CHAPTER

Articles

CONTENTS

- Intro
- A workflow for collecting and understanding stories at scale -- Summary (eval2025)
- AI-assisted causal mapping -- Summary (validation)
- Causal mapping for evaluators -- Summary (eval2024)
- KLAR Outcome Harvesting AI pilot (DEZIM) -- Summary (book chapter draft)
- Qualitative causal mapping in evaluations (health) -- Summary (book chapter)
- ToC and causal maps in Ghana -- Summary (book chapter)

Intro

Some brief one-page bullet-point summaries of some of our key published papers.

A workflow for collecting and understanding stories at scale -- Summary (eval2025)

(Powell et al., 2025)

Source: *Evaluation* 31(3), 394–411 (2025).

- **What problem this paper solves**
- Evaluations often start with a ToC and then collect evidence for each link (e.g. Contribution Analysis), but in many real settings the ToC is uncertain, contested, or incomplete.
- The paper proposes collecting evidence about **structure/theory** (what influences what) and **contribution** simultaneously, using a scalable workflow that stays open-ended.

- **Core idea: "AI-assisted causal mapping pipeline"**

- Treat causal mapping as **causal QDA**: each coded unit is an ordered pair **(influence → consequence)** with provenance, rather than a theme tag.
- Use AI as a **low-level assistant** for interviewing + exhaustive extraction, leaving high-level judgement (prompt design, clustering choices, interpretation) with the evaluator.

- **Pipeline (end-to-end)**

- **Step 1 — AI interviewer**: a single LLM "AI interviewer" conducts semi-structured, adaptive chat interviews at scale.
- **Step 2 — Autocoding causal claims**: an LLM is instructed (radical zero-shot) to list *each* causal link/chain and to ignore hypotheticals.
- **Step 2c — Clustering labels**: embed factor labels and cluster them; then label clusters and optionally do a second "deductive" assignment step to ensure cluster cohesion.
- **Step 3 — Analysis via maps/queries**: produce overview maps, trace evidence for (direct/indirect) contributions, compare subgroups/timepoints.

- **Demonstration study (proof-of-concept)**

- Respondents: online workers recruited via Amazon MTurk; topic: "problems facing the USA" (chosen to elicit causal narratives without a specific intervention).
- Data collection repeated across **three timepoints**; data pooled.
- This is an analogue demonstration; not intended to generalise substantively about "the USA".

- **Key results (reported metrics)**

- **AI interviewing acceptability (proxy)**: 78.5% of interviewees did not ask for changes to the AI's end-of-interview summary; 4.29% asked for changes; 15.3% had no summary (drop-off).
- **Autocoding effort/cost**: ~5 hours to write/test coding instructions; ~$20 API cost (in the reported experiment set-up).
- **Autocoding recall/precision**:

    - Ground-truth link count (authors' assessment): 1154 links.
    - AI-identified links: 1024 (≈ 89%) before precision screening.
    - Precision scoring (0–2 on four criteria: correct endpoints; true causal claim; not hypothetical; correct direction): 65% perfect; 72% dropped only one point.
- **Overview-map "coding coverage"**
    - An 11-factor overview map (plus filters) covered ~42% of raw coded claims while remaining readable.
    - Coarse clustering can collapse opposites/valence (e.g. "military strengthening" and "military weakening" both under "International conflict").
- **Interpretation claims**
- The approach is good for sketching **"causal landscapes"** and triaging hypotheses; it is not reliable enough for high-stakes single-link adjudication without human checking.
- Many outputs depend on **non-automated clustering decisions** (number of clusters, labelling intervention), analogous to researcher degrees of freedom in variable construction.
- **Caveats / ethics**
- Not suitable for **sensitive data** when using third-party LLM APIs; risks of bias and hegemonic worldviews are highlighted.
- Differential response/selection into AI interviewing may not be random.
- Causal mapping shows **strength of evidence**, not **effect size**; forcing magnitudes/polarity is risky.

