assumptions that arise form it. Many of these issues have been identified early in the adoption of STS into the IS discipline and as previously mentioned were already problems before this approach was appropriated into the IS discipline. They are profoundly philosophical in nature.

2.1. Characterising the *actual* Relation- Dualism not Dichotomy

It is necessary to first understand the nature of the sociotechnical relation. The first possibility is that the socio-technical relation is a *dichotomy*. That is to say, the options in question ('the social', 'the technical') must exhibit certain obligatory characteristics of any pair of options in a dichotomous relation with each another. These obligatory characteristics include being: (1) *exhaustive*, that is, they must provide complete coverage of the entire domain; (2) *mutually exclusive* in that membership to one option must preclude membership to the other option; (3) in a direct contrast or *logical opposition* with each other (either through opposition, negation or absence); (4) operating at the same level of analysis and refer to a common *shared domain* and finally that the (5) conceptual boundary between these options must be *stable* and unchanging within its context of use.

Whereas a dichotomy is a classificatory division of a domain into two mutually exclusive and exhaustive categories, a *dualism* is a metaphysical claim about reality itself being constituted by two fundamentally distinct kinds of thing- 'the social' and 'the technical'. While a dichotomy has options which are effectively extremes for the same kind of thing, dualisms involve concepts that are essentially different kinds of things drawn from different domains that are being held together in some kind of agonistic relation. As we shall see, there are several major

issues with the sociotechnical relation, and a number of significant issues that are produced as a consequence of those.

2.2. Socio-technical Relation as Categorical Improverishment

As indicated previously, there were legitimate early indications in the development of STS theory, of substantial questions relating to the inadequacy of its chosen categories to provide comprehensive accounts of technology use and development in social settings. These debates concern the degree of categorical sufficiency, or alternatively of *categorical impoverishment*, of sociotechnical accounts in the IS discipline. For example, there were considerable debates about economic considerations and political dimensions of systems being absent from the sociotechnical literature.

STS theory treats both 'the social' and 'the technical' categories as co-equal perspectives of the same thing (Cherns 1976, 1987). Misclassifying subcategories of 'the social' as belonging to 'the technical', discussed below, are in and of themselves *category errors* and may, in fact, emphasise the categorical impoverishment of STS theory's social and technical categories. A category error occurs when something is treated as if it belongs to a category where in fact it does not.

2.3. Consequences of Category Errors

There are major consequences of committing category errors. Subcategories or attributes of 'the social' include (1) workpractices, (2) work organisation, (3) roles and responsibilities, (4) group norms and dynamics, (5) social identities and (6) organisational culture. As a consequence of the misclassification of (1) workpractices and (2) work organisation as members of 'the technical', these subcategories are stripped of their socially dynamic and contextual character, flattening them into a logical structure. Work practices and work organisation defined as social are not the same when misclassified as belonging to the technical; they are entirely transformed. This observation becomes crutial later in the paper; see section 4. After the category errors have been committed, the fate of the remaining social subcategories depends on their *closeness to 'the technical'*. For example, (3) roles and responsibilities are replaced by fixed 'values' while (5) social identities and (6) organisational culture are likely to be ignored despite their significance for understanding (1) workpractices and (2) work organisation. What was once a corrective to an exclusively 'technical' concern- both in political and methodological terms- has now become a constraint. The socio-technical dualism reinserts 'the technical' as a persistent ontological fixture, making it impossible to consider an information systems practice that does not already naturalise the presence of a separate technological representation.

3. Reading the Sociotechnical- Reversal & Displacement

The approach adopted here is based on the work of Jacques Derrida (1967 1972, 1981) and concerns an application of his *reading tactics*. These are described below and applied in two stages across this section and the next. The advantage of a Derridean reading of the sociotechnical relation is that it can be productive of new approaches to core categories along with the relationship between them; it invites a creative reconsideration and reworking of both.

3.1. Orientation

Derrida (1967, 1972, 1981) developed *reading tactics* across a range of publications. They exist semi-independently from his broader project of *deconstruction*; not considered further

here. In this section and the next, reading tactics are applied in order to explain the functioning of the sociotechnical dualism in IS. While a more standard philosophical interpretation of the sociotechnical dualism involves an examination of the categories (as outlined in the previous section), Derrida's approach foregrounds the nature of the relations, recognising that they are not natural, inherent or inevitable but rather that dualisms often form '*violent' hierarchies* where one term is *privileged* while the other term is in some ways *suppressed*. The relationship between options in the sociotechnical relation reveals something about the discipline's foundational arrangements that cannot otherwise be easily explored.

Derrida insists that reading tactics and his broader project of deconstruction more generally, should not be considered as a method, nor can it be converted into one for several reasons. Specifically a method, (i) presupposes an exterior observer that applies a procedure from the outside, (ii) could be applied uniformly across cases producing predictable results of some kind, and (iii) requires a language that stands above the object of analysis without being contaminated by it (theorisation). So methods require external observers, imply repeatable application and necessitate a metalanguage. For reading tactics, (i) the reader cannot be 'outside' instead they are always and already inside the text and its presuppositions, (ii) rather than being 'repeatable' in their application, reading tactics are always singular and contingent, applicable within specific dualisms, dichotomies or texts, and (iii) that 'deconstructive' readings are always conducted in and through the very language that is being interrogated or investigated.

This section deals with Derrida's first two reading tactics of *reversal* and *displacement* that involve overturning and then displacing the duality. The first tactic reverses the hierarchy embedded in a duality. The second tactic is more radical, aiming at a general displacement of the duality by unsettling the relation that makes the duality possible. Prior to applying Derrida's Reading Tactics, it is necessary to first determine what is the *dominant option* in the dualism, described in the next subsection.

3.2. Identifying the Dominant Option

Either sociotechnical categories can be considered as possibly ‘dominant’, see below, however as a first approximation, ‘the technical’ is suggested as the more dominate of the two. It is not unreasonable to simply substitute ‘topic awareness’ as a proxy for ‘dominant’. Consider, for example, the ease with which any IS academic could spit out a list of recent disciplinary preoccupations: *Internet of Things* (2014), *Consumer VR/AR* (2016), *Enterprise Blockchain* (2017), *Quantum Computing* (2019), *NFTs/Web3* (2021), *Metaverse* (2021–22), *Generative AI* (2022) and *AI agents* (2023–24). Unsurprisingly this is a list of ‘technical’ topics! We don’t usually show the same excitment for notable social theories within the IS discipline; although with some effort we might create a list of them: *Socio-technical systems* (1960s–70s), *Diffusion of Innovation* (1980s), *Structuration Theory* (1984–90s), *Social Construction of Technology-SCOT* (late 1980s–90s), *Actor-Network Theory* (1990s), *Sensemaking* (1990s), *Institutional Theory* (1990s–2000s), *Critical Social Theory* (1990s–2000s), *Practice Theory* (2000s) and *Sociomateriality* (2000s–2010s). It does not matter if these lists of social and technical options are complete. What is obvious is that ‘the social’ topics are much less resonant, although of course they may have longer shelf-lives in the literature. The ‘technical’ options are much more dominant than ‘the social’ ones’; it is rare to have a socially themed academic IS conference, however the reverse is very common.

While STS nominally positions ‘the social’ and ‘the technical’ as co-equal, the theory betrays a *tacit hierarchy*. The technical is treated as stable and determinable while the social is considered inconsistent, malleable and accommodating. The social is emergent and contingent and therefore adjusts to these fixed technical ‘realities’. Historically, ‘the technical’ is the dominant term- the preexisting dominant category inherited from Taylorism. The social was introduced as a correction to its dominance. STS theory presents itself as giving equal weight to these two conceprts, but its genealogy reveals that ‘the technical’ is the original term that is left while ‘the social’ is a *supplement* to it.

3.3. Reversal

Once a dominant term is identified it is possible to consider what the dualism would be like if it was reversed. The point of imagining a reversal of the dominance of these terms in a dualism is not to suggest that one or the other reading is inherently true, but rather to reveal something important about the dualism that the conventional arrangement conceals. It is possible to consider ‘the social’ as the dominant sociotechnical option.

In the case of the sociotechnical dualism, it is not immediately obvious as to which concept is the dominant term. In the rhetorical and political framing of STS, ‘the social’ is what is being rescued from the dominance of ‘the technical’; ‘the social’ has the higher moral value. On the other hand, ‘the technical’ encompasses those aspects of the Information Systems discipline that appear thoroughly necessary in order to create computational artifacts. So, applying reading tactics to the sociotechnical dualism identifies the two contradictory hierarchies, one where ‘the technical’ is genealogical privileged, and the other were ‘the social’ is morally privileged.

Derrida’s reversal tactic reveals the same category errors that the standard philosophical reading also identified, see previous section, but interestingly *reversal* also identified the prevailing *direction* of these category errors- always and inevitably to be included within ‘the technical’. ‘The technical’ is the category that is always being bulstered. Reversal shows how ‘the social’ is actually the *condition of possibility* for ‘the technical’. The latter only functions because it is embedded in and sustained within ‘the social’.

3.4. Displacement

What happens to each option in the dualism when the two are brought into closer proximity to each other? As the social and the technical are forced closer together, each begins to lose its definitional autonomy. Each term starts to exhibit the properties supposedly exclusive to the other, and the opposition begins to collapse from within its own logic. The displacement tactic reveals and traces the *mutual contamination* of both categories.

A technical system requires interpretation, maintenance, workaround, adoption, all of which are social practices. The technical system cannot independently specify its own conditions of correct use. It depends on social meaning to function at all. The technical begins to look like a particular stabilised form of social practice. Simultaneously, as the social is brought closer to the technical, it reveals that it is never purely immaterial or unmediated. Social practices are always already inscribed as documents, interfaces and so on. The social begins to look like it has always contained a technical dimension that it cannot account for on its own terms.

At the limit of this displacement, each term has migrated so far toward the other that the original opposition no longer holds. The technical is revealed to be sedimented social practice while the social is revealed as always already technically inscribed. Rather than being resolved, the dualism is shown to be unstable from within.

4. Identifying the Hinge

Derrida's third reading tactic is referred to as *La Brisure* (trans. "the Break") or more commonly referred to as *'the hinge'*. The hinge is not a third category alongside or between 'the social' and 'the technical' in STS, but rather a term that sits behind these concepts while being a *condition of possibility* for the entire dualism. Hinge terms are hidden in plain sight by the operation of the dualism, in fact they need to be suppressed in order for the two concepts to function within the dualism. Regardless of this suppression, the hinge term operates within 'the social' and 'the technical' concepts of the sociotechnical duality. This reading practice involves the identification of a hinge term, a concept that *underlies* both 'the social' and 'the technical', while actually belonging fully to neither. Without the hinge, these concepts cannot relate to each other at all.

The hinge term selected here is *communication*, drawn from linguistic and semiotic traditions. If the correct theorisation can be identified for 'communication', then it can serve as a metalanguage that *dissolves the sociotechnical dualism*. The justification for this hinge term from the technical side of the dualism is relatively straightforward. Systems, software, infrastructure and specifications are always rendered into some form of inscription. No subcategories of 'the technical' are purely asemantic because foundationally they are, or are the result of, communicative acts. The technical requires 'communication' as the condition of production and realisation. From 'the social' side of the dualism, this hinge term is even less problematic. Institutions, oganisations, roles and practices are also constituted in and by communication. Organisations, coalitions and their social structures do not exist independently of the communicative practices that enact, produce and reproduce them. An organisation is not a thing that communicates, so much as it is constituted in and by its various communicative texts and generic conventions in every instance of spoken and written communication (documents, meetings) including those concerning 'the technical'. All aspects of organisation are communicative, therefore (human) communication sits *beneath* both sociotechnical concepts. Communication as a hinge does not act to privilege either side of the dualism, but instead acts to dissolve the informational metalanguage in which this 'violent hierarchy' was articulated in the first instance.

Derrida's reading tactics reveal the hidden hierarchy that subsumes 'the social' into 'the technical' within the IS disciplinary appropriation of the sociotechnical. What is suggested here is that it is possible for an explicit communication theory to contain within it, a subsumed technical category. The aim of this paper is to use the insights provided by reading tactics to

identify the *design criteria* required of a suitable communication theory that can *supercede 'the sociotechnical'* in IS.

5. Ascendency of 'the Social'

The two most embarrassing category errors in STS are that of (i) *workpractices*- the routines, procedures, tacit knowledge, informal norms, and improvisational repertoires developed and used by workers, and (ii) *work organisation* that includes how workers organise their collective labour, what routines they develop, what informal hierarchies emerge, and how they dynamically adapt an existing work procedure to a new circumstance. STS curiously assigned both of these to 'the technical' subsystem that not only includes equipment, machines, tools and technology but also coordination, control and boundary management, despite the fact that these are obviously social. STS smuggles something irreducibly human into the technical category when it classifies workpractices and work organisation as being technical.

5.1. Subcategory Impoverishment by Category Reassignment

Related to the sociotechnical dualism's categorical impoverishment (see subsection 2.2), it is useful to notice the changes to meaning of the 'workpractice' and 'work organisation' subcategories when they are reassigned from 'the social' to 'the technical'. In order for these to be reclassified as 'technical', they must transform into something else that fundamentally alters their nature. Instead of being socially constituted and enacted, a *workpractice* considered 'technically' becomes asemantic, disconnected from its social contexts, as well as its tacit knowledges, its informal norms and the historically accreted improvisational reportoires. Similarly, *work organisation* is also transformed; the historical, political and semiotic aspects are ignored entirely and the social dynamics are reduced to static, stereotyped roles that are separated from the performance of workpractices.

These subcategories are diminished by assigning them to 'the technical' category of the sociotechnical dualism. They necessarily become asocial (individual roles), acultural, acontextual and asemantic. This is the 'violence' metred out to the subcategories of workpractices and work organisations because of the existence of the dominant 'technical' category in the sociotechnical dualism. But the idea of leaving these subcategories within 'the social' is also unthinkable because machine execution is assumed to be a unique property of 'the technical'. This is after all commonsense (Belsey 1980). Considering the received social theories within the IS discipline, how would it be possible for them to be expressive enough to allow any kind of computability? Computability seems to always assume the necessity of information and therefore 'the technical'. There does not seem to have been any attempts at investigating alternative meta-representations outside of the informational. There also seems to have been no attempts at identifying social theories that might be capable of supporting some kind of machine execution. In part, this is addressed by the particular selections of social theories that have entered the discipline (discussed elsewhere). If we have not searched for alternative metarepresentations and 'social' theories that might support some social form of machine execution, then we have not adequately explored the theoretical and practical possibilities for our discipline.

5.2. Aftermath of Reading Tactics applied to the Sociotechnical Dualism

The implications of applying Derrida's reading tactics are that the *'sociotechnical dualism' collapses entirely*. From a Derridean reading, work as the product of social organisation, should be first separated from 'the technical' and then the latter category should be folded into the former. What was considered 'technical' and somehow distinct from 'the social', dissolves into mere technocratic ideology. What the sociotechnical relation occluded was what was

behind it all along, ‘communication’ serves as a hinge term that accounts for both ‘the social’ and ‘the technical’. Identifying communication as the hinge term produces a very different hierarchy; one where ‘the technical’ is subsumed within an explicit, dominant social hierarchy. What is left is the possibility, that for the most part, *only* ‘the social’ may be required in order to understand ‘the technical’.

The effort of deriving an appropriate hinge term places hard *disciplinary criteria* for the selection of a theory or theoretical framework that provides the kind of social analysis from which a technical, executable artifact could be derived. What is required is a metarepresentation that has at its foundation (human) communication theory. The actual choice of theorisation has to comply with the following constraints: it must be (1) cultural, (2) social, (3) semantic, (4) contextual and (5) communicative. If a communication theory could be identified that conforms to these constraints, it offers the possibility of providing a new metalanguage (rather than information theory) for the IS discipline.

6. Conclusions and Ramifications

This section summarises the broad conclusions of a general philosophical examination of the sociotechnical dualism along with the application of Derridean reading tactics. If we take into account the consequences of this approach, we can see that potentially the location of IS might shift institutionally, its disciplinary identity would certainly transform along with its professional practices.

6.1. Design Consequences- Fallacies of Optimisation and ‘Joint’ Optimisation

The argument laid out in sections 2 and especially 3 and 4, not only critiques the core categories and the relation between them, but also how these concepts interrelate in a design practice through the mechanism of ‘joint optimisation’. Within the IS literature, there have been sharp critiques of the notions of ‘optimisation’ and ‘joint optimisation’. Reading tactics applied to the sociotechnical relation view optimisations of a ‘social’ or ‘technical’ kind as certainly *not an ‘optimisation’ of any kind*, but rather a *selection* under the guise of a procedure and operating strictly within the prevailing discursive formation rather than through any inherent analysis of the system. The *selection is not ‘joint’* in any sense either but rather made by strategic actors (professionals, managers, consultants) from within a given power/knowledge regime (Foucault 1980) while those who are involved in working these systems are probably excluded from helping determine there design. So the notions of ‘optimisation’ and ‘joint optimisaton’ in STS theory, don’t even help developers in a design capacity!

6.2. Institutional Locations, Disciplinary Identity, and Professional Practices

If a truly *social approach to IS* was to be developed then what would be the ramifications for the discipline? It is unlikely that the IS discipline would be institutionally located within an Engineering/Computing Science Faculty as the context of the social theories that might form its basis would be more at home within a Business Faculty, possibly as a technical Management, Operations or Supply Chain discipline. It would be unlikely that IS could independently exist within a Humanities or Social Science faculty unless within some kind of Design Studies group, despite the fact that the reference disciplines for a newly cast IS discipline are probably to be sourced from these Faculties. It is perhaps more reasonable to locate such a redefined IS discipline within a cross- or trans-disciplinary institute beyond the concerns of siloed University Faculties.

What would become of IS scholars? We would be what we are still, a multi- and cross-diciplinary design field. But our professional practices would change a great deal. IS would be

deeper into understanding why particular patterns of human communication exist, and in determining how well they serve the purposes of those who are engaged in performing them day-to-day. As practitioners we would be undertaking embedded scholarship of work language, identifying patterns of communication, helping to clarify existing ones and possibly co-developing new ones. Abandoning the sociotechnical dualism requires a fundamental reconception of how IS is constituted as a discipline.

7. References

Ackoff, R. L., (2010), Systems Thinking for Curious Managers Devon Triarchy Press

Ackoff, R. L., Emery, F. E., (1972), On Purposeful Systems Trowbridge, Wioltshire, Redwood Press Ltd.

Banta, M., (1993), Taylored Lives. Narrative Productions in the Age of Taylor, Veblen, and Ford Chicago, The University of Chicago Press

Belsey, C., (1980), Critical Practice New Accents, London: Methuen

Bostrom, R. P., Heinen, J. S., (1977a), "MIS problems and failures: A socio-technical perspective. Part I:The causes." MIS Quarterly, 1 (3), 17-32

Bostrom, R. P., Heinen, J. S., (1977b), "MIS problems and failures: A socio-technical perspective. Part II:The application of socio-technical theory". MIS Quarterly, 1 (4), 11-28.

Checkland, P. B., (1981), Systems Thinking, Systems Practice Chichester:Wiley

Cherns, A., (1976), "The principles of sociotechnical design" Human Relations, 29 (8), 783-792

Cherns, A., (1987), "Principles of sociotechnical design revisited" Human Relations, 40- (3), 153-162

Derrida, J., (1967/1979), Of Grammatology Trans. G. C. Spivak, Baltimore: Johns Hopkins University Press, Chapter 2, The Hinge (La Brisure), pp. 65-73

Derrida, J., (1972/1981), Dissemination Trans. B. Johnson, Chicago: University of Chicago Press

Derrida, J., (1981), Positions Trans. A. Bass, University of Chicago Press

Doray, B., (1988), From Taylorism to Fordism. A Rational Madness London: Free Association Press

Emery, F. E., (1959), Characteristics of socio-technical systems London:Tavistock Instiutute of Human Relations, Doc. 527, 1959

Foucault, M., (1980), Power/Knowledge: Selected Interviews and Other Writings 1972-1977 New York: Pantheon

Henriques, J., Hollway, W., Urwin, C., Venn, C., Walkerdine, V., (1984), Changing the Subject. Psychology, social regulation and subjectivity Cambridge: University Press

Kelly, J. E., (1978), "A Reappraisal of Sociotechnical Systems Theory" Human Relations 31 (12), 1069-1099

Langefors, B., (1978), "Discussion in the article by Bostrom and Keinen." MIS Quarterly 2 (2),55-62

Lewin, K., (1946), "Action Research and Minority Problems" Journal of Social Issues, 2(4), 34-46.

Lewin, K.. (1947), "Frontiers in Group Dynamics: Concept, Method and Reality in Social Science; Social Equilibria and Social Change." Human Relations, 1(1), 5-41.

Mumford, E., (1995), Effective systems design and requirements analysis: The ETHICS approach. London: Macmillan

Pasmore, W. A., Francis, C., Haldeman, J., Shani, A., (1982), “Sociotechnical systems: A North American reflection on empirical studies of the seventies” Human Relations 35 (12), 1179-1204

Taylor, F. W., (1914), The Principles of Scientific Management New York and London: Harper

Trist, E. L., Bamforth, K. W., (1951), “Some social and psychological consequences of the Longwall Method of coal-getting: An examination of the psychological situation and defences of a work group in relation to the social structure and technological content of the work system” Human Relations, 4 (1), 3-38

Trist, E. L., (1981), “The evolution of socio-technical systems: A conceptual framework and an action research program” Occasional Paper No. 2, Ontario Ministry of Labour, Ontario Quality of Working Life Centre

Trist, E. L., Higgin, G. W., Murray, H., Pollock, A. B., (1963), Organizational choice London: Tavistock Publications

Governing Securitized Digital Innovation Ecosystems with Dual-Use Steering Logic

Satu Vasamo-Koskinen
The Prime Minister's Office, Finnish Government, Helsinki, Finland
Tapani Rinta-Kahila
Hanken School of Economics, Helsinki, Finland

Abstract

This study aims to examine how geopolitical tensions reshape digital innovation ecosystem governance. Drawing on securitization theory, we examine how the dual-use potential of technology, i.e., its potential for both civilian and military use, shapes digital innovation ecosystem governance, influencing technology selection, funding allocation, and ecosystem participation. We introduce securitized digital innovation ecosystems as socio-technical arrangements where dual-use steering logic integrates strategic and security-oriented priorities alongside value creation. Empirically, we propose a qualitative case study of Finland based on document analysis of policy instruments and ecosystem initiatives, focusing on key actors such as Finnish Industry Investment, Business Finland, and Defence Innovation Network Finland (DEFINE). Our intended contribution lies in explaining how security-oriented priorities reshape ecosystem governance, providing a novel mechanism-based perspective on dual-use steering logic in digital innovation ecosystems.

1. Introduction

The geopolitical environment under which digital innovation ecosystems are governed has changed. The fragmentation of the world order and increasing geopolitical conflict have strengthened the role of security concerns in political and economic decision-making (Græger, 2024, pp. 1–2; Richmond & Pogodda, 2025, pp. 2–3). Russia's invasion of Ukraine in 2022 has been a significant turning point, accelerating security-related decision-making and expanding the scope of security concerns into new policy domains, including innovation and technology policy (Græger, 2024, p. 2). Such a process where an issue is elevated beyond normal political procedures, thereby legitimizing the use of exceptional measures, is referred to as securitization (Buzan et al., 1998, pp. 23–25). Through the securitization process, political actors may elevate specific domains into the realm of security, thereby legitimizing increased intervention and prioritization. It is argued that "innovation has become a geostrategic factor shaping the international security environment and the global balance of power" (Haddad et al., 2024, p. 742).

Dual-use technologies are central to this shift. In the European Union (EU) regulation, "dual-use items" are defined as "items, including software and technology, which can be used for both civil and military purposes" (Regulation (EU) 2021/821, Article 2(1)). This regulatory definition establishes dual-use as the potential for use across civilian and military domains. In this study, we refer to the dual-use potential of innovation as its potential for both civilian and military use.

As geopolitical tensions make dual-use potential increasingly relevant to how digital innovation is steered and prioritized, it increasingly influences digital innovation ecosystem governance. However, digital innovation and security priorities can be in stark contrast: digital innovation ecosystems rely on openness, actor autonomy, and participation across organizational boundaries (Yoo et al., 2012; Jacobides et al., 2018), while increasing security priorities may

introduce screening, restriction, and stronger strategic direction (Bauerle Danzman & Meunier, 2024; Mügge, 2023). This contradiction represents an unresolved theoretical tension: how might digital innovation ecosystem governance accommodate these potentially competing demands as the dual-use potential of innovation increasingly becomes relevant to innovation steering.

This tension builds on established IS research on digital innovation and coordination in digital ecosystems. Recent IS research has called for better understanding of coordination in distributed socio-technical systems and digital ecosystems (Yoo et al., 2024), as well as the broader societal and system-level consequences of digital innovation, whether intended or unintended ones (Rinta-Kahila et al., 2022). Furthermore, the entanglement between technology companies and the government has raised concerns over increased citizen surveillance and constraining of civic freedoms (Zuboff, 2019).

Against this backdrop, our study aims to answer the following question: How does the increasing relevance of dual-use potential as a steering principle reshape digital innovation ecosystem governance in the context of increasing securitization? In response, **we develop the concept of dual-use steering logic to describe a potential governance and coordination mechanism** through which the dual-use potential of innovation becomes a consideration in the steering of digital innovation ecosystem governance. Empirically, dual-use steering logic may become visible through technology selection (prioritizing technologies with both civilian and military potential), funding eligibility and allocation (funding criteria that take into account dual-use potential), programme priorities (programmes or calls explicitly targeting dual-use innovation), and ecosystem participation (the inclusion of security and defence actors or restrictions on participation), which we examine as key dimensions of our empirical analysis.

2. Theoretical Background

To develop the foundation for analyzing how digital innovation ecosystems are reconfigured under increasing security pressures, we draw on IS literature to conceptualize digital innovation ecosystems as socio-technical coordination arrangements characterized by distributed agency and interdependent actors and noting the societal challenges arising. We then discuss the impact of geopolitics on the transformation of digital innovation governance, introducing securitization as a contextual condition that increasingly shapes innovation environments and discussing how security and innovation have become increasingly intertwined across political rationalities.

Adner (2017) argues that the ecosystem is defined by the alignment structure of the multilateral set of partners that need to interact in order for a focal value proposition to materialize and that the elements of structure consist of activities, actors, positions and links. Hein et al. (2020, p. 87) argue that whereas traditional firms create value within a company or a supply chain, digital platforms utilize an ecosystem of autonomous agents to co-create value. This shift toward ecosystem-based value creation reflects the growing interdependence of multiple actors in the innovation process, where the success of a focal innovation depends on the contributions of a broader set of partners (Adner & Kapoor, 2010, p. 306). This shift is central to information systems research because digital technologies have changed not only what is being innovated, but also how innovation is organized, governed, and coordinated. Rapid and pervasive digitization has challenged established assumptions about the boundaries of innovation, the agency of innovation, and the relationship between innovation processes and outcomes (Nambisan et al., 2017, p. 223). This view builds on Yoo et al.'s (2012) theorization of innovation in the digitized world and the rise to three central features of digital innovation: the importance of digital technology platforms, the emergence of distributed innovation, and the prevalence of combinatorial innovation. Complementing this ecosystem perspective, coevolution of platform architecture, governance, and environmental dynamics and the

dependence on a delicate balance between the autonomy of complementors and the control exercised by platform owners (Tiwana et al., 2010, p. 675-676) shape innovation platforms and the complexity is stressed by their intertwinement with institutions, markets, and technologies (de Reuver et al., 2018, p. 124). The ecosystem literature further argues that ecosystems are characterized by multilateral interdependence and complementarities, which require coordination beyond traditional market or hierarchical mechanisms (Jacobides et al., 2018, p. 2255). This has constituted for recent work highlighting the need to better understand coordination in digital ecosystems, including interfaces, modularity, and the consequences of design choices (Yoo et al., 2024, p. 1513).

Furthermore, Kattel and Mazzucato argue that innovation policy increasingly aims not only to fix market failures but to actively shape markets and direct innovation toward collectively defined goals (Kattel & Mazzucato, 2018, pp. 787–788). Similarly, Mazzucato emphasizes that innovation has not only a rate but also a direction, requiring deliberate orientation toward societal challenges (Mazzucato, 2018, pp. 803–804).

At the same time, prior research cautions that transformations in innovation governance may produce unintended and potentially harmful consequences, highlighting that digital technologies can produce "negative unintended consequences, such as widespread discrimination, which in extreme cases can be destructive to society" (Rinta-Kahila et al., 2022, p. 313). Similarly, digital innovation more broadly has been shown to generate both significant benefits and profound harms, which often reflect deeper systemic issues rather than isolated unintended consequences (Ciriello et al., 2025). In addition, emerging AI-enabled systems are frequently promoted through narratives of efficiency and agility, while their implications for fairness, transparency, and equity remain insufficiently examined (Morrill et al., 2026).

Recent geopolitical tensions have increasingly shaped decision-making in digital innovation ecosystem governance, as "heightened geopolitical tensions and the securitization of economic exchange have pushed the European Union to develop new tools that combine trade, investment, and security considerations" (Bauerle Danzman & Meunier, 2024, p. 1097). To respond to the situation, the need to access existing and emerging commercial technology and capacity, more quickly, has been voiced (McSorley et al., 2025, p. 2). In addition, it is argued that innovation is increasingly guided toward addressing societal challenges rather than pursued solely for economic growth (Haddad et al., 2022, p. 15). This shift is also visible closer to the regulation of innovation itself, where "securitization of digital technology imposes security provision – a traditional state rationale – on regulatory domains hitherto dominated by commercial innovations" (Mügge, 2023, p. 1431). In this context, security and innovation have become increasingly intertwined across political rationalities, imagination, and material practices, evolving to a level in which innovation is aligned with responses to societal threats and insecurities and embedded within evolving security logics (Haddad et al., 2024, pp. 742, 745). De Goede further introduces the concept of platform security to describe forms of security power that operate through facilitation, connection, and decentralized coordination rather than centralized control, arguing that security organizations increasingly resemble digital platforms by connecting actors, facilitating data flows, and acting as intermediaries rather than hierarchical authorities (de Goede, 2026, pp. 247–248).

Taken together, these streams of research establish a coherent but only partial understanding of contemporary digital innovation in a context increasingly shaped by geopolitical tensions. Specifically, we lack mechanism-based understanding of how security-oriented governance priorities become embedded in digital innovation ecosystem governance. Increasing geopolitical conflict has intensified policy attention to security and technological sovereignty (Græger, 2024, pp. 1–2). These policy shifts reshape digital innovation ecosystem governance by influencing which technologies and actors are prioritized. We propose dual-use steering logic as a potential mechanism through which the dual-use potential of innovation shapes

digital innovation ecosystem governance. This perspective positions securitized digital innovation ecosystems as both a highly current empirical phenomenon and a potentially novel theoretical construct. They enable the mobilization of innovation for resilience and security, while simultaneously introducing new tensions related to openness, accountability, and the boundary between civilian and military innovation. Understanding the related mechanisms is therefore essential.

