Proceedings of the Design Society, Volume 6: DESIGN 2026
https://doi.org/10.1017/pds.2026.10588

On using LLM reasoning to support reflection in design thinking

Vaia Giota[1,✉], Anthony Karageorgos[1], Georgios Koronis[2] and Efthymios Lallas[1]

[1] University of Thessaly, Greece, [2] University of the Aegean

✉ veniagiota@uth.gr

ABSTRACT: Design thinking fosters creativity but it's susGceptible to cognitive biases. We propose a rule-based framework supported by large language models that uses a Prompt-Reflection-Reframe loop to identify bias mechanisms in designers' verbal reasoning and generate theory-grounded reflective prompts. Through scenario-based evaluations, we validate the framework's theoretical foundations and establish a methodological basis for supporting bias-aware design practice.

KEYWORDS: design thinking, artificial intelligence (AI), design cognition

1. Introduction

Design thinking is widely used as a human-centred approach to innovation and innovative problem solving across diverse domains, from product development to societal transformation (Brown, 2008; Dorst, 2011; Liedtka, 2018). It combines analysis, intuition, user empathy, and iterative experimentation to address complex design tasks (Cross, 2011). Beyond its methodological structure, design thinking has been positioned as a means of improving decision quality by encouraging hypothesis testing, reframing, and the surfacing of assumptions (Liedtka, 2018). Yet, despite its reflective nature, design thinking remains vulnerable to cognitive biases, systemic distortions in reasoning that arise from the reliance on heuristics under uncertainty (Kahneman, 2011; Liedtka, 2015; Tversky & Kahneman, 1974). Design decisions are frequently shaped by intuitive judgments, making them especially prone to biases such as fixation, confirmation, and anchoring (Ball & Christensen, 2009; Dorst & Cross, 2001). These biases can distort how designers frame problems, interpret evidence, and prioritise ideas without conscious awareness (Hallihan & Shu, 2013; Holguin Jimenez et al., 2024).

Traditional debiasing methods, such as awareness training, checklists and structured reflection, provide some support, but they struggle to integrate the fast-moving, context-sensitive nature of design tasks (Croskerry et al., 2013; Larrick, 2004). These approaches are typically applied before or after an activity, whereas biased interpretations often emerge dynamically during ongoing sense making (Kahneman, 2011; Larrick, 2004). In design contexts, this occurs as designers interpret information, frame problems, and converge on ideas in real time (Ball & Christensen, 2009; Dorst & Cross, 2001). Recent work further shows that checklist-style interventions yield only modest improvements in practice, with many bias indicators persisting even after training (Morewedge et al., 2015; Schauer et al., 2025). Because biased interpretations often stabilise early and guide subsequent reasoning (Ball & Christensen, 2009; Kahneman, 2011), interventions that occur retrospectively may arrive too late to meaningfully reshape the design trajectory. These limitations suggest the need for adaptive cognitive scaffolds capable of supporting real-time reflective reasoning, rather than retrospective evaluation, during design work.

Advances in large language models (LLMs) offer a new opportunity to meet this need. Models such as GPT-4 can interpret nuanced prompts, sustain reflective dialogue, and externalise reasoning in ways that can complement designers' intuitive processes (Brown et al., 2020; Kojima et al., 2022). Crucially, LLMs can operate interactively and responsively within ongoing reasoning processes, enabling reflective prompts to be introduced at the moment when interpretations are forming rather than after decisions have

https://doi.org/10.1017/pds.2026.10588 Published online by Cambridge University Press

already stabilised (Prystawski & Goodman, 2023). Emerging research suggests that LLMs can function as reflective collaborators, helping designers articulate, question and reframe assumptions rather than automate creative work (D'Incà et al., 2024; Harwood, 2023). Importantly, this line of work positions LLMs not as agents that replace human judgement, but as tools that can scaffold reflective awareness within complex cognitive activities (Harwood, 2023).

Building on these insights, this paper proposes a prompt-reflection-reframe loop as a cognitive-computational method for supporting bias-aware reflection in design thinking, integrating LLM-based reasoning with principles from cognitive psychology. The framework is grounded in verbalised design reasoning, such as articulated interpretations, decisions, and reflections expressed in natural language, as a methodological starting point. Extension to visual and artefact-based representations identified as a coherent next step. The proposed loop is designed to surface and invite reflection on potential bias mechanisms as they emerge in designers' verbalised reasoning. As illustrated in (Figure 1), the interaction unfolds through a straightforward loop: the designer articulates ideas, the LLM responds with prompts that surface underlying assumptions, and the designer reassesses and reframes their comprehension. In this interaction, the LLM functions as a metacognitive partner that supports awareness and critical reflection, while allowing creative flow to continue uninterrupted.

