AI as a Manager Co-Pilot: Augmenting Leadership Without Replacing Judgment

Executive summary

The strongest case for an AI manager co-pilot is not that management is "administrative" and can therefore be automated. It is the opposite: management quality has outsized organizational impact, while the work itself is cognitively fragmented, socially exposed, and emotionally demanding. Gallup[1] estimates that managers account for at least 70% of the variance in team engagement, while recent Gallup research shows managers are disproportionately burned out, undertrained, and squeezed between rising employee expectations and top-down execution pressure. At the same time, management research summarized in MIT Sloan Management Review[2] and Harvard Business Review[3] argues that AI is increasingly useful as a thought partner for coaching, reflection, and structured problem-solving. [4]

The research does **not** support building an AI that tells managers what to do in consequential people decisions. Trust in automation is a calibration problem, not an adoption problem. Classic work shows that under-trust leads to disuse, over-trust leads to misuse, and both are costly. More recent human-AI studies show that human-AI combinations improve on average over unaided humans, but often still underperform the better solo performer; explanations alone frequently increase reliance on both correct and incorrect AI outputs; and users make better decisions when systems provide sources, visible uncertainty, an easy way to correct outputs, and friction that prompts reflection when stakes are high. [5]

The product implication is clear: the right design is a **suggestion-based, non-directive, manager-invoked co-pilot**. It should synthesize context, generate options and coaching questions, help managers rehearse difficult conversations, surface missing evidence and bias risks, and capture follow-through after the human interaction. It should **not** issue final ratings, promotion recommendations, compensation decisions, or termination decisions. Those belong to accountable humans, with escalation to HR, skip-level leaders, and legal review when stakes are high. This approach aligns with the NIST[6] AI Risk Management Framework and with the stage-setting work from the MIT Center for Information Systems Research[7], which emphasizes acceptable-use policies, data accessibility, and identifying where humans must remain in the loop. [8]

If built this way, a manager co-pilot should be evaluated less on raw usage and more on **appropriate reliance**, coaching quality, manager burden, employee clarity, development follow-through, psychological safety, and regretted attrition. The design target is not "AI adoption." It is better management behavior at scale, without eroding judgment, trust, or accountability. [9]

Scope and assumptions

This report assumes a B2B knowledge-work environment with roughly 500 to 10,000 employees, front-line and middle managers with six to twelve direct reports, English-language data, and an existing stack that includes calendar, documents, HRIS, and performance artifacts. The goal is to augment coaching, feedback, 1:1 preparation, goal alignment, and post-conversation follow-through. It is not to automate employment decisions.

Where assumptions matter, the report favors conservative design choices: manager-invoked workflows over always-on nudging, recommendations over prescriptions, auditable sources over hidden model reasoning, and explicit escalation for high-stakes use cases.

Evidence base

Trust in automation

The foundational trust finding is that people do not simply "trust" or "not trust" automation. They rely on it more or less appropriately depending on context, perceived reliability, goals, and their own mental model of the system. Lee and See's classic review argues that trust should be designed for **appropriate reliance**, because complex tasks make full understanding of automation impractical and people therefore use trust as a shortcut for deciding when to rely. Parasuraman and Riley similarly distinguish between use, misuse, disuse, and abuse of automation; in practice, that means the same system can fail because people defer too much, defer too little, or are forced into badly designed automation they do not control. [10]

Modern evidence adds an important nuance: managers may show both **algorithm aversion** and **algorithm appreciation**, depending on task design. One line of work shows that people become more likely to reject even superior algorithms after seeing them make mistakes. Another shows that in some numerical-estimation tasks, people adhere more to advice when they believe it comes from an algorithm than from another person. A third finding is especially relevant for product design: people are much more willing to use an imperfect algorithm when they are allowed to modify it, even slightly. In management settings, that argues for editable suggestions, visible room for human override, and interaction patterns that preserve agency without pretending the human is starting from scratch. [11]

The practical design consequence is not "make the AI more persuasive." It is "make the AI easier to calibrate." That means showing what the system can do, how well it may do it, why it produced a suggestion, and making invocation, dismissal, and correction easy. Those are central recommendations in the human-AI interaction guidelines from Microsoft[12] researchers. [13]