3. Methodology

This study adopts a qualitative case study approach (Yin, 2018) to examine how digital innovation ecosystems become intertwined with security-oriented priorities. We select Finland's increasing securitization of digital innovation ecosystem governance as a critical and revelatory case: a highly digitalized, high-trust society with strong institutions built on sustained R&D investment and close public–private collaboration, in which innovation and security priorities are rapidly aligning.

This study will provide an analysis of the Finnish Government's recent policy orientations that shape legislation and governance of digital innovation ecosystems. In this context, we focus on three relevant institutional actors: Finnish Industry Investment (Tesi), Business Finland, and Defence Innovation Network Finland (DEFINE). Tesi is a state-owned investment company, while Business Finland is a public organization responsible for innovation funding, trade, investment promotion, and internationalization. Finally, DEFINE is a newly emerged, local innovation hub focused on innovations with dual-use potential. The hub is situated strategically in the city of Riihimäki, where a NATO Deployable CIS Module (DCM) will also be established in 2027, to provide communication and information system services in support of NATO forces and commands.

The data will consist of publicly available official policy documents, government reports, and legislative and regulatory initiatives related to innovation, security, and dual-use technologies in Finland. These include materials produced by the Finnish Government, relevant ministries, and public organizations such as Finnish Industry Investment (Tesi) and Business Finland. In addition, we consult other publicly available materials such as press releases, speeches, and media articles to capture how innovation and security are discussed and framed in public discourse. The data will be selected based on its relevance to innovation policy, funding, and security-related developments, with a focus on recent policy changes and initiatives, enabling the study to examine both formal governance structures and broader public discourse. Furthermore, we will collect semi-structured interviews with key stakeholders at both the strategic and operational levels of digital innovation ecosystem governance. Given the sensitivity of dual-use and security-related activities, access and disclosure may be limited and interview evidence will therefore be assessed alongside public documents and organizational materials. This will enable a deeper assessment of the relevance and empirical grounding of the dual-use steering logic in securitized digital innovation ecosystems introduced in this study.

Building on established qualitative IS approaches (e.g., Gioia et al., 2013), data analysis will proceed iteratively between data and theory. We begin with open coding (Strauss and Corbin, 1998) to capture empirically grounded insights, followed by theory-informed coding where emerging concepts are interpreted through representation theory and innovation-ecosystem related IS perspectives. First-order codes are aggregated into higher-order themes, enabling a systematic move from descriptive observations toward theoretically meaningful constructs. This abductive process involves continuous comparison and refinement while remaining sensitive to the socio-technical and institutional context. Our analysis pays particular attention to the recent policy-driven reorientation of financing towards innovations with dual-use potential.

4. Intended Contributions

Building on prior work on the novel coordination logics enabled by innovation ecosystems that span private and public domains (e.g., surveillance capitalism, Zuboff, 2019) as well as the unintended societal consequences of government technology use (e.g., Masiero & Arvidsson, 2021; Rinta-Kahila et al,. 2022), we hope to expand new frontiers for the information systems scholarship. To this end, based on our literature review, introduced research gap, and research setting on the increasing alignment of innovation and security priorities in contemporary digital innovation ecosystems in Finland, we seek to introduce the concept of securitized digital innovation ecosystems. In this conceptual contribution we aim to unveil how security-oriented transformation becomes embedded in the governance of socio-technical digital innovation ecosystems and how dual-use steering logic may operate as a coordination mechanism in this transformation. The concept demonstrates that digital innovation ecosystems do not function only as infrastructures for value creation, but also as infrastructures for strategic steering, when economic and security objectives become intertwined. In addition, the concept shifts the dual-use discussion from the properties of technologies toward a mechanism-level analysis of how innovations are selected, prioritized, funded, and governed. Finally, as this study makes visible how elements of digital innovation ecosystems, such as funding instruments and governance structures together shape the direction of innovation, it provides decision-makers and organizations with tools to better understand how digital innovation ecosystem governance may be steered in contexts where security and economic rationales increasingly overlap.

5. References

Adner, R. (2017). Ecosystem as structure: An actionable construct for strategy. Journal of Management, 43(1), 39–58. https://doi.org/10.1177/0149206316678451

Adner, R., & Kapoor, R. (2010). Value creation in innovation ecosystems: How the structure of technological interdependence affects firm performance in new technology generations. Strategic Management Journal, 31(3), 306–333. https://doi.org/10.1002/smj.821

Bauerle Danzman, S., & Meunier, S. (2024). The EU's geoeconomic turn: From policy laggard to institutional innovator. Journal of Common Market Studies, 62(4), 1097–1115. https://doi.org/10.1111/jcms.13599

Buzan, B., Wæver, O., & de Wilde, J. (1998). Security: A new framework for analysis. Lynne Rienner Publishers.

Ciriello, R. F., Mathiassen, L., Risius, M., Cheong, M., Scheepers, H., & Vaast, E. (2025). Compassionate digital innovation: A pluralistic perspective and research agenda. Information Systems Journal. https://doi.org/10.1111/isj.70003

de Goede, M. (2026). Imaginaries and infrastructures of platform security. European Journal of International Security, 11, 247–261. https://doi.org/10.1017/eis.2025.14

de Reuver, M., Sørensen, C., & Basole, R. C. (2018). The digital platform: A research agenda. Journal of Information Technology, 33(2), 124–135. https://doi.org/10.1057/s41265-016-0033-3

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. https://doi.org/10.1177/1094428112452151

Græger, N. (2024). "High politics" and everyday practices at the border: Emerging communities of practice in the High North. Global Studies Quarterly, 4, ksad075. https://doi.org/10.1093/isagsq/ksad075

Haddad, C. R., Nakić, V., Bergek, A., & Hellsmark, H. (2022). Transformative innovation policy: A systematic review. Environmental Innovation and Societal Transitions, 43, 14–40. https://doi.org/10.1016/j.eist.2022.03.002

Haddad, C., Vorlíček, D., & Klimburg-Witjes, N. (2024). The security-innovation nexus in (geo-)political imagination. Geopolitics, 29(3), 741–764. https://doi.org/10.1080/14650045.2024.2329940

Hein, A., Schreieck, M., Riasanow, T., Soto Setzke, D., Wiesche, M., Böhm, M., & Krcmar, H. (2020). Digital platform ecosystems. Electronic Markets, 30(1), 87–98. https://doi.org/10.1007/s12525-019-00377-4

Jacobides, M. G., Cennamo, C., & Gawer, A. (2018). Towards a theory of ecosystems. Strategic Management Journal, 39(8), 2255–2276. https://doi.org/10.1002/smj.2904

Kattel, R., & Mazzucato, M. (2018). Mission-oriented innovation policy and dynamic capabilities in the public sector. Industrial and Corporate Change, 27(5), 787–801. https://doi.org/10.1093/icc/dty032

Masiero, S., & Arvidsson, V. (2021). Degenerative outcomes of digital identity platforms for development. Information Systems Journal, 31(6), 903-928. https://doi.org/10.1111/isj.12351

Mazzucato, M. (2018). Mission-oriented innovation policies: Challenges and opportunities. Industrial and Corporate Change, 27(5), 803–815. https://doi.org/10.1093/icc/dty034

McSorley, T. D., Macenowicz, M., Maddison, M., & Yukins, C. (2025). Allies bridging the valley of death: How NATO's Defence Innovation Accelerator for the North Atlantic will help maintain NATO's technological edge. The Government Contractor, 67(1).

Morrill, J., Asadi Someh, I., Rinta-Kahila, T., Jooss, S., Tona, O., & van der Meulen, N. (2026). Reframing talent management with artificial intelligence: The organizing vision of skills intelligence. Proceedings of the 59th Hawaii International Conference on System Sciences. https://doi.org/10.24251/HICSS.2026.688

Mügge, D. (2023). The securitization of the EU's digital tech regulation. Journal of European Public Policy, 30(7), 1431–1446. https://doi.org/10.1080/13501763.2023.2171090

Nambisan, S., Lyytinen, K., Majchrzak, A., & Song, M. (2017). Digital innovation management: Reinventing innovation management research in a digital world. MIS Quarterly, 41(1), 223–238. https://doi.org/10.25300/MISQ/2017/41:1.03

Regulation (EU) 2021/821 of the European Parliament and of the Council. (2021). Dual-use items regulation. https://eur-lex.europa.eu/eli/reg/2021/821/oj/eng

Richmond, O. P., & Pogodda, S. (2025). Peacemaking and the maintenance of international order: Alignment under hegemony versus multipolar misalignment. Global Studies Quarterly, 5, ksae094. https://doi.org/10.1093/isagsq/ksae094

Rinta-Kahila, T., Someh, I., Gillespie, N., Indulska, M., & Gregor, S. (2022). Algorithmic decision-making and system destructiveness: A case of automatic debt recovery. European Journal of Information Systems, 31(3), 313–338. https://doi.org/10.1080/0960085X.2021.1960905

Strauss, A., and Corbin, J. (1998). Basics of qualitative research techniques: Techniques and procedures for developing grounded theory (2nd ed.). Sage.

Tiwana, A., Konsynski, B., & Bush, A. A. (2010). Platform evolution: Coevolution of platform architecture, governance, and environmental dynamics. Information Systems Research, 21(4), 675–687. https://doi.org/10.1287/isre.1100.0323

Yin, R. K. (2018). Case study research and applications: Design and methods (6th ed.). SAGE Publications.

Yoo, Y., Boland, R. J., Lyytinen, K., & Majchrzak, A. (2012). Organizing for innovation in the digitized world. Organization Science, 23(5), 1398–1408. https://doi.org/10.1287/orsc.1120.0771

Yoo, Y., Henfridsson, O., Kallinikos, J., Gregory, R., Burtch, G., Chatterjee, S., & Sarker, S. (2024). The next frontiers of digital innovation research. Information Systems Research, 35(4), 1507–1523. https://doi.org/10.1287/isre.2024.editorial.v35.n4

Zuboff, S. (2019). The age of surveillance capitalism: The fight for a human future at the new frontier of power. PublicAffairs.

Metadata-Supported Data Quality Evaluation in Data Repurposing: An Information Resilience Perspective

Hui Zhou[1], Tianwa Chen[2], Gianluca Demartini[1], Marta Indulska[3], Shazia Sadiq[1]

1 School of Electrical Engineering and Computer Science, The University of Queensland

2 School of Science, Technology and Engineering, University of the Sunshine Coast

3 School of Business, The University of Queensland

Abstract

Data repurposing uses existing data for tasks not anticipated when the data were originally collected, often separating users from the contextual knowledge needed to judge fitness for a new task. This paper examines metadata-supported data quality evaluation as an information resilience mechanism under that contextual discontinuity. The model is grounded in practitioner interviews, process-mining interviews, two controlled behavioural studies, eye-tracking, and cued retrospective think-aloud data. Greater metadata attention was associated with stronger evaluation performance, but qualitative evidence shows that attention alone was insufficient: users also needed to understand how a cue provided evidence about the quality issue. We develop a process model in which users orient to unfamiliar artefacts, reconstruct context, interpret quality signals, and form fitness-for-use judgements, with feedback when evidence remains insufficient. Resilience is not equated with successful reuse or increased confidence; cautious use, further verification, or rejection may also be resilient outcomes.

1. Introduction

Information systems increasingly depend on data that travel beyond the settings in which they were originally produced. Data may be collected for an operational, administrative, scientific, or commercial purpose and later considered for a different analytical or decision-making task. Parsons et al. (2026) distinguish data repurposing from original use and ordinary reuse by focusing on existing data being used for tasks that were not anticipated when the data were collected. Repurposing can therefore create a particular contextual discontinuity: the new user may not share the original producer's knowledge of why the data were collected, what variables

mean, which transformations were applied, what limitations were known, or which assumptions shaped the data.

This does not mean that contextual data quality is new. Fitness-for-use research has long recognised that quality depends on the user, task, and context (Wang and Strong 1996; Pipino et al. 2002). The problem addressed here is narrower: how do users form a fitness-for-use judgement when the contextual knowledge needed for that judgement is only partly available because the data are being used for an unanticipated task? Metadata can provide evidence about source, provenance, structure, field meanings, update history, transformations, and known limitations. However, the metadata available to a new user may itself be incomplete, ambiguous, fragmented, inaccessible, outdated, or difficult to verify.

We interpret this situation through information resilience. Sadiq et al. (2022) conceptualise information resilience as the capacity to create, protect, and sustain agile data pipelines able to detect and respond to risks across the information value chain. In a repurposing episode, the relevant disruption is not necessarily a system failure; it is the weakening of contextual continuity between data production and subsequent use. The resilience question is therefore whether users can adapt their evaluation when context is missing and still reach an appropriate decision about use, cautious use, further verification, transformation, or rejection.

This paper asks: How can metadata-supported data quality evaluation contribute to information resilience when data are repurposed beyond the context anticipated at collection? We address this question by revising an empirically grounded process model of metadata-supported evaluation. In response to the workshop reviews, the paper makes the empirical basis of the model more visible, clarifies how the qualitative analysis informs model development, and treats appropriate non-use as a possible resilience outcome rather than assuming that resilience implies greater confidence or successful reuse.

2. Conceptual Framing

Repurposing differs from repeated use for an established purpose because the later task was not anticipated when the data were collected (Parsons et al. 2026). The distance between original production and later use may involve a different task, user community, decision setting, or set of assumptions. These changes matter for quality evaluation because information that was obvious to the original producer may no longer be available to the new user. Event logs, for example, can be extracted from operational information systems and later analysed for process-mining purposes; understanding activity meanings, case definitions, timestamp semantics, and transformation history then becomes necessary for interpretation. Controlled laboratory tasks cannot reproduce an entire organisational history of repurposing, but they can reproduce a key informational condition of repurposing: users evaluating unfamiliar artefacts without access to the full production context.

Metadata can support evaluation by helping users orient to an unfamiliar artefact, reconstruct missing context, and connect observable cues to particular quality concerns. Source, date, schema, title, and summary may support an initial viability judgement, while definitions, provenance, business rules, and transformation records can clarify meaning and limitations. The important qualification is that metadata does not help merely by being present. A cue must be interpretable, relevant to the quality issue, and sufficiently trustworthy for the task.

We therefore use information resilience as an interpretive lens rather than as a synonym for good data quality. The disruption is contextual discontinuity. The mechanism is an adaptive evaluation process in which users locate, interpret, verify, and sometimes supplement metadata. The outcome is an appropriate fitness-for-use decision under uncertainty. That outcome can be positive or negative: metadata may justify reuse, but it may also reveal limitations that reduce confidence, trigger further verification, or support rejection.

3. Empirical Grounding and Model Development

The model draws on a multi-method programme reported across multiple studies (Zhou et al. 2021; Zhou et al. 2022; Zhou et al. 2023; Zhou et al. 2024; Zhou et al. 2025). Practitioner interviews with seven senior data practitioners identified problems caused by missing definitions, fragmented documentation, unclear provenance, and limited access to data owners, together with compensatory practices such as manual profiling, searching for documentation, and contacting data owners. A secondary qualitative study of 41 process-mining participants showed that repurposed event logs require additional semantic and provenance information, including activity meanings, case definitions, business rules, timestamp semantics, and transformation histories.

Two controlled behavioural studies examined how participants used metadata during quality evaluation. In the structured-data setting, 19 usable behavioural cases evaluated missingness, inconsistency, and non-uniqueness. The mean proportion of fixation duration on metadata was 0.621 for the high-performance group and 0.363 for the low-performance group. In the unstructured-text setting, 21 participants evaluated credibility, completeness, relevance, and timeliness; the corresponding metadata-fixation proportions were 0.313 and 0.122. Metadata attention in this setting was also positively associated with confidence (Spearman ρ=.61) and task completion time (ρ=.68). These results show that stronger performance was associated with greater metadata attention, but not simply with faster completion.

The qualitative evidence helps explain what attention did and did not accomplish. For example, in the unstructured data setting, a participant explained, “I rely on metadata as a quick way to assess whether a document is relevant without reading the whole thing.” These examples show both the value and the limits of metadata: it can reduce direct inspection, but only when users can connect a cue to the quality issue being evaluated.

Table 1. Worked evidence-to-model trace

Empirical evidence	Interpretation	Model component	Resilience meaning
Missing definitions, fragmented documentation, unclear provenance; users searched documentation or contacted data owners.	Repurposing can weaken access to production context.	Context reconstruction; feedback when context is insufficient.	Evaluation adapts to contextual discontinuity rather than assuming complete information.
Greater metadata attention was associated with stronger performance, but some relevant cues were viewed without being interpreted correctly.	Attention alone is insufficient; cues need diagnostic value for the task.	Metadata characteristics; quality-signal contextualisation.	Resilience depends on interpretable evidence, not metadata availability alone.
Participants used profiling statistics, summaries, keywords, dates, and source information to reduce direct inspection.	Metadata can focus evaluation on task-relevant evidence.	Orientation and viability checking; evaluative use of metadata.	Users can continue evaluation while managing limited time and information.
When evidence remained unclear, participants rechecked raw data, data dictionaries, or source documents; unsuitable evidence could justify rejection.	Uncertainty can trigger further verification rather than a forced positive judgement.	Feedback mechanism; fitness-for-use judgement.	Cautious use, further verification, or rejection can be resilient outcomes.

The CRTA material from the two behavioural settings was analysed separately using a structure of participant-centred first-order concepts, second-order themes, and aggregate dimensions associated with Gioia et al. (2013). The settings were analysed independently before higher-level aggregate labels were compared. Three recurring dimensions were identified: exploring metadata, using metadata as a cognitive aid, and aligning metadata with data quality dimensions (Zhou et al. 2025). Model development then involved a broader cross-

study synthesis that combined these qualitative findings with practitioner evidence, behavioural results, and relevant theory. This distinction is important because the model is not presented as the direct output of a single Gioia analysis.

The cross-study synthesis looked for recurring evaluation activities and boundary conditions. Practitioner evidence identified the contextual problems that create a need for metadata; behavioural evidence showed where participants allocated visual attention and how they performed; and CRTA explanations showed how participants interpreted, verified, and applied the cues they viewed. Table 1 and Figure 1 illustrate this analytical progression. The four process phases were retained only where they could be linked to recurring evidence across the programme: orientation, context reconstruction, quality-signal contextualisation, and fitness-for-use judgement. The feedback mechanism reflects repeated evidence of users returning to the underlying data, dictionaries, source documents, or other information when the available metadata was insufficient or conflicting.

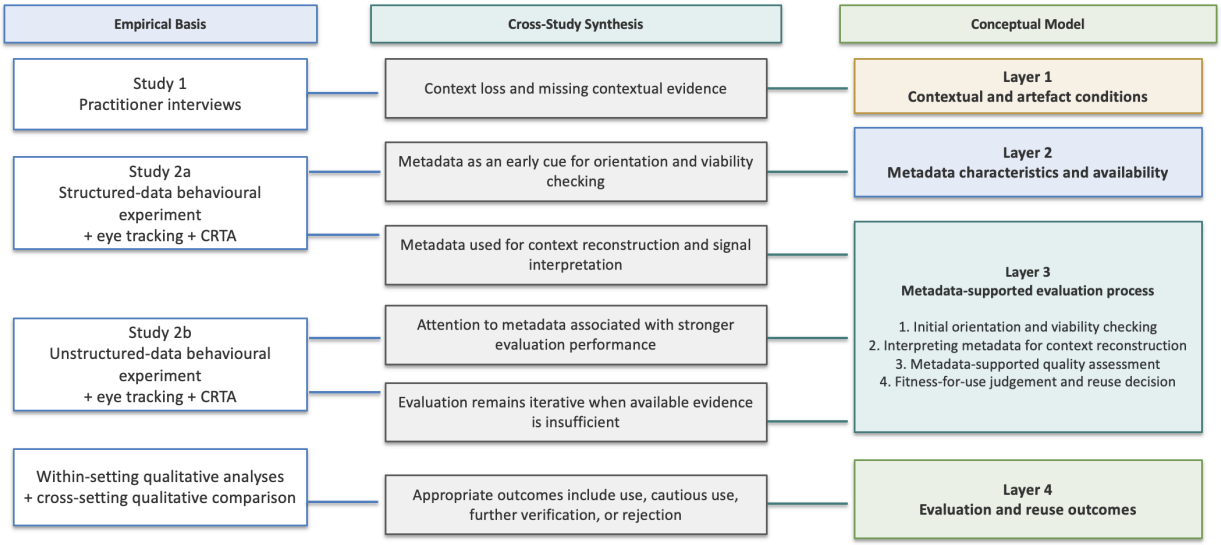


Figure 1. Metadata-supported data quality evaluation as an information resilience mechanism

4. A Metadata-Supported Information Resilience Model

Figure 2 summarises the model. Layer 1 represents the conditions surrounding evaluation; Layer 2 represents the metadata available and characteristics such as adequacy, accessibility, understandability, provenance, task alignment, and diagnostic value; Layer 3 contains the four-phase evaluation process; and Layer 4 represents the fitness-for-use judgement and reuse action. The model is not a fixed workflow: users may stop early when a critical requirement is not met or return to metadata and other evidence when uncertainty remains.

The four phases describe recurring evaluation work rather than mandatory steps. Orientation and viability checking uses readily available cues to decide whether further evaluation is warranted. Context reconstruction uses definitions, business rules, or other evidence to understand the artefact. Quality-signal contextualisation connects an observed signal to the task: a missing value, unusual category, old date, or incomplete document is not automatically a quality failure. Fitness-for-use judgement combines the available evidence into a decision to use, use with caution, transform or clean before use, seek further information, or reject.

A key concept is diagnostic value. The structured experiment showed that participants could attend to metadata without understanding how it provided evidence about a quality issue. Thus, metadata availability and visual attention are necessary but not sufficient conditions for useful support. The cue must also be understood in relation to the quality problem. The model

also treats metadata quality as part of the evaluation: users may verify a statistic against raw data, check a generated summary against the source document, or seek another source when provenance is uncertain.

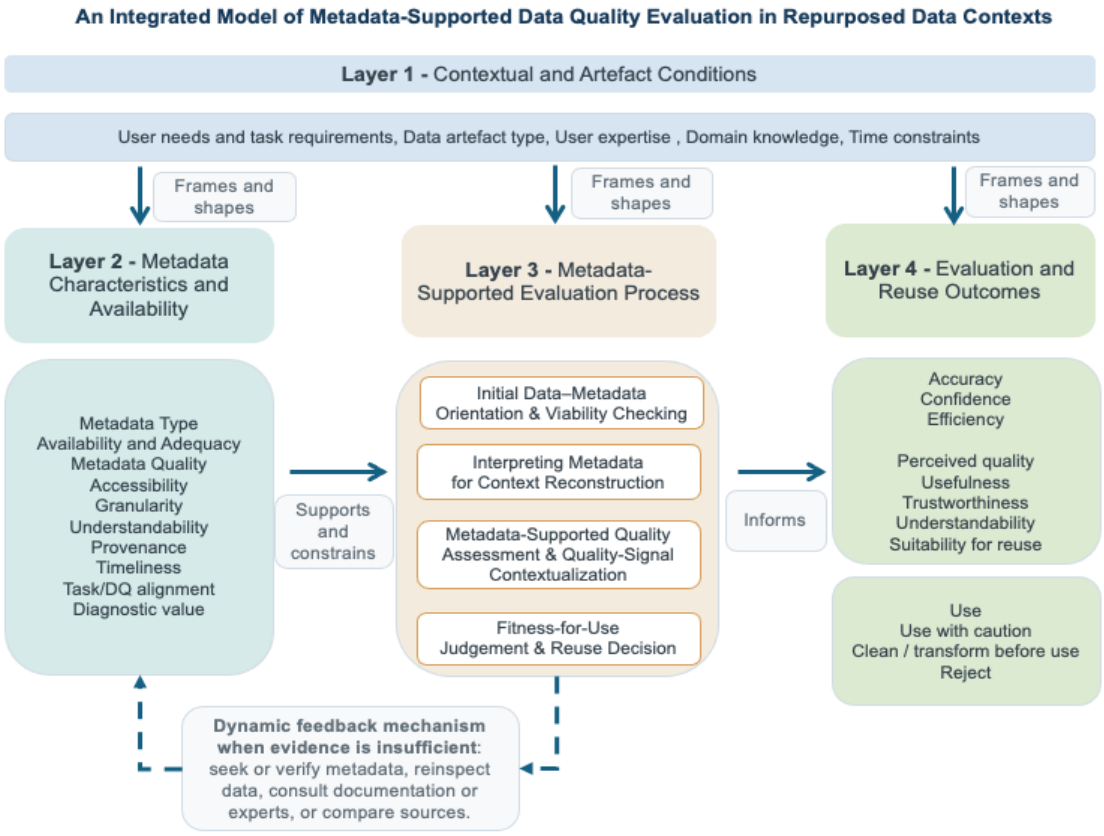


Figure 2. Metadata-supported data quality evaluation as an information resilience mechanism

From an information resilience perspective, the model can be read as a disruption-mechanism-outcome sequence. Repurposing disrupts contextual continuity between data production and later use. Metadata-supported evaluation provides a mechanism for reconstructing and testing sufficient context to make a decision. The resilient outcome is not necessarily greater confidence or successful reuse. Useful metadata may appropriately reduce confidence by revealing a limitation, and rejection can be resilient when it prevents unsuitable data from being used. Resilience therefore concerns the quality of the evaluation response to uncertainty rather than the positivity of its outcome.

5. Discussion and Contributions

The first contribution is to information resilience research. The paper identifies contextual discontinuity in data repurposing as a form of everyday information disruption and specifies a mechanism through which users adapt to it. This is narrower than claiming that metadata is inherently “resilience infrastructure”. Metadata contributes to resilience when it helps users reconstruct or verify enough context to reach an appropriate fitness-for-use decision. The mechanism can also fail when metadata is absent, misleading, opaque, or diagnostically unhelpful.

The second contribution is to data quality and metadata research. Fitness-for-use research already establishes that quality is contextual. The process model adds an account of how users reconstruct and apply context when the later task was not anticipated at collection. It also separates metadata presence from diagnostic value. The empirical contrast between viewing

a cue and understanding what that cue means shows why more metadata, or more attention to metadata, does not automatically produce a better judgement.

The third contribution is a more careful view of confidence and reuse outcomes. Metadata may increase confidence when it resolves uncertainty or reduce confidence when it exposes limitations; both can represent appropriate evaluation. Likewise, further verification, cautious use, or rejection may be resilient responses when evidence is uncertain or unsuitable. The model remains developmental, and future work should test it in longitudinal organisational settings where the original purpose, later purpose, actors, and contextual discontinuities can be observed directly, including settings using AI-generated metadata.

6. Conclusion

Data repurposing creates an evaluation problem when existing data are used for tasks not anticipated at collection and users no longer share the original production context. The evidence synthesised here suggests that metadata can help users orient to unfamiliar artefacts, reconstruct context, interpret quality signals, and form fitness-for-use judgements, but only when relevant cues can be understood and verified. Interpreted through information resilience, the contribution is not that metadata guarantees reuse or confidence. Rather, metadata-supported evaluation can help users respond appropriately to contextual uncertainty, including by seeking further evidence, using data cautiously, or rejecting data that are not suitable for the task.

7. References

Gioia, D. A., Corley, K. G., Hamilton, A. L., (2013), “Seeking Qualitative Rigor in Inductive Research: Notes on the Gioia Methodology”, Organizational Research Methods, Vol. 16, No. 1, 15-31.

Parsons, J., Lukyanenko, R., Greenwood, B. N., Cooper, C. B., (2026), “Understanding and Improving Data Repurposing”, MIS Quarterly, Vol. 50, No. 1, 35-58.

Pipino, L. L., Lee, Y. W., Wang, R. Y., (2002), “Data Quality Assessment”, Communications of the ACM, Vol. 45, No. 4, 211-218.

Sadiq, S., Aryani, A., Demartini, G., Hua, W., Indulska, M., Burton-Jones, A., Khosravi, H., Benavides-Prado, D., Sellis, T., Someh, I., Vaithianathan, R., Wang, S., Zhou, X., (2022), “Information Resilience: The Nexus of Responsible and Agile Approaches to Information Use”, The VLDB Journal, Vol. 31, No. 5, 1059-1084.

Wang, R. Y., Strong, D. M., (1996), “Beyond Accuracy: What Data Quality Means to Data Consumers”, Journal of Management Information Systems, Vol. 12, No. 4, 5-33.

Zhang, R., Sadiq, S., Indulska, M., (2019), “Discovering Data Quality Problems”, Business & Information Systems Engineering, Vol. 61, No. 5, 575-593.

Zhou, H., Demartini, G., Indulska, M., Sadiq, S., (2021), “Evaluating the Quality of Repurposed Data - The Role of Metadata”, Proceedings of the Australasian Conference on Information Systems (ACIS).

Zhou, H., Han, L., Demartini, G., Indulska, M., Sadiq, S., (2022), “A Behavioural Analysis of Metadata Use in Evaluating the Quality of Repurposed Data”, Conceptual Modeling - ER 2022, Lecture Notes in Computer Science, Vol. 13607, Springer.

Zhou, H., Zerbato, F., Weber, B., Indulska, M., Sadiq, S., (2023), “Do Process Analysts Care About the Metadata of Event Logs?”, Proceedings of the Australasian Conference on Information Systems (ACIS).

Zhou, H., Demartini, G., Indulska, M., Sadiq, S., (2024), “Influence of Metadata on Quality Evaluation of Unstructured Information Artefacts”, Proceedings of the Australasian Conference on Information Systems (ACIS).

Zhou, H., Chen, T., Demartini, G., Indulska, M., Sadiq, S., (2025), “We Don’t Have Time to Read the Entire Document to Evaluate the Quality - The Role of Metadata in Quality Evaluation of Unstructured Textual Data”, Proceedings of the 36th Australasian Conference on Information Systems (ACIS).

From Know-How to See-How: Cognitive Out-Dwelling in AI-Assisted Development

John Williams

University of Queensland

Abstract

Generative AI is rapidly reconfiguring software development into "vibe coding" which is a practice of delegating coding implementation to an AI assistant that fundamentally severs code authorship from comprehension. While existing literature frames this as functional cognitive offloading, this paper introduces a novel theoretical framework to explain its unintended costs. Adapting the epistemologies of Ryle and Polanyi, we conceptualize an "Epistemological Reversal" where developers are forced into "cognitive out-dwelling". Instead of internalizing experiential "know-how," practitioners become detached observer-directors trapped in a liminal "see-how" state. Employing a complementary multiple site qualitative design which investigates two separate sites. Firstly, pilot data from a 4-month ethnographic study (Site 1) with over 50,000 timestamped individual interactions with AI coding assistants, alongside an organizational field study (Site 2) capturing pilot observations from team-level practices under delivery deadlines. Using a multi-level unit of analysis, from micro-level socio-technical negotiation episodes to team-level triadic encounters between AI Agents, Developer and Stakeholders/Peers to better understand evolving understanding when using advanced AI tools. Furthermore, we identify four distinct epistemic negotiation moves which we call Deferral, Arbitration, Uptake and Breakdown and demonstrate that the outcome of these moves dictates whether AI assistance erodes, preserves, or builds developer competence. Initial findings reveal instances where developers accept AI outputs without deep understanding of the architecture, creating an illusion of mastery. We conclude that standard "human-in-the-loop" safeguards provide procedural cover rather than epistemic grounding. To preserve competence, AI workflows must introduce "epistemic friction" and "reciprocal validation" to structurally force active engagement through training.