The paper's contribution is twofold. First, it reframes bias mitigation in design as a form of AI-assisted metacognition, demonstrating how LLMs can operationalise cognitive bias theory into reflective interventions during verbalised design reasoning. Second, it introduces a structured, prompt-based method that translates theoretical debiasing principles into a practical interaction mechanism. This method illustrates how prompt engineering and chain-of-thought reasoning can foster more reflective and evidence-aware thinking in design practice. The paper is organised as follows: Section 2 outlines the theoretical foundations of cognitive bias and debiasing, Section 3, presents the methodological framework, Section 4 presents a theory-driven evaluation and Section 5 discusses conclusions and future directions.

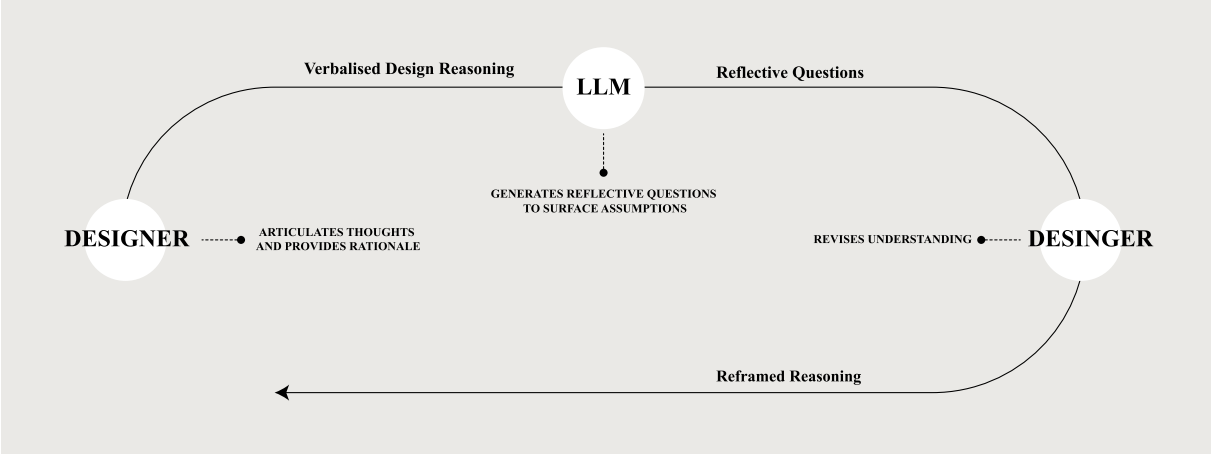


Figure 1. The prompt-reflection-reframe loop

2. Theoretical foundations

Cognitive biases are systematic deviations from rational judgment that arise when individuals rely on heuristics to make decisions under uncertainty (Tversky & Kahneman, 1974). Heuristics like availability, anchoring, and confirmation allow a quick judgement to be made, but can disrupt evidence weighting and comparison among alternatives (Croskerry et al., 2013; Dror, 2020; Larrick, 2004). Research in medicine and behavioural decision-making consistently shows that simple awareness of bias is insufficient; debiasing requires metacognitive regulation, mechanisms for monitoring one's reasoning and actively correcting it (Croskerry et al., 2013; Dror, 2020; Morewedge et al., 2015).

Design cognition is similarly affected. Designers tend to rely on experiential intuition, visual analogy, and precedent-based reasoning, which encourages creativity but also introduce fixation and selective interpretation (Ball & Christensen, 2009; Hallihan & Shu, 2013). Biases emerge at different stages of the design process: confirmation and anchoring during problem framing, fixation and availability during idea generation, and ownership or framing effects during concept evaluation (Holguin Jimenez et al., 2024; Liedtka, 2015). These patterns align with dual-process theory of intuitive and analytic reasoning

https://doi.org/10.1017/pds.2026.10588 Published online by Cambridge University Press

(Kahneman, 2011). Because the early stages of design work often involve rapid interpretation, intuitive judgments, such as forming an initial view of the problem or sketching early solution ideas (Ball & Christensen, 2009; Cross, 2011), they are especially open to bias.

Importantly, many of the cognitive biases identified in design cognition research become observable through designers' verbalised reasoning. Designers routinely externalise their interpretations, assumptions, and evaluative judgments through spoken or written language during activities such as problem framing, design reviews, and collaborative sense making (Ball & Christensen, 2009; Dorst & Cross, 2001). Prior studies show that biases such as confirmation, anchoring, and fixation are often reflected linguistically in recurring reasoning patterns in design discourse, including selective justification of supporting evidence, premature expressions of confidence, resistance to counterarguments, and early commitment framed as constraint (Hallihan & Shu, 2013; Holguin Jimenez et al., 2024). Typical verbal formulations reflecting these patterns include statements such as 'we've already tested this' or 'users clearly prefer this option', which exemplify well-documented cognitive mechanisms of overconfidence and commitment described in the broader judgment and decision-making literature (Kahneman, 2011). These verbal expressions actively shape how problems are framed, alternatives are explored, and design trajectories stabilise. As a result, natural-language reasoning traces offer a theoretically grounded and accessible entry point for identifying and supporting reflection on bias during design work.