Decision support systems

The most relevant decision-support lesson is that “human in the loop” is not automatically better. A 2024 meta-analysis covering 106 experiments found that human-AI combinations outperformed humans alone on average, but did **not** outperform the better of the human-only or AI-only baselines on average. Performance was more promising when humans were stronger than AI and when tasks involved content creation; it was worse when AI was already stronger, because humans often struggled to determine when to trust the model and when to trust themselves. [14]

That matters because much of managerial work is not pure prediction. It mixes diagnosis, narrative framing, conflict handling, developmental judgment, and moral trade-offs. In these settings, AI is most useful when it improves **human preparation and reflection**, not when it substitutes for judgment. HBR’s recent management writing points in the same direction: AI can reduce the burden of coaching preparation and act as a thought partner for weighing pros and cons, but the final managerial act remains human. [15]

Research on explainability is also sobering. In one large human-AI study, explanations did not increase complementary team performance beyond simpler confidence-based approaches; instead, they increased the chance that humans accepted the AI’s recommendation **regardless of whether the AI was correct**. More recent work on LLMs reaches a similar conclusion: explanations increase reliance on both correct and incorrect responses, whereas sources and visible inconsistencies reduce reliance on incorrect outputs. A related study found that “cognitive forcing” interventions can reduce overreliance, but at a cost in user satisfaction. The lesson is that a manager co-pilot should not present a single polished recommendation early and expect the manager to remain independent. It should structure reflection. [16]

A particularly useful design refinement comes from work showing that trust calibration improves when systems estimate not just AI confidence but the likely strength of the human’s own judgment for the specific case. In managerial terms, the product should behave differently when the evidence is thin, the context is ambiguous, or the manager has domain knowledge the model lacks. In those cases, the system should shift from “suggest” to “ask” mode. [17]

Emotional labor of management

Management is emotional labor. Research reviews define emotional labor as regulating emotions to meet role requirements, and leadership research argues that leaders use emotional labor to influence followers’ moods, motivation, and performance. The distinction between **deep acting** and **surface acting** matters: trying to genuinely align one’s emotions with one’s role can support performance, while “faking it” more often carries wellbeing costs. [18]

This is directly relevant to management product design because many manager tasks are not analytically difficult but **interpersonally difficult**: giving candid feedback without

humiliation, holding boundaries without cruelty, communicating uncertainty without spreading panic, and sustaining fairness when one employee is struggling and another is quietly excelling. HBR's work on leadership emotional labor describes leaders as repeatedly projecting optimism, steadiness, or confidence even when they themselves are uncertain or skeptical. Gallup's current research shows the operational cost of that burden: managers are more likely than non-managers to be disengaged, burned out, looking for a new job, and feeling unsupported. [19]

The implication is that AI's most defensible role is **load reduction and rehearsal**, not emotional substitution. A co-pilot can help a manager prepare for a hard conversation, identify likely emotional triggers, generate respectful wording, summarize past commitments, and remind the manager what not to ignore. It cannot authentically repair a trust breach, absorb moral responsibility, or bear the social consequences of a bad call. Where management is relationship work, AI can support the manager's preparation but cannot become the relationship. [20]

Limits of AI decision-making

The strongest reason not to automate managerial judgment is not merely that models can be wrong. It is that consequential management decisions are **socio-technical**, value-laden, and often underdetermined by the available data. NIST's trustworthiness criteria explicitly include validity, reliability, safety, accountability, transparency, explainability, privacy, and fairness. Those are not optional extras for people decisions. They are prerequisites. [21]

Empirical work adds two more cautions. First, users systematically overestimate the accuracy of LLM-generated answers, especially when those answers include default explanations. Second, AI behavior in moral scenarios can shape human moral decisions and alter felt responsibility. Meanwhile, work highlighted by Harvard Business School[22] emphasizes that human experience and judgment remain critical because AI cannot reliably distinguish strong ideas from weak ones or set long-term strategic direction by itself. In manager workflows, those findings argue against delegating promotions, layoffs, performance ratings, or disciplinary decisions to a model, even when the model can summarize evidence or simulate trade-offs. [23]

A manager co-pilot should therefore treat sensitive inferences cautiously. It should summarize observed facts, ask about missing context, and flag risks. It should not assert that an employee is "unmotivated," "high potential," "low empathy," or "flight risk" as if those were objective truths. In people management, categories become labels, and labels become outcomes. That is precisely where judgment has to remain accountable and human. [24]