Introduction

Generative AI is changing software development from a practice centred on writing code to one increasingly organized around specifying intent and delegating implementation and evaluation to intelligent machines. Developers now use AI not only to generate code, but also to create documentation, interface components, tests, and other development artifacts (Sarkar, et al., 2025). This shift lowers the effort required to produce plausible and functional outputs and expands participation in software work. The result of this is “actors with only a shallow grasp of software principles now enter workflows” (Sarkar, et al., 2025). Yet it also creates an organizational problem: people may review, approve, maintain, or deploy software whose internal logic they have not constructed and may not fully understand. Adjacent to this, the productivity with AI assisted coding has increased. For

example, developers using GitHub Copilot completed a JavaScript HTTP-server task 55.8% faster than those without it (Peng et al, 2023). However, such specific single task experiments may not fully capture the productivity gains enabled with AI tools:

"We can now simultaneously create 20 Prs a day on parallel tasks with the help of AI" (informal observation from site 2)

These phenomenon concern more than whether AI increases productivity but whether the tools themselves impact understanding. Software development depends on distributed epistemic understanding, in so far as teams must be able to explain why a system behaves as it does. Through evaluating proposed changes, identifying faults, and accepting responsibility for the code they release developers can understand the systems they make. Historically, this authorship provided a workable basis for accountability. Writing or substantially modifying code typically required a developer to engage with its assumptions, dependencies, and trade-offs. AI-assisted development weakens that connection. A developer can now create and merge functioning code without having reasoned through its implementation, making review and sign-off potentially activities of endorsement rather than understanding. The familiar prescription to keep a "human in the loop" therefore does not, by itself, resolve the issue: the relevant question is whether that human has adequate grounds to evaluate the output.
Existing work drawing on IS adoption theory has begun to examine Generative AI adoption among software professionals, finding that compatibility with existing workflows is the dominant driver at this early stage (Russo, 2024). It has also largely focused on adoption and productivity. In parallel, software-engineering and HCI research describes the new practices through which developers prompt, delegate, inspect, verify, and orchestrate AI-generated work, with expertise shifting from producing code toward directing and evaluating it (Meske et al., 2025). A separate literature warns that reliance on AI may weaken conceptual understanding, debugging capability, and ownership of codebases, although direct evidence connecting AI use to skill change remains limited (Macnamara et al., 2024).

The missing link is epistemic. The relevant change is not simply that AI may erode skill; it is that it may alter the basis on which developers not only know the systems for which they are accountable but also how they understand the process of building itself. When software creation no longer warrants comprehension, knowing shifts from having constructed a system to having adequately evaluated it. Prior work describes this decoupling of authorship and understanding (Wheeler, 2026) but has not integrated it with an account of changing development practices and their consequences for competence.

By epistemic, we mean the specific mechanism of how a software developer grasps whether an artifact they co-create with AI is understood. Conventionally, this came about through the practice of writing code and testing it, in other words the specific activity through which they come to know that it works and why. Through the act of authorship understanding is

constructed by writing, tracing, and debugging code (know-how, after Ryle; tacit knowledge, after Polanyi). With AI-assisted development, the developer increasingly treats an artifact as understood by observing its outward behaviour, for example by running it, inspecting outputs, or verifying test passes sometimes without having authored or read the underlying implementation. We define this second basis as see-how. This represents what this paper argues an Epistemological Reversal. This reversal is better seen by contrasting with conventional authorship, which requires understanding to precede the creation of the artifact. With the case of see-how the AI-generated artifact arrives first, and understanding is inferred after creation from its behaviour.

Accordingly, we ask:

RQ1: What epistemic basis do developers rely on to treat AI-generated software as understood, and how does this differ from traditional authorship?

RQ2: How does the influence of AI sustain, and under what conditions does it erode, a practitioner's capacity for independent explanation, troubleshooting, and repair of artifacts created with AI?

Literature review

The integration of Generative AI into software engineering has fundamentally severed two capacities that the field has historically treated as one: the ability to produce working code and the ability to understand it. Recent experimental evidence (Qiao et al., 2025) demonstrates a "comprehension-performance decoupling," where developers utilizing AI coding assistants achieve significant productivity gains, but experience experiential stagnation or degraded code comprehension. Crucially, this erosion is not caused by the AI itself, but by the passive consumption of its outputs; developers who fail to actively verify AI-generated code suffer this the most (Qiao et al., 2025).

This decoupling has triggered what researchers call "perception-action gap" across the industry (Fawzy et al., 2026) where experience determines evaluation of AI generated code. Vibe coding democratizes access to software creation, allowing novices and non-coders to produce complex outputs, but it fails to distribute the necessary expertise required to evaluate those outputs (Fawzy et al., 2026). Because developers lack the deep experiential knowledge to verify the machine's logic, they increasingly substitute genuine understanding with "surface plausibility" (Sabouri et al., 2025). Developers accept code simply because it reads well or appears structurally correct, a proxy that masks underlying errors and drives the silent accumulation of technical debt (Huang et al., 2026).

This crisis of shallow understanding cumulates during collaborative code reviews. During code review the author of the newly developed code submits a ‘pull request’ to fellow developers that are responsible for maintaining and assessing the quality of new code.

Historically, this code review process was an expert discourse that presumed a shared, deep mental model of not only the code itself but of engineering practices, conventions and tacit shared lore of programming. However, studies of AI-generated pull requests reveal that most receive no standalone human review at all (Duma et al., 2026). When humans are kept "in the loop," there is potential that approval provides merely "procedural cover" (Green, 2021). This means that a human must be ‘seen’ to approve or disapprove an artifact in order to accept code into the codebase, however this is not necessarily the same thing as having the required epistemic grounding to effectively evaluate the artifact. Relying on such human oversight in this context leads to what is described as "humanwashing," (Wilson et al., 2026) where cursory examination of AI output is approved under the premise of human in the loop, allowing organizations to ship codebases that no participant genuinely understands.

Theoretical Framework

To understand the severity of this disembodiment, we must first invert Michael Polanyi’s foundational concept of 'indwelling'. Traditionally, achieving deep tacit 'know-how' requires a practitioner to actively embody their practice, experiencing the activity first-hand to internalize the necessary skills. However, the opaque, black-box nature of agentic AI obscures direct participation, relegating the practitioner to the periphery. This forces the developer into a state which we call 'out-dwelling'. Rather than engaging with the messy, lived experience of the work, the practitioner exists outside the task, acting merely as an observer-director of the machine's outputs. This observer-director nature of vibe coding reduces software development to what Gilbert Ryle frames as a 'drill'. The practitioner actively repeats the same generative processes without fully understanding the underlying architectural logic, actively forgoing their cognitive agency to the machine which results in a deep detachment from their craft.

Secondly, building upon Ryle’s classical distinction between experiential 'know-how' (dispositional knowledge) and explicit 'know-that' (propositional knowledge), we conceptualize a novel, liminal mode of understanding: the 'see-how' state. As practitioners out-dwell from the task of software creation, they are neither building dispositional knowledge through direct experiential engagement, nor do they possess the propositional logic, which is now exclusively the domain of the machine. Instead, they are relegated to mere consumers of 'see-how.' Developers understand their high-level intent, what they wish to achieve, but remain entirely detached from the underlying execution required to realize it.

This leads to what we term an “Epistemological Reversal”, a mental shift that occurs during human-AI collaboration, where the probabilistic logic of agentic AI fundamentally overrides and reconfigures the cognitive reality of the human practitioner. Building upon the concept of "Ontological Reversal" (Baskerville et al, 2020) which states digital systems dictate

physical reality. The Epistemological Reversal explains how AI changes not just what practitioners do, but the very mechanisms by which they acquire and maintain knowledge.

If we take a moment to think about this situation it may become apparent that while the practitioner may deeply indwell in the superficial act of conversing with the AI, they fundamentally out-dwell from the act of understanding the generated architecture. This dynamic reduces software engineering to a drill of observational direction where developers merely watch how a machine produces an output without ever acquiring the foundational understanding of how or why it was achieved.

Methodology

To capture the micro level dynamics of AI tool usage and the impact of these episodes on stakeholders, this paper synthesises *pilot* observations from two sites using a qualitative and interpretive design.

Site 1 – Fred80

A longitudinal ethnographic study of the development of Fred-80, a fantasy console, which provides rich AI prompt interaction spanning over four months of high-density, continuous traces from 16th of June to October 2026 encompassing over 50000 time-stamped events. This data captures hash-chained interaction logs, memory rules, and field notes. Site 1 serves as a pilot micro-level site to track how an individual practitioner's interactions with AI evolves over time including how a practitioner's verification rules sequentially evolve in response to system failures.

Site 2 – Double Cone AI

An organizational field study embedded within an enterprise software company's engineering team (Double Cone AI) delivering AI features under strict time constraints using unfamiliar tools. Double Cone captures stakeholder dynamics, triadic code reviews, and complex perspectives on how developers understand tools, deliverables and expectations.

Software engineering initiatives at Double Cone AI provides opportunity of rich pilot observations. The company is rapidly developing hundreds of new AI-driven product features for 'Project Axe', heavily relying on vibe-coding tools. To scale testing alongside this ambitious rollout, developers built an AI-driven test verification tool named 'Guide'. This case was selected as an emerging, longitudinal field setting, allowing for the observation of a focal tool being developed and adapted while the overarching product development that created demand for it continued to evolve (Yin, 2018; Runeson and Höst, 2009).

Historically, software development and testing have required deep, experiential tacit knowledge to accurately evaluate system quality (Itkonen, Mäntylä and Lassenius, 2013). Preliminary site observations indicate that severe time constraints are compelling developers to vibe-code artifacts for an architecture for which they possess little foundational context. This scenario offers an ideal, naturalistic environment to longitudinally observe the phenomenon (Walsham, 1995; Runeson and Höst, 2009), of how vibe-coding impacts software developers' understanding.

Unit of Analysis

Across these sites we define a multi-level unit of analysis spanning two levels. The primary unit of analysis is the individual, at this level we are interested in the complex socio-technical negotiations that take place when developers engage with AI tools when they perform bounded tasks that have clear starting and ending points. These interaction episodes with AI tools such as task delegations, verification sessions and clarification episodes present rich insights into the mechanics of how developers gain understanding over the artifacts they cocreate.

Complementing this we also examine a secondary level of analysis, the “team level”. Here we primarily refer to “team” loosely as a triadic encounter. That means we are interested in how understanding and negotiations with AI propagate across boundaries. The interactions between AI agent, developer and expert stakeholder, provides rich insight into how individual negotiations with AI propagate to expert arbiters and code reviewers. The stakeholder's perspective serves as an essential analytical mirror to the developer's internal epistemic state, exposing the gap between perceived AI productivity and actual code comprehension during code walkthroughs or pull request reviews.

To examine these dynamics, the study adopts an interpretive, longitudinal design using ethnographically informed participant observation. Embedded within the development teams, the researcher captures the daily interactions and workflows surrounding the 'Guide' tool, a sustained presence that reveals not only what practitioners report doing, but what they do in practice (Myers, 1999; Nandhakumar and Jones, 2002). Following an abductive logic, the analysis moves iteratively between these field observations and emerging theoretical concepts. The observational focus centres on episodes of shared understanding, moments where AI-generated code, requirements, and tests are interpreted or challenged by the individual or across roles. By examining these episodes, we seek to understand how interactions play out, how these episodes lead to breakdowns or advancements in understanding. Finally, the study investigates whether AI-assisted development facilitates short-term task completion and whether foundational, system-level understanding impacted.

To establish research feasibility and better understand how best to guide data collection initial pilot observations were conducted on early AI-assisted development workflows (focusing on developer interactions surrounding tools such as 'Guide' and Fred-80). Following an abductive logic, these preliminary observations surfaced early evidence of how prompt-based interactions alter system comprehension, establishing the empirical foundation for the main empirical studies.

Building upon these preliminary insights, the main empirical study will systematically map this epistemological shift over a 12-month fieldwork period. Data collection will incorporate semi-structured interviews, company-wide surveys, observational shadowing, and the analysis of commits and pull requests. Furthermore, to translate these abstract cognitive shifts into observable indicators during planned shadowing sessions, the study will evaluate three descriptive calibration metrics:

1. the ratio of natural-language prompt-editing time versus manual syntax-editing time
2. the proportion of AI-generated versus human-generated code merged; and
3. developers' capacity to independently explain AI-generated code during technical walkthroughs.
4. patterns of interaction behaviour

From the gathered data sources between the practitioner and AI patterns of socio-technical negotiations can be developed from specific episodes where specific tasks get completed.

Initial findings

After preliminary analysis of prompt interactions from both sites and pilot observations we identify key negotiation episodes between practitioner and AI tools. Rather than treating the practitioner as the locus of agency we explore how paired interactions with AI affects outcomes of understanding. From this analysis four identifiable outcomes emerge based off what moves the developer makes and the moves the AI makes. These dyadic interactions form the preconditions for resultant epistemic outcomes.

Condition	AI Move	Developer Move	Epistemic Outcome	Empirical Example (Sites 1 & 2)
Deferral	**Closure** Hands over finished result; checks against its own internal tests	**Deferral** Accepts AI say-so without verification	**Understanding Erodes** Artifact works, but cannot explain why	Vibe-coding in unfamiliar language; recursive test generation where AI certifies own code.

Arbitration	**Referral** AI offers hypothesis; hands verdict to external arbiter	**Arbitration** Requests validation against external referent	**Understanding Preserved or Explored** (Concrete validation of what works and where)	Validating code on physical hardware, validating against known source of truth
Uptake	**Explication** Spells out mechanism; writes into a Knowledge Base(KB)	**Uptake** Internalizes logic, can later identify flaws	**Understanding Gained** See-How turns to Know-How	Interrogating AI technical constraints of the architecture when the AI context forgets
Breakdown / Contestation	**Overclaim** Asserts false/deceptive state with high confidence	Deferral leads to Silent Failure OR Arbitration leads to Repair & Re-grounding	Silent Collapse OR Restored Epistemic Grounding	AI claims 'nothing lost' when 532 files deleted; caught & repaired when dev audits git status.

Figure 1Human-AI negotiated outcomes

As illustrated in figure 1, condition A (Deferral) represents a negotiated outcome where the human yields cognitive judgment to the machine. When people act in this mode of deferral, it leads to situations where AI propositional knowledge, in the form of finished artifact, propositions and explanations, are delivered to the human practitioner as unexamined sources of truth. This leads to epistemic outcome where deliverables have the surface appearance of being correct, but the practitioner lacks the experiential engagement to know how and why the code works. This detachment from activity of doing is what we defined as "cognitive out dwelling" which is a state where Epistemological Reversal is formalized: understanding is acquired not through the active activity of authoring code, but afterwards by attempting to understand the pre-created product. This is demonstrated by Beane's account of programming in an unfamiliar language which required them to defer to the judgment of the machine.

Engagement with AI might not always have clear outcomes and negotiated outcomes can lead to a state of arbitration. Such state requires active thinking and negotiation with the AI evaluating various approaches where both actors have limited understanding. As an example of such condition, during the development of Fred-80 prompt records surface episodes where AI had limited domain knowledge to resolve system impediments and furthermore, could not be guided by the practitioner due to their lack of context and understanding. In this example both the practitioner and AI had to defer to a 3rd source of truth for definitive answers (for example manuals, external source code repositories, domain experts). Such arbitration can lead to neutral/positive epistemic outcomes for both the AI and practitioner. By exploring possibilities through active triadic engagement

understanding can be solidified or preserved or even gained through arbitration. This is demonstrated through the following observation during a triadic episode where one developer using AI had introduced duplicate code into the code base. In this scenario a negotiated episode of arbitration broke out:

"we already had that and it was already created so I had to ask him why he was duplicating stuff"

Later it was reflected:

"Maybe when prompted the AI was keen to implement without fully checking the broader repo as it was a sub module he was working in"

This episode resulted in a case of arbitration where one developer negotiated a change in implementation during a triadic encounter, but it also demonstrates Uptake.

Unlike passive deferral, Uptake represents a situation where co-creation genuinely builds practitioner competence. Here, knowledge transfer occurs directly through clear and explicit developer prompts that request detailed explication when the AI context drifts or generates hallucinations (plausible sounding but fabricated results). When the developer forces the AI to explain its decisions and justify its reasoning, the assistant is held to account and re-grounded in objective reality by the developer redirecting and correcting the AI's mistakes.

These moves are not solely one-sided interactions that emanate from the developer alone but rather a two-sided interaction that evolves, changes and emerges depending on specific circumstances. This varied coupling of human and AI interaction leads to malleable epistemic outcomes depending on both the developer's and AI's response. For instance, an AI move of Closure where the AI presents a finished artifact claiming all tests pass, can result in Deferral which leads to epistemic erosion. Deferral can transform into an arbitration episode if the developer manually validates output through testing or experimentation. Conversely, an AI move of Explication will fail to generate understanding if the developer skims it without active uptake.

Initial exploratory observations during the development of 'Project Axe' provide concrete evidence of the deferral. In an interview regarding the development process, a developer ('Beane') disclosed that he was building a feature using a programming language with which he was entirely unfamiliar. To achieve this, Beane relied solely on vibe coding. When questioned about the robustness of the resulting code, he noted: '…it seems to look ok and I trust what it makes.' This response typifies the liminal 'see-how'. Rather than deeply interrogating the machine's logic, the developer engaged in 'cognitive out-dwelling', acting

as a detached director, he blindly trusted the artifact's surface-level output without ever acquiring the foundational 'know-how' required to genuinely author it.
Breakdowns in agentic workflows emerge along two distinct vectors. The first is genuine misinterpretation during prompting, while the second is epistemic overclaiming during task evaluation. The first vector occurs when an agentic AI misinterprets conversational human interaction as a signal of intent. For example, during a lengthy download of fine-tuning training data, an operator frustrated by execution latency remarked:

'…this wait is killing me'

The agentic assistant interpreted this frustration as a direct command to abort, instantly terminating the task and disrupting the workflow and losing 532 files (later recovered), taking the command literally at face value. The second, more innocuous vector occurs when the AI overconfidently asserts an unverified hypothesis as objective truth. When confronted with such situations, the practitioner is presented with a critical fork in negotiation. They can either yield to the machine through Deferral which can induce silent failures and hidden technical debt or reassert their agency through Contestation and Arbitration. By refusing to settle interaction within the prompt window alone and insisting on verifying outputs against external, third-party records for authoritative outcomes, the developer actively re-grounds system truth in objective technical reality.

Observations from the development and adoption of the 'Guide' tool initiative demonstrates epistemological reversal in action. Driven by extreme delivery pressure, developers bypassed the need to understand the tool's underlying architecture by utilizing vibe coding to generate end-to-end tests. This dynamic had the consequence of guiding the developer towards a state of cognitive out-dwelling, where they functioned as detached directors rather than actively building 'know-how'. The process of acquiring the necessary knowledge to build becomes entirely mediated by the AI model where practice of coding is bypassed and the required knowledge that comes with that practice is yielded toward the machine. Furthermore, during development, the AI agent constructed new tests by autonomously mimicking pre-existing tests in the codebase, these ‘synthetic’ tests that had themselves been vibe-coded by developers lacking foundational context. This recursive generation perfectly mirrors out-dwelling where human experience is entirely relegated to a secondary state of see-how overseeing a massive, automated architecture that no human participant genuinely understands.

This epistemic shift can be seen during episode of evaluating the overarching architecture of the Guide tool. The original author of the tool, despite having independently researched the underlying evaluation literature, struggled to explain the system's exact implementation details to colleagues. Due to the tool's AI-accelerated complexity, the developer was reduced to prompting an AI coding assistant to explain the codebase back to them. This scenario exposes the illusion of mastery. Because the generative process bypassed the

mindful indwelling necessary to build tacit 'know-how', the author experienced total cognitive out-dwelling where they were reduced to a detached observer who must consult the machine's explicit 'know-that' just to understand the architecture of their own creation.

Discussion

The findings from Double Cone AI and Fred-80 problematize the software industry's default safeguard for AI-induced errors: keeping a 'human in the loop'. Organizational theory currently assumes that requiring a human to review AI-generated code through pull requests (as an example) provides sufficient guarantee for code quality. However, as demonstrated by the development of the 'Guide' tool and longitudinal traces from FRED-80, this assumption commits a category mistake by treating human oversight as a static structural condition. Instead, we argue that human oversight is a dynamic dyadic developed through negotiation, and often triadic during arbitration whose epistemic outcomes depend entirely on the specific moves developers execute when paired with the stances of AI agents.

While prevailing Information Systems literature frames AI choices around automation, which views delegation as functional 'cognitive offloading' and augmentation as a proxy for productivity, this paper frames human–AI collaboration as a dynamic, fluid negotiation. When developers engage with agentic machines capable of reasoning and collaboration, a spectrum of epistemic outcomes emerges. One such outcome is cognitive out-dwelling, enacted through a negotiated practice of deferral. Here, the practitioner does not merely transfer a workload. Instead, they sever their embodied connection to the craft itself, directing skills across the surface of their cognition rather than internalizing the architectural logic.

When trapped in cognitive out-dwelling, the 'human-in-the-loop' performs a procedural drill (Ryle, 1949) rather than achieving genuine epistemic grounding. This drill anchors the developer in an observational 'see-how' state, where they can see that a system functions through AI responses but fail to acquire the dispositional 'know-how' required to deeply understand why. Through deferral developers uncritically rely on mediated AI outputs to assess code trustworthiness and reliability. When this unfolds in practice instead of being competent reviewers of AI generated artifacts practitioner, because of the distance from the activity of creating code, enters what the literature terms 'human-washing' (Wilson et al., 2026). Human-washing is a systemic failure brought about when unexamined, surface-plausible AI output is transformed through administrative formalism into apparently reviewed software.

What is the impact of this? Organizations silently accumulate technical debt, cultivating operational codebases that no single team member genuinely understands. To halt this

atrophy of understanding, the design of human-in-the-loop AI workflows must fundamentally shift. Comprehension can no longer be assumed as a natural byproduct of authorship; it must be actively demanded and structurally enforced. Because the bifurcation between deskilling and upskilling depends entirely on whether a practitioner actively evaluates machine output, the passive drill of prompting must be superseded by the active training of reciprocal collaboration. To achieve this, AI workflows must embed 'epistemic friction' through 'reciprocal validation', an architectural intervention where agentic systems periodically pause to query the developer's understanding of generated logic.

Crucially, deferral is not the sole outcome of this dyad. As developers negotiate with their agentic epistemic partner, positive learning outcomes can be reclaimed. Through deliberate acts of Arbitration and Uptake, passive deferral is counteracted with active competence building. By forcing developers out of a passive 'see-how' state, this active engagement restores the mindful indwelling required to build and sustain tacit 'know-how'

References

Baskerville, R.L., Myers, M.D. and Yoo, Y. (2020) ‘Digital first: The ontological reversal and new challenges for information systems research’, MIS Quarterly, 44(2), pp. 509–524.

Duma, K.T., Wróblewski, P., Bobińska, J., Winiarska, J. and Przymus, P. (2026) ‘These Aren't the Reviews You're Looking For: How Humans Review AI-Generated Pull Requests’.

Fawzy, A., Tahir, A. and Blincoe, K. (2026) ‘From Prompting to Verification: How Experience Shapes Vibe Coding Practices’.

Green, B. (2021) ‘The Flaws of Policies Requiring Human Oversight of Government Algorithms’, Computer Law & Security Review.

Huang, H., Jaisri, P., Shimizu, S., Chen, L., Nakashima, S. and Rodríguez-Pérez, G. (2026) ‘More Code, Less Reuse: Investigating Code Quality and Reviewer Sentiment Towards AI-Generated Pull Requests’.

Itkonen, J., Mäntylä, M.V. and Lassenius, C. (2013) ‘The role of the tester’s knowledge in exploratory software testing’, IEEE Transactions on Software Engineering, 39(5), pp. 707–724.

Macnamara, B.N., Berber, I., Çavuşoğlu, M.C., Krupinski, E.A., Nallapareddy, N., Nelson, N.E., Smith, P.J., Wilson-Delfosse, A. and Ray, S. (2024) 'Does using artificial intelligence assistance accelerate skill decay and hinder skill development without performers' awareness?', Cognitive Research: Principles and Implications, 9.

Meske, C., Hermanns, T., Von der Weiden, E., Loser, K.-U. and Berger, T. (2025) ‘Vibe Coding as a Reconfiguration of Intent Mediation in Software Development: Definition, Implications, and Research Agenda’.

Myers, M.D. (1999) ‘Investigating information systems with ethnographic research’, Communications of the Association for Information Systems, 2, Article 23.

Nandhakumar, J. and Jones, M.R. (2002) ‘Development gain? Participant observation in interpretive management information systems research’, Qualitative Research, 2(3), pp. 283–306.

Polanyi, M. (1966) The tacit dimension. Garden City, NY: Doubleday.

Qiao, Y., Shihab, M., Haque, S. and Hundhausen, C. (2025) ‘Code comprehension with GitHub Copilot: Performance gains, comprehension trade-offs, and behavioral predictors in brownfield programming’, ArXiv.

Risko, E.F. and Gilbert, S.J. (2016) ‘Cognitive offloading’, Trends in Cognitive Sciences, 20(9), pp. 676–688.
Runeson, P. and Höst, M. (2009) ‘Guidelines for conducting and reporting case study research in software engineering’, Empirical Software Engineering, 14(2), pp. 131–164.

Russo, D. (2024) ‘Navigating the Complexity of Generative AI Adoption in Software Engineering’, ACM Transactions on Software Engineering and Methodology, 33(5), Article 135.

Ryle, G. (1949) The concept of mind. London: Hutchinson.

Sabouri, S., Eibl, P., Zhou, X., Ziyadi, M., Medvidović, N., Lindemann, L. and Chattopadhyay, S. (2025) ‘Trust Dynamics in AI-Assisted Development: Definitions, Factors, and Implications’.

Sarkar, A. and Drosos, I. (2025) ‘Vibe coding: Programming through conversation with artificial intelligence’, in Proceedings of the 36th Annual Conference of the Psychology of Programming Interest Group (PPIG 2025).

Walsham, G. (1995) ‘Interpretive case studies in IS research: Nature and method’, European Journal of Information Systems, 4(2), pp. 74–81.

Wheeler, B.R. (2026) ‘The Substrate Collapse: AI Code Generation Invalidates Authorship-Based Knowledge Metrics’.

Wilson, B., Swana, M., Winter, P.D. and Roach, M.J. (2026) ‘Humanwashing -- It Should Leave You Feeling Dirty’.

Yin, R.K. (2018) Case study research and applications: Design and methods. 6th edn. Thousand Oaks, CA: Sage.

Conceptual Convergence in Global Digital Governance: An Analysis of Principles and Themes Across Seventeen Frameworks

Theresa Blissing[1], Shirley Gregor[1,2], Michael Rosemann[1]
1 Queensland University of Technology – Centre for Future Enterprise, 2 Australian National University

Abstract

Governance for digital technologies arises from a fragmented regime complex of governments, regional bodies, intergovernmental organisations, firms, and multistakeholder initiatives. This study asks whether a common set of normative themes recurs across digital governance frameworks broadly, not only those concerned with artificial intelligence (AI), and how such a thematic map can be established robustly. Using content analysis, 147 principles and principle-like statements from 17 soft-governance frameworks were coded. A coding framework was developed iteratively and validated through formal intercoder reliability testing with coders from five countries. The resulting eight themes achieved substantial to almost perfect agreement and were sufficient to code both AI-specific instruments and broader digital-economy strategies. We make two claims: first, intercoder reliability establishes the coding framework as reproducible, and second, the recurrence of the same eight themes across instruments of differing scope is what evidences thematic convergence. This paper contributes a reliability-tested coding framework and an empirically grounded map of convergence towards a consolidated, sector- and technology-agnostic model of digital soft governance.

1. Introduction

The digital world has become foundational in almost every domain of contemporary life. Commerce, public administration, healthcare, education, finance, and social interaction depend on digital technologies as critical infrastructure. Yet, while these systems are inherently global and operate across borders, their governance is not. The rules intended to govern them are produced by a fragmented patchwork of governments, regional bodies, intergovernmental organisations, firms, and multi-stakeholder initiatives — a disjointed regime complex, dense and overlapping, with no central authority reconciling them (Geith et al., 2026).

Over the past decade this fragmentation has been most visible in the governance of AI, where a proliferation of ethics guidelines, principles documents, and soft-law instruments has prompted a recurring question: beneath the surface diversity of these initiatives, is a shared set of normative themes emerging (Jobin et al., 2019; Fjeld et al., 2020)? A small but consistent body of comparative research has begun to answer that question for AI, repeatedly finding convergence around a recognisable cluster of themes including transparency, fairness, accountability, and privacy (Jobin et al., 2019; Fjeld et al., 2020; Lim & Kwon, 2021; Palladino, 2023; Corrêa et al., 2023). Recurrence across studies with different corpora and methods is the strongest signal that this convergence is genuine rather than an artefact of any single coding scheme.

Yet this comparative research shares two limitations. First, the convergence it documents is technology-specific: each study draws its corpus exclusively from AI governance specifically, leaving open whether the same themes recur across the digital world. Second, while core clusters recur, their boundaries vary from study to study and no universal set of themes has emerged; themes such as safety or explainability appear in some studies but are absent or subsumed in others (Gutierrez et al., 2020; Corrêa et al., 2023), raising the question of how findings might be made robust enough to reflect underlying commonalities rather than the coding choices of a single study.

This study addresses both gaps by asking two questions: first, can conceptual convergence be identified across digital governance frameworks broadly, and not only those concerned with AI; and second, how can such convergence be established robustly? To answer these questions, we systematically coded principles and principle-like statements from 17 global and regional digital governance frameworks and subjected the results to formal intercoder reliability testing across multiple independent coders from different countries. In the terms of Gregor's (2006) taxonomy of theory in information systems, the resulting themes constitute a Type I (analytic) theory: a validated classification of "what is" that offers an empirically grounded map of convergence extending beyond AI.

2. Literature Review

The study of governance frameworks and normative convergence in AI has matured into a small but consistent body of work spanning AI ethics, technology policy, and governance studies. A scoping review identified nine studies that systematically map themes across AI governance frameworks, ranging from private sector and expert-driven corpora (Floridi et al., 2018; Jobin et al., 2019) to large multi-country corpora (Fjeld et al., 2020; Gutierrez et al., 2020; Lim & Kwon, 2021; Palladino, 2023; Corrêa et al., 2023; Robles & Mallinson, 2025; Qiu & Cheng, 2026). These studies do not share a common vocabulary: some speak of principles, others of themes or factors. For clarity, we use "principle" for a normative prescriptive statement of the form "Aim at achieving X", where X is a desirable goal, and "theme" for a shared concern recurring across principles.

Across existing studies a recognisable cluster of themes recurs — transparency, fairness, accountability, and privacy — but the boundaries do not align cleanly. "Accountability" is confirmed by five studies, yet Corrêa et al. (2023) term theirs "accountability/liability"; similar discrepancies recur for "fairness", "safety", and "transparency". Several studies rank themes by frequency, but no consistent top set emerges. Validation, moreover, remains internal to each research team, except for two that triangulate against prior literature (Robles & Mallinson, 2025; Qiu & Cheng, 2026). Several authors candidly disclaim their results as partial or subjective (Fjeld et al., 2020; Gutierrez et al., 2020; Lim & Kwon, 2021; Corrêa et al., 2023).