References

Powell, Cabral, & Mishan (2025). *A Workflow for Collecting and Understanding Stories at Scale, Supported by Artificial Intelligence*. SAGE PublicationsSage UK: London, England. https://doi.org/10.1177/13563890251328640.

AI-assisted causal mapping -- Summary (validation)

(Powell & Cabral, 2025)

- **Goal / research question**
- Test whether an **untrained LLM** can **identify and label causal claims** in qualitative interview "stories" well enough to be useful, compared with **human expert coding** (a criterion study).
- Focus is on **validity/usefulness of causal-claim extraction**, not causal inference.

- **Core framing: causal mapping vs systems modelling**

- In systems mapping, an edge (X \rightarrow Y) is often read as "(X) causally influences (Y)".
- In causal mapping (as used here), an edge means **there is evidence that (X) influences (Y) / a stakeholder claims (X) influences (Y)**.
- Output is therefore a **repository of evidence with provenance**, not a predictive system model.

- **"Naive" (minimalist) causal coding definition**

- Deliberately avoids philosophical detail; codes **undifferentiated causal influence** only.
- Does **not** encode effect size/strength; does **not** do causal inference; does **not** encode polarity as a separate field (left implicit in labels like "employment" vs "unemployment").
- Coding decision reduced to: **where is a causal claim, and what influences what?**

- **Data and criterion reference**

- Corpus from a **QuIP** evaluation (2019) of an "Agriculture and Nutrition Programme".
- Dataset previously hand-coded by expert analysts (used as a **criterion study**).
- Validation subset: **3 sources**, **163 statements**, **~15 A4 pages**.

- **Extraction procedure (AI as low-level assistant)**

- Implemented via the **Causal Map** web app using **GPT-4.0**.
- Temperature set to **0** for reproducibility.
- AI instructed to produce an **exhaustive, transparent** list of claims with **verbatim quotes**; synthesis is done later by causal mapping algorithms.
- Exclusions: **ignore hypotheticals/wishes**.
- Output per claim: statement ID + quote + influence factor + consequence factor.

- **Two validation variants**

- **Variant 1 — open coding ("radical zero-shot")**
  - No codebook; includes an "orientation" so the AI understands the research context.
  - Uses a multi-pass prompting process (initial extraction + revision passes).
- **Variant 2 — codebook-assisted ("closed-ish")**
  - Adds a partial codebook (most-used top-level labels from the human coding).

    - Uses hierarchical labels `general concept; specific concept`.
- **Validation metrics and headline results**
- **Precision** (human-rated, four criteria): correct endpoints; correct causal claim; not hypothetical; correct direction.
    - Variant 1: 180 links; perfect composite score (8/8) for **84%** of links.
    - Variant 2: 172 links; perfect composite score (8/8) for **87%** of links.
- **Recall (proxy)**: compared link counts vs the human-coded set (acknowledging no true ground truth because granularity is underdetermined).
- **Utility check (overview-map similarity)**
- Detailed maps differ (expected in qualitative coding).
- When zoomed out to top-level labels and filtered to the most frequent nodes/links, AI and human overview maps are **broadly similar**.
- **Scope limits / risks**
- Small sample; single (relatively "easy") dataset; informal rating process.
- Label choice/consistency remains a major source of variation; batching can introduce inconsistency across prompts.
- Suitable for mapping "how people think" and building auditable evidence sets; not suitable for high-stakes adjudication of specific links without checking.

References

Powell, & Cabral (2025). *AI-assisted Causal Mapping: A Validation Study*. Routledge. https://www.tandfonline.com/doi/abs/10.1080/13645579.2025.2591157.