Product blueprint

The product strategy can be summarized in one sentence: **build a reflective coach with evidence-backed options, not an automated manager.** The co-pilot should begin from

manager intent, then help the manager think better, communicate better, and follow through better. It should never bypass the manager's role in sensemaking and accountability. [25]

Design pattern	What it does	Strengths	Risks	Recommendation
Directive next-best action	Tells the manager what to say or decide	Fast, simple to use	Overreliance, anchoring, authority laundering, loss of accountability	Do not use for people decisions
Evidence-backed option set	Presents 3–4 options with rationale, uncertainty, and sources	Preserves agency, supports calibration, more auditable	Slightly slower; requires good source quality	**Core pattern**
Reflective coaching mode	Asks questions, challenges assumptions, suggests missing evidence	Best for judgment, fairness, and emotional preparation	Can feel less "efficient"	**Core pattern**
Live in-meeting nudges	Provides silent prompts during conversations	Useful in rehearsed, consent-based contexts	Social disruption, authenticity concerns, surveillance risk	Use sparingly; opt-in only
Autonomous people decisioning	Drafts ratings or recommendations as defaults	High throughput	Unacceptable legal, ethical, and trust risk	Prohibit

Note: This comparison synthesizes the trust-calibration literature, algorithm-control findings, explanation/overreliance studies, and human-AI interaction design guidelines. [26]

The recommended product has five modules.

Context pack builder. Before a 1:1, feedback discussion, or staffing decision, the co-pilot should synthesize recent commitments, goals, relevant artifacts, and open questions. It should always label the provenance of each fact and ask the manager to correct the record before generating suggestions. This is where "show the source," "make correction easy," and "learn cautiously" matter most. [27]

Option canvas. The system should generate multiple ways forward, not a single answer: for example, “coach now,” “gather more evidence,” “clarify expectations,” or “escalate for support.” Each option should show expected upside, downside, assumptions, and the evidence it rests on. When confidence is low, the product should explicitly say so and shift to a question-led format. [28]

Conversation rehearsal. For emotionally charged interactions, the system should simulate likely employee reactions, propose opening questions, help the manager phrase candid feedback respectfully, and identify what emotional labor the interaction may require. This is one of the clearest ways AI can save time while improving managerial quality without impersonating empathy. [29]

Bias and fairness check. Before any consequential call, the co-pilot should ask structured questions such as: What evidence would reverse this judgment? Have peers in similar roles been evaluated on the same standard? Are you reacting to recent salience rather than the full period? What context might this employee or team add that is missing from the record? These prompts are more defensible than predictive “talent scores.” [30]

Follow-through tracker. After the human interaction, the system should draft notes, capture commitments, suggest the next check-in interval, and create a manager reflection prompt. This is crucial because managerial effectiveness often fails less from poor intent than from inconsistent follow-through. Project Oxygen at Google[31] emphasizes coaching, communication, development, empowerment, inclusion, and decision quality; those are behaviors the product can reinforce through repetition and reminders. [32]

Concrete prompt patterns for the co-pilot should sound like coaching, not commands. Useful examples include:

- “Give me three opening questions for a developmental 1:1 with a high performer who seems disengaged.”
- “Summarize the objective evidence I should reference in this feedback discussion, and separate it from my interpretations.”
- “What am I not seeing that could explain this employee’s recent performance dip?”
- “Draft two ways to communicate a firm boundary without sounding punitive.”
- “Role-play this conversation as the employee. What objections or emotions are most likely?”
- “Challenge my current view. What evidence would support a different conclusion?”
- “What assumptions in my draft feedback could sound identity-based rather than behavior-based?”
- “Convert this conversation into three follow-up commitments with owners and dates.”
- “What should I ask before I make a judgment on promotion readiness?”