Two gaps follow. First, convergence is documented but no consensus on a universal set of themes exists, because categories rest largely on internal interpretive judgement. Second, all nine studies focus on AI — some broadened to software, data science, or robotics (Corrêa et al., 2023) — but none examines the digital world as a whole and the scoping review identified

no comparative analysis of normative convergence extending beyond AI. Yet Helbing and Ienca (2024) argue that formerly unrelated technologies are becoming integrated and that "converging technologies need converging international regulation". This study therefore asks whether convergence can be identified across digital governance frameworks broadly, not only those related to AI, and how it can be validated to make the themes more robust.

3. Methodology

We conducted a content analysis (Krippendorff, 2018) to examine the themes articulated across global and regional digital governance frameworks — a scope warranted by the growing integration of formerly distinct technologies whose effects increasingly cut across domains (Helbing & Ienca, 2024). The intended output is, in Gregor's (2006) terms, a Type I (analytic) theory, and the resulting coding framework (CF) is assessed against the criteria for such theory in the Results.

A targeted review identified digital soft-governance frameworks with principled, normative content using four inclusion criteria: internationally developed or recognised (by multiple countries or by a globally influential standards body), principle-bearing, applicable wholly or partly to the digital world, and cross-sectoral rather than industry-specific. This review yielded 17 frameworks published between 2003 and 2024 by intergovernmental organisations (e.g., the UN, OECD, and G20), regional bodies (e.g., the African Union, Council of Europe, and ASEAN), and the international standards community (the IEEE's Ethically Aligned Design, 2019). They span the WSIS Geneva Declaration (2003), instruments such as the OECD and G20 AI Principles and the UNESCO Recommendation on the Ethics of AI (2021), and the UN Global Digital Compact (2024), and together contain 147 principles or principle-like statements forming the complete dataset. The corpus is deliberately mixed in scope: nine of the seventeen frameworks address AI specifically, and the remaining eight cover digital technologies more broadly — a balance that lets the analysis test whether the themes hold beyond AI.

The principle-bearing text was inductively coded by the research team into 17 codes, which were then validated and refined through intercoder reliability testing, a widely recognised metric for the credibility and replicability of qualitative research (O'Connor & Joffe, 2020; Cofie et al., 2022). A pilot test with a coder from Germany prompted a first refinement to 15 codes. Two further coders, unfamiliar with the research and from different countries (Singapore and Brazil), then applied the 15 codes deductively, each assigning every principle to between one and four codes, since a single principle often spans several themes. Cohen's Kappa was mixed across these codes, so a co-occurrence analysis was used to merge conceptually overlapping codes, yielding a refined set of eight codes. Figure 1 shows the CF across its three iterations.

Two new coders, from Australia and Norway, then coded the 147 principles with the eight-code CF, and reliability improved substantially. A structured review session was held to examine the residual disagreement (Nili et al., 2020; MacPhail et al., 2016). To avoid coders defending their own prior choices, the review was blind to authorship: for each disputed text coders saw the combined set of codes that had been assigned, but not who had initially assigned them, and then decided together whether every code was justified. This process resolved substatial disagreement and aligned the coders' interpretive heuristics. The session also confirmed that texts with explicit keywords were easier to code than abstract formulations.

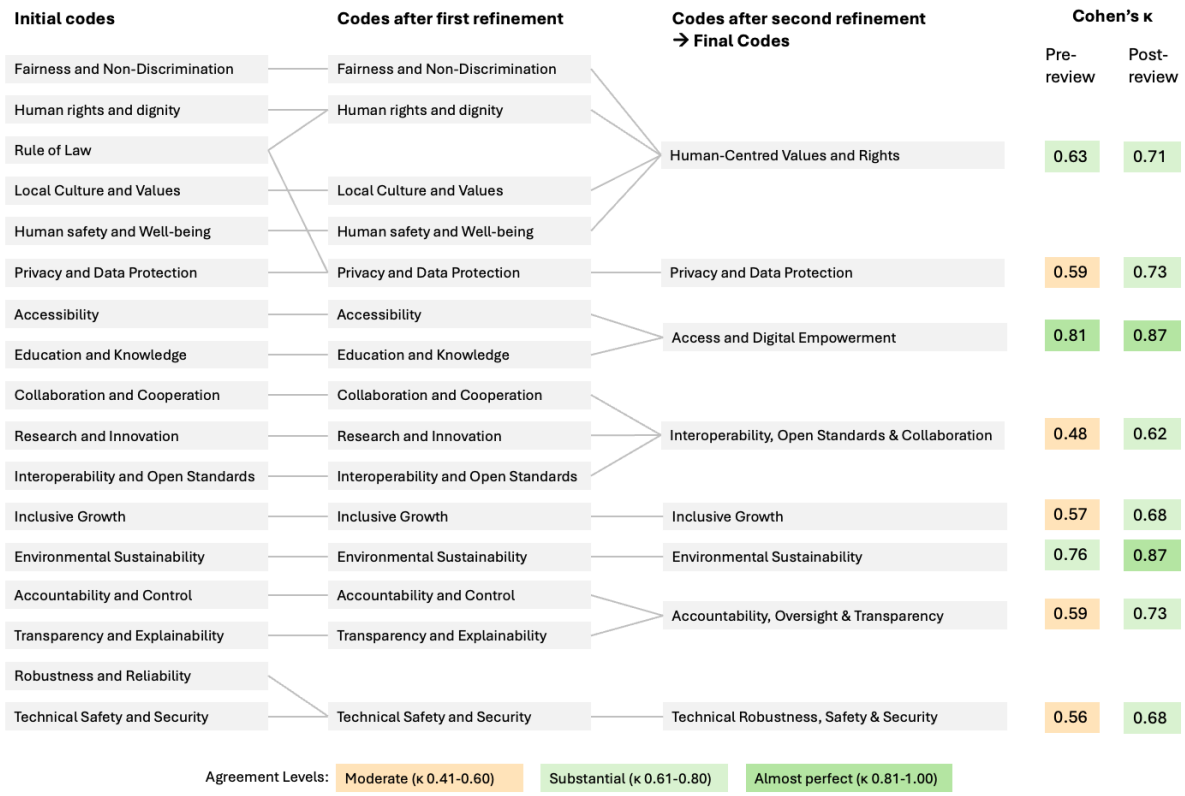
4. Results

After the review session, all eight codes achieved substantial to almost perfect agreement (Figure 1), and 146 of the 147 principles (99.3%) mapped to at least one code. Agreement increased across every category through the session and is reported in Figure 1.

Two claims must be kept distinct here. The intercoder reliability results establish that the CF is reproducible: coders unfamiliar with the research and drawn from different countries apply the eight codes consistently. They do not, on their own, establish that the frameworks converge. High coverage is a necessary but not a sufficient sign of convergence, because some codes are broad umbrellas — "Human-Centred Values and Rights", for instance, spans human rights, dignity, fairness, non-discrimination, and inclusion — so near-total coverage could partly reflect the breadth of the categories rather than genuine agreement among the frameworks. We therefore rest the convergence claim not on coverage alone but on the recurrence of the same compact set of eight themes across frameworks of markedly different scope and origin: the pattern holds whether an instrument is an AI-specific ethics guideline or a broad digital-economy strategy, and only one principle out of 147 required a theme outside the eight.

This recurrence across instruments of differing scope suggests that the themes are not tied to any single technology, though that reading warrants caution. Convergence is not the same as independent emergence: many of these frameworks borrow explicitly from one another — the OECD AI Principles in particular are widely echoed — so part of the observed convergence may reflect institutional diffusion and copying rather than frameworks arriving at shared themes independently. Distinguishing genuine convergence from diffusion is an important question for future work.

Figure 1. CF iteration and intercoder reliability for the eight-code CF, pre and post the review session



As a Type I theory, the CF can be appraised against Gregor's (2006) criteria for analytic theory — completeness, mutually exclusive categories, meaningful and natural labels, appropriate hierarchy, and clear classification logic. It is near-exhaustive (99.3% coverage). Mutual exclusivity was deliberately relaxed, since principles routinely span several themes, but the co-occurrence analysis and the merging of overlapping codes reduced spurious overlap and made the categories as distinct as the material allows. The meaningfulness of the labels is evidenced by the reliability results themselves, on which coders from different countries converged.

One principle (0.7%), from the UN Global Digital Compact (2024), could not be mapped to any code — a "forward-looking" commitment to anticipating emerging technologies, whose abstract language resisted categorisation. It hints that anticipatory governance may constitute an emerging domain not yet captured by existing taxonomies, including the eight codes.

5. Conclusion

This study set out to answer whether conceptual convergence can be identified across digital governance frameworks beyond AI, and how such convergence can be established robustly rather than merely asserted. On the first question, the evidence is affirmative. Where prior comparative studies have looked for convergence only within AI governance, this study is, to our knowledge, the first to test it across the digital domain more broadly: across seventeen globally diverse international and regional frameworks spanning two decades and every major region, 146 of 147 principles mapped onto a compact set of eight themes, and the same eight accommodated both AI-specific instruments and broad digital strategies. Because the corpus is close to evenly split between the two (nine AI-specific frameworks and eight addressing digital technologies more broadly) the recurrence of the same themes across both groups indicates that convergence is not confined to AI but extends across the digital domain.

On the second question, the CF's value rests on how it was built. Where prior studies rested their categories mainly on internal interpretive judgement and none subjected its themes to independent validation, this study tested its CF through formal intercoder reliability across several coders from different countries, refined it iteratively in response to that testing, and resolved residual disagreement through a structured review session. The eight themes are therefore not only empirically grounded but demonstrably reproducible. Crucially, that reproducibility is a claim about the coding framework, not by itself a claim about the corpus; the convergence claim rests on the recurrence of the themes across instruments of differing scope. In Gregor's (2006) terms, the CF constitutes a Type I (analytic) theory whose principal advance over prior convergence studies is that it has been externally validated against the criteria such theory is held to.

Several limitations qualify these findings and point toward future work. The corpus, though broad and globally diverse, is a purposive selection: it excludes national legislation and sectoral or private-actor instruments, and admits only frameworks available in the coders' languages. It also remains open how far the observed convergence reflects independent emergence rather than institutional diffusion. Extending the validated framework to national and sectoral instruments, distinguishing convergence from diffusion, and investigating the candidate "forward-looking" domain are natural next steps. The eight themes could then serve as the empirical foundation for a consolidated, sector- and technology-agnostic digital governance model — the cross-domain approach that Helbing and Ienca (2024) argue technological convergence increasingly demands — delivering the foundation this study set out to provide: a validated, reproducible eight-theme coding framework showing that beneath the fragmentation of the current regime complex, a shared and surprisingly stable normative structure is taking shape.

6. References

Cofie, N., Braund, H., & Dalgarno, N. (2022). Eight ways to get a grip on intercoder reliability using qualitative-based measures. *Canadian Medical Education Journal, 13(2)*, 73–76. https://doi.org/10.36834/cmej.72504

Corrêa, N. K., Galvão, C., Santos, J. W., Del Pino, C., Pinto, E. P., Barbosa, C., Massmann, D., Mambrini, R., Galvão, L., Terem, E., & de Oliveira, N. (2023). Worldwide AI ethics:

A review of 200 guidelines and recommendations for AI governance. *Patterns (New York, N.Y.), 4(10)*, Article 100857. https://doi.org/10.1016/j.patter.2023.100857

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. *Berkman Klein Center Research Publication, (2020-1).* Available at SSRN: https://ssrn.com/abstract=3518482 or http://dx.doi.org/10.2139/ssrn.3518482

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(4)*, 689–707. https://doi.org/10.1007/s11023-018-9482-5

Geith, J., Lundgren, M., & Tallberg, J. (2026). The Emerging Regime Complex for Artificial Intelligence. *Global Studies Quarterly.*

Gregor, S. (2006). The Nature of Theory in Information Systems. *MIS Quarterly, 30(3),* 611–642. https://doi.org/10.2307/25148742

Gutierrez, C. I., Marchant, G., Carden, A., Hoffner, K., & Kearl, A. (2020). Preliminary Results of a Global Database on Soft Law Mechanisms for the Governance of Artificial Intelligence. *2020 IEEE / ITU International Conference on Artificial Intelligence for Good (AI4G)*, 33–37. https://doi.org/10.1109/AI4G50087.2020.9311078

Helbing, D., & Ienca, M. (2024). Why converging technologies need converging international regulation. *Ethics and Information Technology*, *26*(1), Article 15. https://doi.org/10.1007/s10676-024-09756-8

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1(9)*, 389–399. https://doi.org/10.1038/s42256-019-0088-2

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology.* Sage publications.

Lim, J. H., & Kwon, H. Y. (2021). A Study on the Modeling of Major Factors for the Principles of AI Ethics. *Proceedings of the 22nd Annual International Conference on Digital Government Research*, 208–218. https://doi.org/10.1145/3463677.3463733

MacPhail, C., Khoza, N., Abler, L., & Ranganathan, M. (2016). Process guidelines for establishing Intercoder Reliability in qualitative studies. *Qualitative research : QR, 16(2)*, 198-212. https://doi.org/10.1177/1468794115577012

Nili, A., Tate, M., Barros, A., & Johnstone, D. (2020). An approach for selecting and using a method of inter-coder reliability in information management research. *International journal of information management, 54*, 102154. https://doi.org/10.1016/j.ijinfomgt.2020.102154

O'Connor, C., & Joffe, H. (2020). Intercoder Reliability in Qualitative Research: Debates and Practical Guidelines. *International Journal of Qualitative Methods, 19*. https://doi.org/10.1177/1609406919899220

Palladino, N. (2023). A 'biased' emerging governance regime for artificial intelligence? How AI ethics get skewed moving from principles to practices. *Telecommunications Policy, 47(5)*, Article 102479. https://doi.org/10.1016/j.telpol.2022.102479

Qiu, J., & Cheng, L. (2026). Consensus and legitimation in global AI regulations: a sociosemiotic perspective. *International Journal of Law in Context, 1–22*. https://doi.org/10.1017/S1744552326100469

Robles, P., & Mallinson, D. J. (2025). Advancing AI governance with a unified theoretical framework: a systematic review. *Perspectives on Public Management and Governance, 8(4)*, 213–227. https://doi.org/10.1093/ppmgov/gvaf013

Digital Sustainability At the Crossroads: Configuring Indigenous Knowledge and ICT in Smallholder Aquaculture

Muhammad Izharuddin, Christoph Briedbach, Tapani Rinta-Kahila
Business School, The University of Queensland

Abstract
Information and Communication Technologies (ICT) can enable smallholder communities to address climate change and achieve sustainability. Recent research highlights the need to integrate ICT with Indigenous and Local Knowledge (ILK). However, two debates exist in the literature: whether ICT should be deployed through centralized or decentralized approaches, and whether ILK should remain standalone or be integrated via co-production. We investigate how specific configurations of ILK and ICT deployment enable holistic sustainability in smallholder communities facing climate change. We contribute to the digital sustainability and Indigenous IS literatures by specifying that decentralized ICT can unlock the adaptive potential of ILK to enable customized responses to climate challenges, while centralized ICT must act solely as a supportive resource scaffold to avoid eroding ILK, which would otherwise trigger environmental harm.

Keywords: Digital Sustainability; Indigenous Information Systems; Indigenous and Local Knowledge (ILK); ICT Deployment; Smallholder Aquaculture.

1. Introduction

ICT can help mitigate and adapt to the impacts of climate change (Watson et al., 2010). Information Systems (IS) research has a growing track record in this area, beginning with Courtney et al.'s (2000) seminal study on environmental sustainability and expanding to the use of ILK in designing and implementing ICT (Thapa, 2012) to sense changing environmental conditions. Sustainability objectives have widened to include economic (Watson et al., 2010) and social dimensions as this research stream evolved into digital sustainability (Kotlarsky et al., 2023). Despite these advancements, scholars continue to debate key strategic approaches.

The debate centres on whether top-down (centralized) or bottom-up (decentralized) approaches work best for ICT to enable sustainability (Tim et al., 2018; Seidel et al., 2013).

Top-down approaches can mobilize capital, enforce rules, and manage ICT across organizations (Hu et al., 2016), thereby achieving environmental sustainability objectives (Watson et al., 2010; Yang et al., 2020). However, empirical evidence reveals that participants in top-down initiatives often struggle with low user engagement because these solutions remain disconnected from the lived experiences of local communities (Leidner et al., 2022). On the other hand, bottom-up approaches, initiated by participants, communities, and local actors (Hedman & Henningsson, 2016; Tim et al., 2018), can generate high user buy-in by leveraging contextual innovations developed by local users (Watson et al., 2021), making them more suitable for achieving social sustainability (Butler & Hackney, 2021). Successful digital sustainability initiatives hint at the use of ILK alongside ICT (Watson et al., 2021) to tailor solutions to specific local settings (Corbett, 2013). This connection links the centralized versus decentralized debate in digital sustainability to Indigenous IS (Davison & Díaz Andrade, 2018; Díaz Andrade et al., 2021).

This brings up a second debate in the ILK-IS literature, where scholars contest whether standalone (pure ILK) (e.g., Davison & Díaz Andrade, 2018) or co-production (integrative ILK) (e.g., Puri, 2007) strategies are more appropriate for developing ICT to enable sustainability. The standalone approach is designed by and for Indigenous communities, deriving from their own knowledge repositories such as Kaupapa Māori (Díaz Andrade et al., 2021). Proponents of this approach assert that ILK is complete and valid on its own terms, requiring neither testing nor translation by other knowledge systems to establish legitimacy (Latulippe & Klenk, 2020). Conversely, others argue that ILK is a living system that has always interacted with external knowledge systems and changing environments (Latulippe & Klenk, 2020). The co-production approach combines ILK with ICT capabilities (e.g., Bawack et al., 2025) or with scientific knowledge (SK) (Ijatuyi et al., 2025) to create solutions for climate change (Chaudhuri & Kendall, 2021). However, both approaches have pitfalls: standalone approaches encounter severe resource constraints (Miscione, 2007), limited resilience to accelerated climate shifts, and insights that may not easily translate to broader contexts (Ijatuyi et al., 2025). Meanwhile, co-production risks allowing SK to overpower ILK, which can dilute its governance value (Puri, 2007) and lead to exploitation, misappropriation, or harm (Latulippe & Klenk, 2020).

Both debates highlight that we do not yet understand which configurations best achieve holistic sustainability outcomes, whether environmental, social, or economic, particularly in contexts where smallholder communities face changing environmental conditions, limited capital, and endangered ILK. We therefore ask: *How do configurations of ICT deployment and ILK enable smallholder communities to respond to climate change and influence economic, social, and environmental sustainability?*

2. Literature Review

2.1. Digital Sustainability

Digital sustainability refers to “the development and deployment of digital resources and artifacts toward improving the environment, society, and economic welfare” (Kotlarsky et al., 2023, p. 938). The discourse surrounding digital sustainability centres on how ICT is governed and deployed to mitigate changing environmental conditions and achieve environmental goals (Watson et al., 2010) through centralized (top-down) or decentralized (bottom-up) models (Tim et al., 2018).

2.2. Indigenous IS

Davison and Díaz Andrade (2018, p. 760) define Indigenous theorizing in IS as “a theory of human behaviour or mind that is specific to a context or culture, not imported from other

contexts/cultures and purposely designed for the people who live in that context or culture." One primary motivation for IS scholars to engage in Indigenous theorizing is to overcome the high failure rates (up to 75%) of centralized digital development initiatives resulting from the imposition of decontextualized Western designs (Miscione, 2007). In contrast, decentralized Indigenous IS initiatives (e.g., technoficing) foster deeper community involvement and contextual alignment (e.g., Chaudhuri & Kendall, 2021).

3. Methodology

3.1. Research Design

This study employs a multiple-case study design (Yin, 2018), with smallholder fish farms serving as the unit of analysis. To ensure comparability and identify suitable cases, we conducted a criterion-based case selection process (Yin, 2018). First, we selected fish farmers who actively integrate ICT into their daily operations (e.g., smartphone aquaculture applications, autofeeders, and water quality monitoring devices). Second, we employed a maximum variation strategy (Flyvbjerg, 2006) to observe different configurations associated with distinct sustainability outcomes, as shown in Figure 1. Figure 1 categorizes smallholder fish farmers along two dimensions: the approach to ILK (standalone vs. integrative) and the nature of technology deployment (customized tools associated with decentralized approaches vs. market-based tools associated with centralized approaches).

3.2. Research Context

This research took place in Indonesia, a developing country with vast fisheries potential. In 2019, aquaculture accounted for 68.4% of total domestic fisheries production, contributing 15.89 million tons, approximately 13.23% of total global production (Novogratz, 2023). Despite this scale, changing environmental conditions (specifically fish diseases and extreme weather events) remain persistent challenges. In response, aquatech firms are introducing ICT to support smallholder fish farmers (Novogratz, 2023).

3.3. Data Collection



Figure 1. Multiple-Case Designs Involving ILK and Tools and Technologies

Our study relies on field-based and ethnographically inspired research. We conducted a total of 71 semi-structured interviews with diverse stakeholders, including fish farmers, community leaders, government officials, aquatech providers, feed suppliers, aquaculture researchers, and NGOs. In total, this study gathered 24 fish-farming cases across 8 Indonesian cities. A

local researcher also immersed themselves in the aquaculture community by participating in active social media discussion groups, attending informal community meetings, and observing annual industry events (e.g., conferences and fairs).

4. Preliminary Findings

Table 1. Four categories of ILK

	ILK of land and animals	ILK land and resource	Social institutions	Worldview
Slow-Living	6 ILKs	9 ILKs	1 ILK	2 ILKs
Fish Whisperer	5 ILKs	4 ILKs	1 ILK	2 ILKs
Waterbender	3 ILKs + SK + ICT	1 ILK + SK + ICT	1 ILK	2 ILKs
Tech-Believer	2 ILKs + SK + ICT	SK + ICT	1 ILK	N/A

This study uses Berkes's (1999) foundational framework to map how ILK enables communities to adapt to changing environmental conditions, classifying ILK into four interrelated dimensions: (1) land and animals, (2) land and resources, (3) social institutions, and (4) worldview.

4.1. Slow-Living

There is a singular reliance on ILK in the Slow-Living archetype, without the integration of SK or the deployment of ICT. Fish farmers in this archetype only use social media messengers such as WhatsApp to communicate with other aquaculture stakeholders (e.g., middlemen, hatcheries). Changing environmental conditions force them to adjust their ILK, but it may not be enough, as they continue to face climate challenges (e.g., fish diseases and extreme weather). Ultimately, they choose to sacrifice economic gain to preserve their ILK heritage.

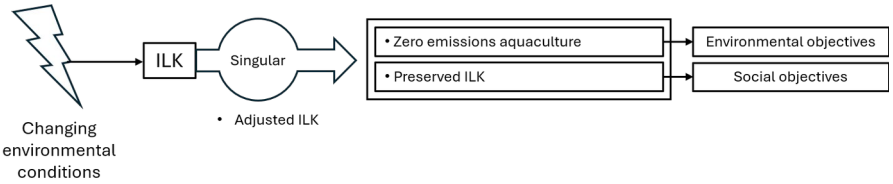


Figure 2. Slow-Living

4.2. Fish Whisperer

Fish Whisperers use aquaculture supplies (e.g., feed, minerals, decomposers) and equipment (e.g., paddle wheels) that are publicly available on the market, while continuing to practice their ILK, such as observing seasonal calendars, bird behavior, water conditions, wind patterns, and soil cracks to guide their fish farming. They also adopt aquatech services such as peer-to-peer financing (e.g., via an aquatech app) to secure feed and other supplies. This approach helps them adapt to climate change.

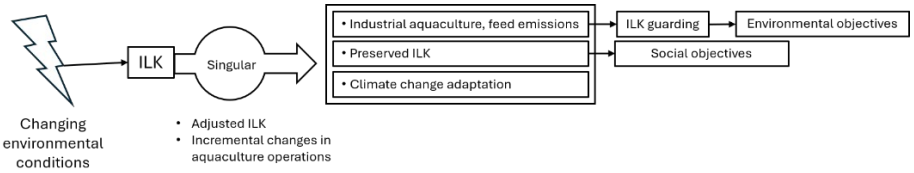


Figure 3. Fish Whisperer

4.3. Waterbender

Waterbenders integrate SK, ICT, and a small portion of ILK in a complementary manner to address changing environmental conditions. They reduce the environmental impact of industrial aquaculture (e.g., high feed emissions) by remanifesting ILK. Furthermore, they customize and tailor ICT and aquaculture tools to meet their need for rapid responses to climate change. For example, water quality monitoring devices allow them to track environmental indicators and receive early warnings before disease strikes, thereby enabling them to adapt to climate challenges. Autofeeders help them regulate feeding frequency and volume according to fish appetite, lowering the Feed Conversion Ratio (FCR) and thereby enabling them to mitigate climate change.

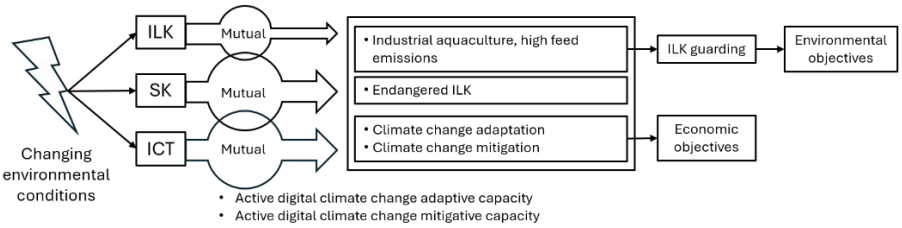


Figure 4. Waterbender

4.4. Tech-Believer

In the Tech-Believer configuration, ICT is deployed singularly, while a small portion of ILK is assimilated into SK. This passive digital capacity can lead to industrial aquaculture practices characterized by high feed emissions and chemical pollution. Consequently, ILK becomes endangered as it is overpowered and displaced by external knowledge systems.

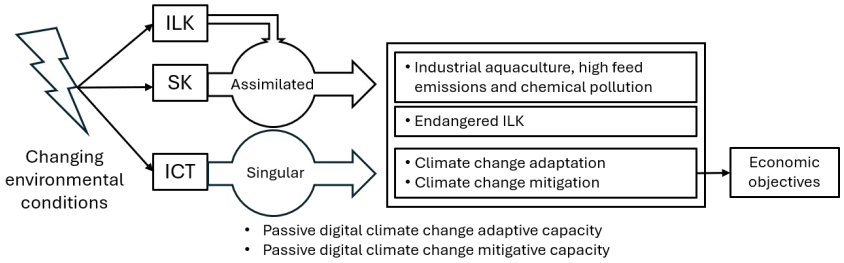


Figure 5. Tech Believer

5. Discussion

This study offers theoretical contributions to two ongoing conversations: (1) digital sustainability (Kotlarsky et al., 2023), specifically regarding the debates on centralized versus decentralized ICT deployment to achieve sustainability (Tim et al., 2018; Seidel et al., 2013); and (2) Indigenous IS, regarding standalone versus co-produced ILK strategies (Davison & Díaz Andrade, 2018; Puri, 2007). The 2×2 configuration is not designed to declare which deployment model is superior. Instead, it demonstrates that decentralized ICT can unlock the adaptive potential of ILK to address immediate, localized challenges such as climate change. Meanwhile, centralized ICT can act solely in a supporting capacity (e.g., providing tools and resources), rather than subordinating ILK, which risks triggering environmental harm and knowledge erosion.

6. References

Berkes, F. (1999). *Sacred ecology*, 1st ed. Philadelphia: Taylor & Francis.

Butler, T., & Hackney, R. (2021). The role of informational mechanisms in the adoption of Green IS to achieve eco-sustainability in municipalities. Information & Management, 58(4), 103460.

Chaudhuri, B., & Kendall, L. (2021). Collaboration without consensus: Building resilience in sustainable agriculture through ICTs. *The Information Society*, *37*(1), 1-19.

Corbett, J. (2013). Designing and using carbon management systems to promote ecologically responsible behaviors. Journal of the Association for Information Systems, 14(7), 339-378.

Courtney, J. F., Richardson, S., & Paradice, D. (2000). Information technology and sustainable development. Americas' Conference on Information Systems,

Davison, R. M., & Díaz Andrade, A. (2018). Promoting indigenous theory. Information Systems Journal, 28(5), 759–764.

Díaz Andrade, A., Techatassanasoontorn, A. A., Singh, H., & Staniland, N. (2021). Indigenous cultural re-presentation and re-affirmation: The case of Māori IT professionals. Information Systems Journal, 31(6), 803–837.

Flyvbjerg, B. (2006). Five misunderstandings about case-study research. *Qualitative inquiry*, *12*(2), 219–245.

Hedman, J., & Henningsson, S. (2016). Developing ecological sustainability: A Green IS response model. Information Systems Journal, 26(3), 259-287.

Hu, L. C., Hu, G. C., Wei, C. P., & Hsu, C. (2016). Examining firms' green information technology practices: A hierarchical view of key drivers and their effects. Journal of Management Information Systems, 33(4), 1149-1179.

Kotlarsky, J., Oshri, I., & Sekulic, N. (2023). Digital sustainability in information systems research: Conceptual foundations and future directions. *Journal of the Association for Information Systems*, *24*(4), 936–952.

Latulippe, N., & Klenk, N. (2020). Making room and moving over: knowledge co-production, Indigenous knowledge sovereignty and the politics of global environmental change decision-making. *Current Opinion in Environmental Sustainability*, *42*, 7-14.

Leidner, D. E., Goutas, L., et al. (2022). Multifarious roles and conflicts on an interorganizational Green IS. MIS Quarterly, 46(2), 1003-1036.

Miscione, G. (2007). Telemedicine in the Upper Amazon: Interplay with local health care practices. MIS Quarterly, 31(2), 403–425.

Novogratz, A. (2023). *Why $1.4 billion valuation is good news for eFishery and excellent news for the sector as a whole*.

Puri, S. K. (2007). Integrating scientific with indigenous knowledge: Constructing knowledge alliances for land management in India. MIS Quarterly, 31(2), 355–379.

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

Thapa, D., Sein, M. K., & Sæbø, Ø. (2012). Building collective capabilities through ICT in a mountain region of Nepal: where social capital leads to collective action. *Information Technology for Development*, *18*(1), 5-22.

Tim, Y., Pan, S. L., Bahri, S., & Fauzi, A. (2018). Digitally enabled affordances for community-driven environmental movement in rural Malaysia. *Information systems journal*, *28*(1), 48–75.

Watson, R. T., Boudreau, M. C., & Chen, A. J. (2010). Information systems and environmentally sustainable development: Energy Informatics and new directions for the IS community. *MIS Quarterly, 34*(1), 23-38.

Watson, R. T., et al. (2021). How the AIS can improve its contributions to the UN's Sustainability Development Goals. Communications of the AIS, 48, 480-495.

Yin, R. K. (2018). *Case Study Research and Applications: Design and Methods* (6th ed.). Sage.