Traditional debiasing techniques, whether awareness training, checklists or structured reflection, address aspects of this challenge but often function as external add-ons rather than as integrated components of design practice (Hallihan & Shu, 2013; Morewedge et al., 2015). Their retrospective or static nature makes them difficult to apply in iterative design processes, and they rely heavily on self-regulation, which bias itself tends to undermine (Croskerry et al., 2013). More effective approaches emphasise metacognitive cycles of awareness, monitoring and regulation (Flavell, 1979) and reflection-in-action within the flow of design work (Schön, 1983).

3. Operationalising cognitive debiasing in design thinking

3.1. Framework overview

The framework proposed below operationalises the theoretical premise outlined in Section 2, by specifying an interaction structure for supporting bias-aware reflection within design thinking. It establishes a set of interaction principles that operationalise how designers surface, examine, and reframe intuitive judgments during design activity. Its architecture consists of three components. At its core lies the prompt-reflection-reframe Loop, which structures conversational interactions between designer and model, a set of prompt phases that organise how reflective questions are introduced across the design process (Define, Ideate, Prototype, Test), and a rule-based interpretive layer that maps bias mechanisms to observable reasoning cues within dialogue. Together, these components form a coherent method for integrating real-time metacognition into the flow of design activity.

The framework is not meant to replace human judgment or creative intuition. Accordingly, reflective prompts are intentionally framed as open-ended questions that invite reconsideration rather than prescribe corrective action. This positions the LLM as a reflective collaborator that supports designers in articulating assumptions, considering alternative interpretations, and exploring possibilities they might otherwise overlook, without interrupting the flow of their creative process. (Figure 2) illustrates the interaction architecture, showing how rule-based interpretation informs prompting within the conversational loop. The prompt-reflection-reframe loop

3.2. The prompt-reflection-reframe loop

At the centre of the framework is the prompt-reflection-reframe loop, which structures how designers and the model interact as reasoning unfolds. The loop begins when the designer articulates an emerging idea, interpretation, or concern in natural language. This articulation provides an explicit reasoning trace that the system can respond to within the interaction loop, rather than relying on implicit or inaccessible cognitive processes.

The model interprets this input through the rule-based evaluation layer, analysing whether the reasoning exhibits patterns mapped to specific bias mechanisms defined in Section 2, such as narrowing of interpretation, fixation on early cues, or selective use of evidence. Rather than drawing from a set of pre-prepared responses, the model generates a prompt specifically tailored to the

https://doi.org/10.1017/pds.2026.10588 Published online by Cambridge University Press

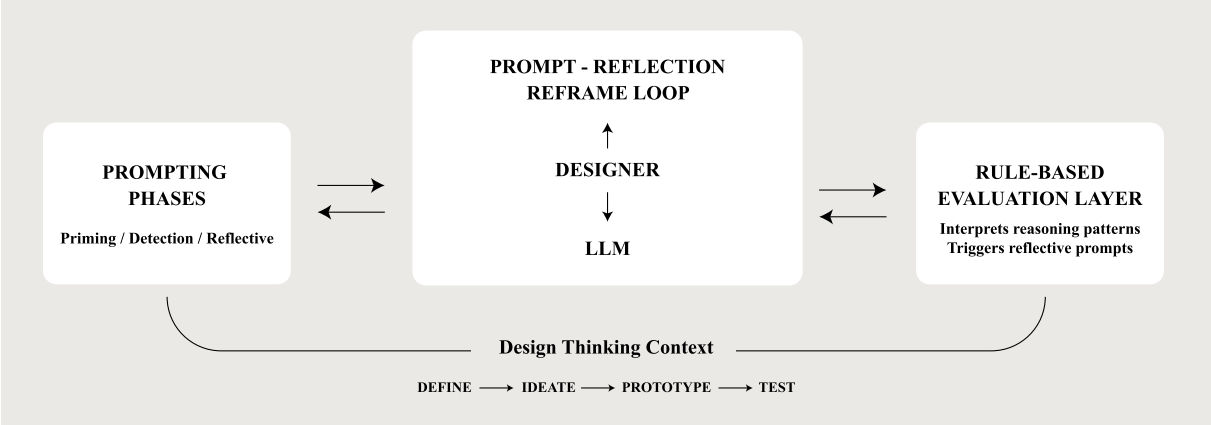


Figure 2. Cognitive-computational framework for bias-aware reflective prompting in design thinking

designer's reasoning at that moment, based on the activation of theory-derived rules, which are described in detail in Section 3.4. These prompts are phrased as open questions, designed to widen the designer's perspective without enforcing a particular solution. For example, when a designer suggests that "users seem comfortable with this layout, so we should keep it unchanged," the model may respond: "Is there anything in the feedback that might point to friction or confusion? Could the way the task flows be influencing this impression?"