These prompt patterns align with the most coach-like behaviors in management research while avoiding direct substitution of human judgment. [33]

Collaboration model

```
flowchart TD
    A[Manager intent and case context] --> B[AI context pack: facts, history,
sources, uncertainties]
    B --> C[Manager reviews and corrects context]
    C --> D[AI generates options, coaching questions, and risk checks]
    D --> E{Risk tier}
    E -->|Informational or developmental| F[Manager selects and edits plan]
    E -->|Consequential people decision| G[Escalate to HR, skip-level, or
legal review]
    F --> H[Manager holds conversation and makes judgment]
    G --> H
    H --> I[AI drafts notes, commitments, and follow-up]
    I --> J[Manager approves, sends, and schedules next step]
    J --> K[Outcomes logged for audit, learning, and drift monitoring]
    K --> L[Governance review of usage, bias, and policy compliance]
```

This model intentionally keeps the AI in an assistive role and the manager in the accountable role. It also separates **developmental support** from **consequential decisions**. That distinction is critical both for trust and for governance maturity: early-stage enterprise AI programs need explicit acceptable-use policies and clear identification of where humans must stay in the loop. [34]

A practical operating rule is to use three risk tiers. **Informational** tasks include summarization and preparation. **Developmental** tasks include coaching prompts, rehearsal, and follow-up planning. **Consequential** tasks include formal ratings, promotions, compensation recommendations, discipline, and exits. The first two tiers can be AI-assisted with manager control; the third should trigger extra review and must not default to model recommendations. [24]

Failure modes and safeguards

Failure mode	What it looks like in practice	Safeguard
Over-reliance	Manager follows polished AI advice without checking context	Sources, uncertainty labels, required context confirmation, post-hoc audit of appropriate reliance
Under-trust	Managers ignore useful support because early outputs disappoint	Editable outputs, transparent limits, training, low-stakes early use cases
Anchoring	First AI suggestion narrows the manager's thinking	Show multiple options; collect manager view before revealing AI recommendation in higher-stakes workflows
Authority laundering	"The AI recommended it" becomes rhetorical cover	UI language that forbids defaulting to AI authority; accountability banner naming human owner

Failure mode	What it looks like in practice	Safeguard
Empathy theater	AI-generated language sounds caring but inauthentic	Rehearsal support only; manager edits everything sent or spoken
Surveillance chill	Employees feel watched if meetings are analyzed by default	Explicit consent, visible recording state, narrow data retention, customer-controlled invocation
Bias amplification	Historical inequities shape prompts, summaries, or risk flags	Fairness testing, demographic review, counterfactual prompts, no opaque talent scores
Context rot	Model uses stale goals, old notes, or incomplete evidence	Freshness indicators, source timestamps, manager confirmation before use
De-skilling	Managers outsource difficult thinking over time	Reflection prompts, skill-building feedback, periodic “no AI” calibration exercises
Metric gaming	Managers optimize usage or template completion rather than quality	Reward outcomes and audited quality, not tool clicks
Privacy leakage	Sensitive personal or health data enter prompts or summaries	Strict policy controls, redaction, purpose limitation, audit logs

Note: These safeguards follow directly from the trust-calibration literature, overreliance studies, fairness guidance, and emotional-labor research. [35]

The most important operational safeguard is to measure **appropriate reliance** rather than raw acceptance. A good system is not one that managers accept frequently. It is one that managers follow when the AI is genuinely helpful and override when local knowledge, nuance, or evidence justify human judgment. [36]

KPI mapping

The table below translates product capabilities into measurable management outcomes. The links are hypotheses grounded in the manager-effectiveness, engagement, development, and turnover literature.

AI feature	Leading manager KPI	Team and retention KPI	Why it should move
Context pack builder	Prep time per 1:1, meeting readiness score, note completeness	Clear expectations, employee confidence in manager, reduced follow-up slippage	Better preparation improves conversation quality and consistency

AI feature	Leading manager KPI	Team and retention KPI	Why it should move
Coaching-question generator	Frequency of meaningful 1:1s, coaching quality rubric, manager confidence before difficult conversations	Perceived support for growth, internal mobility, reduced turnover intent	Developmental support from managers is closely tied to growth and retention
Feedback rehearsal	Specificity and respect scores in written or observed feedback, appeal/rework rate	Fairness perceptions, psychological safety, performance improvement success	Rehearsal lowers avoidable friction in hard conversations
Option canvas with trade-offs	Appropriate reliance rate, decision-cycle quality, escalation accuracy	Better promotion calibration, fewer regretted people calls, stronger trust	Structured options reduce both snap judgments and blind AI deference
Bias and fairness prompts	Evidence completeness, comparator use, override rationale quality	Reduced disparity in outcomes, higher perceived fairness	Counterfactual prompting improves review discipline
Follow-through tracker	Commitment completion, overdue action closure, next-step clarity	Stronger development follow-through, engagement, lower regretted attrition	Many manager failures are execution failures, not intention failures
Reflection summaries	Manager learning loop completion, burnout check-ins, peer-consult rate	Higher manager engagement and retention	Reflection routines reduce the “manager squeeze” and improve support quality