Causal mapping for evaluators -- Summary (eval2024)

(Powell et al., 2024)

Source: (DOI: 10.1177/13563890231196601)

- **History / lineage (why this isn't "new", just under-used in evaluation)**: Causal mapping (diagramming "what causes what" using directed links between factors) has been used since the **1970s** across disciplines (e.g. Axelrod-style **document coding** of causal assertions; management/OR traditions emphasising maps for **decision support**; comparative methods like Laukkanen's work on **standardising factor vocabularies** and combining maps). The evaluation literature has relatively sparse, inconsistent "causal mapping" usage; this paper synthesises the wider literature and re-specifies it for evaluators.
- **How we pitch it to evaluators (the niche)**: treat causal mapping as a **discrete evaluation task**: (i) systematically **assemble causal evidence from narrative sources** into an explicit link database with provenance, then (ii) separately use that assembled evidence to make evaluative judgements about "what is really happening". This is positioned as a way to work with large bodies of **messy, heterogeneous** qualitative causal data (different boundaries, contexts, specificity, and ambiguity) without forcing early convergence on a single prior ToC.
- **How causal mapping differs from adjacent approaches**:
- **Primary object is evidence-with-provenance**: causal mapping is explicitly about *who/what source said what link*, not a modeller's best estimate of system structure.
- **Epistemic first, ontic later**: unlike approaches mainly aimed at simulation/prediction (e.g. SD/BBNs/CLDs/FCMs as typically used), causal mapping foregrounds **organising claims/evidence**; inference about reality is a later step.
- **Lightweight causal typing**: it usually does not require consistent weights/functional forms/necessity-sufficiency labels; it can incorporate them when elicited, but warns about spurious precision.
- **How causal-mapping approaches differ among themselves (key axes)**:
- **Mode of construction**: coding **documents** vs coding **interviews** vs **group** map-building (consensus/problem-structuring) vs hybrids.
- **Elicitation openness**: **closed** (pre-specified factor lists) vs **open** (respondent-generated factors), with chaining variants (forward/back).
- **Single-source vs multi-source & context handling**: idiographic maps vs aggregated multi-source maps; whether and how you track **case/context metadata** to avoid invalid transitive inferences.
- **Coding philosophy**: "factors as variables" vs "factors as **changes**" (e.g. QuIP-style); whether polarity/opposites are represented as separate factors/links or handled differently; extent of factor-name **standardisation/merging/nesting**.

- **Problem / motivation**: Evaluators need to represent (a) what causally influences what **in the world**, and (b) what different stakeholders **claim/believe** causally influences what. Causal mapping—defined as the **collection, coding, and visualisation of interconnected causal claims** with explicit **provenance**—is widely used outside evaluation, but under-specified in evaluation practice/literature.
- **Core argument (the "Janus" dilemma + resolution)**:
- **Janus dilemma**: Causal mapping faces two directions—maps can be read as **models of beliefs** or as **models of causal reality**; in practice these get blurred unless source information and analysis steps are explicit.
- **Resolution**: Treat causal maps primarily as **repositories of causal evidence** (epistemic objects), not as direct models of either beliefs or reality. Maps then support structured questions like: *Is there evidence X influences Z? Directly/indirectly? How much evidence? How many sources? How reliable?* The *evaluation* step that judges "what is really happening" is distinct and subsequent.
- **What causal maps encode (and don't)**:
- **Epistemic content**: Map elements are claims/perceptions/evidence, not facts.
- **Causal semantics are usually coarse**: ordinary language claims typically encode **partial influences**, not total/necessary/sufficient causation; coding a link need not assert evidence quality (though you may later weight/filter by quality).
- **Multiple sources + contexts**: maps may be single-source or multi-source; inference across sources requires care about **which case/context** each link refers to.
- **Boundaries are often messy**: system boundaries are frequently loose/implicit; mapping can proceed, but ambiguity must be managed rather than hidden.
- **Causal mapping in evaluation = 3 tasks (workflow)**:
- **Task 1 — Gather narrative causal material**: interviews, open-ended survey questions, document/literature review, archives/secondary text, or consensus processes (e.g., Delphi, participatory systems mapping). Elicitation may use **back-chaining** ("what influenced X?") and **forward-chaining** ("what followed/could follow?"). Question framing affects factor semantics (e.g., QuIP tends to elicit **changes** like "reduced hunger" rather than variables).
- **Task 2 — Code causal claims ("causal QDA")**: unlike standard thematic QDA (codes = concepts), causal QDA codes **links**: each highlighted quote yields an **influence factor → consequence factor** pair; factors mainly exist as endpoints of links. Labelling can be **exploratory/inductive** (curate a common vocabulary across sources) or **confirmatory** (codebook from a ToC/prior work), with sequencing cautions to reduce framing/bias. Manual coding is costly; partial automation via NLP/ML is possible but not the focus.
- **Task 3 — Answer evaluation questions using the link database**: global maps become "hairballs", so analysis should generate **selective maps** aligned to questions (e.g., consequences of an intervention; causes of a valued outcome). Techniques include bundling **co-terminal links** (thickness/count), producing frequency-based overview maps (caution:

rare-but-important links), rolling-up hierarchical factor taxonomies (with caveats), and limited quantitative summaries (warning: sensitive to coding granularity).

- **Limits / risks**:

- **Inference depends on source credibility**: stronger conclusions require explicit, context-specific **rules of inference** (e.g., independent mentions threshold + theoretical plausibility + bias-mitigation steps).
- **Effect strength/type is hard to capture**: respondents rarely provide consistent magnitudes/necessity/sufficiency/certainty; forcing weights risks **spurious precision**.
- **Transitivity is both payoff and trap**: inferring (C \rightarrow E) from (C \rightarrow D) and (D \rightarrow E) is powerful for indirect effects, but can be invalid when links come from **non-overlapping contexts**; valid inference requires attention to the **intersection of contexts**.

- **Concrete analytic contributions highlighted**:

- Treat diagrams as an **index into the underlying corpus**: tool support should allow tracing from any link/factor back to transcript excerpts + source metadata.
- Quantify robustness of evidence-based "arguments" along paths using **maximum flow / minimum cut** on the causal-claim network (how many claims would need removal to eliminate all paths between (C) and (E)), plus **source thread count** (how many distinct sources each provide a complete path).

- **Conclusion / evaluator-facing payoff**:

- Helps evaluators (i) assemble narrative evidence about intervention and contextual influences (direct/indirect, intended/unintended), (ii) search/summarise/select quotations systematically, (iii) increase transparency/peer-reviewability of qualitative causal reasoning, (iv) communicate complexity with readable graphics.
- Key discipline is a **two-step separation**: first assemble and organise causal evidence; then judge what is actually happening—avoiding premature constraint of data collection to fit a prior ToC that stakeholders may not share.

References

Powell, Copestake, & Remnant (2024). *Causal Mapping for Evaluators*. https://doi.org/10.1177/13563890231196601.

KLAR Outcome Harvesting AI pilot (DEZIM) -- Summary (book chapter draft)

Source: draft chapter in `content/000 Articles/020 !! dezim klar book chapter (DRAFT).md`.

- **Purpose**
- Pilot an AI interviewer (“Harvest Assistant”) for Outcome Harvesting (OH) in the KLAR! programme, focusing on scalability, inclusion, and democratic evaluation value.

- **Key method move**

- AI-led interviewing + AI post-hoc transcript analysis to draft a structured OH outcome table.
- Strong emphasis on **traceability**: verbatim short citations + page references; explicit missing-information prompts.
- Human validation checks accuracy and de-duplicates overlapping outcomes across sources (triangulation).

- **Results highlights (as reported)**

- 39 invited; 19 responded; 38 outcome statements; 6 met SMART criteria; others retained as leads.
- Real-time outcome summaries enable respondent validation and transparency.

- **Operational insights**

- Prompt simplicity improves adherence; model choice matters; version prompts/models for comparability.
- Scaling shifts bottlenecks to analysis unless the end-to-end workflow is designed.