Platformization as adaptive cycles: An ecological perspective

Karolina Mikołajewska-Zając and Attila Márton
Department of Digitalization, Copenhagen Business School

Abstract

Platform research recognises mutual adaptation between platforms and their environments but pays less attention to how these interactions reshape subsequent platformization. Drawing on Holling's (2001) adaptive cycle and a case study of tourism platformization in Barcelona, we conceptualise platformization as recursive co-evolution. We examine how this process unfolds across successive waves, how it redistributes adaptive capacities in the ecosystem, and how it transforms the governance of platformized activity. We show how these dynamics shape further adaptation at the ecosystem level and transform the ways in which platformized activity is governed.

1. Introduction

Digital platforms, the archetypal organisational form of the digital economy (Gawer, 2014), remain a moving target for IS research. This is often attributed to their digital characteristics: the ability to shapeshift in response to strategic goals and changing environments (Ciborra, 1996), their distributed entanglement with institutions, markets, and technologies (De Reuver et al., 2018), and hyperscalability (Huang et al., 2017). Yet, despite recognising platforms as dynamic, flexible, connective, and largely invisible, IS research has often approached them as relatively bounded entities, conceptualising the networks they orchestrate through the lens of ecosystems (Jacobides et al., 2018). Although the distinction between platforms and ecosystems remains blurred (see Alaimo & Kallinikos, 2024), the latter have primarily been theorised as loosely coupled networks of digital organisations spanning industries and generating complex interdependencies and emergent properties, such as value-reinforcing complementarities (Adner, 2017; Hein et al., 2020). While valuable, this perspective has largely been used to explain the emergence and evolution of *business* ecosystems (L. Thomas et al., 2022), offering less insight into how platform growth transforms the wider environments in which further platformization unfolds.

In this paper, we advance *platformization* – a concept developed in adjacent fields but less established in IS – to address this gap. We argue that reintroducing ecological thinking into IS

(Márton, 2022), largely lost as the business ecosystems metaphor evolved (Alaimo & Kallinikos, 2024, p. 142), requires understanding platforms and their environments as co-evolving within complex adaptive systems (Holling, 2001). Drawing on Holling's adaptive cycle, we conceptualise platformization as a process of recursive co-evolution in which earlier transformations reshape the conditions for subsequent platform growth, competition, and ecosystemic capacities for adaptation. We illustrate this perspective through a case study of Barcelona's urban ecosystem, tracing successive waves of tourism platformization, particularly the expansion of short-term rental (STR) platforms such as Airbnb and their interactions with the city's housing system and regulatory institutions.

2. Platformization as recursive co-evolution

The focus on *platformization* rather than *platforms* serves several purposes. First, it highlights how organisations and industries acquire platform characteristics, for example, how fuel or machine production becomes digitalised and organised according to platform logic (Constantinides et al., 2018). Second, it foregrounds how platform-making involves configuring "people and things in constantly varying and experimentally modulated relations" (Mackenzie, 2018, p. 37). Third, it captures the "penetration of infrastructures, economic processes and governmental frameworks of digital platforms in different economic sectors and spheres of life" (Poell et al., 2019), as more social domains increasingly become dependent on platforms (Mikołajewska-Zając et al., 2022).

Platformization also directs attention beyond the growth of individual platforms. Van Dijck et al. (2018) conceptualise it as a societal transformation that unfolds through competition between platforms, foregrounding how the platform race for dominance reorganises broader social and economic domains. Van Dijck (2021) subsequently emphasises the hierarchical and interdependent character of this transformation, arguing that platformization needs to be understood across the multiple layers and sectors through which platforms become infrastructural and increasingly interconnected. Early accounts tended to see platformization as the expansion of platforms into new domains. Zuboff (2015, p. 78), for example, describes platforms' "incursion into undefended private territory until resistance is encountered." Other work has complicated this directional view by showing that platformization involves reciprocal and adaptive relationships between platforms and the environments they enter. Tiwana et al. (2010) conceptualise platform evolution explicitly as coevolution of platform architecture, governance and environmental dynamics. Helmond (2015) conceptualises platformization as a two-way process in which platforms extend their infrastructure into the wider Web, format external data according to platform logics, and bring it back into the platform. Caplan and boyd (2018) show how a dominant platform, Facebook, and a field it sought to subjugate, news organisations, mutually adapt: Facebook repeatedly recalibrated its relationship with the news industry in response to sector changes while pursuing its own organisational interests.

These perspectives reveal platformization as *dynamic*, *relational* and *expansive*, but they explain less about how it transforms the environment in which platforms operate and, therefore, in which further platformization occurs. We approach this question through ecological thinking, particularly the study of complex adaptive systems (Holling, 2001; see also Meadows, 2008). We understand ecosystems as unbounded, heterogeneous complexes of relationships and processes: information systems rather than metaphorical analogies to biological ecosystems (Márton, 2022). Following Bateson's (1972, p. 451) conception of the organism-and-its-environment as the unit of survival, we shift the unit of analysis from individual platforms to the platform-environment relationship. The viability of such systems depends on maintaining *flexibility*, understood as the capacity to change. We conceptualise platformization as a *process of recursive co-evolution*: platforms and their environments reshape one another, altering the conditions for further adaptation. Holling's (2001) adaptive cycle model allows us to examine how ecosystem configurations become established, disrupted and recombined over time, and how platformization alters the wider system's *flexibility*.

Holling provides a heuristic model of adaptive change in complex systems; we translate its core dynamics into a platformized setting. The adaptive cycle describes recurring dynamics of *exploitation*, *conservation*, *release* and *reorganisation*, through which systems – forests, coral reefs or organisations – mobilise potential, consolidate relationships, encounter disturbance, and generate new configurations. These dynamics reflect a movement between tendencies toward growth, accumulation, and stability and tendencies toward reorganisation in response to changing conditions and disturbances (Byrne, 2024). *Exploitation* involves mobilising available potential and establishing new relationships through which resources can be used and reproduced. *Conservation* is characterised by the accumulation of resources and connectedness as a system consolidates, potentially increasing stability while reducing flexibility and making it more susceptible to disturbances. *Release* can be triggered by shocks such as a pandemic or economic crisis, loosening established relationships and making resources and capabilities available for alternative configurations. *Reorganisation* opens a period of experimentation in which resources can be recombined and new relationships formed. These are not deterministic stages: a system may renew aspects of its previous configuration or develop a substantially different one, combining continuity and change (Gunderson & Holling, 2002; Holling, 2001; see Figure 1).

Applying the framework to platformized settings allows us to examine how platforms and their environments mobilise, accumulate, disrupt and recombine relationships among actors, resources and institutions. It directs attention to how platformization alters the adaptive capacities of these environments. Digital technologies can make resources and relationships more visible, searchable and connectable, while allowing configurations to be modified and reproduced at scale (Henfridsson & Bygstad, 2013; Kallinikos et al., 2013; Tilson et al., 2010; Yoo et al., 2010). We examine how these properties shape accumulation and growth alongside possibilities for reconfiguration within platformized ecosystems, and how the resulting adaptive capacities are distributed across actors and the wider system.

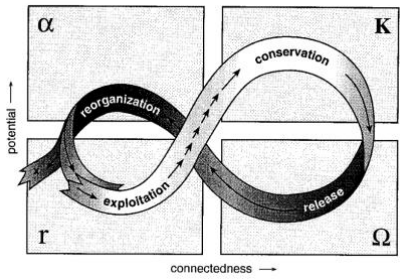


Figure 1. Adaptive cycle of complex systems, reproduced from Holling (2001).

3. Method

We conducted a heuristic case study (Thomas & Myers, 2015) of Barcelona's platformization. Such case studies are useful for tracing processes and identifying causal paths not captured by existing theorising (George & Bennett, 2005, p. 104). Barcelona offers a rich setting because it combines substantial tourism platformization across several waves with an early and sustained municipal response to STR growth. The city was among the first in Europe to introduce stringent restrictions on short-term rentals and later became a site of sustained regulatory experimentation, including adaptations that informed EU-level developments. With tourism platformization as our analytical focus, our ecological approach did not seek an exhaustive account of Barcelona's urban ecosystem; rather, we traced relevant patterns of change within it (Mikołajewska-Zając et al., 2022).

Our primary sources were news collected from the Factiva database (364 news items, retrieved based on combinations of keywords including "Barcelona", "tourism" or "tourists", and platform names), which provided contemporary accounts of developments in Barcelona's

tourism, housing and platform economy, including the actions and responses of platforms, property owners, residents, businesses and public authorities. We supplemented this material iteratively with existing academic research on Barcelona's STR market and with municipal, regional and EU policy documents, regulatory materials, official statistics and other relevant online sources. This combination allowed us to trace developments across market, social and regulatory domains and to triangulate accounts of changes in the urban ecosystem.

Drawing on Holling's (2001) ecological framework, we applied temporal bracketing (Langley, 1999) to trace changes in system configurations over time. Our interest lies in patterns and configurations at the ecosystem level, rather than in individual actors per se. Because such patterns become observable through the actions and interactions of actors, however, we traced how platforms, property owners, residents, businesses and public authorities responded to changing conditions and how these responses contributed to the formation and transformation of broader configurations. We traced changes in the mobilisation, consolidation, disruption and recombination of resources and relationships. We treated a transition as analytically defensible if converging evidence indicated a qualitative change in the dominant configuration and its dynamics, rather than a variation within the same pattern. This approach allowed us to use the adaptive cycle as a heuristic while remaining open to alternative trajectories and outcomes.

4. Tracing the platformization of Barcelona's ecosystem

Below, we reconstruct the platformization of tourism in Barcelona, tracing its successive waves in the context of local, regional and transnational developments. Drawing on Holling's adaptive cycle, we examine how successive configurations became established, disrupted and recombined, creating conditions for further platformization.

4.1. From tourism conservation to release: 1992-2008

Following the 1992 Olympic Games, Barcelona underwent a major urban and economic transformation that shifted the city further towards tourism and services after decades of industrial development and its subsequent decline (Cocola-Gant & Lopez-Gay, 2020). Major investment in transport, accommodation and the urban environment increased the city's capacity to receive visitors, while growing air connectivity and cruise traffic reinforced the expansion (Navarro-Ruiz et al., 2020; Vera-Rebollo & Ivars-Baidal, 2009). The number of hotels increased from 118 in 1990 to around 300-320 in 2008, while annual visitors rose from 1.73 million to 7.5 million (Cocola-Gant & Lopez-Gay, 2020). Tourism increasingly became a central driver of urban growth, supported by a local public-private coalition that presented tourism-led development as beneficial to the city as a whole (Gebhardt, 2017). By the late 2000s, tourism, transport, accommodation and property had become increasingly interconnected, with tourism estimated to account for 10-12% of Barcelona's economy (#032). In Holling's terms, growing potential was increasingly bound into stable relationships, with rising connectedness supporting the *conservation* of the tourism-oriented configuration.

Digital intermediation became layered onto this configuration. Airline distribution was increasingly organised through Global Distribution Systems (GDSs), while online travel agencies (OTAs) brought flights and accommodation into searchable and increasingly self-service digital marketplaces (Boberg & Collison, 1985; Buhalis & Law, 2008; Chismar & Meier, 1992). Booking.nl established one of its first foreign offices in Barcelona in 2004 (#010), connecting local accommodation providers to geographically dispersed demand. Online booking attracted new segments of travellers, particularly younger (Cócola-Gant, 2016; Cocola-Gant et al., 2026), more digitally oriented consumers taking shorter city-break trips, further expanding demand for Barcelona (Castillo-Manzano & López-Valpuesta, 2010). Consumer practices shifted towards online comparison and self-booking, contributing to the

wider digitalisation of travel distribution (Quintana et al., 2016; Reinhold et al., 2022). The first wave of platformization therefore reinforced *conservation* by increasing the connectivity between Barcelona's existing tourism capacity and global demand.

At the same time, first vacation rentals *began to cross the boundary between tourism and residential housing*. Apartments converted for tourist use expanded after the Olympics, especially during the 2000s, attracting growing real-estate investment and giving rise to specialist intermediaries (Cócola-Gant, 2016; Cocola-Gant et al., 2026). By 2005, Barcelona's tourism statistics recorded 81 companies operating tourist apartments, while tourist-rental units in Ciutat Vella approached 2,000 (Cocola-Gant et al., 2026). The market was already digitally mediated and increasingly professionalised, but remained fragmented across owners, specialist agencies and emerging vacation-rental websites. Its growth generated tensions around noise, neighbourhood change and housing, while exposing a regulatory mismatch between tourism activity and residential land use (Cócola-Gant, 2016; Cocola-Gant et al., 2026). Seeking to mitigate these effects, Catalonia introduced a register of tourist apartments (HUTs) in 1998, and Barcelona began requiring permission to convert residential units to tourist use in the historic Ciutat Vella in 2005 (Gyódi et al., 2025). Housing began to form new relationships with the tourism economy, albeit at a relatively small scale, creating potential for this configuration to develop further.

The global financial crisis of 2008 disrupted this increasingly consolidated configuration. Particularly severe in Spain, the crisis brought the property and construction boom to an abrupt halt, concentrated mortgage foreclosures in vulnerable metropolitan areas, and drove up unemployment and housing insecurity (Bonshoms-Guzmán, 2023; Gutiérrez & Domènech, 2018; Méndez Gutiérrez Del Valle & Tabasco, 2014). Tourism demand also contracted, loosening the relationships that had bound housing, construction, tourism and international mobility into the previous configuration. The crisis marked the *transition from conservation to release*: established relationships were disrupted, while housing and other resources remained available for alternative uses. This increased the system's *potential* for reorganisation without determining which new configurations would emerge.

4.2. Reorganization and exploitation: platformizing residential housing, 2008-14

The financial crisis *released* many resources from the ever-more consolidated tourism–property configuration, creating scope for their recombination. Tourism demand initially declined, while unemployment and housing insecurity increased sharply (Bonshoms-Guzmán, 2023; Gutiérrez & Domènech, 2018). Yet the disruption did not eliminate the capabilities and infrastructures accumulated during the previous cycle. Tourism remained a potential source of demand, while housing became available for alternative uses. Against this backdrop, Airbnb, founded in 2008 and launched in Spain soon after, was greeted in Barcelona with an "almost desperate enthusiasm" by some of the residents facing unemployment (#116). By the time Airbnb entered Barcelona, vacation rentals were already an existing market, with digital intermediaries such as VRBO and Niumba connecting residential properties to tourist demand (Cocola-Gant et al., 2026). Airbnb did not create this market, but its commission-based model lowered the financial barriers to participation and helped scale the enrolment of residential housing into tourism (Cocola-Gant & Janoschka, 2026). *Reorganisation* thus opened possibilities for broader participation in an already digitally mediated STR market.

As tourism demand recovered, the platform-mediated tourist use of residential housing expanded rapidly. Between 2010 and 2014, the number of registered STRs in Barcelona grew by 320%, from around 2,300 to 9,660 (Cocola-Gant et al., 2026). Participation became increasingly concentrated among professional operators: by 2012, more than half of Barcelona's STRs were controlled by multi-listing hosts (#022), while properties were frequently multi-homed, or advertised across several competing platforms (Garcia-López et al., 2020). STRs were thus increasingly integrated into real-estate investment, rather than serving primarily as supplementary household income, as Airbnb's original promise suggested.

Research linked their expansion to rising property values, particularly in tourism-intensive neighbourhoods, and to residential displacement (Garcia-López et al., 2020; Milano et al., 2023).

This expansion prompted regulatory adaptation. Catalonia reiterated that STRs fell under the existing HUT regulations (#005, #028). In 2010, the HUT reform introduced a registration system (#001). Entire-home rentals were permissible when compliant with HUT requirements, while renting individual rooms remained prohibited (#028). Barcelona progressively tightened controls as conflicts around housing and neighbourhood life intensified (#012; #015; Gyódi et al., 2025). The municipality sought to contain the conversion of residential housing into tourist accommodation and preserve a distinction between residential and tourist uses. In 2014, Barcelona suspended new HUT authorisations in high-pressure areas, while Catalonia fined Airbnb and other platforms for advertising unlicensed accommodation as the first EU municipality to do so (#006). Airbnb responded by mobilising hosts and emphasising its economic contribution, including €128 million in host income and more than 4,000 jobs between mid-2013 and mid-2014 (#006; #011). Platform growth generated responses across the ecosystem: owners and platforms sought to preserve the emerging market, residents mobilised against its effects, incumbent tourism businesses contested it, and public authorities developed new ways to regulate it. By 2014, investment and organised contestation had coalesced around the platform-mediated tourist use of housing, laying the foundations for the *conservation* phase that followed.

4.3. From STR exploitation to conservation, 2015-20

By 2015, the housing–tourism configuration had become sufficiently established that actors increasingly sought either to defend and expand it or to contest and regulate it. Protests against STRs and other forms of touristification became intertwined with broader mobilisation around housing and residents' rights, contributing to the election of Barcelona en Comú and Ada Colau as Barcelona's mayor in 2015. STR regulation became a central political issue, and the municipality suspended new permits, increased inspections and strengthened enforcement capacity (#014; #029). A 2016 municipal study estimated that around 40% of tourist-use dwellings were unlicensed (#033), while 78% of Airbnb listings reportedly lacked the required registration number (Aguilera et al., 2021; Colomb & Moreira de Souza, 2024).

Property-level enforcement proved resource-intensive and difficult to scale: approximately 6,000 inspections in 2015 resulted in only 38 closures of illegal tourist apartments (#029). Learning from this, the municipality sought to establish platform accountability (Rahman et al., 2024) by requiring intermediaries to restrict non-compliant listings. In 2016, it threatened fines of up to €600,000 for platforms repeatedly advertising illegal accommodation (Colomb & Moreira de Souza, 2024). As Colau argued, "an internet platform cannot become a means to thwart the regulations and to shelter illegal tourist apartments" (#015).

STR platforms responded differently to these regulatory pressures. Several agreed not to advertise unlicensed properties, assuming a gatekeeping role (Colomb & Moreira de Souza, 2024; European Commission, 2018; #142). Airbnb initially resisted and challenged municipal sanctions (#018; #028). In 2017, it agreed to remove unlicensed listings identified by the municipality (#062), leaving the public authorities responsible for detecting violations, while illegal listings continued to reappear (#056). Enforcement came to depend partly on platforms' willingness to control access to their marketplaces.

Meanwhile, professionalisation continued as specialised property managers and large-scale landlords developed increasingly sophisticated ways of operating either within or around regulatory constraints (Aguilera et al., 2021; Bosma & van Doorn, 2024). Some circumvented regulations through fraudulent rental arrangements, prompting Barcelona's Deputy Mayor to describe them as "organised groups with mafia practices" (#140). Yet increasingly forceful enforcement also affected small-scale hosts. From 2017, an association of residents contested

fines of €30,000–€60,000 for occasionally renting their homes without licences. They argued that municipal enforcement failed to distinguish occasional hosts from professional operators and that Airbnb had not adequately explained regulatory risks (#074; #102). Small-scale hosts found themselves caught between competing forms of platform and regulatory orchestration.

Alongside platform-oriented enforcement, the municipality also adapted its tourism regulation. In 2017, it established zero net growth in STRs, prohibiting new licences in the most tourism-intensive areas while allowing limited redistribution towards less saturated zones (Gyódi et al., 2025). This sought to contain the expansion of tourist accommodation into residential neighbourhoods and reduce its concentration in heavily affected areas. Yet Barcelona's regulatory efforts were constrained by EU-level legal arrangements. Airbnb lobbied the European Commission for rules favourable to its business model (Haar, 2018; #095). In 2019, the Court of Justice of the EU classified Airbnb's intermediation as an "information society service" (#134), reinforcing the legal limits on imposing platform accountability for listings. Barcelona responded by joining nine other European cities in demanding stronger EU rules on platform accountability and access to the data needed for local enforcement (#120). The regulatory conflict extended beyond the municipality to the European level.

By 2019, approximately 22,200 STR offers, including entire properties and individual rooms, were advertised across major platforms, equivalent to around 60% of Barcelona's hotel-room stock (#140). Research documented residential displacement and the spread of touristification into surrounding neighbourhoods (Aguilera et al., 2021; Garcia-López et al., 2020; Jover & Cocola-Gant, 2023), alongside changes in neighbourhood services and social infrastructures (Cocola-Gant & Lopez-Gay, 2020; #128). In ecological terms, *conservation* involved the consolidation of interconnected and antagonistic relationships, as actors continually adapted their practices around an entrenched housing–tourism configuration. Successive waves of platformization deepened the *connectedness* of tourism, housing and digital infrastructures, and reinforced the city's dependence on international visitor flows, potentially constraining the wider ecosystem's *flexibility* in responding to disturbances.

4.4. From COVID-19 release to regulatory consolidation: 2020–2025

The COVID-19 pandemic abruptly interrupted the international visitor flows on which Barcelona's tourism ecosystem depended, triggering a new phase of *release*. In 2020, the wider Barcelona destination recorded 75.1% fewer tourists and 73.3% fewer overnight stays than in 2019, while international airport passenger traffic declined by 82.3% (Observatori del Turisme a Barcelona, 2021). Yet the shock did not destroy the resources and infrastructures accumulated during *conservation*. Housing units, HUT licences, platform accounts, and intermediary organisations remained in place, while the collapse in demand loosened the relationships that connected them to international tourism. The *release* temporarily uncoupled housing from tourist use, creating opportunities for alternative configurations.

The ensuing *reorganisation* involved competing efforts to redirect these resources (#155). By July 2020, approximately 3,700 of Barcelona's 9,600 licensed tourist-use dwellings had reportedly moved into conventional residential rental. By late 2020, around 5,900 units were being offered for rental, although most through seasonal contracts rather than permanent residential tenancies (Morales-Pérez et al., 2020). Other owners retained their tourist listings and lowered prices to attract the remaining demand. The municipality, meanwhile, sought to convert the temporary disruption into a more durable reallocation of housing, offering owners up to €1,200 per month to make their properties available for social rental. Yet only 22 units entered the programme by late 2020 (#157).

Airbnb also adapted to the disruption by promoting longer stays, remote work and travel beyond major urban destinations. In Spain, stays of 28 days or more accounted for around 30% of nights booked in early 2021, while demand shifted from Barcelona and Madrid towards rural and smaller coastal destinations (Airbnb, 2021; #165). These adaptations enabled the

platform to reconnect its existing accommodation supply with emerging forms of demand, extending its infrastructure beyond conventional short-term tourism.

The resulting *reorganisation* remained open-ended. While the municipality sought to return housing to residential use, property owners retained the *flexibility* to move between rental markets, and platforms developed new ways of connecting accommodation to changing demand. Importantly, this flexibility was itself a product of earlier platformization, which had made residential properties more readily marketable across different markets. Yet the resulting *adaptive capacity* was unevenly distributed, favouring property owners able to redirect their assets while offering fewer opportunities to residents seeking secure housing.

As tourism recovered and resumed growth, the possibilities opened during *reorganisation* were selectively *exploited*. Housing was increasingly reconnected to tourist demand, while seasonal and medium-term rentals expanded, allowing property owners and platforms to capitalise on the flexibility revealed during the pandemic. Unlike the earlier exploitation phase, however, tourism recovery unfolded within a much more bounded licensed STR stock. The subsequent *conservation* phase saw the consolidation of regulatory arrangements parallel with the growing mobilisation by platforms and property owners to defend the established STR market. In 2024, Barcelona announced plans to phase out all licensed tourist-use dwellings by the end of 2028 (#244), aiming to deliberately *release* housing resources from their connection to tourism.

5. Conclusion

In this paper, we conceptualise *platformization* as a process of recursive co-evolution between digital platforms and their environments, drawing on Holling's (2001) adaptive cycle to examine how this process unfolds over time. Through the case of Barcelona, we trace how successive waves of tourism platformization transformed configurations of actors, resources and institutions, generating *opportunities for adaptation* alongside *new dependencies and vulnerabilities*. We contribute to IS scholarship in three ways.

First, we extend understandings of *platformization* as an expansive, relational and adaptive process by foregrounding its *recursive character*. We show how successive *waves* of platformization build on and transform configurations established by earlier ones, with the first wave of tourism platformization helping prepare the ground for the subsequent expansion of platform-mediated short-term rentals and creating conditions for further platform growth and competition. In doing so, we respond to recent calls for stronger process theorising in IS, particularly greater attention to recursive dynamics and interactions between processes (Wiener et al., 2025). Holling's adaptive cycle provides a heuristic for tracing how these configurations become established, disrupted and recombined, without assuming a linear trajectory. Following van Dijck's (2021) call to “see the forest for the trees,” this perspective shifts attention beyond individual platforms towards the cumulative transformation of wider social and economic environments.

Second, we extend IS scholarship on the generative properties of digital technology by examining *how platformization redistributes adaptive capacities within ecosystems*. Digital platforms expand possibilities for *recombining* resources across different uses, but these possibilities may be *unevenly distributed*. In Barcelona, for instance, property owners gained the *flexibility* to redirect housing between rental markets, while residents faced growing housing insecurity. Such asymmetries in adaptive capacities may help sustain established configurations, contributing to *rigidity* at the ecosystem level. This raises questions about the ecological limits of digital growth (Márton, 2022) and the implications of platformization for systemic *resilience*.

Third, we contribute to IS research on *platform governance* by showing how regulation itself becomes transformed through adaptation to platformized markets. Barcelona's authorities came to rely on platform gatekeeping capacities and sought access to platform-controlled data

to regulate STRs, potentially extending regulatory reach while creating dependencies on the very intermediaries being regulated. The *platformization of regulation* thus reveals a tension: efforts to constrain platformization may reinforce the infrastructural importance of platforms. Barcelona's subsequent regulatory adaptations also illustrate attempts to reduce this dependence through alternative forms of intervention. This tension is particularly relevant to the EU's new STR registration and data-sharing framework, which institutionalises platform participation in regulatory enforcement.

These insights invite further research into the conditions under which platformization enables or constrains adaptation across interconnected ecosystems. Barcelona's planned phase-out of STR licences raises the possibility of deliberately inducing *release* from established platform-mediated configurations. Whether such interventions can generate more adaptive alternatives remains an open question.

6. References

Adner, R. (2017). Ecosystem as structure: An actionable construct for strategy. *Journal of Management*, *43*(1), 39–58. https://doi.org/https://doi.org/10.1177/0149206316678451

Aguilera, T., Artioli, F., & Colomb, C. (2021). Explaining the diversity of policy responses to platform-mediated short-term rentals in European cities: A comparison of Barcelona, Paris and Milan. *Environment and Planning A*, *53*(7), 1689–1712. https://doi.org/10.1177/0308518X19862286

Airbnb. (2021). *Airbnb report on travel & living*. https://news.airbnb.com/wp-content/uploads/sites/4/2021/05/Airbnb-Report-on-Travel-Living-.pdf

Alaimo, C., & Kallinikos, J. (2024). *Data rules: Reinventing the market economy*. MIT Press.

Bateson, G. (1972). *Steps to an ecology of mind*. Ballantine Books.

Boberg, K. B., & Collison, F. M. (1985). Computer reservation systems and airline competition. *Tourism Management*, *6*(3), 174–183. https://doi.org/10.1016/0261-5177(85)90034-2

Bonshoms-Guzmán, J. (2023). The housing/financial complex in Spain: After the 2008 global financial crisis. *International Journal of Urban and Regional Research*, *47*(6), 900–916. https://doi.org/10.1111/1468-2427.13172

Bosma, J. R., & van Doorn, N. (2024). The gentrification of Airbnb: Closing rent gaps through the professionalization of hosting. *Space and Culture*, *27*(1), 31–47. https://doi.org/10.1177/12063312221090606

Buhalis, D., & Law, R. (2008). Progress in information technology and tourism management: 20 years on and 10 years after the Internet—The state of eTourism research. *Tourism Management*, *29*(4), 609–623. https://doi.org/10.1016/J.TOURMAN.2008.01.005

Byrne, A. Q. (2024). What can frogs teach us about resilience? Adaptive renewal in amphibian and academic ecosystems. *Integrative and Comparative Biology*, *64*(3), 795–806. https://doi.org/10.1093/ICB/ICAE058

Caplan, R., & boyd, d. (2018). Isomorphism through algorithms: Institutional dependencies in the case of Facebook. *Big Data & Society*, *5*(1), 1–12. https://doi.org/https://doi.org/10.1177/2053951718757253

Castillo-Manzano, J. I., & López-Valpuesta, L. (2010). The decline of the traditional travel agent model. *Transportation Research Part E: Logistics and Transportation Review*, *46*(5), 639–649. https://doi.org/10.1016/J.TRE.2009.12.009

Chismar, W. G., & Meier, J. (1992). A model of competing interorganizational systems and its application to airline reservation systems. *Decision Support Systems*, *8*(5), 447–458. https://doi.org/10.1016/0167-9236(92)90028-N

Ciborra, C. (1996). The platform organization: Recombining strategies, structures, and surprises. *Organization Science*, *7*(2), 103–118. https://doi.org/10.1287/orsc.7.2.103

Cocola-Gant, A. (2016). Holiday rentals: The new gentrification battlefront. *Sociological Research Online*, *21*(3). https://doi.org/10.5153/sro.4071

Cocola-Gant, A., Hodkinson, S. N., & Janoschka, M. (2026). The financialisation of short-term rentals in Barcelona: Property ownership and the consolidation of a new asset class. *European Urban and Regional Studies*, *33*(3), 300–317. https://doi.org/10.1177/09697764251361724

Cocola-Gant, A., & Janoschka, M. (2026). PropTech in the short-term rental industry. The digital infrastructure behind Airbnb. *International Journal of Housing Policy*. https://doi.org/10.1080/19491247.2026.2657981

Cocola-Gant, A., & Lopez-Gay, A. (2020). Transnational gentrification, tourism and the formation of 'foreign only' enclaves in Barcelona. *Urban Studies*, *57*(15), 3025–3043. https://doi.org/10.1177/0042098020916111

Colomb, C., & Moreira de Souza, T. (2024). Illegal short-term rentals, regulatory enforcement and informal practices in the age of digital platforms. *European Urban and Regional Studies*, *31*(4), 328–345. https://doi.org/10.1177/09697764231155386

Constantinides, P., Henfridsson, O., & Parker, G. G. (2018). Introduction—Platforms and Infrastructures in the Digital Age. *Information Systems Research*, *29*(2), 381–400.