Note: The management-outcome logic is anchored in Gallup’s findings on manager influence; Project Oxygen’s behavioral model at Google; Gallup’s evidence that supervisor support is central to development and turnover intent; and Gallup’s meta-analysis linking engagement to turnover and performance outcomes. [37]

A few metric choices deserve special emphasis. First, include a **manager burden index**: time spent preparing for 1:1s, time spent writing feedback, and end-of-week manager energy. The product should improve quality without silently shifting cognitive load elsewhere. Second, include **employee-side** measures, not just manager self-reports: clarity of expectations, feeling cared about, support for development, fairness, and trust in the manager. Third, track **regretted attrition**, not just gross turnover, because the product aims to preserve valuable employees rather than merely reduce exits in aggregate. [38]

Validation plan

The validation strategy should use a sequence of experiments, each designed to answer a different question.

Product A/B test.
Use this to validate interaction design, not business impact. Compare a baseline interface that gives a polished recommendation plus explanation against the recommended interface: multiple options, sources, uncertainty, and editable context. Primary outcomes should be appropriate reliance on benchmark managerial scenarios, suggestion dismissal/correction rates, manager-rated usefulness, and time-to-decision. Under reasonable assumptions for a session-level binary outcome with baseline helpfulness around 60% and a minimum detectable effect of 4 percentage points, an illustrative sample is about **2,400 sessions per arm** for 80% power at a two-sided 5% alpha. Assumptions should be reported openly. The design question here is whether the system makes managers think better, not whether it feels smoother. [39]

Manager-level randomized controlled trial.
Use this to test whether the co-pilot changes manager behavior over one performance cycle. Randomize managers, not employees, to avoid contamination within teams. Compare the full co-pilot against an active control such as note summarization only. Run for 10 to 12 weeks. Primary outcomes should include employee-rated coaching quality, clarity of expectations, support for development, 1:1 completion rates, and manager burden/burnout. For a standardized effect size of **0.20** on manager-level survey outcomes, a practical rule of thumb is roughly **400 managers per arm**; allowing 15% attrition suggests planning for about **470 managers per arm**. If employee responses are analyzed individually, report team size and ICC and adjust standard errors or sample calculations accordingly. [40]

Longitudinal rollout study.
Use this to test whether behavior change translates into retention and talent outcomes. If pure randomization is politically difficult, use a stepped-wedge or matched difference-in-differences rollout by business unit over 12 to 18 months. Primary outcomes should be regretted attrition, internal mobility, promotion calibration, employee engagement, and manager engagement. As an illustrative assumption set, detecting an absolute drop in regretted attrition from **12% to 10%** with 80% power requires about **3,800 employees per arm** in a simple two-arm design; with average team size of eight and ICC around 0.02, that becomes about **4,300 employees** or approximately **540 managers per arm**. Because retention moves slowly, treat this as a second-order validation, not a launch gate. [41]

Safety and fairness review.
Before broad rollout, run structured red-team evaluations on realistic cases: bias-sensitive feedback, poor-quality documentation, hybrid-work misunderstandings, burnout signals, and borderline promotion or performance cases. Review outputs across demographic slices and role levels, and audit whether the system overstates confidence, introduces

unsupported trait language, or nudges toward premature judgment. This should be a standing governance process, not a one-time prelaunch checklist. [42]

Across all experiments, the north-star metric should be **better management behavior with preserved accountability**. Usage, satisfaction, and time saved matter, but they are secondary. A manager co-pilot is successful if employee experience improves, managers remain the accountable decision-makers, and the organization sees stronger coaching, clearer expectations, and better retention of valued talent. [43]