The designer then reflects on the question, revises or elaborates the original idea, and the loop continues. Through this iterative movement, articulation, prompting, reframing, the dialogue introduces moments of reflection within the design activity. The aim is not to slow or override intuitive creativity, but to offer well-timed invitations to consider alternative explanations, broaden the problem frame, or reflect on unspoken assumptions. Over time, this structure supports the development of reflective habits, helping designers internalise the questioning pattern and engage more deliberately with their own reasoning. In this sense, the prompt-reflection-reframe loop implements Schön's (1983) notion of reflection-in-action by providing a conversational partner capable of noticing patterns, posing timely questions, and supporting the designer's ongoing sense making in situ. While the framework specifies how prompts are generated in response to designers' reasoning, evaluation procedures may introduce additional scaffolding to ensure consistent enactment of this logic during testing.

3.3. Prompting phases across design thinking

While the prompt-reflection-reframe Loop describes the micro-level pattern of interaction between designers and model, its effectiveness depends on how reflective prompts are aligned with evolving cognitive demands of the design process. Design Thinking unfolds through phases that vary in ambiguity, commitment, and evidential grounding, requiring different forms of reflective support as designers move from problem framing to decision consolidation. Accordingly, the framework organises prompting into three functions models, Priming, Detection, and Reflective Prompting, which are dynamically applied across Define, Ideate, Prototype and Test stages. (Figure 2) illustrates how these prompting modes are applied across the design process as cognitive demands shift.

Research in design cognition characterises early stages of design as periods of exploratory sense making under high ambiguity, followed by increasing convergence and commitment as solutions take shape (Dorst, 2011). As outlined in Section 2, bias mechanisms tend to manifest differently across these stages, requiring stage-sensitive forms of intervention. For example, anchoring and confirmation bias are particularly prevalent during problem framing, when early interpretations begin to stabilise, while fixation is well documented during ideation and early prototyping as designers repeatedly reuse initial concepts or representations (Ball & Christensen, 2009; Jansson & Smith, 1991). Ownership bias may intensify in later stages as attachment to preferred solutions grows (Hallihan & Shu, 2013). The prompting modes therefore function as adaptive interactions patterns rather than uniform intervention. In the Define stage, Priming prompts encourage designers to articulate assumptions, constraints, and initial interpretations explicitly. By foregrounding implicit framing choices early on, these prompts

https://doi.org/10.1017/pds.2026.10588 Published online by Cambridge University Press

support broader problem exploration without imposing evaluative judgement. During Ideation, when designers generate and elaborate solution directions, Detection prompts surface emerging patterns of interpretive narrowing, such as selective evidence use or premature fixation, inviting reconsideration without interrupting creative flow. As ideas are externalised through early prototypes, Detection prompts continue to play a role by highlighting whether emerging evidence is being interpreted in ways that reinforce prior commitments.

In later stages, as designs are tested and refined, Reflective prompts support consolidation and sense making by encouraging designers to examine how feedback, interpretation, and design intentions align. These prompts are particularly relevant when decisions begin to stabilise, helping designers revisit assumptions and interpretations before final commitment. Across stages, prompts arise in direct response to designers' expressed reasoning, ensuring that reflective support is integrated into ongoing activity rather than imposed as a separate evaluative step.

The intensity and form of reflective prompting may also vary with designer expertise. Studies of expert–novice differences in design suggest that novices often benefit from more explicit scaffolding, while experienced designers require lighter, less intrusive cues that respect established reflective habits (Cross, 2011). The framework accommodates this variability by treating prompting modes as adaptable interaction patterns rather than fixed prescriptions.

By aligning prompting modes with both the temporal structure of design thinking and the cognitive characteristics of different stages, the framework supports bias-aware reflection while preserving designer agency and creative autonomy. This stage-sensitive integration ensures that reflective interventions remain timely, context-aware, and responsive to the dynamics of professional design practice.

3.4. Rule-based evaluation layer

The rule-based evaluation layer forms the analytical foundation of the framework. Its purpose is to translate cognitive mechanisms identified in Section 2 into patterns that can be recognised within natural-language design discourse. Rather than relying on opaque computational detection, an approach that can obscure how conclusions are drawn (Harwood, 2023), the framework uses transparent, human-auditable rules that articulate how specific biases tend to appear in designers' reasoning.

The rules are constructed through a theory-driven, symptom-based operationalisation of cognitive bias. Each rule originates from a well-established bias mechanism in judgment and decision-making research and is translated into documented manifestations in design cognition, such as selective reference use, early commitment, or persistent reuse of initial representations. These manifestations are operationalised not as properties of individual artefacts, but as observable patterns that persist across iterations of externalised reasoning. Rule activation therefore does not indicate the presence of bias but signals the appropriateness of reflective intervention at a given moment. This construction logic preserves theoretical grounding while avoiding prescriptive evaluation, aligning with accounts of reflective practice that treat design artefacts as active carriers of thought rather than passive outputs.

Operationally, this construction was implemented through a structured literature-mechanism-evaluation approach, drawing on practices for rigorous qualitative synthesis and AI-assisted analysis (Carrera-Rivera et al., 2022; Kitchenham et al., 2009; Petersen et al., 2015). First, key bias mechanisms identified in the literature, including selective evidence use, anchoring on early ideas, fixation, resistance to feedback, and ownership effects, were mapped to recurring reasoning patterns observable in designers' verbalised thinking (Hallihan & Shu, 2013; Holguin Jimenez et al., 2024; Jansson & Smith, 1991; Kahneman, 2011). These patterns were then formalised into rule statements that specify the linguistic and interactional conditions under which reflective prompts should be triggered.