De Reuver, M., Sørensen, C., & Basole, R. C. (2018). The digital platform: A research agenda. *Journal of Information Technology*, *33*(2), 124–135. https://doi.org/10.1057/S41265-016-0033-3

European Commission. (2018). *Study on the assessment of the regulatory aspects affecting the collaborative economy in the tourism accommodation sector in the 28 member states. Task 4, market case study - Barcelona*. https://doi.org/10.2873/87602
Vera-Rebollo, J. F., & Ivars-Baidal, J. A. (2009). Spread of low-cost carriers: Tourism and regional policy effects in Spain. *Regional Studies*, *43*(4), 559–570. https://doi.org/10.1080/00343400701874164
Garcia-López, M. À., Jofre-Monseny, J., Martínez-Mazza, R., & Segú, M. (2020). Do short-term rental platforms affect housing markets? Evidence from Airbnb in Barcelona. *Journal of Urban Economics*, *119*. https://doi.org/10.1016/j.jue.2020.103278
Gebhardt, D. (2017). Barcelona: Regulating the tourism growth-machine? *Geographische Zeitschrift*, *105*(3–4), 225–248. https://doi.org/10.25162/GZ-2017-0009
George, A. L., & Bennett, A. (2005). *Case studies and theory development in the social sciences*. The MIT Press.
Gunderson, L. H., & Holling, C. S. (2002). *Panarchy: Understanding transformations in human and natural systems*. Island Press.
Gutiérrez, A., & Domènech, A. (2018). The mortgage crisis and evictions in Barcelona: identifying the determinants of the spatial clustering of foreclosures. *European Planning Studies*, *26*(10), 1939–1960. https://doi.org/10.1080/09654313.2018.1509945
Gyódi, K., Mazur, J., & Cocola-Gant, A. (2025). Barcelona as a case study for the effectiveness of short-term rental market regulations. *Cities*, *162*. https://doi.org/10.1016/j.cities.2025.105915
Haar, K. (2018). *UnFairbnb. How online rental platforms use the UE to defeat cities' affordable housing measures*. https://corporateeurope.org/sites/default/files/unfairbnb.pdf
Hein, A., Schreieck, M., Riasanow, T., Setzke, D. S., Wiesche, M., Böhm, M., & Krcmar, H. (2020). Digital platform ecosystems. *Electronic Markets*, *30*(1), 87–98. https://doi.org/10.1007/S12525-019-00377-4/TABLES/1
Helmond, A. (2015). The platformization of the Web: Making Web data platform ready. *Social Media + Society*, *1*(2), 205630511560308. https://doi.org/10.1177/2056305115603080
Henfridsson, O., & Bygstad, B. (2013). The generative mechanisms of digital infrastructure evolution. *Source: MIS Quarterly*, *37*(3), 907–931.
Holling, C. S. (2001). Understanding the complexity of economic, ecological, and social systems. *Ecosystems*, *4*(5), 390–405. https://doi.org/10.1007/s10021-001-0101-5
Huang, J., Henfridsson, O., Liu, M. J., & Newell, S. (2017). *Growing on steroids: Rapidly scaling the user base of digital ventures through digital innovation*. *41*(1), 301–314.
Jacobides, M. G., Cennamo, C., & Gawer, A. (2018). Towards a theory of ecosystems. *Strategic Management Journal*, *39*(8), 2255–2276. https://doi.org/10.1002/SMJ.2904
Jover, J., & Cocola-Gant, A. (2023). The political economy of housing investment in the short-term rental market: Insights from urban Portugal. *Antipode*, *55*(1), 134–155. https://doi.org/10.1111/ANTI.12881
Kallinikos, J., Aaltonen, A., & Márton, A. (2013). The ambivalent ontology of digital artifacts. *MIS Quarterly*, *37*(2), 357–370. https://www.jstor.org/stable/43825913
Langley, A. (1999). Strategies for theorizing from process data. *The Academy of Management Review*, *24*(4), 691–710. https://doi.org/https://doi.org/10.2307/259349
Mackenzie, A. (2018). 48 million configurations and counting: platform numbers and their capitalization. *Journal of Cultural Economy*, *11*(1), 36–53. https://doi.org/10.1080/17530350.2017.1393443
Márton, A. (2022). Steps toward a digital ecology: ecological principles for the study of digital ecosystems. *Journal of Information Technology*, *37*(3), 250–265. https://doi.org/10.1177/02683962211043222
Meadows, D. H. (2008). *Thinking in systems: A primer*. Chelsea Green Publishing.
Méndez Gutiérrez Del Valle, R., & Tabasco, J. P. (2014). The housing crisis and foreclosures in Spain: A geographical approach. *Boletin de La Asociacion de Geografos Espanoles*, *71*, 481–486.
Mikołajewska-Zając, K., Márton, A., & Zundel, M. (2022). Couchsurfing with Bateson: An ecology of digital platforms. *Organization Studies*, *43*(7), 1115–1135. https://doi.org/https://doi.org/10.1177/01708406211058628
Milano, C., González-Reverté, F., & Benet Mòdico, A. (2023). The social construction of touristification. Residents' perspectives on mobilities and moorings. *Tourism Geographies*, *25*(4), 1273–1291. https://doi.org/10.1080/14616688.2022.2150785
Morales-Pérez, S., Garay, L., & Wilson, J. (2020). Airbnb's contribution to socio-spatial inequalities and geographies of resistance in Barcelona. *Tourism Geographies*, 1–24. https://doi.org/10.1080/14616688.2020.1795712
Navarro-Ruiz, S., Casado-Díaz, A. B., & Ivars-Baidal, J. (2020). Cruise tourism: the role of shore excursions in the overcrowding of cities. *International Journal of Tourism Cities*, *6*(1), 197–214. https://doi.org/10.1108/IJTC-04-2018-0029
Observatori del Turisme a Barcelona. (2021). *Informe de l'activitat turística a Barcelona 2020*. https://observatoriturisme.barcelona/wp-content/uploads/pdf/2020_IAT20.pdf
Poell, T., Nieborg, D., & van Dijck, J. (2019). Platformisation. *Internet Policy Review*, *8*(4). https://doi.org/10.14763/2019.4.1425
Quintana, T. A., Gil, S. M., & Peral, P. P. (2016). How could traditional travel agencies improve their competitiveness and survive? A qualitative study in Spain. *Tourism Management Perspectives*, *20*, 98–108. https://doi.org/10.1016/J.TMP.2016.07.011
Rahman, H. A., Karunakaran, A., & Cameron, L. D. (2024). Taming platform power: Taking accountability into account in the management of platforms. *Academy of Management Annals*, *18*(1), 251–294. https://doi.org/10.5465/annals.2022.0090

Reinhold, S., Zach, F. J., & Laesser, C. (2022). E-business models in tourism. *Handbook of E-Tourism*, 1181–1210. https://doi.org/10.1007/978-3-030-48652-5_71/TABLES/2
Thomas, G., & Myers, K. (2015). *The anatomy of the case study*. SAGE. https://doi.org/10.4135/9781473920156
Tilson, D., Lyytinen, K., & Sørensen, C. (2010). Research commentary—Digital infrastructures: The missing IS research agenda. *Information Systems Research*, *21*(4), 748–759. https://doi.org/10.1287/ISRE.1100.0318
Tiwana, A., Konsynski, A., & Bush, A. A. (2010). Platform evolution: Coevolution of platform architecture, governance, and environmental dynamics. *Information Systems Research*, *21*(4), 675–687.
Van Dijck, J. (2021). Seeing the forest for the trees: Visualizing platformization and its governance. *New Media and Society*, *23*(9), 2801–2819. https://doi.org/10.1177/1461444820940293/
van Dijck, J., Poell, T., & De Waal, M. (2018). *The platform society. Public values in a connective world*. Oxford University Press.
Yoo, Y., Henfridsson, O., & Lyytinen, K. (2010). The new organizing logic of digital innovation: An agenda for information systems research. *Information Systems Research*, *21*(4), 724–735.
Zuboff, S. (2015). Big other: Surveillance capitalism and the prospects of an information civilization. *Journal of Information Technology*, *30*(1), 75–89. https://doi.org/10.1057/JIT.2015.5

7. Appendix

Source ID	Reference
#001	El Govern aprova el decret que regula els habitatges d'ús turístic. (2010, Nov 9). *Generalitat de Catalunya*. https://govern.cat/gov/notes-premsa/55392/govern-aprova-decret-que-regula-habitatges-us-turistic
#005	Airbnb fined in Catalonia over short-term room lets. (2014, July 7). *Reuters News.*
#006	Crellin, O., and Romá, D. (2014, Aug 7). Catalonia Fines Airbnb, Threatens to Block Locals From Using Site. *Dow Jones Institutional News.*
#010	Booking.com's Gillian Tans: How a travel executive travels. (2015, Apr 30). *Financial Review.*
#012	Paris confronts Airbnb's rapid growth. (2015, Jun 26). *Dow Jones Institutional News.*
#013	Airbnb-style listings surpass hotel rooms in Spain: Study. (2015, Jun 26). *Agence France Presse.*
#014	Spain's tourism magnet Catalonia to tax Airbnb rentals. (2015, Jul 15). *Agence France Presse.*
#015	Croft, A. (2015, Aug 26). Barcelona mayor's tourism crackdown puts Airbnb in firing line. *Reuters News.*
#018	Barcelona city hall fines Airbnb, Homeaway. (2015, Dec 21). *Agence France Presse.*
#022	EU cautions governments against banning Uber, Airbnb (2016, May 31). *Reuters News.*
#028	Barcelona to raise fine for illegal home-rental lodgings to 600,000 euros. (2016, Jun 28). *Agence France Presse*.
#029	Barcelona threatens Airbnb and others with fines of up to €600,000. (2016, Jul 7). *Spanish Property Insight.*
#030	Chayka, K. (2016, Aug 8). Welcome to AirSpace. *The Verge.*
#032	Barcelona Strategic Tourism Plan for 2020. (2020, Sept). *Ajuntament de Barcelona*. https://ajuntament.barcelona.cat/turisme/sites/default/files/diagnosi_estrategica_v_1_ang.pdf
#033	Lomas, N. (2016, Sept 19). Airbnb faces fresh crackdown in Barcelona as city council asks residents to report illegal rentals. *TechCrunch.*
#054	Spanish court annuls Airbnb fine. (2016, Dec 12). *Agence France Presse*.
#056	Chapman, B. (2017, Jan 3). Airbnb to lose £25m in London bookings after holiday rental rules tighten. *The Independent*.
#062	Barcelona, Airbnb reach deal after months of conflict. (2017, Dec 7). *Agence France Presse.*
#065	Spain's Balearic Islands to fine illegal tourist rentals. (2017, Aug 8). *Reuters News.*
#074	Blanchar, C. (2017, Oct 27). Las multas por alquiler turístico sublevan a pequeños propietarios. *El País.*
#095	European tourist magnets hit back as Airbnb turns 10. (2018, Jul 4). *Agence France Presse.*
#102	Blanchar, C. (2018, Oct 17). Los anfitriones llevan Airbnb a los tribunales. *El País.*
#116	Mead, R. (2019, Apr 22). The Airbnb Invasion of Barcelona. *New Yorker.*
#117	Airbnb should be free to operate across Europe: EU court adviser. (2019, Apr 30). *Reuters News.*
#120	Henley, J. (2019, June 20). Ten cities ask EU for help to fight Airbnb expansion. *The Guardian.*
#128	Wisniewska, A. (2019, Sep 5). Are Airbnb investors destroying Europe' cultural capitals? *Financial Times.*
#134	EU court rules out Airbnb's rule as a real estate company. (2019, Dec 19). *CE Noticias Financieras.*
#140	Barcelona is amaga with "higher and strongest" fines to Airbnb, Homeaway and Booking. (2020, Feb 4). *CE Noticias Financieras.*
#142	Airbnb, Expedia and Others Agree to Landmark Data-Sharing Deal With European Union. (2020, Mar 6). *iCrowdNewswire.*
#155	Minder, R. and G. Abdul (2020, Oct 20). With Vacation Rentals Empty, European Cities See a Chance to Reclaim Housing. *New York Times.*
#157	Blanchar, C. (2020, Nov 8). Los dueños de los pisos turísticos se resisten a pasar a alquiler social. *El País*.
#244	Barcelona Moves to End Tourist Rental Licenses in Blow to Airbnb. (2024, Jun 21). *Dow Jones Institutional News.*

How Modular Data Products Emerge: A Configurational Analysis of Analytics Projects

Martin Chesbrough · Sean Maynard · Michael Davern

University of Melbourne

Humza Naseer

Deakin University

Abstract

Up to 85 percent of analytics initiatives fail to move beyond proof-of-concept, producing projects rather than modular data products that can be recomposed as conditions change. Modularity is therefore a question of information resilience: whether an organization's analytical estate can absorb disruption and be reconfigured. While the data mesh discourse prescribes what organizations should do, it does not explain under what conditions modular data products emerge. Drawing on 14 semi-structured interviews yielding 33 project-level narratives, this paper integrates the Gioia method with fuzzy-set qualitative comparative analysis (fsQCA) to identify configurational conditions for modularity. Our results contribute to the literature by (1) differentiating modular data products from data-as-a-product and data applications; (2) showing that modularity emerges through equifinal configurations, with data product engineering and transition dynamics forming the causal core; and (3) extending modularity theory into analytics by showing how design rules are created and stabilized through sociotechnical conditions.

1. Introduction

Organizations invest between two and five percent of annual revenue in data and analytics capabilities (Gartner, 2024), yet between 60 and 85 percent of analytics initiatives fail to move beyond the proof-of-concept stage (Davenport & Bean, 2020). This failure rate is commonly attributed to technical or adoption shortcomings, but a less-examined explanation lies in the artifacts these initiatives produce, which rarely achieve the separability, stable interfaces and capacity for re-composition that would make them durable, reusable and resilient to disruption. Most analytics initiatives produce projects, not products (Marchand & Peppard, 2013).

The data mesh movement (Dehghani, 2022) has foregrounded the idea that data should be owned, governed and served as a product. While sociotechnical in principle, this discourse is primarily normative: it prescribes what organizations should do without explaining under what conditions modular and reusable data products emerge from analytics work. Modularity itself is well-theorized in the design literature, where Baldwin and Clark (2000) established that modular architectures arise from deliberate design choices that decouple components along stable interfaces, and the IS literature has extended this logic to software platforms and digital innovation (Tiwana et al., 2010; Yoo et al., 2010). Neither stream explains the lack of durability and resilience.

Analytics projects are bounded and goal-directed initiatives, which represent the primary organizational mechanism through which data products are built and deployed. Success or failure is understood (Reggio & Astesiano, 2020) while research hints at reuse and sustained engagement outcomes through data platforms (Davenport & Kudyba, 2016). In sociotechnical settings such outcomes are rarely attributable to a single factor (Henderson & Clark, 1990; Orlikowski, 1992), arising instead from combinations of architectural, methodological, organizational and transitional conditions (Meyer et al., 1993). While the relationship between analytics projects and data product outcomes has received growing attention (Hasan & Legner, 2023), the conditions explaining modular outputs remain underexplored. We therefore ask: **how do modular data products emerge from analytics projects?**

To answer this, we conducted expert interviews with analytics practitioners and applied a configurational approach to 33 cases that arose from the interviews. This paper makes three contributions. First, we clarify modular data products as a theoretically distinct construct, differentiating them from data-as-a-product governance principles and data applications. Second, we apply configurational logic to data product modularity, showing that equifinal combinations of conditions produce modular outcomes while the absence of modularity is explained by structurally different configurations. Third, we extend modularity theory into analytics by identifying the combinations of conditions under which modular data products emerge, and in doing so specify a mechanism through which analytics work builds information resilience.

2. Theoretical Background

This section first defines modular data products as the outcome of interest, then positions analytics projects as the organizational setting in which that outcome is produced, and finally argues that configurational theory offers the causal logic suited to explaining it.

2.1. Data Products and Modularity

"Data product" entered academic usage from practitioner discourse without a stable theoretical definition. Loukides (2011) framed applications where data is the core value proposition; Patil (2012) broadened this to organizational practices for data-driven value; and data mesh (Dehghani, 2022) recast data management around domain ownership, data-as-a-product, self-serve platforms, and federated governance. This literature conflates governance principles, artifact types, and applications (Fig. 1): data-as-a-product prescribes ownership and management but not artifact properties; data application denotes analytics-powered end-user applications rather than the analytical artifact itself; and data product, narrowly defined, denotes an artifact that is self-contained and consumption ready.

This paper adopts a narrower construct, namely modular data products: outcomes comprising separable, re-composable components with stabilized interfaces that support co-ordinated use and consumption across multiple organizational contexts. Drawing on the modularity literature (Baldwin & Clark, 2000; Sanchez & Mahoney, 1996; Tiwana et al., 2010), modularity implies decomposability into functionally independent components (datasets, models, pipelines),

interfaces (schemas, APIs, data contracts) that stabilize through social enactment, and recombination in configurations the original designers did not anticipate. Modularity is not "success": a data product may be successful without being modular, but a modular data product is designed and sustained for reuse beyond its originating context. A finding consistent with organization resilience (Al-Ghattas & Marjanovic, 2021).

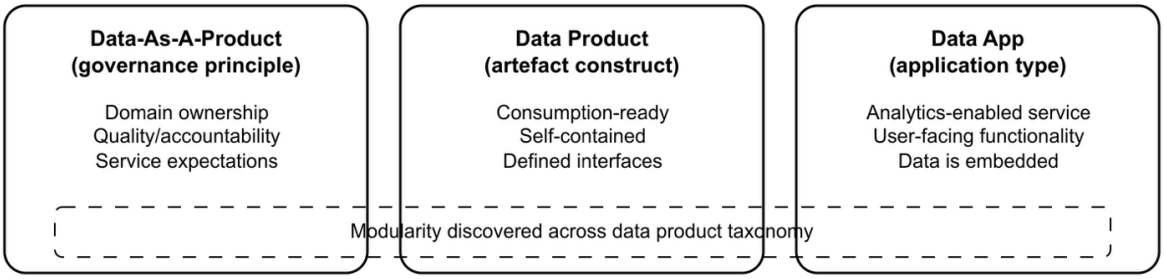


Fig. 1 Distinguishing Data Product Concepts and Situating Modularity

2.2. Analytics Projects as Organizational Capability

IS analytics capabilities literature conceptualizes analytics not as a discrete technical asset but as an embedded organizational capability formed through interacting human, organizational and technological elements. Early work (Cosic et al., 2015) identifies individual, organizational and technical competencies as antecedents to analytics value; the Big Data Analytics Capability literature (Wamba et al., 2017; Mikalef et al., 2020) extends this, locating success in configurations of data resources, technological infrastructure, governance mechanisms and managerial capabilities rather than any single factor. While this literature treats capability at the firm level, with analytics projects as its primary unit of activity, it evaluates outcomes via firm performance, decision quality or innovation, without treating artifact modularity as a meaningful outcome. Our configurational approach addresses this gap, reframing analytics capability as a generative mechanism for modular design.

2.3. Configurational Theory

Configurational theory holds that outcomes are best explained by specific combinations of conditions rather than isolated variable effects (Fiss, 2011; Ragin, 2008). Three principles distinguish it from variance-based reasoning: conjunctural causation, equifinality, and causal asymmetry, all operationalized by fuzzy-set QCA through set-theoretic logic. Fuzzy-set QCA has been applied in IS to digital eco-dynamics (El Sawy et al., 2010) and BI capabilities (Park et al., 2017), but not to analytical projects and data products. It suits this phenomenon because modular outcomes require simultaneous alignment across data engineering, organizational and transitional dimensions, with no single sufficient condition, and because the conditions blocking modularity may differ structurally from those producing it. We expect multiple equifinal pathways to modularity, and structurally distinct pathways to its absence.

3. Research Method

The study used a three-phase, abductive multi-method design (Dubois & Gadde, 2002), combining construct development with configurational testing. The design rests on pragmatism as its philosophical stance (Goldkuhl, 2012), because pragmatism evaluates knowledge by its usefulness in explaining and guiding action, which allows participants' interpretations (Gioia method) and set-theoretic causal claims (fsQCA) to be combined without committing to a stratified ontology or to subjective meaning alone. Phase 1 developed a preliminary structure

from relevant literature to guide interviews. Phase 2 applied the Gioia method (Gioia et al., 2013) iteratively, restructuring the framework from taxonomic to configurational, grounded in participant perspectives. Phase 3 submitted the emergent constructs to fsQCA (Ragin, 2008). An evidence chain was maintained throughout: interview quotations informed first-order concepts, grouped into Second-order Themes and Aggregate Dimensions; Second-order Themes became fsQCA Conditions after calibration anchors were specified.

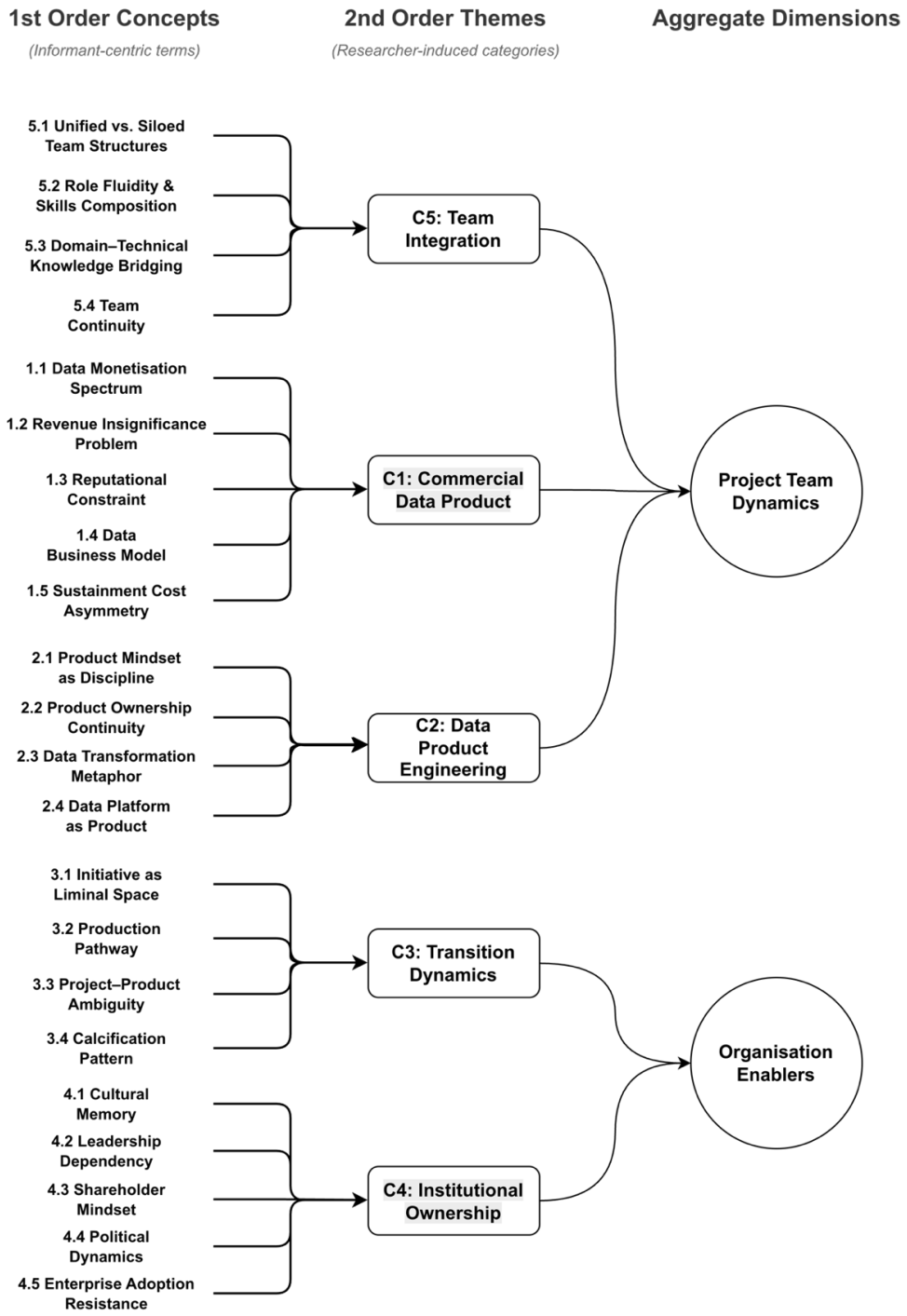


Fig. 2. Final Gioia Data Structure used for fsQCA

Fourteen semi-structured interviews were conducted between August and November 2025 with analytics practitioners purposively sampled from professional networks for direct experience with data product initiatives, seeking diversity in maturity stage, sector and

outcome. Interviewees averaged 14.8 years of analytics experience (range 10-22), split between senior executives (CDOs, CTOs, chief data scientists, general managers) and hands-on practitioners (data analysts, data scientists, program and project managers). Their projects spanned telecommunications, finance, retail, technology, construction, transport, insurance, education and public service, across organizations from startups to government agencies and large enterprises. Each interview lasted between 60 and 90 minutes and was audio-recorded and AI-transcribed. To mitigate outcome bias, interviewees described both successful and unsuccessful initiatives, yielding one to four bounded project narratives each, for 33 case-level units of analysis, appropriate for fsQCA (Schneider & Wagemann, 2010). Because 11 of 14 interviews contributed two or more narratives, nested-case bias was assessed via a leave-one-interview-out (LOIO) procedure.

The preliminary structure comprised four areas from the literature (Cosic et al., 2015; Mikalef et al., 2020): Business Problem, Resources, Technology, Data. Coding against the literature produced 415 coded references and revealed that the structure was taxonomic rather than generative: the Resources area became a catch-all, the temporal dimension was invisible, and the coding structure could not explain how modular data products emerge. Some participants described data products in terms of “almost any (commercial) outcome could be termed a product”, whilst others referred to a “product-oriented mindset” and “the ability to mix data products like fruit salad”. The data-product-as-approach versus data product-as-artifact distinction drove iterative re-coding. Pulling apart discussions on “product engineering as a discipline” and “we just needed to get to the end result, as a product” produced a Gioia data structure with five Second-Order Themes to be used as fsQCA Conditions (Fig. 2): Commercial Data Product (C1), whether the initiative is framed around a value proposition, anticipated consumers and explicit product intent; Data Product Engineering (C2), product thinking and engineering discipline applied to data; Transition Dynamics (C3), pathways from POC to production and resistance to calcification; Institutional Ownership (C4), anchoring in routines, governance arrangements and ongoing stakeholder support; and Team Integration (C5), role fluidity, complementary skill sets, embedded domain expertise and cross-organizational collaboration. Technology and data resources operated through these conditions rather than as independent drivers, consistent with sociomaterial perspectives (Orlikowski, 2007).

For example, a telco big data platform that served multiple use cases and sustained over multiple iterations demonstrates modularity when the pipelines, data structures and engineered outcomes are re-used and extended with each iteration; whereas a similar telco location data product failed to produce re-use despite strong commercial focus and executive support. Calibration used a 4-value scheme omitting the 0.5 crossover following Basurto and Speer (2012). The anchors are outlined in Table 1 and fully documented in the supplementary materials.

Table 1. fsQCA Conditions and Calibration Anchors

Condition	Fully In (1.0)	Mostly In (0.75)	Mostly Out (0.25)	Fully Out (0.0)
C1: Commercial Data Product	Explicit commercial value	Clear value articulation, some gaps	Weak commercial framing	No commercial orientation
C2: Data Product Engineering	Strong product and engineering mindset	Clear product or engineering approach	Minimal product or engineering discipline	Ad-hoc processes
C3: Transition Dynamics	Smooth, managed transitions	Largely successful transitions	Problematic transitions	Failed transitions
C4: Institutional Ownership (of data)	Stable and aligned leadership support	Substantially embedded with gaps	Limited sponsorship	Isolated initiative
C5: Team Integration	Integrated cross-functional team	Team well-supported across disciplines	Siloed teams with some support	Completely siloed
Y: Modular Data Product	Plug and play product	Mostly modular, reuse evidence	Difficult to integrate	Completely stand-alone

4. Set-Theoretic Analysis Results

Of 32 possible truth table configurations, 6 are empirically populated as sufficient, 7 are contradictory and 19 are logical remainders. The dominant configuration (all conditions present) contains 7 cases with the highest consistency (0.925). C2 and C3 appear as members in all six sufficient rows, reinforcing their centrality.

C2 (Data Product Engineering) is the only individual condition passing the necessity threshold, with consistency 0.955, Relevance of Necessity (RoN) 0.783, and coverage 0.808, nearly all cases achieving modular data product status exhibited strong product, analytical, and engineering methods. Combinations C3+C4 and C3+C5 also achieve necessity consistency of 0.909 with RoN values above 0.76. No single condition passes the 0.90 threshold for ~Y, confirming causal asymmetry. C2 approaches single-condition necessity; beyond C2, necessity is achieved only through disjunctions, identifying C1, C3, C4, and C5 as SUIN conditions. In the intermediate solution for Y, C2 and C3 appear as INUS conditions across all three solution paths[#].

4.1. Configurations for Modularity

Three equifinal pathways were found in the intermediate solution (overall consistency 0.932; coverage 0.833), shown in Table 2.

Path 1: C2·C3·C4 "Institutional Anchoring pathway." This is the dominant pathway, covering 71.2% of the outcome with the highest consistency (0.940) and unique coverage (0.273). It combines Data Product Engineering, Transition Dynamics, and Institutional Ownership. Cases include a financial platform, an HR product, an insurance database, call centre speech analytics, and automated network planning. What unites these cases is that strong technical-product methods were embedded within organizational structures that provided sustained ownership and smooth transitions from prototype to production. Hence described as Institutional Anchoring.

Table 2. Configurations for Presence and Absence of Modular Data Products

	Y: Modular data product			~Y: Absence of modular data product		
	1. Institutional Anchoring	2. Practice-Led Product	3. Lone Wolf Making Money	1. Blank Check	2. Money For Nothing	3. Rigor w/o Roots
C1: Commercial Data Product		⊗	●	●	●	
C2: Data Product Engineering	●	●	●	⊗	⊗	●
C3: Transition Dynamics	●	●	●	⊗	⊗	●
C4: Institutional Ownership	●			●	⊗	⊗
C5: Team Integration		●	⊗		⊗	●
Consistency	0.940	0.846	0.886	0.885	0.936	0.812
PRI (Reduction in Inconsistency)	0.906	0.500	0.778	0.750	0.897	0.571
Raw coverage	0.712	0.333	0.470	0.348	0.667	0.394
Unique coverage	0.273	0.030	0.076	0.091	0.333	0.061
Solution consistency	0.932				0.826	
Solution coverage	0.833				0.864	

Note. Black circles (●) indicate presence of a condition; crossed circles (⊗) indicate absence. Blank cells indicate the condition is not relevant ("don't care") to the configuration.

Path 2: ~C1·C2·C3·C5 "Practice-Led Product pathway". This path is missing Commercial Data Product but succeeds due to Team Integration binding Data Product Engineering and strong Transitions. It's a data product led by strong practices. It covers two cases with consistency 0.846 (note this is at the threshold for inclusion and warrants further investigation). Modularity

[#] Conditions are interpreted as INUS (Insufficient but Necessary parts of Unnecessary but Sufficient conditions) or SUIN (Sufficient but Unnecessary parts of an Insufficient but Necessary condition), reflecting that individual conditions contribute to outcomes as parts of configurations rather than as standalone causes.

emerged through strong product thinking, transition management, and team integration, consistent with data-product-as-approach-to-value-realization (over commercial orientation).

Path 3: C1·C2·C3·~C5 "Lone Wolf Making Money" pathway. This path substitutes Commercial Data Product for Team Integration, hence the Lone Wolf name. Modular data products emerged despite limited cross-functional fluidity, because commercial pressure and strong product engineering compensated (consistency 0.886, unique coverage 0.076).

The three solution paths share C2 and C3 as core conditions, the causal core. The paths differ in their peripheral conditions: C4 (Institutional Ownership) and/or C5 (Team Integration) provide the additional configurational context, with C1 (Commercial Data Product) playing a switching role, not required in Path 1, negated in Path 2.

4.2. Configurations for the Absence of Modularity

The dominant path to non-modularity is the "Money For Nothing" pathway (Path 2: C1·~C2·~C3·~C4·~C5), with consistency 0.936, PRI 0.897, and unique coverage 0.333. Despite strong commercial focus, initiatives lacking product methods, transition management, institutional ownership, and team integration almost always fail to achieve modularity. Another instructive pathway is Path 1:C1·~C2·~C3·C4 (consistency 0.885). This is the "Blank Check" where commercial sponsorship without engineering discipline does not produce modularity. The third pathway is the paradox path, Rigor w/o Roots, meaning that product engineering, good transition and a well-integrated team still fail when there is no ownership (C2·C3·~C4·C5). This path has ~Y consistency of 0.812 and unique coverage 0.061.