- **Risks / responsible scaling**

- GDPR/legal basis, consent, third-party naming pathways; document data flow and model/prompt versions; data sovereignty (EU-hosted inference where possible); attention to equity/digital divides.

Qualitative causal mapping in evaluations (health) -- Summary (book chapter)

(Remnant et al., 2025)

Source: book chapter draft in `content/000 Articles/020 !! health book chapter.md`.

- **Purpose**
- Position QuIP + causal mapping as a credible, cost-effective way to elicit and analyse perceived drivers/barriers in complex interventions (including health services evaluations).

- **Data collection stance**

- QuIP focuses on *changes* that matter to respondents, and the perceived causes of those changes.
- Goal-free / blindfolded questioning is used to reduce pro-project bias; unprompted mention is treated as important evidence.
- Not designed to estimate effect sizes; complements (rather than replaces) quantitative inference and other theory-based approaches.

- **Coding stance ("natively causal")**

- Coding is not thematic tags that are linked later; coding is **pairs/chains of cause→effect factors** ("causal nuggets").
- Coding is parsimonious: only causal claims are coded; non-causal descriptive text is not.
- Inductive label harmonisation across sources is expected; analyst should manage positionality and avoid over-fitting to prior ToC.

- **Use**

- Compare empirical causal maps against ToCs; compare groups (e.g. men/women; staff cadres) and pathways.
- Keep a traceable link from visual summaries back to underlying quotes for verification/peer review.

- **Relationship to realist ideas**

- Affinity to mechanism/context thinking (multiple pathways), but with broader open capture rather than only a few "hotspots".

References

Remnant, Copestake, Powell, & Channon (2025). *Qualitative Causal Mapping in Evaluations*. In *Handbook of Health Services Evaluation: Theories, Methods and Innovative Practices*. https://doi.org/10.1007/978-3-031-87869-5_12.

ToC and causal maps in Ghana -- Summary (book chapter)

(Powell et al., 2023)

Source: book chapter draft in `content/000 Articles/020 !! toc book chapter.md`.

- **Purpose**
- Show how QuIP-style causal mapping can compare an official programme ToC ("their theory") with empirically coded beneficiary narratives ("our theory"), as a disciplined way to revise ToCs and "middle-level theory".

- **Minimal definition (what a causal map is)**

- A causal map is nodes + directed links, where a link means (at minimum) *someone believes C influenced E*.
- Links need not encode necessity/sufficiency, nor quantified strength/polarity (though those are sometimes added in other approaches).

- **QuIP as causal mapping**

- Goal-free / (partially) blindfolded elicitation of stories of change reduces confirmation bias.
- "Causal back-chaining" elicits causes, causes-of-causes, etc.
- Coding is inductive and multi-source; maps are then filtered/queried to answer evaluation questions.

- **How analysis is actually done**

- Global maps are too large; use filters (e.g. theme/keyword searches, distance steps, frequency thresholds).
- **Hierarchical coding / zooming out**: encode subfactors in factor labels so detailed factors can be rolled up into higher-level factors for readable summary maps.
- Evidence strength is often shown with counts on links (mentions / sources).

- **Interpretation pitfalls (explicitly listed)**

- **Beliefs about causation are not facts about causation**: evaluator judgement remains separate.
- **Absence of a mentioned link is not evidence of absence** (random-walk conversations; negative cases).
- **Transitivity trap / context overlap**: stitching A→B (source 1) and B→C (source 2) does not justify A→C unless contexts overlap.
- Aggregation/generalisation is non-trivial; counts support confidence but don't convert to "truth percentages".

References

Powell, Larquemin, Copestake, Remnant, & Avard (2023). *Does Our Theory Match Your Theory? Theories of Change and Causal Maps in Ghana*. In *Strategic Thinking, Design and the Theory of*

Change. A Framework for Designing Impactful and Transformational Social Interventions.