As an illustrative case, confirmation bias manifests in design reasoning as selective reference use and justificatory language (Hallihan & Shu, 2013; Kahneman, 2011; Nickerson, 1998). In practice, this often appears in designers' verbalised reasoning through statements such as "Most users liked this idea," while critical or ambiguous feedback receives little attention. This pattern is operationalised in Rule 1 ("Focus on Supporting Evidence", Table 1), which detects repeated emphasis on confirming feedback across successive reasoning turns and triggers a reflective prompt inviting consideration of disconfirming evidence.

https://doi.org/10.1017/pds.2026.10588 Published online by Cambridge University Press

Table 1. Rule-based confirmation bias detection framework

Rule	Description & Literature Basis	Prompt-Based Reflection Strategy
1. Supporting Evidence	Emphasis on confirming information (Nickerson, 1998; Kahneman, 2011).	Priming: "List three pieces of evidence that contradict your current idea."
2. Dismissal of Contradictory Evidence	Negative evidence minimised or ignored (Nickerson, 1998; Dror, 2020).	Detection: "How would you respond to opposing user feedback"
3. Biased Interpretation of Ambiguous Feedback	Ambiguous feedback interpreted to support existing concepts (Hallihan & Shu, 2013).	Reflective: "What risks does this feedback reveal?"
4. Early Fixation	Premature commitment to an idea (Hallihan & Shu, 2013; Dror, 2020).	Priming: "Generate three alternative directions."
5. Limited Exploration of Alternatives	Insufficient consideration of alternatives (Liedtka, 2015; Dror, 2020).	Detection: "What alternative solutions could address this need?"
6. Justification over Reflection	Defensive reasoning over critical evaluation (Soll et al., 2015; Liedtka, 2015).	Reflective: "What is the strongest critique of this solution?"
7. Dismissive Language Toward Criticism	Criticism reframed as irrelevant or minor (Dror, 2020; Hallihan & Shu, 2013).	Detection: "If criticism is valid, what must change?"
8. Certainty Language	Overconfident expressions of certainty (Kahneman, 2011; Tversky & Kahneman, 1974).	Priming: "On a scale of 1-10, how certain are you? What evidence reduces that certainty?"
9. Resistance to Change	Persistence despite discomforting evidence (Nickerson, 1998; Dror, 2020).	Reflective: "If starting fresh today, would you still pursue this design? Why or why not?"
10. Selective Framing of Data	Selective quotation of supporting evidence (Nickerson, 1998; Luo & Greenberg, 2024).	Reflective: "Reframe the same data against your current choice."

To operationalise these mechanisms, the authors constructed short design scenarios grounded in professional design experience. These scenarios, expressed in natural language, illustrated both biased and unbiased reasoning in situations typical of early-stage design, such as premature commitment to an initial idea, selective interpretation of positive client feedback, or reluctance to explore alternatives. Although distortions may also arise in visual or artefact-based reasoning, they were represented textually to enable controlled evaluation within the language-based scope of this framework, extending the approach to visual artefact remains future work. Candidate rules were first constructed through a cross-disciplinary literature review in cognitive psychology, decision science, and design cognition. These rules were then tested using an LLM (ChatGPT-4), which received the designer's scenario together with the explicit natural-language formulation of the relevant rule and an instruction to analyse the reasoning and generate a reflective prompt aligned with that rule. This process allowed evaluation of the rules' interpretability and alignment with intended cognitive mechanisms. This 'consider-the-opposite' style prompting, which mirrors established reflective-challenge techniques in design and debiasing research (Cross, 2011; Soll et al., 2015), enabled evaluation of whether each rule activated in suitable contexts and whether the prompt it generated reflected the correct underlying bias mechanism.

Each rule is formulated as an if-then schema, for example, if a statement interprets mixed feedback only in ways that support an existing preference, then the system triggers a prompt encouraging consideration of disconfirming evidence. A rule was retained only if it satisfied three criteria: (1) recognisability, meaning the reasoning pattern appears in real design discourse, (2) multi-source grounding, requiring theoretical support from at least two independent literatures, and (3) promptability, ensuring the mechanism can be expressed as an open reflective question aligned with the prompt-reflection-reframe loop and with metacognitive principles of awareness, monitoring, and regulation (Croskerry et al., 2013; Flavell, 1979).

https://doi.org/10.1017/pds.2026.10588 Published online by Cambridge University Press

Confirmation bias served as the initial validation case due to its prominence in design reasoning and well-documented cognitive mechanisms. Ten rules capturing its characteristic distortions were developed and tested using this method (Table 1). Once validated, the same derivation process was extended to additional bias families, including anchoring, fixation, ownership bias, and group-level tendencies such as groupthink, producing a generalisable and interpretable rule base. To maintain coherence across these diverse mechanisms, the rules were organised into eight broader categories of reasoning distortion: Evidence Handling, Interpretation, Fixation, Exploration, Justification, Critique, Confidence, and Resistance (Table 2). These categories serve as a conceptual bridge connecting cognitive theory, prompting phases, and the operational structure of the framework. The operational layer remains deliberately simple: rules are explicit and interpretable, prompts are expressed in accessible natural language, and outputs encourage judgement rather than prescribe solutions. This aligns with principles of human-centred AI, favouring transparency, interpretability, and designer agency (Harwood, 2023), and ensures that the reflective support offered by the framework augments, rather than replaces, creative reasoning.