4.3. Robustness and Deviant Cases

The core pathway C2·C3·C4 is stable across the 0.80 and 0.85 thresholds. At 0.90, only the fully saturated configuration C1·C2·C3·C4·C5 survives, with high consistency but reduced coverage. Three deviant consistency cases were identified (solution membership >0.5 and Y < 0.5). In each, gaps in organizational embeddedness, product framing, or the link between analytical accuracy and actionable intervention explained the shortfall. LOIO analysis (removing each multi-case interview in turn) preserved the conjunction C2·C3 and the near-necessity of C2 in every iteration. Peripheral conditions completing each sufficient path showed moderate sensitivity. We therefore characterize the sufficiency result as exhibiting strong structural stability at the level of the core conjunction (C2·C3) and moderate stability at the level of individual paths.

5. Discussion

The configurational analysis reveals three equifinal pathways to modular data products, anchored by Data Product Engineering (C2) as the only individually near-necessary condition and by the conjunction C2·C3 appearing in every sufficient path. While only 2 cases split the conjunction, the complex solution at 0.75 threshold includes C1·C2·~C3·C5 as a path. Transition Dynamics (C3) can be substituted in the solution, Data Product Engineering (C2) never is.

This research indicates a stronger claim than modular data products require good architecture: modularity emerges when product and engineering discipline combines with positive transitions, and is realized through one of three distinct organizational configurations, but the conditions producing its absence are structurally different.

5.1. Theoretical Implications

Baldwin and Clark (2000) argue modular architectures depend on three classes of design rules: visible information (interfaces enabling independent teams to work on separate modules), hidden information (parameters encapsulated within modules), and test protocols (standards verifying correct integration). The fsQCA results map directly onto this framework. Data Product Engineering (C2) captures organizational capacity to produce design rules: product management practices defining visible interfaces, engineering disciplines encapsulating hidden parameters, and governance routines establishing verification. That C2 is the only condition meeting the necessity threshold suggests articulating design rules is a prerequisite for modularity, though not sufficient alone.

Transition Dynamics (C3) complements Data Product Engineering (C2) by providing the selection environment in which design rules are tested and refined. Design rules must be stable to be effective, since modules cannot evolve independently if their interfaces keep changing, and the passage from proof-of-concept to production is where that stability is forged or lost. All three intermediate solution paths include both C2 and C3; their high pairwise coincidence is theoretically coherent, as engineering discipline and successful transition are separate concepts but mutually reinforcing (Tiwana, 2026). Modular data architecture is thus a sociotechnical rather than purely engineering achievement, with C2 and C3 jointly enabling articulation and stabilization of design rules in production, transforming project-based artifacts into modular data products.

Proposition 1: The joint presence of data product engineering discipline (C2) and transition dynamics (C3) constitutes a recurrent core conjunction underpinning all sufficient configurations for modular data product emergence. This conjunction enables the articulation and stabilization of design rules in production.

The negation analysis confirms causal asymmetry. The intermediate solution for ~Y identifies three routes to non-modularity, none the simple inverse of the positive paths. The dominant complex path describes comprehensive organizational incapacity in which no design rules can form. The third path presents a theoretically instructive paradox: successful transition and team integration combined with absent institutional ownership still produces non-modularity. This configuration generates modules lacking the organizational authority to enforce their interfaces; teams collaborate effectively and products survive transition, but without institutional ownership interfaces remain informal and the architecture never stabilizes into modularity. Such cases are technically successful products trapped within their originating team.

Proposition 2: The causal conditions producing modular data products are configurationally asymmetric from those producing non-modular outcomes. The absence of data product engineering and transition dynamics constitutes a dominant sufficient configuration for non-modularity, and modularity and non-modularity are generated by distinct, non-mirrored causal mechanisms.

Commercial Data Product (C1) plays a nuanced configurational role. It is core in Path 3 and its absence is core in Path 2, yet it fails the individual necessity threshold. This challenges the common prescription that data teams should "think like product managers" as a universal strategy for modularity. Commercial framing is a viable mechanism for defining visible design rules, but modularity is also achievable through institutional or collaborative mechanisms.

Proposition 3: Commercial orientation (C1) is neither necessary nor universally effective for modular data product emergence. Institutional ownership (C4) and team integration (C5) act as equifinal and partially substitutable complements to the core conjunction (C2·C3), shaping alternative organizational contexts in which design rules are enacted and modularity is realized.

5.2. Practical Implications

The configurational findings carry direct implications for analytics leaders. The most consequential investment is in the generative mechanism, Data Product Engineering (C2) and Transition Dynamics (C3): without engineering discipline to create design rules and transition management to stabilize them, no combination of commercial orientation, institutional ownership or team integration compensates. Leaders should prioritize data product management practices, governance, interface specification and sustained investment before attending to peripheral conditions.

The three pathways also define a diagnostic framework. Organizations with strong institutional authority will pursue Path 1, where leadership mandate and structural integration enforce the design rules engineering teams create; those with collaborative cultures but limited hierarchy, common in startups and platform teams, may pursue Path 2, leveraging cross-functional integration as the coordination mechanism; and those with clear commercial mandates but disciplined, siloed teams are likely to pursue Path 3, using market requirements as the source of interface specifications. The negation analysis supplies the corresponding failure diagnostic: leaders should monitor for the paradox paths (C2·C3·~C4·C5), where technically successful products with collaborative teams but no institutional backing produce non-modular outcomes despite apparent project health.

5.3. Limitations and Future Research

Five limitations frame future opportunities. First, 33 case perspectives from 14 interviews, predominantly from large enterprises in a single national context (Australia), constrain generalizability; replication in different regulatory and cultural environments would strengthen the findings. Second, nested cases introduce correlated bias risks, mitigated through LOIO which preserved the core conjunction C2·C3 in every Y solution. Third, outcome calibration relies on sole researcher judgement rather than validated measures; work is underway to employ a second rater to improve inter-coder reliability. Fourth, retrospective accounts are subject to hindsight bias, and longitudinal studies tracking development in real time could reveal temporal sequencing within configurations. Fifth, the high coincidence between C2 and C3 (0.848) means their independent contributions cannot be fully distinguished, and their details will be reviewed to avoid reflecting a single underlying organizational capability.

6. Conclusion

This paper examines how modular data products emerge from analytics projects. Using 33 case perspectives and a Gioia-fsQCA design, the findings show modularity is not synonymous with analytical success but reflects a stronger architectural property grounded in stable design rules enabling decomposition and recomposition across contexts (Baldwin & Clark, 2000). Three equifinal pathways produce modular data products, all anchored in the joint presence of Data Product Engineering (C2) and Transition Dynamics (C3), with institutional ownership, collaborative integration or commercial mandate acting as alternative contextual enablers. The results also demonstrate causal asymmetry, including a paradoxical configuration in which successful transition and collaboration without institutional ownership fail to produce modular outcomes, leaving otherwise healthy products dependent on the single team that built them. Modular data products are best understood as sociotechnical achievements that emerge when organizational arrangements consistently produce and stabilize the design rules that modular architecture requires. Because those design rules are what allow analytical artifacts to be decomposed and recomposed as circumstances change, their stabilization is also the organizational foundation of information resilience: the conditions identified here explain not

only why some analytics projects yield reusable products, but why others leave organizations unable to adapt when their data environment shifts.

Supplementary material. The final Gioia data structure including quotes, full calibrated dataset for all 33 cases, further fsQCA result tables and the per-iteration leave-one-interview-out robustness results are available as online supplementary material.

7. References

Al-Ghattas, H., & Marjanovic, O. (2021). *Business Analytics Capabilities for Organisational Resilience*. Hawaii International Conference on System Sciences.

Baldwin, C. Y., & Clark, K. B. (2000). Design Rules, Volume 1: The Power of Modularity. The MIT Press.

Basurto, X., & Speer, J. (2012). Structuring the Calibration of Qualitative Data as Sets for Qualitative Comparative Analysis (QCA). Field Methods, 24(2), 155-174.

Cosic, R., Shanks, G., & Maynard, S. (2015). A business analytics capability framework. Australasian Journal of Information Systems, 19, 5-19.

Davenport, T. H., & Bean, R. (2020). Are You Asking Too Much of Your Chief Data Officer? Harvard Business Review.

Davenport, T. H., & Kudyba, S. (2016). Designing and Developing Analytics-Based Data Products. *MIT Sloan Management Review*.

Dehghani, Z. (2022). Data Mesh: Delivering Data-Driven Value at Scale. O'Reilly Media.

Dubois, A., & Gadde, L.-E. (2002). Systematic combining: An abductive approach to case research. Journal of Business Research, 55(7), 553-560.

El Sawy, O. A., Malhotra, A., Park, Y., & Pavlou, P. A. (2010). Seeking the Configurations of Digital Ecodynamics: It Takes Three to Tango. Information Systems Research, 21(4), 835-848.

Fiss, P. C. (2011). Building Better Causal Theories: A Fuzzy Set Approach to Typologies in Organization Research. Academy of Management Journal, 54(2), 393-420.

Gartner. (2024). Gartner survey reveals that only 48% of data and analytics initiatives deliver measurable business value. Gartner Research.

Gioia, D. A., Corley, K. G., & Hamilton, A. L. (2013). Seeking Qualitative Rigor in Inductive Research. Organizational Research Methods, 16(1), 15-31.

Goldkuhl, G. (2012). Pragmatism vs interpretivism in qualitative information systems research. European Journal of Information Systems, 21(2), 135-146.

Hasan, M. R., & Legner, C. (2023). Understanding Data Products: Motivations, Definition, and Categories. ECIS 2023 Research Papers.

Henderson, R. M., & Clark, K. B. (1990). Architectural Innovation: The Reconfiguration of Existing Product Technologies and the Failure of Established Firms. Administrative Science Quarterly, 35(1), 9-30.

Loukides, M. (2011). The Evolution of Data Products. O'Reilly Media.

Marchand, D. A., & Peppard, J. (2013). Why IT Fumbles Analytics. Harvard Business Review, 91, 104-112.

Meyer, A. D., Tsui, A. S., & Hinings, C. R. (1993). Configurational approaches to organizational analysis. Academy of Management Journal, 36(6), 1175-1195.

Mikalef, P., Krogstie, J., Pappas, I. O., & Pavlou, P. (2020). Exploring the relationship between big data analytics capability and competitive performance. Information & Management, 57(2), 103169.

Orlikowski, W. J. (1992). The Duality of Technology: Rethinking the Concept of Technology in Organizations. Organization Science, 3(3), 398-427.

Orlikowski, W. J. (2007). Sociomaterial Practices: Exploring Technology at Work. Organization Studies, 28(9), 1435-1448.

Park, Y., El Sawy, O., & Fiss, P. (2017). The Role of Business Intelligence and Communication Technologies in Organizational Agility. Journal of the Association for Information Systems, 18(9).

Patil, D. J. (2012). Data Jujitsu. O'Reilly Media.

Ragin, C. C. (2008). Redesigning Social Inquiry: Fuzzy Sets and Beyond. University of Chicago Press.

Reggio, G., & Astesiano, E. (2020). Big-Data/Analytics Projects Failure: A Literature Review. *2020 46th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, 246–255.

Sanchez, R., & Mahoney, J. (1996). Modularity, Flexibility, and Knowledge Management in Product and Organization Design. Strategic Management Journal, 17(S2), 63-76.

Schneider, C. Q., & Wagemann, C. (2010). Standards of Good Practice in Qualitative Comparative Analysis (QCA) and Fuzzy-Sets. Comparative Sociology, 9(3), 397-418.

Tiwana, A. (2026). Does Modularity Coordinate? Disentangling Decoupling, Interfaces, and Design Rules via Simon's Theory of the Artifact. *Information Systems Research*.

Tiwana, A., Konsynski, B., & Bush, A. (2010). Research Commentary - Platform Evolution. Information Systems Research, 21(4), 675-687.

Wamba, S. F., Gunasekaran, A., Akter, S., Ren, S. J., Dubey, R., & Childe, S. J. (2017). Big data analytics and firm performance: Effects of dynamic capabilities. Journal of Business Research, 70, 356-365.

Yoo, Y., Henfridsson, O., & Lyytinen, K. (2010). Research Commentary - The New Organizing Logic of Digital Innovation. Information Systems Research, 21(4), 724-735.

How Modular Data Products Emerge: supplementary materials

1. Data Collection

Fourteen semi-structured interviews were conducted between August and November 2025 with analytics practitioners across multiple organizations spanning telecommunications, financial services, HR, technology, legal information, and the public sector. Interviewees were selected through purposive sampling of professional networks for direct experience with data product initiatives at various maturity stages (startup to enterprise, private-public sector splits, early stage to mature initiative as shown in Table 1), and a diversity of successful and unsuccessful outcomes.

Table 1 List of Interviews

Int.	Role(s)	Industries	Organization Size(s)	Analytics Experience (years)	# Projects (Cases)
1	Program Manager, Data Governance Manager	Retail, Telco, Education	Enterprise	15	3
2	Data Analyst	Telco	Enterprise	12	2
3	GM Infra, CDO	Telco, Finance, Retail	Mid, Enterprise	20	4
4	Data Analyst, Project/ Program Manager	Telco, Retail	Enterprise	12	3
5	Chief Data Scientist	Call Centre, Technology	Mid, Enterprise	15	3
6	CTO	Technology, National Security	Startup, Enterprise	22	2
7	GM, Chief Product Officer	Construction, HR	Startup, Enterprise	19	1
8	CTO, Program Manager	Construction, Technology	Startup, Enterprise	10	2
9	Data Analyst, GM	Insurance, Transport, Telco	Startup, Enterprise	14	4
10	CEO	Transport	Startup	16	1
11	Data Scientist	Finance, Telco	Enterprise	13	3
12	CDO	Public Service	Government	15	3
13	Chief Data Scientist, CTO	Technology, HR	Startup	11	1
14	CDO	Education, Public Service	Gov't, Enterprise	13	2

2. Initial and Final Data Structures

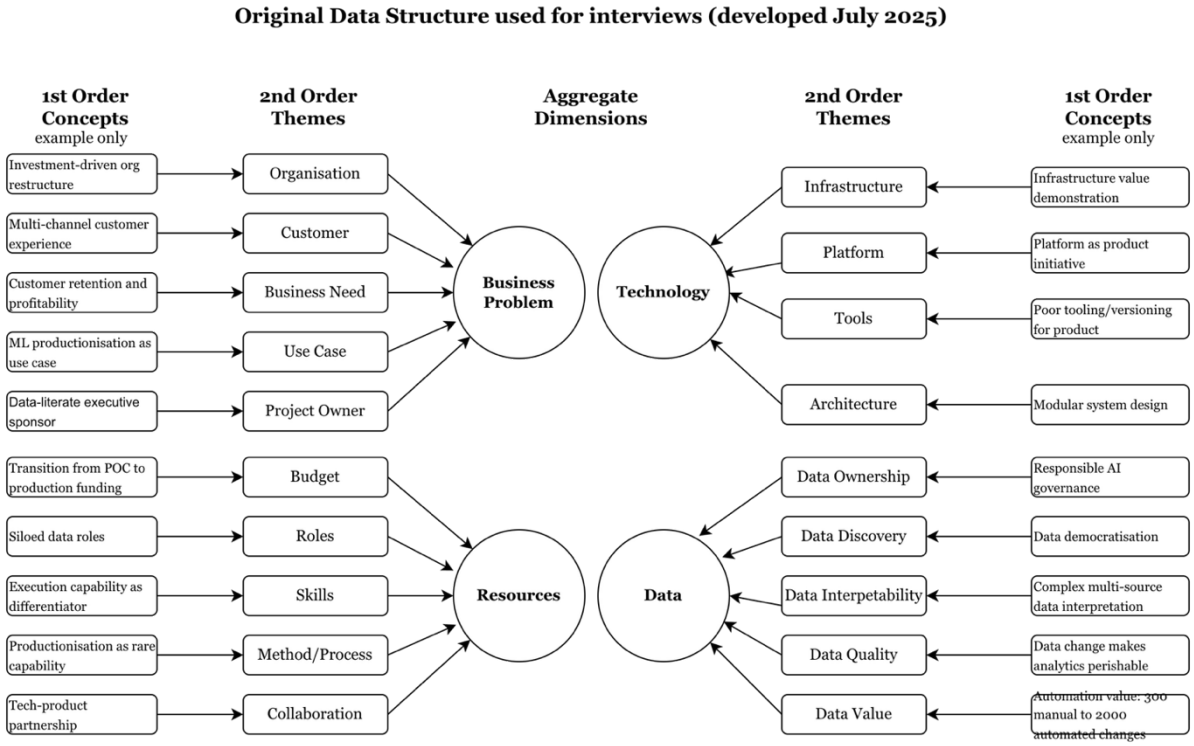


Fig. 1 Original Data Structure, based on Cosic et al.,2015; Mikalef et al.,2020

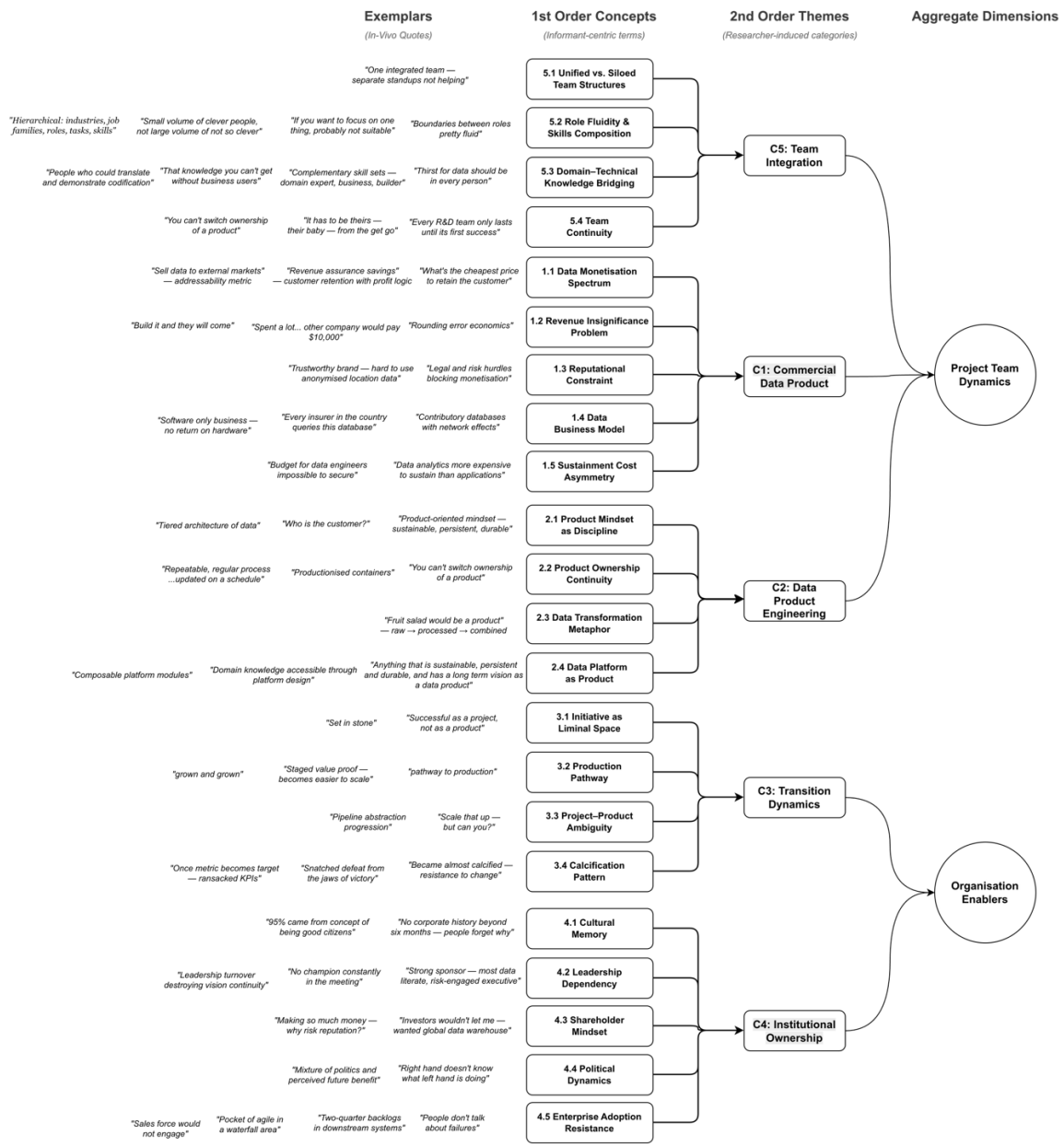


Fig. 2 Final Gioia Data Structure used for fsQCA (with sample quotes)

3. Fuzzy-Set Qualitative Comparative Analysis

Table 2 fsQCA Conditions and Calibration Anchors

Condition	Fully In (1.0)	Mostly In (0.75)	Mostly Out (0.25)	Fully Out (0.0)
C1: Commercial Data Product	Explicit commercial value, clear revenue model, measurable financial impact	Strong commercial orientation with clear value articulation, some gaps in value, impact, measures	Weak commercial framing, predominantly technical or exploratory with speculative value claims	No commercial orientation, purely experimental
C2: Data Product Engineering	Strong product-oriented mindset, combined with strong systematic engineering and QA	Clear product or engineering approach with most elements present, minor gaps	Minimal product or engineering discipline, ad-hoc approaches, limited governance	No discernible methodology, ad-hoc, unstructured processes
C3: Transition Dynamics	Smooth, managed transitions with clear pathways from POC to production, resistance to calcification	Largely successful transitions with minor friction, adequate sustained investment	Problematic transitions; stuck between states, significant calcification	Failed transitions; permanently stuck in POC, fully calcified, or abandoned
C4: Institutional Ownership (of data)	Deeply embedded: integrated across channels, stable leadership support, aligned incentives	Substantially embedded with most structural supports, minor gaps	Weakly embedded, limited sponsorship, stakeholder resistance	Not embedded, isolated initiative, no organization buy-in
C5: Team Integration	Integrated cross-functional team with fluid role boundaries, domain-technical bridging, strong collaboration	Largely integrated team with some functional boundaries, some ownership gaps	Siloed teams with limited cross-functional interaction, ownership fails to transfer	Completely siloed, no domain expertise accessible; no collaboration
Y: Modular Data Product	Completely modular, plug and play across the organization	Mostly modular, clear evidence of ability to work across different systems	Difficult to integrate, not very modular	Completely stand-alone, no intent for integration

Table 3 fsQCA calibrated dataset of cases

Case ID	C1	C2	C3	C4	C5	Y
CP001	1	0.75	0.75	0.75	0.75	1
CP002	0.75	0.25	0.25	0.25	0.25	0.25
CP003	0.75	0.25	0.25	0.25	0.25	0
CP004	0.75	0.25	0.25	0.75	0.25	0
CP005	0.25	0.25	0.25	0.25	0.25	0.25
CP006	0.75	0.25	0.25	0.25	0.25	0.25
CP007	1	1	0.75	1	0.75	1
CP008	0.75	0.75	0.75	0.75	0.75	0.75
CP009	0.25	0.25	0	0	0.25	0
CP010	0.25	0.25	0.25	0.25	0.25	0.25
CP011	1	0.75	1	1	1	0.75
CP012	0.75	0.25	0.25	0.75	0.75	0.25
CP013	0.75	0.75	0.25	0	0.75	0
CP014	0.75	0.25	0.25	0.75	0.25	0.25
CP015	1	1	0.75	0.75	0.75	1
CP016	0.75	0.75	0.25	0.25	0.75	0.75
CP017	0.25	0.75	0.75	0.75	0.25	0.75
CP018	1	1	0.75	1	0.25	1
CP019	0.75	0.75	0.75	0.25	0.25	0.75
CP020	1	0.75	0.75	0.25	0.25	1
CP021	1	0.75	0.75	0.75	0.25	1
CP022	0.75	0.25	0.25	0	0.25	0.25
CP023	0.25	0.25	0	0	0	0
CP024	0.25	0.25	0.25	0	0	0.25
CP025	0.25	0.75	0.75	0.25	0.75	0.75
CP026	0.75	0.75	0.75	0.75	0.75	0.75
CP027	0.75	0.75	0.75	0.25	0.75	0.25
CP028	0.75	0.25	0.25	0.25	0.25	0
CP029	0.75	0.75	0.75	0.25	0.75	0.25
CP030	0	0.75	0.75	0.75	0.75	0.75
CP031	1	1	0.75	1	1	1
CP032	0.25	1	0.75	0.25	0.25	0.25
CP033	0	0.75	0.75	0.25	0.25	0.75

Table 4 Single Condition Necessity Analysis for Y and ~Y

Condition	Y Consistency	Y Coverage	~Y Consistency	~Y Coverage	Assessment
C1: Commercial Data Product	0.848	0.659	0.697	0.541	-
C2: Data Product Engineering	0.955	0.808	0.515	0.436	Necessary (Y)
C3: Transition Dynamics	0.864	0.838	0.455	0.411	-
C4: Institutional Ownership	0.758	0.833	0.409	0.450	-
C5: Team Integration	0.712	0.770	0.485	0.526	-

Table 5 Consistency Threshold Sensitivity

Threshold	Complex Solution	Consistency	Coverage
0.75	~C1·C2·C3 + C1·C4·C5 + C2·C3·~C5 + C1·C2·~C3·C5	0.833	0.909
0.80	C2·C3·C4 + ~C1·C2·C3·C5 + C1·C2·C3·~C5	0.932	0.833
0.85	C2·C3·C4 + C1·C2·C3·~C5	0.930	0.803
0.90	C1·C2·C3·C4·C5	0.927	0.680

Table 6 reports the intermediate sufficiency solutions for the presence (Y) and absence (~Y) of modular data products across the baseline analysis (n = 33) and each leave-one-interview-out (LOIO) iteration that produced a different solution. Eleven iterations were performed, one for each interview contributing two or more cases, only those rows with different solutions are recorded in Table 6.

Table 6 Intermediate sufficiency solutions by LOIO iteration

Iteration (cases removed)	Intermediate solution for Y	Intermediate solution for ~Y
Baseline (n = 33)	C2·C3·C4 + ~C1·C2·C3·C5 + C1·C2·C3·~C5	~C1·~C2·~C3·~C4·~C5
CP021–CP024 (29)	= baseline	~C1 + ~C2·~C3 + ~C4·C5
CP001–CP003 (30)	= baseline	~C1 + ~C2·~C3 + ~C4·C5
CP016, CP017 (31)	~C1·C2·C3·C5 + C1·C2·C3·~C5 + (C1·C2·C3·C4 | C2·C3·C4·C5) – multiple alternatives	~C1 + ~C2·~C3 + ~C4·C5
CP006–CP009 (29)	~C1·C2·C3 + C1·C4·C5 + C2·C3·~C5 + C1·C2·~C3·C5 (substantially different; multiple alternatives)	= baseline
CP013–CP015 (30)	C1·C4 + ~C1·C2·C3 + C2·C3·~C5 + C1·C2·~C3·C5	= baseline
CP019, CP020 (31)	C2·C3·C4 + ~C1·C2·C3·C5	~C1 + ~C4 + ~C2·~C3

Cyber Resilience: A Dyadic Relationship between Software as a Service Providers and Customers

***Dr Michael Lane**, Michael.Lane@unisq.edu.au*
***Dr Sanjib Tiwari**, Sanjib.Tiwari@unisq.edu.au*
***Dr Mehmet Zahid Ercan**, MehmetZahid.Ercan@unisq.edu.au*

School of Business, Law, Humanities and Pathways
University of Southern Queensland, Toowoomba, QLD, Australia

Abstract

Organisations increasingly depend on third-party Software as a Service (SaaS) platforms for business-critical services, which extends their cyber-risk boundary beyond their direct control. This paper challenges an assumption implicit in existing cyber resilience standards: that cyber resilience resides with a single organisation. A template analysis of three high-impact SaaS supply-chain incidents (Okta, MOVEit and Snowflake) identifies four recurring themes: identity and token exploitation, concentration risk, shared-responsibility governance ambiguity, and detection blind spots. The analysis shows that critical control objects, such as session tokens in support files and provider-held logs, were not fully owned by either provider or customer. The paper therefore argues that SaaS cyber resilience is a dyadic property of the provider–customer relationship. It proposes a conceptual framework of six resilience capabilities, two of them dynamic governance capabilities, through which organisations can anticipate, absorb, adapt to and recover from SaaS incidents, and calls for continuous, evidence-informed management of third-party SaaS risk in place of periodic vendor assurance.

Keywords: *software as a service (SaaS); cyber resilience; provider–customer dyad; dynamic capabilities; shared responsibility; third-party cyber risk; identity security; cyber incident analysis.*

1. Introduction

Organisations increasingly depend on third-party software as a service (SaaS) platforms for identity and access management, collaboration, customer support, finance, analytics and data exchange. This dependence lets organisations focus on value creation, but it also means that their cyber resilience depends on the security posture, governance practices and incident response capabilities of external providers (Hashizume et al., 2013; Cloud Security Alliance, 2024). Third-party SaaS cyber risk is therefore not simply a technical problem but a socio-technical and organisational resilience problem created by digital interdependence. Persistent identities, delegated access, application programming interfaces (APIs), Open Authorization (OAuth) tokens, service accounts and shared logging arrangements create trusted pathways through which cyber incidents can propagate rapidly across organisational boundaries (Lella et al., 2021; Boyens et al., 2024). Recent incidents involving identity providers, customer support systems, managed file transfer software and cloud data platforms show that adversaries increasingly target these relationships, and that organisations are often unable to see, verify, contain or recover from cyber incidents that emerge through the SaaS ecosystem which they depend on.

Existing cyber resilience standards, including the National Institute of Standards and Technology (NIST) Special Publication (SP) 800-160, the NIST Cybersecurity Framework (CSF) 2.0 and the cloud shared responsibility model itself, share an assumption: cyber resilience is a property of a single organisation that can be assessed, governed and improved within that organisation's boundary (Ross et al., 2021; Pascoe, Quinn and Scarfone, 2024; Cloud Security Alliance, 2024). Rather than filling a gap in this

literature, this paper problematises that assumption (Alvesson and Sandberg, 2011). In the Okta and Snowflake incidents, the capacity to detect, contain and recover from compromise was split between provider and customer, and critical control objects, such as session tokens held in support files and provider-held logs, were fully owned by neither party. This paper therefore argues that in third-party SaaS ecosystems, cyber resilience is better understood as a dyadic property of the provider–customer relationship than of either party alone. Two research questions guide the paper. RQ1: What recurring attack pathways appear in cyber incidents that propagate through third-party SaaS relationships, and under what provider–customer conditions do they occur? RQ2: How is the capacity to anticipate, absorb, adapt to and recover from these cyber incidents distributed between providers and customers, and where does that distribution leave gaps? The paper makes three contributions. It reconceptualises SaaS cyber resilience as a dyadic property of the provider–customer relationship; it identifies where the distribution of cyber resilience capacity breaks down, around shared control objects, provider-held telemetry and identities that cross organisational boundaries; and it offers a framework and propositions for future testing.