Open questions and limitations

Several questions remain open. First, the evidence base for trust calibration is strong, but much of it comes from broad human-AI decision settings, not manager-specific deployments. Second, “emotional labor” is culturally variable, so prompt and language design will need localization and fairness testing. Third, live conversation support raises unresolved consent, privacy, and authenticity questions that make pre-meeting and post-meeting workflows safer starting points. Fourth, some interventions that reduce overreliance also reduce user satisfaction, so product teams should expect a real tension between friction and safety. Finally, long-term effects on managerial capability need monitoring: a co-pilot that saves time in quarter one may still be harmful if it quietly weakens managerial judgment in year two. [44]

---

[1] [34] https://cisr.mit.edu/publication/2024_1201_EnterpriseAIMaturityModel_WeillWoernerSebastian

https://cisr.mit.edu/publication/2024_1201_EnterpriseAIMaturityModel_WeillWoernerSebastian

[2] [9] [17] [28] [36] https://mingyin.org/paper/CHI-23/who_to_trust_camera.pdf

https://mingyin.org/paper/CHI-23/who_to_trust_camera.pdf

[3] [8] [12] [21] [22] [24] [30] [31] [42] https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf

https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf

[4] [6] [7] [37] [40] [43] https://news.gallup.com/businessjournal/182792/managers-account-variance-employee-engagement.aspx

https://news.gallup.com/businessjournal/182792/managers-account-variance-employee-engagement.aspx

[5] [10] [35] https://journals.sagepub.com/doi/10.1518/hfes.46.1.50_30392

https://journals.sagepub.com/doi/10.1518/hfes.46.1.50_30392

[11] https://marketing.wharton.upenn.edu/wp-content/uploads/2016/10/Dietvorst-Simmons-Massey-2014.pdf

https://marketing.wharton.upenn.edu/wp-content/uploads/2016/10/Dietvorst-Simmons-Massey-2014.pdf

[13] [27] https://www.microsoft.com/en-us/research/wp-content/uploads/2019/01/Guidelines-for-Human-AI-Interaction-camera-ready.pdf

https://www.microsoft.com/en-us/research/wp-content/uploads/2019/01/Guidelines-for-Human-AI-Interaction-camera-ready.pdf

[14] https://www.nature.com/articles/s41562-024-02024-1

https://www.nature.com/articles/s41562-024-02024-1

[15] [20] [29] [33] https://hbr.org/2023/06/how-ai-can-help-stressed-out-managers-be-better-coaches

https://hbr.org/2023/06/how-ai-can-help-stressed-out-managers-be-better-coaches

[16] [39] https://idl.cs.washington.edu/files/2021-AIExplanationsTeamPerformance-CHI.pdf

https://idl.cs.washington.edu/files/2021-AIExplanationsTeamPerformance-CHI.pdf

[18] https://oveislab.com/s/grandeysayre2019.pdf

https://oveislab.com/s/grandeysayre2019.pdf

[19] https://hbr.org/2022/11/the-emotional-labor-of-being-a-leader

https://hbr.org/2022/11/the-emotional-labor-of-being-a-leader

[23] https://www.nature.com/articles/s42256-024-00976-7

https://www.nature.com/articles/s42256-024-00976-7

[25] https://hbr.org/2025/02/how-ai-can-help-managers-think-through-problems

https://hbr.org/2025/02/how-ai-can-help-managers-think-through-problems

[26] https://faculty.wharton.upenn.edu/wp-content/uploads/2016/08/Dietvorst-Simmons-Massey-2018.pdf

https://faculty.wharton.upenn.edu/wp-content/uploads/2016/08/Dietvorst-Simmons-Massey-2018.pdf

[32] https://rework.withgoogle.com/intl/en/guides/following-the-data-the-research-behind-great-managers

https://rework.withgoogle.com/intl/en/guides/following-the-data-the-research-behind-great-managers

[38] https://www.gallup.com/workplace/510326/manager-squeeze-new-workplace-testing-team-leaders.aspx

https://www.gallup.com/workplace/510326/manager-squeeze-new-workplace-testing-team-leaders.aspx

[41] https://www.workcompprofessionals.com/advisory/2016L5/august/MetaAnalysis_Q12_ResearchPaper_0416_v5_sz.pdf

https://www.workcompprofessionals.com/advisory/2016L5/august/MetaAnalysis_Q12_ResearchPaper_0416_v5_sz.pdf

[44] https://arxiv.org/abs/2102.09692

https://arxiv.org/abs/2102.09692