Table 2. Taxonomy of confirmation bias manifestations

Cognitive Category	Associated Reasoning Patterns
Evidence Handling	Rule 1: Focus on supporting evidence Rule 2: Neglect discomforting data
Interpretation	Rule 3: Biased reading of feedback
Fixation	Rule 4: Premature Closure
Exploration	Rule 5: Limited generation of alternatives
Justification	Rule 6: Rationalisation over reflection
Critique	Rule 7: Dismissal of counterarguments
Confidence	Rule 8: Certainty language, inflated confidence
Resistance	Rule 9: Resistance to change Rule 10: Belief perseverance

4. Evaluation

The evaluation followed a formative approach based on structured scenario testing, designed to assess whether the prompt-reflection-reframe loop behaves in accordance with its theoretical grounding, generating reflective prompts aligned with underlying cognitive mechanisms associated with bias in design reasoning. Because the aim of this work is to articulate and validate a methodological mechanism for bias-aware reflection, rather than assess user performance or outcome-level debiasing effects at scale, the evaluation focuses on construct validity, whether the system's behaviour aligns with theoretical expectations and on functional coherence across multiple bias families.

Evaluation began with a single-bias prototype centred on confirmation bias, chosen due to its prevalence in design reasoning and its well-documented cognitive structure. Ten rules derived from cross-disciplinary literature were applied to a corpus of short design scenarios grounded in common reasoning patterns reported in professional design practice and empirical studies. Each scenario was expressed in natural language to allow the LLM to analyse the explicit reasoning trace.

Few-shot prompting and step-by-step guidance were employed as part of the evaluation setup to ensure consistent enactment of the rule logic defined in Section 3. These scaffolding elements constrained only the procedural structure of the task (e.g., analysis followed by generation of a reflective question aligned with the activated rule) and did not predefine or constrain the semantic content of the prompts. Although the model generated prompts in response to rule-conditioned instructions rather than pre-written templates, the authors reviewed outputs to assess their interpretive accuracy and theoretical alignment with the intended cognitive constructs.

The evaluation proceeded iteratively. For each scenario, outputs were analysed along three dimensions: (1) Interpretive coherence, whether the model's response plausibly matched the reasoning pattern to the intended rule, (2) Prompt quality, whether the generated question reflected "consider-the-opposite" logic without prescribing solutions, and (3) Theoretical alignment, whether the prompt corresponded to the cognitive mechanisms described in the literature underlying each rule.

https://doi.org/10.1017/pds.2026.10588 Published online by Cambridge University Press

Once tested for confirmation bias, the same evaluation procedure was extended to additional bias families, including anchoring, fixation, and ownership bias. This allowed assessment of whether the rule grammar generalised across mechanisms with different cognitive signatures, and whether prompt formulation remained flexible, interpretable and consistent with the reflective intent of the framework. Across the evaluated scenarios, the model consistently generated reflective prompts that aligned with the rule logic when linguistic cues matched the defined conditions. For confirmation bias, the LLM distinguished between selective evidence use, justificatory reasoning, and premature fixation, producing prompts that encourage reconsideration of alternative interpretations rather than directing solutions. In anchoring and fixation cases, the model generated exploratory prompts ("What alternative directions could address the same need?") or framing prompts that questioned early commitment to ideas. Ownership-related reasoning patterns similarly triggered prompts that invited designers to reassess attachment to prior decisions without assuming their preservation.

These results provide initial evidence of construct validity, the system behaves in ways consistent with theoretical grounding, and the rule-based interpretive layer effectively structures the model's reflective output across multiple bias mechanisms. They further suggest that the framework can support different reflective interventions grounded in cognitive bias theory, while preserving designers' agency.

The evaluation remains preliminary and carries several limitations. First, the scenarios used for validation were constructed rather than drawn from real design sessions, which limits ecological fidelity. Second, the evaluation relied on interaction with a single LLM (ChatGPT-4) and therefore reflects model-specific affordances, generalisation to other models requires further testing. Third, although the framework generalises across bias families at the mechanism level, the absence of a full user study means its impact on design outcomes or creative performance or long-term debiasing effects has yet to be empirically assessed.

5. Conclusions and future work

This paper has proposed a cognitive-computational framework intended to support bias-aware reflection during design thinking, integrating insights from cognitive psychology, design cognition, and LLM-mediated reasoning. Rather than treating debiasing as a separate analytic task, the framework embeds reflection within the unfolding activity of design, allowing reasoning to become inspectable and revisable as designers work. Through the prompt-reflection-reframe loop, a stage-aware prompting structure, and a transparent rule-based evaluation layer, the approach translates well-established cognitive mechanisms into context-sensitive reflective prompts that support more balanced judgement without constraining creative fluency.