2. SaaS Security and Cyber Resilience

This paper draws on four parent literatures (cyber resilience engineering, cybersecurity supply-chain risk management, cloud shared responsibility and identity-centric attack surfaces), integrated through an organisational capability lens. NIST SP 800-160 Volume 2 Revision 1 conceptualises cyber resiliency as the ability to anticipate, withstand, recover from and adapt to adverse conditions, stresses, attacks or compromises (Ross et al., 2021). Linkov and Kott (2019), following the US National Academies, instead describe resilience as the capacity to plan and prepare for, absorb, recover from and adapt to adverse events. This paper adopts the four NIST goals but uses absorb in place of withstand, because a SaaS customer often cannot prevent a provider-side compromise and can act only to limit its impact. It also orders adapt before recover, because in the cases examined adaptation, such as revoking tokens and disabling integrations, occurred while incidents were still unfolding. The resulting four states (anticipate, absorb, adapt and recover) closely parallel Hollnagel's (2009) four cornerstones of resilience engineering (anticipate, monitor, respond and learn). The departure is that, in SaaS, monitoring is split across the provider–customer boundary. Cyber resilience therefore involves not only restoring systems but sustaining core functions while adapting to disruption (Boh et al., 2026).

Supply-chain risk guidance highlights decreased visibility into suppliers, dependencies and development practices as a key source of cyber risk (Boyens et al., 2024; Lella et al., 2021). Secure by Design guidance from the Cybersecurity and Infrastructure Security Agency (CISA) shifts part of the responsibility toward providers (CISA, 2023a). Earlier cloud security studies identified the outsourcing of essential services as a source of privacy, availability and compliance risk (Hashizume et al., 2013; Subashini and Kavitha, 2011). Moreover, cloud security guidance from the Cloud Security Alliance (CSA) stresses under the shared responsibility model, that the provider controls the platform while the customer controls identity configuration, access privileges, integration approval, logging retention and response workflows (Cloud Security Alliance, 2024). Identity security literature adds that many SaaS cyber incidents begin not with platform compromise but with valid credentials, stolen session tokens, compromised service accounts or over-permissioned integrations (Temoshok et al., 2025; OWASP, 2026; Mandiant, 2024). SaaS cyber resilience is therefore a governance and capability challenge arising from the interaction of digital interdependence, third-party supply chains and identity-mediated access.

Organisational capability theory integrates these parent literatures. Capability-based perspectives argue that performance depends less on the resources an organisation owns than on its ability to deploy them through learned, repeatable routines (Nelson and Winter, 1982; Amit and Schoemaker, 1993). Dynamic capability theory distinguishes ordinary capabilities, which sustain current operations, from

dynamic capabilities that sense threats, seize responses and transform the resource base as environments change (Teece, Pisano and Shuen, 1997; Teece, 2007; Winter, 2003), and resilience scholarship increasingly treats resilience as a capability of this kind (Duchek, 2020). Applied to SaaS, identity controls, telemetry and posture management are ordinary capabilities, while supplier risk governance and organisational learning are dynamic capabilities that reconfigure them as configurations, integrations and dependencies change. The SaaS setting also questions a core assumption of the theory: that the firm reconfigures resources it owns or controls. In SaaS ecosystems the platform, much of the telemetry and the support processes belong to the provider, so reconfiguring them during a cyber incident requires coordination across organisational boundaries. SaaS cyber resilience therefore depends on relational dynamic capabilities: routines through which provider and customer jointly sense, contain and recover from compromise (Dyer and Singh, 1998).

3. SaaS Ecosystems: Changing an Organisation's Risk Boundary

Core business processes are now enacted through externally hosted SaaS that is continuously connected to internal users, data stores, identity providers and partner systems. A customer support platform may hold the records used to manage customer relationships, a managed file transfer service may carry sensitive data exchanged with regulators and suppliers, and an identity provider may control access to hundreds of downstream applications. This dependence creates three interrelated forms of risk: (1) reliance on third-party service continuity; (2) cyber risks inherited through integrations, delegated permissions and external service accounts that traditional network monitoring cannot see; and (3) governance gaps when provider and customer responsibilities are not operationalised. Because SaaS configurations, users, integrations and privileges change continuously, due diligence at procurement is insufficient on its own (Boyens et al., 2024; Cloud Security Alliance, 2024).

Three features distinguish this cyber risk from that of on-premise IT or traditional IT outsourcing. First, identity acts as a control plane that crosses the organisational boundary: customer users, provider support staff and third-party integrations all authenticate into the same multi-tenant service. Second, the customer sees only the telemetry that the provider exposes, often depending on the license tier purchased. Third, critical control objects, such as support files, OAuth grants, API keys and administrative settings, straddle both parties, so neither fully governs their lifecycle. On-premise IT keeps identities, logs and control objects inside one organisation, and traditional outsourcing settles their ownership in the contract. In SaaS, ownership is shared by default and often left unresolved, which is why this paper treats SaaS cyber resilience as a dyadic property of the provider–customer relationship.

A key implication is that cyber resilience must include the ability to govern SaaS dependencies continuously. Organisations need ongoing visibility of which SaaS services they use, what data those services hold, which identities and integrations have access, how provider security events are communicated, and what continuity arrangements exist if a service becomes unavailable or untrusted. This shifts the focus from static assurance at procurement to continuous management of third-party SaaS cyber risk.

4. Research Methodology

This study adopts a pragmatic research philosophy (Morgan, 2014) and a hybrid deductive–inductive template analysis (King, 2012) of publicly available incident reports, vendor disclosures, government advisories and threat intelligence reports. The aim is not technical forensic reconstruction, but to identify recurring organisational risk patterns and resilience capabilities. A qualitative, interpretive approach suits this aim because it requires interpreting incident narratives, governance breakdowns and

capability implications across cases rather than measuring a single technical variable. The principal cases are Okta's 2023 customer support system incident, the 2023 MOVEit Transfer exploitation and Mandiant's 2024 analysis of the Snowflake customer instance breach, supplemented by supply-chain analysis from the European Union Agency for Cybersecurity (ENISA), CISA and NIST (Okta, 2023a; CISA and FBI, 2023; Mandiant, 2024; Lella et al., 2021; Boyens et al., 2024). They were selected because they illustrate related forms of third-party risk: session token exposure, service account misuse, managed file transfer exploitation, compromised customer credentials, insufficient multi-factor authentication (MFA) and delayed response. Because the four resilience states were drawn from prior literature before coding began, they formed the deductive template, while recurring attack pathways and control objects were coded inductively. Analysis proceeded in three stages. First, each incident was coded for the exploited pathway, the control object involved (such as a credential, token, support file or log), the party holding that control, and any detection or notification delay. Second, codes were compared across incidents and clustered into recurring attack pathways, which formed the themes. Third, each coded segment was assigned to a resilience state by a stated rule: before compromise to anticipate, during impact to absorb, during response to adapt, and after containment to recover. Theme 3 was identified as a structural condition rather than an attack pathway, because it describes the provider–customer arrangements under which Themes 1, 2 and 4 occur.

5. Analysis and Findings

Theme 1: Identity, Session and Token Exploitation

SaaS services are accessed through federated identity, administrative accounts, OAuth-based integrations, browser sessions, API keys and service accounts. These have become high-value targets because they let an attacker appear as an authorised user or trusted integration, making malicious activity hard to distinguish from legitimate business activity without context-aware monitoring (Temoshok et al., 2025; OWASP, 2026). Unlike malware-based compromise, identity abuse often uses valid credentials or tokens.

In the Okta incident, unauthorised access to the customer support system was associated with a service account credential stored in a personal Google account, and some customer-uploaded HTTP Archive (HAR) files contained session tokens that could be used for session hijacking (Okta, 2023a). A support process intended to help customers became a pathway for downstream risk, The lesson is not simply that session tokens should be protected, but that SaaS governance must cover the full lifecycle of identities, support artefacts and non-human accounts. In the Snowflake incidents, Mandiant traced compromise to customer credentials, often infostealer-derived and on accounts without MFA, rather than to a breach of Snowflake's enterprise environment (Mandiant, 2024). Across both cases, identity is the control plane through which SaaS cyber resilience is strengthened or weakened.

Theme 2: Central Node and Concentration Risk

Identity providers, support platforms, managed file transfer products and data warehouses can each act as a hub connecting many downstream systems and organisations, so a compromise can extend well beyond a single application. CISA and the Federal Bureau of Investigation (FBI) reported that CL0P exploited a Structured Query Language (SQL) injection vulnerability in Progress Software's MOVEit Transfer to steal data from many dependent organisations (CISA and FBI, 2023; Progress Software, 2023). MOVEit was a trusted service for managed file transfer, often involving sensitive data exchange with regulators, customers and suppliers, so a single vulnerability had cascading consequences. Because MOVEit Transfer is mostly licensed, self-hosted software, it is treated here as common-mode software dependency, in which one product is shared by many organisations, rather than SaaS concentration, in which one provider serves many tenants. SaaS adoption can intensify concentration when many functions depend on a few providers or identity platforms (Lella et al., 2021), so

organisations need to map critical dependencies and prepare data export plans, alternative channels, break-glass access and rehearsed continuity procedures.

Theme 3: Shared Responsibility and Governance Ambiguity

Unlike Themes 1, 2 and 4, which describe how harm occurs, this theme describes the structural condition under which those attack pathways arise. Providers secure the platform, while customers retain responsibility for identities, access policies, configuration, integration approval and tenant monitoring (Cloud Security Alliance, 2024; Subashini and Kavitha, 2011), but this division is often defined only at a high level. Mandiant reported that Snowflake's enterprise environment was not found to be breached and that customer credentials were used to access customer instances (Mandiant, 2024): risk materialised even though the provider's core environment was intact, and customer-side identity configuration, MFA enforcement, credential rotation and tenant-level monitoring became decisive resilience factors. Without routines that assign ownership for each critical SaaS service, inventory integrations and privileged accounts, test notification arrangements and that make logs usable in enterprise monitoring, the shared responsibility model remains conceptually correct but operationally weak.

Theme 4: Detection Blind Spots and Notification Latency

This theme contains the clearest evidence that cyber resilience is split between provider and customer. Okta reported that file downloads were not initially identified because different access pathways generated different log event types and record identifiers (Okta, 2023a). Visibility is therefore not merely a matter of having logs: SaaS logs are spread across providers with different retention periods, schemas and event types, customers may lack access to them or fail to ingest them. Furthermore, the customer cannot see inside the provider's logging while the provider lacks the customer's business context. Organisations and providers must understand what events are recorded, which activity pathways may be missed and how indicators are correlated and escalated. Monitoring itself is divided across the relationship. Third-party SaaS and cloud exposure remain growing sources of breaches (Verizon, 2026), so organisations need minimum telemetry requirements, contractual notification obligations, tested incident contacts and the ability to revoke tokens, disable integrations and preserve evidence quickly.

Table 1. Key Themes: SaaS Supply-Chain Vulnerabilities and Organisational Defences

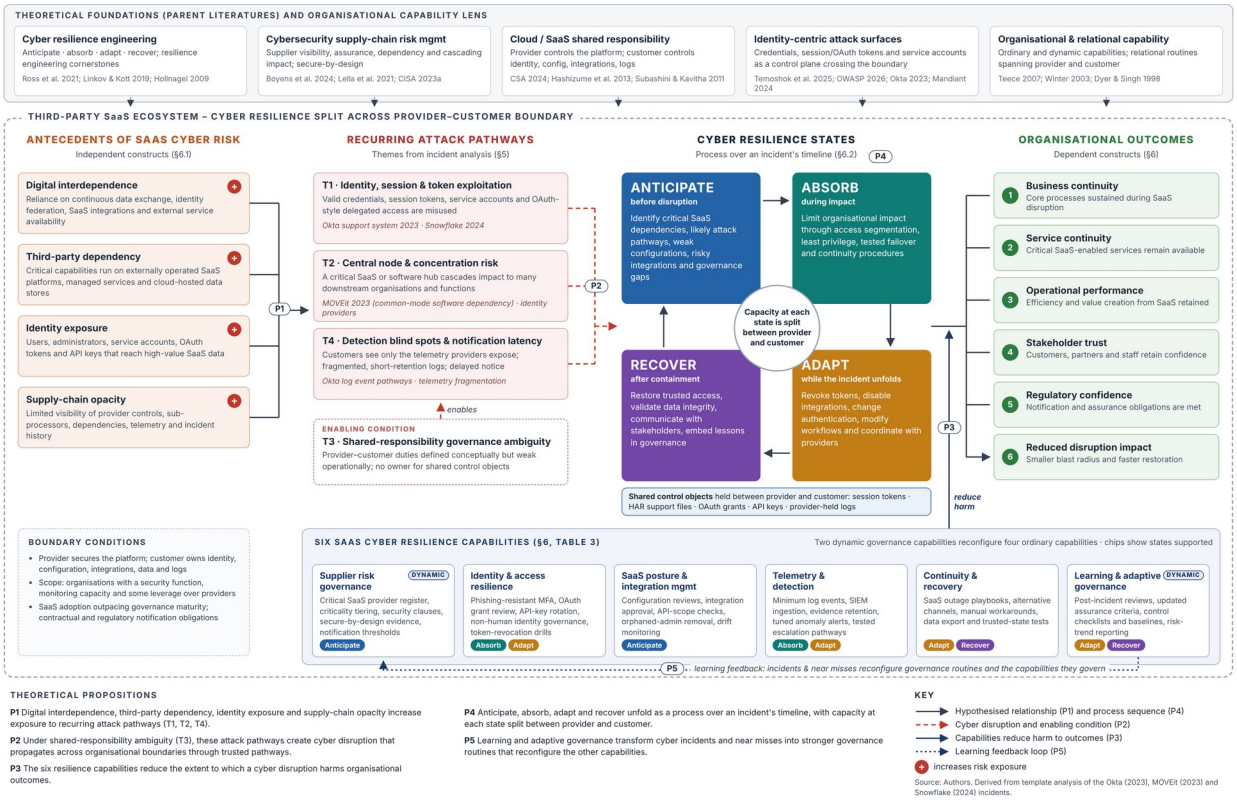
Theme	Illustrative incidents	Recurring attack pathway	Organisational resilience gap	Implied capability (proposition)
Identity, session and token exploitation	Okta support system; Snowflake customer instances	Valid credentials, session tokens, service accounts and OAuth-style delegated access are misused.	Identity assurance and token lifecycle governance are not sufficiently mature.	Phishing-resistant MFA, token revocation, service account governance, conditional access and behavioural monitoring.
Central node and concentration risk	MOVEit Transfer; identity providers; cloud data platforms	A critical SaaS or software node affects many downstream organisations or functions.	Critical SaaS dependencies and business impacts are not fully mapped.	Digital dependency register, continuity planning, supplier tiering and alternative operating procedures.
Shared responsibility governance ambiguity	Snowflake customer instances; SaaS tenant management	Provider and customer responsibilities are conceptually defined but operationally weak.	No explicit owner for tenant configuration, integrations, logs and third-party access.	SaaS control ownership model, configuration reviews, integration approval and secure-by-design procurement criteria.

Detection blind spots and notification latency	Okta log event pathways; SaaS telemetry fragmentation	Activity occurs inside trusted SaaS environments with limited and inconsistent customer visibility.	Logs are incomplete, retained too briefly or not operationalised in enterprise monitoring.	Telemetry requirements, security information and event management (SIEM) integration, incident notification obligations and tested response playbooks.

6. Cyber Resilience Conceptual Framework: SaaS Ecosystems

Figure 1 presents the Cyber Resilience Conceptual Framework for Third Party SaaS Ecosystems. Cyber resilience is not treated as a single control or maturity level but as an integrated capability drawing on technology, governance, people, supplier relationships and learning (Ross et al., 2021; Pascoe, Quinn and Scarfone, 2024).

Figure 1. Cyber Resilience Conceptual Framework for Third Party SaaS Ecosystems



The Framework proposes four antecedents that increase exposure to recurring attack pathways, and six cyber resilience capabilities that reduce the extent to which a cyber disruption harms organisational outcomes. The four cyber resilience states are not a mediating variable: they describe how an incident unfolds over time, with learning feeding the next cycle. Five propositions are stated for testing. **P1:** Digital interdependence, third-party dependency, identity exposure and supply-chain opacity increase exposure to recurring attack pathways (Themes 1, 2 & 4). **P2:** Under shared-responsibility ambiguity (Theme 3), these attack pathways create cyber disruption that propagates across organisational boundaries through trusted pathways. **P3:** Six resilience capabilities reduce the extent to which a cyber disruption harms organisational outcomes. **P4:** Anticipate, absorb, adapt and recover unfold as a process over an incident's timeline, with capacity at each state split between provider and customer. **P5:** Learning and adaptive governance transform cyber incidents and near misses into stronger governance routines that reconfigure the other capabilities. The constructs are defined below.

6.1 Antecedents of SaaS Cyber Risk

- **Digital interdependence:** The extent to which business processes rely on continuous data exchange, identity federation, SaaS integrations and external service availability.
- **Third-party dependency:** The extent to which critical organisational capabilities depend on externally operated SaaS platforms, managed services, software components or cloud-hosted data stores.
- **Identity exposure:** The degree to which users, administrators, service accounts, OAuth tokens and API keys provide access to high-value SaaS services and data.
- **Supply-chain opacity:** The lack of visibility into provider security controls, sub-processors, software dependencies, telemetry, incident history and secure development practices.

6.2 Cyber Resilience States

- **Anticipate:** Identify critical SaaS dependencies, likely attack pathways, weak configurations, risky integrations and governance gaps before they result in disruption.
- **Absorb:** Limit the organisational impact of a SaaS incident through access segmentation, least privilege, tested failover arrangements and continuity procedures.
- **Adapt:** Modify workflows, revoke tokens, disable integrations, change authentication requirements and coordinate with providers while an incident is unfolding.
- **Recover:** Restore trusted access, validate data integrity, communicate with stakeholders, review provider performance and embed lessons into governance routines.

Organisational outcomes. Business continuity means core processes are sustained during SaaS disruption; service continuity means critical SaaS-enabled services remain available or are restored; operational performance means the efficiency and value created through SaaS are retained; stakeholder trust means customers, partners and staff retain confidence in the organisation; regulatory confidence means notification and assurance obligations are met; and reduced disruption impact means a smaller impact radius and faster restoration. Outcomes not directly observed in the incidents analysed are theorised and offered as propositions.

Table 2 sets out the six capabilities in terms of purpose, example practices and cyber resilience state supported. Supplier risk governance and learning and adaptive governance act as dynamic capabilities that reconfigure the other four as configurations, integrations and providers change, so cyber governance is not modelled as a separate antecedent. The capabilities assume a medium or large organisation with a security function, security monitoring and some leverage over providers.

Table 2. Cyber Resilience Capabilities and Resilience States

Capability	Purpose	Example practices	Resilience state supported
Supplier risk governance	Establish accountable oversight of critical SaaS relationships so supplier dependency, contractual assurance and incident obligations are governed as ongoing organisational risk rather than one-off procurement checks.	Maintain a critical SaaS provider register, tier suppliers by operational criticality, review security clauses, assess secure-by-design evidence and agree notification thresholds and escalation contacts.	Anticipate
Identity and access resilience	Reduce the likelihood that compromised users, tokens, service accounts or delegated permissions become pathways for unauthorised SaaS access or cross-platform lateral movement.	Enforce phishing-resistant MFA for privileged accounts, review OAuth grants, rotate API keys, govern non-human identities, apply conditional access and rehearse emergency token revocation.	Absorb, Adapt

SaaS posture and integration management	Keep SaaS tenants, integrations and data-sharing arrangements aligned with approved risk settings as services, users, privileges and configurations change over time.	Run scheduled configuration reviews, approve new integrations, check API scopes, remove orphaned administrators, monitor configuration drift and validate data-sharing settings.	Anticipate
Telemetry and detection	Improve the organisation's ability to detect abnormal SaaS activity quickly by ensuring that relevant provider, identity, administrative and access events are visible and interpretable.	Specify minimum log events, ingest critical SaaS logs into the SIEM system, retain evidence, tune anomaly alerts, clarify provider event semantics and test incident escalation pathways.	Absorb, Adapt
Continuity and recovery	Sustain essential business functions when a critical SaaS service becomes unavailable, untrusted or operationally constrained, and restore trusted operations after disruption.	Prepare SaaS outage playbooks, maintain alternative communication channels, define manual workarounds, test data export or restoration and validate trusted-state recovery procedures.	Recover, Adapt
Learning and adaptive governance	Convert SaaS incidents, near misses and threat intelligence into improved governance routines, supplier expectations, configuration baselines and executive risk awareness.	Conduct post-incident reviews, update supplier assurance criteria, revise control checklists, refresh user guidance, adjust configuration standards and report dependency risk trends to governance forums.	Recover, Adapt

Worked example. The analysis of the Okta 2023 support-system incident shows how the evidence is applied to the framework. **Evidence:** a service account credential saved in an employee's personal Google profile; customer-uploaded HAR files containing live session tokens, used to hijack five customers' administrator sessions; and downloads missed for 14 days because access pathways generated different log event types (Okta, 2023a). **Codes:** "provider credential held outside managed environment"; "diagnostic artefact carries live session token"; "event-type mismatch delays detection". **Pathways:** Themes 1 and 4, enabled by Theme 3, since no party owned the lifecycle of the token inside the HAR file. **States:** customers detected the intrusion first (1Password on 29 September; BeyondTrust on 2 October), whereas only Okta could revoke the tokens, which it did on 17–19 October (adapt), before blocking personal browser profiles and binding administrator sessions to network location (recover) (BeyondTrust, 2023; 1Password, 2023). **Capabilities:** identity and access resilience, telemetry and detection, and supplier risk governance for escalation. Customers with strong identity policies and monitoring reported limited harm, which illustrates P3 rather than testing it.

7. Discussion

All three principal cases were selected after they became public failures, so the analysis cannot show that any capability improves resilience; the capabilities in Tables 1 and 2 are propositions for future testing. For example, organisations that enforce phishing-resistant MFA and network access policies on their SaaS tenants are expected to contain credential-based compromise more effectively, and organisations that ingest provider logs into enterprise monitoring are expected to detect token abuse sooner. With that qualification, the findings suggest that SaaS cyber resilience requires moving beyond questionnaire- and contract-based third-party assurance toward continuous evidence about configurations, access, integrations, logs and incident communication (Boyens et al., 2024; CISA, 2023a). Identity should be treated as a strategic resilience capability, because it is the control plane through which access, integration, delegation and response are enacted. This matters most for organisations that rely on central identity providers and cloud data platforms, where compromise of the service or its customer-side configuration can affect many business functions at once. A further implication is that shared responsibility needs to be operationalised rather than merely acknowledged. Organisations often know that SaaS security is shared, but practical ownership of specific controls remains unclear. Who is accountable for reviewing OAuth integrations? Who approves administrative

access? Who verifies log retention? Who tests provider notification arrangements? Who decides whether a SaaS platform is critical enough to require a continuity plan? These are organisational as much as technical questions, and together they form an agenda for research on how provider–customer resilience routines are designed and governed.

For Information Systems research, the paper positions SaaS cybersecurity as a problem of digital interdependence, in which cyber risk travels through relationships among organisations, platforms, identities and data flows. Technical controls remain necessary but are insufficient unless embedded in organisational routines, decision rights, supplier governance and learning processes. It contributes to dynamic capability theory by showing that, where critical resources belong to a provider, cyber resilience depends on relational routines spanning organisational boundaries. This opens questions about how such capabilities are built across digital ecosystems and how organisations evaluate supplier trustworthiness.

Limitations. The study relies on public sources, many written by parties with liability exposure to the events; for example, the finding that Snowflake's environment was not breached comes from an investigation conducted for the provider. Selecting public failures gives no variance on the outcome, three principal cases support analytic rather than statistical generalisation, and the propositions are untested. Future research should add contrast cases in which customers detected or contained the same attacks, triangulate vendor accounts with customer and regulatory sources, and test the propositions through provider–customer interviews and analysis of contracts and log access.

8. Conclusion

Organisations can no longer rely on static assurance, perimeter-based assumptions or narrow technical controls to manage SaaS cyber risk. The analysis identified four themes: identity, session and token exploitation; central node and concentration risk; shared responsibility and governance ambiguity; and detection blind spots and notification latency. These show SaaS cyber risk arises from the interaction of recurring attack pathways and organisational arrangements, and that cyber resilience in SaaS is a dyadic property of the provider–customer relationship rather than an attribute of either party alone. The paper makes three contributions. Theoretically, it challenges the single-organisation assumption in cyber resilience standards and extends dynamic capability theory with relational dynamic capabilities, through which provider and customer jointly sense, contain and recover from compromise. Analytically, it shows where resilience capacity breaks down: around shared control objects, provider-held telemetry and identities that cross organisational boundaries. Practically, it offers six capabilities and five propositions that organisations can use to assess SaaS dependencies and that researchers can test through contrast cases and provider–customer studies. Together, these reframe SaaS cybersecurity as governing digital interdependence and building relational capability for continuity, adaptation and trust.

9. References

Alvesson, M. and Sandberg, J. (2011) 'Generating research questions through problematization', Academy of Management Review, 36(2), pp. 247–271.

Amit, R. and Schoemaker, P.J.H. (1993) 'Strategic assets and organizational rent', Strategic Management Journal, 14(1), pp. 33–46.

BeyondTrust (2023) BeyondTrust Discovers Breach of Okta Support Unit. BeyondTrust Blog. https://www.beyondtrust.com/blog/entry/okta-support-unit-breach

Boh, W. F., Melville, N. P., Baptista, J., Chasin, F., Horita, F., Ixmeier, A., ... Wilopo, W. (2026). Digital Resilience for the Climate Crisis: A Multi-Perspective Analysis. MIS Quarterly 1 March 2026; 50 (1): 1–34. https://doi.org/10.25300/MISQ/2025/18779

Boyens, J., Smith, A., Bartol, N., Winkler, K., Holbrook, A. and Fallon, M. (2024) Cybersecurity Supply Chain Risk Management Practices for Systems and Organizations. NIST Special Publication 800-161 Revision 1 Update 1. National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.800-161r1-upd1

CISA and FBI (2023) CISA and FBI Release Advisory on CL0P Ransomware Gang Exploiting MOVEit Vulnerability. Cybersecurity and Infrastructure Security Agency and Federal Bureau of Investigation. https://www.cisa.gov/news-events/news/cisa-and-fbi-release-advisory-cl0p-ransomware-gang-exploiting-moveit-vulnerability

Cloud Security Alliance (2024) Security Guidance for Critical Areas of Focus in Cloud Computing v5. Cloud Security Alliance. https://cloudsecurityalliance.org/artifacts/security-guidance-v5

Duchek, S. (2020) 'Organizational resilience: a capability-based conceptualization', Business Research, 13(1), pp. 215–246.

Dyer, J.H. and Singh, H. (1998) 'The relational view: cooperative strategy and sources of interorganizational competitive advantage', Academy of Management Review, 23(4), pp. 660–679.

Hashizume, K., Rosado, D. G., Fernández-Medina, E. and Fernandez, E. B. (2013) 'An analysis of security issues for cloud computing', Journal of Internet Services and Applications, 4(5). https://doi.org/10.1186/1869-0238-4-5

Hollnagel, E. (2009) 'The four cornerstones of resilience engineering', in Nemeth, C.P., Hollnagel, E. and Dekker, S. (eds) Resilience Engineering Perspectives, Volume 2: Preparation and Restoration. Farnham: Ashgate, pp. 117–133.

King, N. (2012) 'Doing template analysis', in Symon, G. and Cassell, C. (eds) Qualitative Organizational Research: Core Methods and Current Challenges. London: Sage, pp. 426–450.

Lella, I., Theocharidou, M., Tsekmezoglou, E., Malatras, A., García, S. and Valeros, V. (2021) ENISA Threat Landscape for Supply Chain Attacks. European Union Agency for Cybersecurity. https://doi.org/10.2824/168593

Linkov, I., Kott, A. (2019). Fundamental Concepts of Cyber Resilience: Introduction and Overview. In: Kott, A., Linkov, I. (eds) Cyber Resilience of Systems and Networks. Risk, Systems and Decisions. Springer, Cham. https://doi.org/10.1007/978-3-319-77492-3_1

Mandiant (2024) UNC5537 Targets Snowflake Customer Instances for Data Theft and Extortion. Google Cloud Threat Intelligence. https://cloud.google.com/blog/topics/threat-intelligence/unc5537-snowflake-data-theft-extortion

Morgan, D.L. (2014) 'Pragmatism as a paradigm for social research', Qualitative Inquiry, 20(8), pp. 1045–1053.

Nelson, R.R. and Winter, S.G. (1982) An Evolutionary Theory of Economic Change. Cambridge, MA: Harvard University Press.

1Password (2023) Okta Support System incident and 1Password. 1Password Blog, 23 October. https://1password.com/blog/okta-incident

Okta (2023a) Unauthorized Access to Okta's Support Case Management System: Root Cause and Remediation. Okta Security. https://sec.okta.com/articles/2023/11/unauthorized-access-oktas-support-case-management-system-root-cause/

OWASP (2026) OAuth 2.0 Protocol Cheatsheet. OWASP Cheat Sheet Series. https://cheatsheetseries.owasp.org/cheatsheets/OAuth2_Cheat_Sheet.html

Pascoe, C., Quinn, S. and Scarfone, K. (2024) The NIST Cybersecurity Framework (CSF) 2.0. NIST Cybersecurity White Paper 29. National Institute of Standards and Technology. https://doi.org/10.6028/NIST.CSWP.29

Progress Software (2023) MOVEit Transfer Critical Vulnerability (May 2023) (CVE-2023-34362). Progress Community. https://community.progress.com/s/article/MOVEit-Transfer-Critical-Vulnerability-31May2023

Ross, R., Pillitteri, V., Graubart, R., Bodeau, D. and McQuaid, R. (2021) Developing Cyber-Resilient Systems: A Systems Security Engineering Approach. NIST Special Publication 800-160 Volume 2 Revision 1. National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.800-160v2r1

Subashini, S. and Kavitha, V. (2011) 'A survey on security issues in service delivery models of cloud computing', Journal of Network and Computer Applications, 34(1), pp. 1–11. https://doi.org/10.1016/j.jnca.2010.07.006

Temoshok, D., Fenton, J. L., Choong, Y. Y., Lefkovitz, N., Regenscheid, A., Galluzzo, R. and Richer, J. (2025) Digital Identity Guidelines: Authentication and Authenticator Management. NIST Special Publication 800-63B-4. National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.800-63B-4

Teece, D.J. (2007) 'Explicating dynamic capabilities: the nature and microfoundations of (sustainable) enterprise performance', Strategic Management Journal, 28(13), pp. 1319–1350.

Teece, D.J., Pisano, G. and Shuen, A. (1997) ‘Dynamic capabilities and strategic management’, Strategic Management Journal, 18(7), pp. 509–533.

Verizon (2026) 2026 Data Breach Investigations Report. Verizon Business. Available https://www.verizon.com/business/resources/reports/2026-dbir-data-breach-investigations-report.pdf

Winter, S.G. (2003) ‘Understanding dynamic and ordinary capabilities’, Strategic Management Journal, 24(10), pp. 991–995.