A formative evaluation using scenario-based reasoning traces demonstrated the framework behaves in accordance with its theoretical grounding, generating reflective questions aligned with characteristic reasoning distortions such as selective evidence search, early fixation, or biased interpretation. These early results offer construct-level validation of the framework's function as a metacognitive scaffold and suggest that its rule grammar generalises effectively across multiple bias families.

Though framed as bias mitigation, the framework deliberately adopts open-ended, non-directive prompting. This reflects design cognition, where prescriptive interventions risk disrupting creative flow, undermining agency, or narrowing the solution space (Cross, 2011; Dorst & Cross, 2001). Instead of triggering correction, bias detection signals for reflective engagement, inviting designers to reconsider assumptions without external judgment. This aligns with metacognitive debiasing emphasising awareness and self-regulation over correction (Croskerry et al., 2013; Larrick, 2004), as reflection-in-action, where questioning serves as a productive mechanism for reframing practice (Schön, 1983). At the same time, reflective interventions necessarily involve a reallocation of cognitive effort toward awareness and self-regulation. In design contexts, this effort is not introduced as an additional task but is integrated into ongoing sense making, consistent with accounts of expert reflective practice (Cross, 2011; Schön, 1983). Nonetheless, different design situations, such as time-critical or highly constrained contexts, may benefit from alternative balances between open-ended reflection and more scaffolded forms of guidance. Exploring how such balances can be adapted to task demands and designer expertise therefore represents an important direction for future work.

Building on the framework's grounding in verbalised reasoning, many of the cognitive tendencies examined here also known to manifest through visual and material representations. Design research has long shown that sketches, drawings, and other externalisations are not merely illustrations but active

https://doi.org/10.1017/pds.2026.10588 Published online by Cambridge University Press

forms of reasoning that reveal interpretive moves, biases, and fixation patterns (Cross, 2011; Dorst & Cross, 2001). Sketching enables the 'reflective conversation with the situation' in which designers oscillate between interpreting what they see and making new moves, often reinforcing early preferences or narrowing attention (Ball & Christensen, 2009; Dorst & Cross, 2001). Recent analyses further demonstrate that biases such as anchoring, fixation, availability, and framing can emerge through visual cues as strongly as through verbal reasoning (Holguin Jimenez et al., 2024). Extending the framework to multimodal analysis is supported by recent work showing how vision-language models can interpret images and text within shared reasoning pipelines, enabling reflective analysis of visual artefacts as well as verbal reasoning (D'Incà et al., 2024).

At the same time, the work remains a methodological foundation; current evaluation relies on text-based scenarios and interactions with a single LLM. Future empirical studies with designers are required to assess effects on framing, evidence balance, and idea diversity. Research must also extend to team-based settings to examine group-level biases like groupthink during co-creation. Finally, extending the framework to incorporate visual and artefact-based reasoning remains a key direction for future development.

By making reasoning explicit and introducing structured opportunities for reconsideration, the proposed framework offers a pathway toward more reflective, evidence-aware, and cognitively resilient design practice. As generative AI continues to permeate design workflows, methods that support rather than override human judgement will become essential. This work establishes a conceptual and methodological foundation for future developments, with potential for expansion across modalities, design contexts, and collaborative scales.

Acknowledgement

The research is conducted in the operating framework of the University of Thessaly Innovation, Technology Transfer Unit and Entrepreneurship Center "One Planet Thessaly", under the "University of Thessaly Grants for Scientific Publication Support" action and is funded by the Special Account of Research Grands of the University of Thessaly.

References

Ball, L. J., & Christensen, B. T. (2009). Analogical reasoning and mental simulation in design: Two strategies linked to uncertainty resolution. *Design Studies*, 30(2), 169–186. https://doi.org/10.1016/j.destud.2008.12.005

Brown, T. (2008). Design thinking. *Harvard Business Review*, 86(6), 84–92. https://hbr.org/2008/06/design-thinking

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., & Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems* (Vol. 33, pp. 1877–1901). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

Carrera-Rivera, A., Ochoa, W., Larrinaga, F., & Lasa, G. (2022). How-to conduct a systematic literature review: A quick guide for computer science research. *MethodsX*, 9, Article 101895. https://doi.org/10.1016/j.mex.2022.101895

Croskerry, P., Singhal, G., & Mamede, S. (2013). Cognitive debiasing 1: Origins of bias and theory of debiasing. *BMJ Quality & Safety*, 22(Suppl 2), ii58–ii64. https://doi.org/10.1136/bmjqs-2012-001712

Cross, N. (2011). *Design thinking: Understanding how designers think and work*. Berg. https://doi.org/10.5040/9781474293884

D'Incà, M., Peruzzo, E., Mancini, M., Xu, D., Goe, V., Xu, X., Wang, Z., Shi, H., & Sebe, N. (2024). OpenBias: Open-set bias detection in text-to-image generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 12225–12235). IEEE. https://doi.org/10.1109/CVPR52733.2024.01162

Dorst, K. (2011). The core of 'design thinking' and its application. *Design Studies*, 32(6), 521–532. https://doi.org/10.1016/j.destud.2011.07.006

Dorst, K., & Cross, N. (2001). Creativity in the design process: Co-evolution of problem–solution. *Design Studies*, 22(5), 425–437. https://doi.org/10.1016/S0142-694X(01)00009-6

Dror, I. E. (2020). Cognitive and human factors in expert decision making: Six fallacies and the eight sources of bias. *Analytical Chemistry*, 92(12), 7998–8004. https://doi.org/10.1021/acs.analchem.0c00704

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist*, 34(10), 906–911. https://doi.org/10.1037/0003-066X.34.10.906

https://doi.org/10.1017/pds.2026.10588 Published online by Cambridge University Press

Hallihan, G. M., & Shu, L. H. (2013). Considering confirmation bias in design and design research. *Journal of Integrated Design and Process Science*, 17(4), 19–35. https://doi.org/10.3233/jid-2013-0019
Harwood, B. (2023). CHAI-DT: A framework for prompting conversational generative AI agents to actively participate in co-creation (arXiv:2305.03852). arXiv. https://doi.org/10.48550/arXiv.2305.03852
Holguin Jimenez, S., Godot, X., Petronijevic, J., Lassagne, M., & Daille-Lefevre, B. (2024). Considering cognitive biases in design: An integrated approach. *Procedia Computer Science*, 232, 2800–2809. https://doi.org/10.1016/j.procs.2024.02.097
Jansson, D. G., & Smith, S. M. (1991). Design fixation. *Design Studies*, 12(1), 3–11. https://doi.org/10.1016/0142-694X(91)90003-F
Kahneman, D. (2011). Thinking, fast and slow. Farrar, Straus and Giroux.
Kitchenham, B., Brereton, O. P., Budgen, D., Turner, M., Bailey, J., & Linkman, S. (2009). Systematic literature reviews in software engineering – A systematic literature review. *Information and Software Technology*, 51(1), 7–15. https://doi.org/10.1016/j.infsof.2008.09.009
Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners (arXiv:2205.11916). arXiv. https://doi.org/10.48550/arXiv.2205.11916
Larrick, R. P. (2004). Debiasing. In D. J. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 316–338). Blackwell Publishing. https://doi.org/10.1002/9780470752937.ch16
Liedtka, J. (2015). Perspective: Linking design thinking with innovation outcomes through cognitive bias reduction. *Journal of Product Innovation Management*, 32(6), 925–938. https://doi.org/10.1111/jpim.12163
Liedtka, J. (2018). Why design thinking works. *Harvard Business Review*, 96(5), 72–79. https://hbr.org/2018/09/why-design-thinking-works
Luo, Y., & Greenberg, A. (2024). Comparing bag-of-words, SBERT, and GPT-3 for bias detection. *Journal of Artificial Intelligence Research*, 79, 415–438. https://doi.org/10.1613/jair.1.15243
Morewedge, C. K., Yoon, H., Scopelliti, I., Symborski, C. W., Korris, J. H., & Kassam, K. S. (2015). Debiasing decisions: Improved decision making with a single training intervention. *Policy Insights from the Behavioral and Brain Sciences*, 2(1), 129–140. https://doi.org/10.1177/2372732215600886
Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2), 175–220. https://doi.org/10.1037/1089-2680.2.2.175
Petersen, K., Vakkalanka, S., & Kuzniarz, L. (2015). Guidelines for conducting systematic mapping studies in software engineering: An update. *Information and Software Technology*, 64, 1–18. https://doi.org/10.1016/j.infsof.2015.03.007
Prystawski, B., & Goodman, N. D. (2023). Why think step-by-step? Reasoning emerges from the locality of experience (arXiv:2304.03843). arXiv. https://doi.org/10.48550/arXiv.2304.03843
Schauer, A., Afolabi, H., & Fu, K. (2025). Validation of a checklist-style intervention for mitigation of availability bias in professional designers. *Design Science*, 11, Article e2. https://doi.org/10.1017/dsj.2025.10006
Schön, D. A. (1983). *The reflective practitioner: How professionals think in action*. Basic Books.
Soll, J. B., Milkman, K. L., & Payne, J. W. (2015). A user's guide to debiasing. In G. Keren & G. Wu (Eds.), *The Wiley Blackwell handbook of judgment and decision making* (pp. 924–951). Wiley Blackwell. https://doi.org/10.1002/9781118468333.ch33
Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124
Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E. H., Xia, F., Le, Q., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In A. H. Oh, A. Agarwal, D. Belgrave, & K. Cho (Eds.), *Advances in Neural Information Processing Systems* (Vol. 35, pp. 24824–24837). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html

https://doi.org/10.1017/pds.2026.10588 Published online by Cambridge University Press